Linear Subspace and Manifold Learning via Extrinsic Geometry

by

Brian St. Thomas

Department of Statistical Science
Duke University

Date: ________________________

Approved:

______________________________
Sayan Mukherjee, Supervisor

______________________________
Li Ma

______________________________
Scott Schmidler

______________________________
Lizhen Lin

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University 2015
Abstract

Linear Subspace and Manifold Learning via Extrinsic Geometry

by

Brian St. Thomas

Department of Statistical Science
Duke University

Date: __________________________
Approved:

______________________________
Sayan Mukherjee, Supervisor

______________________________
Li Ma

______________________________
Scott Schmidler

______________________________
Lizhen Lin

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University 2015
Abstract

In the last few decades, data analysis techniques have had to expand to handle large sets of data with complicated structure. This includes identifying low dimensional structure in high dimensional data, analyzing shape and image data, and learning from or classifying large corpora of text documents. Common Bayesian and Machine Learning techniques rely on using the unique geometry of these data types, however departing from Euclidean geometry can result in both theoretical and practical complications. Bayesian nonparametric approaches can be particularly challenging in these areas.

This dissertation proposes a novel approach to these challenges by working with convenient embeddings of the manifold valued parameters of interest, commonly making use of an extrinsic distance or measure on the manifold. Carefully selected extrinsic distances are shown to reduce the computational cost and to increase accuracy of inference. The embeddings are also used to yield straight forward derivations for nonparametric techniques. The methods developed are applied to subspace learning in dimension reduction problems, planar shapes, shape constrained regression, and text analysis.
To my family
## Contents

Abstract iv

List of Tables ix

List of Figures x

List of Abbreviations and Symbols xiv

Acknowledgements xv

1 Introduction 1

2 Bayesian Inference on Mixtures of Linear Subspaces 4

  2.1 Mixtures of linear subspaces .......................... 7

  2.1.1 Likelihood specification ............................. 7

  2.1.2 Prior specification and the spherical embedding .... 9

  2.2 Posterior sampling .................................... 15

  2.2.1 Sampling $U_{1:K}$ and $\theta_{1:K}$ from the Gibbs posterior 17

  2.2.2 Sampling algorithm ................................ 19

  2.3 Posterior consistency .................................. 21

  2.4 Results on real and simulated data .................... 23

  2.4.1 Simulated data .................................... 24

  2.4.2 Real data ......................................... 25

  2.5 Discussion ............................................ 27
A Appendix to Chapter 2 104
B Appendix to Chapter 3 114
C Appendix to Chapter 4 124
Bibliography 130
Biography 139
List of Tables

2.1 Range of cluster assignment accuracy for the simulated data for the three algorithms: $K$-means, mixture of subspaces, and mixture of normals. ................................................. 25

2.2 Range of cluster assignment accuracy for the three data sets using the three classifiers on the left-out-data. .................................................. 27

2.3 Posterior probabilities for the dimension of the subspace of clusters in each data set. ................................................................. 28

4.1 [1-d functions] Root mean square error for simulated data with $n = 100$ and the results averaged across 50 simulation replicates in each case. ................................................................. 69

4.2 [2-d functions] Root mean square error for simulated data with $n = 100$ and the results averaged across 50 simulation replicates in each case. ................................................................. 71

4.3 [3-d functions] Root mean square error for simulated data with $n = 100$ and the results averaged across 50 simulation replicates in each case. ................................................................. 73

4.4 Empirical coverage of the bootstrap procedure for $F_1$ at evaluated at 10 covariate levels for two one-dimensional functions. ................. 73

5.1 The number of estimated topics from the proposed ensemble method on 15 different synthetic corpora with a true number of topics $k = 40$. The clustering method reliably reduces the number of topics to approximate the true number. ................................................. 93
List of Figures

2.1 An illustration of the spherical embedding for subspaces from $\text{Gr}(1, 2)$ into $\mathbb{R}^2$ on the left, and of $\text{Gr}(k, 2)$, $k = 0, 1, 2$ on the right. Images of the embedding are in black, and the center of the sphere is in red. The coordinates for the embedding into $\mathbb{R}^2$ is the first column of the projection matrix. By including the last entry in the diagonal of the projection matrix, we obtain coordinates for the embedding into $\mathbb{R}^3$. $\text{Gr}(0, 2)$ and $\text{Gr}(2, 2)$ are trivial sets giving $0_3$ and $I_3$ as their projection matrices. They act as poles on the sphere with coordinates $(0, 0, 0)$ and $(1, 0, 1)$. ................................................................. 12

2.2 The true, $K$-means, standard normal, and subspace model classification of data generated from a two-component mixture model, with one-dimensional affine subspace and a two-dimensional affine subspace, creating a line intersecting a plane. ................................................................. 26

3.1 **Left** The training values on the sphere. **Middle** The held out values to be predicted through extrinsic regression. **Right** The extrinsic predictions (blue) plotted against the true values (red). ................................................................. 38

3.2 The performance of extrinsic and intrinsic regression models on 50 test observations from sphere regression models with concentration parameters from 1 to 20. Each color corresponds to a concentration parameter. The extrinsic and intrinsic models have similar performance in predictive MSE with low concentration parameters. However in terms of MSE, the extrinsic model appears to perform better with lower sample sizes even with lower concentration parameters. . . 40

3.3 Speed comparisons between the extrinsic and intrinsic kernel regressions as a function of the number of training observations. The average seconds to produce an estimate for a single test observation are plotted in red for the intrinsic model, and black for the extrinsic model. The multiple between the speed for the intrinsic and extrinsic estimates plotted are also plotted for reference. ................................. 42
3.4 Predicted CC shape for children ages 9, 12, 16, and 19. The black shape corresponds to typically developing children, while the red shape corresponds to children diagnosed with ADHD. Kernel regression allows us to visualize how CC shape changes through development. Here sections of CC appear smaller in ADHD diagnoses than in normal development.

3.5 Pairwise distance between two observations generated by the specified model indexed by \( t \) measured by distance between points in the Conway embedding. This visualization of the extrinsic distance shows the cluster by dimension.

3.6 The estimated dimension and residual for the extrinsic kernel regression estimate at each time point \( t \) from data generated from the specified model. The regression estimate is accurate on the dimension of the subspace and prediction residuals are consistent with a concentration parameter \( \kappa = 1 \).

3.7 The covariance matrices estimated from the daily closing prices from various stock market index funds over various weeks. Covariance between markets change substantially from week to week.

3.8 The distribution of residuals from extrinsic kernel regression compared to the distribution of the distance between observations. In general smaller residuals suggest that the changes in covariance structure is dynamic in a learnable way.

4.1 RMSE for 1-dimensional projection estimates

4.2 RMSE for 2-dimensional projection estimates

4.3 RMSE for 3-dimensional projection estimates

4.4 Bootstrap confidence interval estimates for sinusoidal (\( F_2 \)) and logistic (\( F_6 \)) functions. The black line is the estimate, while the blue lines give the confidence intervals. The true function is given in green, while the training points are given in black.

4.5 Bootstrap confidence interval estimates for \( F_1 \) and \( F_6 \) functions. The estimate is given as the black plane, while the confidence intervals are given by the blue planes. The true function is given in green, while the training data is given in red.
4.6 Bootstrap Estimate and 95% Bootstrap Confidence Interval for DDT concentration ($\mu$mol/L), TiO$_2$ concentration ($\mu$g/L) and the proportion of human hepatic cells exhibiting micronuclei (per 3,000). The point estimate is in red, and bounds are in blue. The observations are in black. ................................. 74

5.1 The prior densities to produce concentration parameters for standard LDA models. The Gamma shape parameter is 50/300 and the rate parameter is 1, to center the topic distribution on the proposed symmetric Dirichlet prior by Steyvers and Griffiths (2007), but still provide a power-law like shape in values. ........................................ 81

5.2 The corpus term frequencies and example document term frequencies generated from the LDA process. Term frequencies are from 5,000 documents of average length 500, with $\alpha_i \overset{iid}{\sim} Ga(50/300, 1)$, and $\beta = 0.1$ for the topic and topic proportion Dirichlet concentration parameters in the asymmetric case, and with $\alpha = 50/300$ and $\beta = 0.1$ in the symmetric case. Power law plots appear as straight lines on log-log plots, so neither realization conforms to Zipf’s law. ....................... 82

5.3 The corpus term frequencies ranked and plotted by the log the sorting index and the log of the term frequency. The black is NSF data, red is the power law synthetic data, and blue is the symmetric dirichlet generated data. ........................................ 84

5.4 Document term frequencies ranked and plotted by the log the sorting index and the log of the term frequency. The black is NSF data while red is power law distributed data. There is notably less variability in the synthetic data, however the shapes of the distributions are similar throughout. ........................................ 85

5.5 The resulting cluster of topics from an ensemble of 100 models. Each model contains 50 topics, so 5000 topics are clustered using $K$-means clustering, with the number of clusters set to 50. Clusters with a number of members far away from 100 members indicate that the member topics are of low quality since they tend to be too generic, or overly specific to a small number of documents. ....................... 90

5.6 Word clouds of large (more than 120 members) clusters. These topics put high probability mass on words that are independent. The prevalence of these words in every document makes the presence of any one word uninformative of the others, i.e. $P(\text{science}|\text{education}) = P(\text{science})$ in a corpus of NSF grant awards. ....................... 91

xii
5.7 The corpus wide topic proportions of the retained and discarded topic. Of the topics, 68% were retained and 32% were discarded as junk topics.

5.8 Examples of quality topics retained by the model. These topics are more related to specific fields of scientific research contained in the NSF grant award data set.

5.9 Topics from LDA applied to 13,092 NSF awards. The size of the terms correspond to the value of the term in the topic.

5.10 Topics for SAM with orthogonal constraints. The size of the terms correspond to the absolute value of the terms in the topics. Since there are both negative and positive values, and since the positive and negative values yield coherent themes, both are shown.

5.11 Distribution on the number of topics as estimated from SAM on the Conway Sphere. The uncertainty in the number of topics seems to be more than in the mixtures of subspaces application of the Conway Sphere, in which the dimensions of each subspace tended to have clear posterior modes.
List of Abbreviations and Symbols

Symbols

Gr(d, m) Grassmannian of d-dimensional subspaces in $\mathbb{R}^m$.
V(d, m) Stiefel manifold of $m \times d$ matrices.
O(d) Orthogonal group of $d \times d$ matrices.
$S^d$ The $d - 1$ dimensional surface of the sphere in $\mathbb{R}^d$.

Abbreviations

SVD Singular Value Decomposition
LDA Latent Dirichlet Allocation model
PCA Principal Components Analysis
HDP Hierarchical Dirichlet Process
SAM Spherical Admixture Model
Acknowledgements

First, I would like to thank my advisor, Sayan Mukherjee, not for just the time he has spent with me on the work for this dissertation, but also for his guidance and advice for professional development. His input has been and will continue to be invaluable to me. In addition I would like to thank the members of my committee, Lizhen Lin, Scott Schmidler, and Li Ma, who have patiently given me their time and support throughout my graduate career.

I would also like to thank my classmates, who struggled with me and showed me new ways of thinking as we prepared for our first year exam. It is a humbling experience to work so closely and learn so much from such dedicated and talented individuals.

I must thank my undergraduate professors Meredith Greer, Pallavi Jayawant, Paula Schlax, and David Haines, as well as my High School teachers Ronald Tiberio and Alan Brazier for inspiring me to pursue a Ph.D. Their love of education and curiosity proved contagious and encouraged me to be a life-long learner.

I’ve been blessed with the most supportive and loving parents and siblings anyone could ask for. I am forever grateful for everything they’ve done for me over the years and filling my life with joy.

Lastly, but certainly not least, I wouldn’t have enjoyed graduate school and Durham nearly as much if I hadn’t met the love of my life, Tina Kelliher.
As data analysis is required for high dimensional and manifold valued data, theory related to Riemannian geometry has become increasingly useful. Intrinsic methods for inference on the parameter space or for prediction of manifold valued data are appealing theoretically, and also from a Bayesian perspective of inferring generative models. However optimization using intrinsic measures can require working through objects like a Frechet mean, which is much less convenient than something like a mean in Euclidean space. In problems where the data or parameters exist in a union of different manifolds, or a stratified space, the manifolds may not share a common intrinsic measure. Thus working with extrinsic geometry, where additional structure is imposed on the data via a carefully chosen embedding, can facilitate both model building and computation.

Chapter 2 describes a finite Bayesian mixture model of linear subspaces that estimates the dimension of the subspaces. To develop this mixture model, an extrinsic distance on linear subspaces is described. This distance, a generalization of the chordal distance to measure distances between subspaces of different dimension, is used to give a joint prior on the dimension of the subspace and the subspace parame-
ters. The chapter and its appendix provide conditions for posterior consistency of the model, and the model is applied to several open source data sets. Although the focus of the chapter is establishing properties of the model, the procedure of developing the model outlines how extrinsic regression can be used to build new models.

Chapter 3 introduces an extrinsic kernel regression technique for manifold valued data. A general procedure is outlined for making kernel estimates given an arbitrary embedding of a manifold and distance in embedded space. The method is then applied to spherical, planar shape, and linear subspace response data. In addition to stating some asymptotic properties of the method, it’s demonstrated that the simplified computations in embedded space lead to estimates that are at least as accurate as their intrinsic counterparts and can be made considerably faster.

Chapter 4 gives a projection approach to multivariate monotone regression. Monotone regression is a special case of shape constrained regression where the regression function is \textit{a priori} strictly increasing. A procedure is given for estimating the monotone function. Starting with an initial estimate of the function, repeated projections onto the space of 1-d monotone functions are applied until the monotonicity criteria is satisfied. This procedure uses the closed-cone geometry of both the multivariate and univariate monotone functions. The performance of this method given different initial estimators is demonstrated. A bootstrapping procedure for point-wise confidence intervals is also given and demonstrated, and the coverage of these intervals is studied. The method is applied to a toxicology data example.

Chapter 5 focuses on a specific type of text analysis called topic modeling. This is a challenging problem as data sets tend to be extremely noisy and high dimensional, but also because it is difficult to measure the performance of models. An example of linking the known power law properties of language to synthetic corpus generation for the purposes of recovering parameters is shown. An ensemble technique for Latent Dirichlet Allocation is given as a strategy to empirically overcome the proven
posterior inconsistency of the model. The extrinsic distance on linear subspaces described in Chapter 2 is then to state a nonparametric version of the Spherical Admixture Model. Both these models are applied to a set of publicly available NSF grant award abstracts.

Chapter 6 gives concluding remarks.
Bayesian Inference on Mixtures of Linear Subspaces

The problem of modeling manifolds has been of great interest in a variety of statistical problems including dimension reduction (Belkin and Niyogi (2003); Donoho and Grimes (2003); Roweis and Saul (2000)), characterizing the distributions of statistical models as points on a Riemannian manifold (Amari (1982); Efron (1978); Rao (1945)), and the extensive literature in statistics and machine learning on manifold learning (Cook (2007); Gine and Koltchinskii (2006); Mukherjee et al. (2010)). A generalization of the manifold setting is to model unions and intersections of manifolds (of possibly different dimensions), formally called stratified spaces (Bendich et al. (2012); Geiger et al. (2001); Goresky and MacPherson (1988)). Stratified spaces arise when data or parameter spaces are characterized by combinations of manifolds such as the case of mixture models. One of the most important special cases arises when the manifolds involved are all affine subspaces or linear subspaces. Mixtures of linear subspaces have been suggested in applications such as tracking images (Haro et al. (2008); Vidal et al. (2005)), quantitative analysis of evolution or artificial selection (Hansen and Houle (2008); Lande (1979)), and applications in communication and coding theory (Ashikhmin and Calderbank (2003); Zheng and
Tse (2002)). In this chapter we provide a model for the simplest instance of inferring stratified spaces, estimating mixtures of linear subspaces of different dimensions.

The idea of dimension reduction via projections onto low-dimensional subspaces goes back at least to Adcock (1878) and Edgeworth (1884), with methodological and foundational contributions by Fisher (1922); see Cook (2007) for an excellent review. It is very interesting that in 1922 Fisher suggested that the statistical setting where the number of variables is greater than the number of observations, \( p \gg n \) could be addressed by reducing the dimension of \( p \) to very few \( p^* \) summaries of the data where \( p^* < n \). The summaries in this setting were linear combinations of the variables. This idea of dimension reduction has been extensively used statistics ranging from classical methods such as principal components analysis (PCA, Hotelling (1933)) to a variety of recent methods, some algorithmic and some likelihood based, that fall under the category of nonlinear dimension reduction and manifold learning (Belkin and Niyogi (2003); Cook (2007); Donoho and Grimes (2003); Gine and Koltchinskii (2006); Mukherjee et al. (2010); Roweis and Saul (2000)). A challenging setting for both algorithmic and probabilistic models in this setting is where the data are being generated from multiple populations inducing a mixture distribution. It is particularly challenging when the mixtures are of different dimensions.

In many applications a useful model for the observed high-dimensional data assumes the data is concentrated around a lower-dimensional structure in the high-dimensional ambient space. In addition, it is often the case that the data is generated from multiple processes or populations each of which has low-dimensional structure. In general, the degrees of freedom or number of parameters of the processes capturing the different populations need not be equal. In this chapter, we address this problem of modeling data arising from a mixture of manifolds of different dimensions for the restricted case where the manifolds are linear subspaces.

The most recent work that offers both estimators and provides guarantees on
estimates for inferring mixtures of subspaces has been limited to equidimensional subspaces (Lerman and Zhang (2010); Page et al. (2013)). A Bayesian procedure for mixtures of subspaces of equal dimensions was developed in Page et al. (2013). A penalized loss based procedure was introduced in Lerman and Zhang (2010) to learn mixtures of $K$-flats. There are significant difficulties in extending either approach to subspaces of different dimensions. The key difficulty in extending either approach is addressing the singularity introduced in moving between subspaces of different dimensions when one parameterizes a subspace as a point on the Grassmann manifold and uses the natural geodesic on this manifold. This difficulty appears in the Bayesian approach as requiring the posterior samples to come from models of different dimensions which will require methods such as reversible jump MCMC which may cause mixing problems. The difficulty is immediate in the penalized loss model as the loss is based on a distance to subspaces and if the dimensions of the subspaces vary the loss based procedure becomes very difficult.

The key idea we develop in this chapter is that subspaces of different dimensions $1, 2, \ldots, m$ can be embedded into a sphere of relatively low dimension $S^{(m-1)(m+2)/2}$ where chordal distances on the sphere can be used to compute distances between subspaces of differing dimensions (Conway et al. (1996)). This embedding removes the discontinuity that occurs in moving between subspaces of different dimensions when one uses the natural metric for a Grassmann manifold. The other tool we make use of is a Gibbs posterior (Jiang and Tanner (2008)) which allows us to efficiently obtain posterior samples of the model parameters.

The structure of the chapter is as follows. In Section 2.1 we state a likelihood model for a mixture of $k$ subspaces each of dimension $d_k$. In Section 2.1.2 we define the embedding procedure we use to model subspaces of different dimensions and specify the model with respect to the likelihood and prior. In Section 2.2 we provide an algorithm for sampling from the posterior distribution. For some of the parameters
standard methods will not be sufficient for efficient sampling and we use a Gibbs posterior for efficient sampling. In Section 2.3 a frequentist analysis of the Bayesian procedure is given that proves posterior consistency of the procedure. In Section 2.4 we use simulated data to provide an empirical analysis of the model and then we use real data to show the utility of the model. We close with a discussion.

Notations

The Grassmann manifold or Grassmannian of $d$-dimensional subspaces in $\mathbb{R}^m$ will be denoted $\text{Gr}(d,m)$. The Stiefel manifold of $m \times d$ matrices with orthonormal columns will be denoted $\text{V}(d,m)$ and when $d = m$ we write $O(d)$ for the orthogonal group. We use boldfaced uppercase letters, e.g., $U$, to denote subspaces and the corresponding letter in normal typeface, e.g., $U$, to denote the matrix whose columns form an orthonormal basis for the respective subspace. Note that $U \in \text{Gr}(d,m)$ and $U \in \text{V}(d,m)$. A subspace has infinitely many different orthonormal bases, related to one another by the equivalence relation $U' = UX$ where $X \in O(d)$. We identify a subspace $U$ with the equivalence class of all its orthonormal bases $\{UX \in \text{V}(m,d) : X \in O(d)\}$ and thereby allowing the identification $\text{Gr}(d,m) = \text{V}(d,m)/O(d)$.

In this chapter, the dimension of the ambient space $m$ will always be fixed but our discussions will often involve multiple copies of Grassmannians $\text{Gr}(d,m)$ with different values of $d$. We will use the term ‘Grassmannian of dimension $d$’ with referring to $\text{Gr}(d,m)$ even though as a manifold, $\dim \text{Gr}(d,m) = d(m-d)$.

2.1 Mixtures of linear subspaces

2.1.1 Likelihood specification

We consider the data $X = (x_1, \ldots, x_n)$ drawn in an independent and identically distributed manner from a mixture of $K$ subspaces where each observation $x_i$ is measured in the ambient space $\mathbb{R}^m$. We assume that each population is concentrated
near a linear subspace $U_k$ which we represent with an orthonormal basis $U_k = \text{span}(U_k)$, $k = 1, \ldots, K$.

We first state the likelihood of a sample conditional on the mixture component. Each mixture component is modeled using a $d_k$-dimensional normal distribution to capture the subspace and a $m - d_k$-dimensional normal distribution to model the residual error or null space:

$$U_k^T x \sim N_{d_k}(\mu_k, \Sigma_k), \quad V_k^T x \sim N_{m-d_k}(V_k^T \theta_k, \sigma_k^2 I),$$

where $U_k$ is the orthonormal basis for the $k$th component and is modeled by a multivariate normal with mean $\mu_k$ and covariance $\Sigma_k$ and $V_k$ is the basis for the null space $\ker(U_k)$ which models the residual error as multivariate normal with variance $\sigma_k^2 I$. We are estimating affine subspaces so the parameter $\theta_k$ serves as a location parameter for the component and by construction $\theta_k \in V_k$. Also note that without loss of generality we can assume that $\Sigma_k$ is diagonal since we may diagonalize the covariance matrix $\Sigma_k = Q_k D_k Q_k^T$ and rotate $U_k$ by $Q_k$ resulting in a parameterization that depends on $U_k$ and a diagonal matrix. Combining the distributions for the null space and subspace gives the observed distribution in ambient space which can be stated as either equivalent distributions

$$x \sim \begin{cases} N_m (U_k \mu_k + \theta_k, U_k \Sigma_k U_k^T + \sigma_k^2 V_k V_k^T) \\ N_m (U_k \mu_k + \theta_k, U_k (\Sigma_k - \sigma^2 I_{d_k}) U_k^T + \sigma_k^2 I_m) \end{cases}. \quad (2.1)$$

It will be convenient for us to use the second parameterization for our likelihood model.

Given the above likelihood model for a component we can specify the following mixture model

$$x \sim \sum_{k=1}^K w_k N_m (U_k \mu_k + \theta_k, U_k (\Sigma_k - \sigma^2 I_{d_k}) U_k^T + \sigma_k^2 I_m), \quad (2.2)$$

8
where \( w = (w_1, \ldots, w_K) \) is a probability vector and we assume \( K \) components. We will use a latent or auxiliary variable approach to sample from the above mixture model and specify a \( K \)-dimensional vector \( z \) with a single entry of 1 and all other entries of zero, \( \delta \sim \text{Mult}(1, w) \). The conditional probability of \( x \) given the latent variable is

\[
x \mid \delta \sim \sum_{k=1}^{K} \delta_k \mathcal{N}_m \left( U_k \mu_k + \theta_k, U_k (\Sigma_k - \sigma^2 I_{d_k}) U_k^T + \sigma^2_k I_m \right).
\]

2.1.2 Prior specification and the spherical embedding

The parameters in the likelihood are for each component \( (\theta_k, \Sigma_k, \sigma_k^2, U_k, \mu_k, d_k) \) and the mixture weights \( w \). Again we fix the number of mixtures as \( K \). We will need to specify priors for each of these parameters. For some of these parameters straightforward conjugate priors exist: the location parameter \( \theta_k \) is normal, the variance terms \( \Sigma_k \) and \( \sigma_k^2 \) are Gamma, and the mixture weights are Dirichlet. A prior distribution for each triple \( (U_k, \mu_k, d_k) \) is less obvious.

The inherent difficulty in sampling the triple is that we do not want to fix the dimension of the subspace \( d_k \), we want to consider \( d_k \) as random. We can state the following joint prior on the triple \( (U_k, \mu_k, d_k) \)

\[
\pi(U_k, \mu_k, d_k) = \pi(U_k \mid d_k) \pi(\mu_k \mid d_k) \pi(d_k).
\]

Given \( d_k \) we can specify \( \mu_k \mid d_k \) as a multivariate normal of dimension \( d_k \). Given \( d_k \) we can also specify a conjugate distribution for \( U_k \) as the von Mises–Fisher (MF) distribution

\[
\text{MF}(U_k \mid A) \sim \text{etr}(A^T U_k),
\]

where \( \text{etr} \) is the exponential trace operator. The matrix von Mises–Fisher distribution is a spherical distribution over the set of all \( m \times d_k \) matrices, also known as the Stiefel manifold which we denote as \( \text{V}(d_k, m) \). A prior on \( d_k \) would take values over...
[0, \ldots, m] and for each value the conditional distributions $\pi(U_k \mid d_k)$ and $\pi(\mu_k \mid d_k)$ need to be specified. For $\mu_k$ a prior distribution of $\mathcal{N}_{d_k}(0, \lambda I)$ seems reasonable since we can assume the mean is zero and the entries independent for any $d_k$. Specifying the conditional distribution for $\pi(U_k \mid d_k)$ is not as clear. As $d_k$ changes the dimension of the matrix $A$ needs to change and one cannot simply add columns of zeroes since columns need to be orthonormal. In addition we would like the priors on $U_k$ to be as close as possible as we change dimension $d_k$ to avoid model fitting inconsistencies. This raises the key difficulty in prior specification over subspaces of different dimension: how to measure the distance between subspaces of different dimensions. Note that we cannot simply integrate out $d_k$ or $U_k$ as nuisance parameters since we have no prior specification.

We will use the geometry of the subspace $U_k$ to specify an appropriate joint prior on $(U_k, d_k)$. Recall that the set of all $d_k$-dimensional linear subspaces in $\mathbb{R}^m$ is the Grassmann manifold $\text{Gr}(d_k, m)$ and that we represent a subspace $U_k \in \text{Gr}(d_k, m)$ with an orthonormal matrix $U_k \in \text{V}(d_k, m)$ from an equivalence class $\{U_k \in \text{V}(k, m) : \text{span}(U_k) = U_k\}$. We need to place priors on Grassmanians of different dimension $d_k$. The key tool we use to specify such a prior is the embedding of $\text{Gr}(d_k, m)$ into $S^{(m-1)(m+2)/2}$, an appropriately chosen sphere\(^1\) in $\mathbb{R}^{m(m+1)/2}$, as proposed in Conway et al. (1996). This embedding allows us to embed subspaces of different dimension into the same space and measure distances between the embedded subspaces as a function of only the ambient (embedded) space. We will use this embedding to place a prior on $U_k$ which implicitly specifies a prior on $d_k$. This embedding will have some very nice properties in terms of prior specification and computation.

The following theorem states that embedding the Grassmannian into a sphere allows us to measure distances between subspaces.

\(^1\) Note that the dimension of a sphere in $\mathbb{R}^d$ is $d - 1$ and that $m(m+1)/2 - 1 = (m - 1)(m + 2)/2$. 

10
**Theorem 1** (Conway–Hardin–Sloane 1996). The representation of a subspace \( U \in \text{Gr}(d, m) \) by its projection matrix \( P_U \) gives an isometric embedding of \( \text{Gr}(d, m) \) into a sphere of radius \( \sqrt{d(m-d)/m} \) in \( \mathbb{R}^{m(m+1)/2} \), with \( d_p(U, V) = \frac{1}{\sqrt{2}} \| P_U - P_V \|_F \), where \( P_V \) is the projection matrix onto \( V \).

The embedding procedure proceeds in the following steps: (1) given a basis \( U_k \) compute the projection matrix \( P_k = U_k^T U_k \), (2) take all the entries of \( P_k \) in the upper triangle (or lower triangle) as well as all the elements in the diagonal except for one as a vector in \( \mathbb{R}^{m(m+1)/2-1} \). The sum of all the entries on the vector will be a constant, this is a result of the orthogonality of \( U_k \), which means that all the subspaces of dimension \( k \) lie on the same sphere. The key observation by Conway et al. (1996) was that if the extra coordinate is included, thus embedding into \( \mathbb{R}^{m(m+1)/2} \), the subspaces are still embedded into spheres and each of these spheres are cross sections of a higher-dimensional sphere which we denote as \( S^{(m-1)(m+2)/2} \). The sphere \( S^{(m-1)(m+2)/2} \) is centered at \( \varphi \left( \frac{1}{2} I_m \right) = \text{vech} \left( \frac{1}{2} I_m \right) \) where \( \varphi(A) \) denotes the embedding of the projection matrix \( A \) and \( \text{vech} \) is the half-vectorization operation

\[
\text{vech} \left( \begin{bmatrix} a & b \\ b & d \end{bmatrix} \right) = \begin{bmatrix} a \\ b \\ d \end{bmatrix}.
\]

The 0-dimensional subspace is embedded at the origin \( 0 \in \mathbb{R}^{m(m+1)/2} \). The radius of \( S^{(m-1)(m+2)/2} \) is \( \sqrt{m}/2 \). In summary,

\[ S^{(m-1)(m+2)/2} = \{ x \in \mathbb{R}^{m(m+1)/2} : \| x - c \|^2 = \sqrt{m}/2 \}, \text{ where } c = \text{vech} \left( \frac{1}{2} I_m \right). \]

Grassmann manifolds are embedded into cross-sections of \( S^{(m-1)(m+2)/2} \) where the projection matrix corresponding to the pre-image has an integer valued trace. The geodesic distance along the surface of the sphere, \( d_{S^{(m-1)(m+2)/2}} \), corresponds to the
projective distance $d_p(\cdot, \cdot)$ between two subspaces $U_1, U_2 \in \text{Gr}(d, m)$,

$$d_p(U_1, U_2) = \left[ \sum_{j=1}^{d} \sin^2(\theta_j) \right]^{1/2},$$

where $\theta_1, \ldots, \theta_d$ are the principal angles between the subspaces. We illustrate the embedding for two projection matrices in Figure 2.1.

**Figure 2.1**: An illustration of the spherical embedding for subspaces from $\text{Gr}(1, 2)$ into $\mathbb{R}^2$ on the left, and of $\text{Gr}(k, 2)$, $k = 0, 1, 2$ on the right. Images of the embedding are in black, and the center of the sphere is in red. The coordinates for the embedding into $\mathbb{R}^2$ is the first column of the projection matrix. By including the last entry in the diagonal of the projection matrix, we obtain coordinates for the embedding into $\mathbb{R}^3$. $\text{Gr}(0, 2)$ and $\text{Gr}(2, 2)$ are trivial sets giving $O_3$ and $I_3$ as their projection matrices. They act as poles on the sphere with coordinates $(0, 0, 0)$ and $(1, 0, 1)$.

The representation of Grassmannians as points on $S^{(m-1)(m+2)/2}$ has several useful properties.

**Sphere interpretation** The sphere $S^{(m-1)(m+2)/2}$ provides an intuitive way to sample subspaces of different dimensions by sampling from $S^{(m-1)(m+2)/2}$. Under the projective distance, the sphere also has an intuitive structure. For example, distances between subspaces of different dimensions can also be computed as the distance between points on the sphere, these points will be on different cross-sections. Under the projective distance, the orthogonal complement of
a subspace $U$ is the point on $\mathbb{S}^{(m-1)(m+2)/2}$ that maximizes the projective distance. Further, the projection matrix is always invariant to the representation $U$.

**Differentiable** The projective distance however is square differentiable everywhere, making it more suitable in general for optimization problems. This is not the case for distances like geodesic distance or the Asimov–Golub–Van Loan distance where maximizing the distance between a set of subspaces will result in distances that lie near non-differentiable points Conway et al. (1996). This numerical instability can lead to sub-optimal solutions.

**Ease of computation** The projective distance is easy to compute via principal angles, which are in turn readily computable with singular value decomposition (Golub and Loan (2013)). Working with the embedding requires only a relatively small number of coordinates — in fact only quadratic in $m$ or $m(m+1)/2$. Furthermore one can exploit many properties of a sphere in Euclidean space in our computations. For example sampling from a sphere is simple. The number of required coordinates is small compared to alternative embeddings of the Grassmannian, see Hamm and Lee (2008). In contrast the usual Plücker embedding requires a number of coordinates that is $\binom{m}{d}$, i.e., exponential in $m$. Moreover the Plücker embedding does not reveal a clear relationship between Grassmannians of different dimensions, as there is using the spherical embedding.

We will place a prior on projection matrices by placing a distribution over the lower half of $\mathbb{S}^{(m-1)(m+2)/2}$, points on $\mathbb{S}^{(m-1)(m+2)/2}$ corresponding to cross-sections where the subspace corresponding to the pre-image has dimension $d < m(m + 1)/4$. We only consider the lower half since we assume the model to be low-dimensional. The prior over projection matrices imples a prior over $U_k$ and $d_k$. A point drawn from
$\mathbb{S}^{(m-1)(m+2)/2}$ may not correspond to a subspace, recall only points with integer trace have subspaces as a pre-image. We address this problem by the following procedure: given a sampled point $q \in \mathbb{S}^{(m-1)(m+2)/2}$ we return the closest point $p \in \mathbb{S}^{(m-1)(m+2)/2}$ that is the pre-image of a subspace. The following theorem states the procedure.

**Theorem 2.** Given a point $q \in \mathbb{S}^\ell$, the point $p$ that minimizes the geodesic distance on $\mathbb{S}^\ell$, $d_{\mathbb{S}^\ell}(q,p)$, subject to

$$\varphi^{-1}(p) \in \bigcup_{d=0}^\ell \operatorname{Gr}(d,\ell)$$

can be found by the following procedure

(i) Compute $Q = \varphi^{-1}(q)$.

(ii) Set the dimension of $p$ to $d = \text{tr}(Q)$.

(iii) Compute the eigendecomposition $Q = AA^{-1}$.

(iv) Set $B$ an $\ell \times d$ matrix equal to the columns of $A$ corresponding to the top $d$ eigenvalues.

(v) Let $p = \varphi(BB^T)$.

**Proof.** In the case where the point $q \in \mathbb{S}^\ell$ is already on a cross-section of the sphere corresponding to $\operatorname{Gr}(d,\ell)$, the eigendecomposition will return exactly $d$ non-zero eigenvalues. The eigenvectors give a basis for the subspace that is embedded into the point $q$. Similarly when the point $q$ is between cross sections corresponding to Grassmannians, the above algorithm minimizes the Euclidean distance between the point $p$ and $q$, and therefore minimizes the distance on $\mathbb{S}^\ell$. \qed
The full model is specified as follows for each $x_i$, $i = 1, \ldots, n$,

$$ w \sim \text{Dir}_K(\alpha), $$

$$ \delta_i \sim \text{Mult}(w), $$

$$ P_k \sim \mathcal{P}(S^{(m-1)(m+2)/2}), \quad U_k U_k^T = P_k, \quad d_k = \text{tr}(P_k), \quad (2.3) $$

$$ \mu_k \mid d_k \sim \mathcal{N}_{d_k}(0, \lambda I), $$

$$ \theta_k \mid U_k \sim \mathcal{N}_m(0, \phi I), \quad U_k^T \theta_k = 0, \quad (2.4) $$

$$ \sigma_k^{-2} \sim \text{Ga}(a, b), $$

$$ \Sigma_{k(j)}^{-1} \mid d_k \sim \text{Ga}(c, d), \quad j = 1, \ldots, d_k, $$

$$ x_i \mid \delta_i \sim \sum_{k=1}^K \delta_{ik} \mathcal{N}_m\left(U_k \mu_k + \theta_k, U_k (\Sigma_k - \sigma_k^2 I_{d_k}) U_k^T + \sigma_k^2 I_m\right), $$

where equation (2.3) corresponds to sampling from a distribution $\mathcal{P}$ supported on the lower half of the sphere $S^{(m-1)(m+2)/2}$ a projection matrix $P_k$ that corresponds to a subspace and computing the dimension $d_k$ as the trace of the subspace and computing the subspace $U_k$ from the projection and equation (2.4) corresponds to sampling from a normal distribution subject to the projection constraint $U_k^T \theta_k = 0$.

### 2.2 Posterior sampling

In this section we provide an efficient algorithm for sampling the model parameters from the posterior distribution. Sampling directly from a joint posterior distribution of all the parameters is intractable and we will use Markov chain Monte Carlo methods for sampling. For most of the parameters we can sample from the posterior using a Gibbs sampler. This is not the case for sampling from the posterior distribution over projection matrices with prior $\mathcal{P}$ on the sphere $S^{(m-1)(m+2)/2}$. The prior $\mathcal{P}$ should place more mass on cross-sections of the sphere corresponding to lower dimensions $d_k$. Sampling efficiently from a joint distribution on $d_k, P_k$ is difficult. We will address this problem by using a Gibbs posterior (Jiang and Tanner (2008)) to
sample the projection matrices. We first state the Gibbs posterior we use to sample $U_k$ and $\theta_k$ efficiently and the rationale for this form of the posterior. We then close with the sampling algorithm for all the model parameters.

It is not obvious how to place a prior on the sphere $S^{(m-1)(m+2)/2}$ that will allow for efficiently sampling. We can however follow the idea of a Gibbs posterior to design an efficient sampler. The idea behind a Gibbs posterior is to replace the standard posterior which takes the form of

$$\frac{\text{posterior}}{\text{prior}} \times \text{likelihood}$$

with a distribution based on a loss or risk function that depends on both the data as well the parameter of interest in our case the loss function is given by

$$L(P_{1:[k]}, \theta_{1:[k]}, X) = \frac{1}{n} \sum_{i=1}^{n} \left[ \min_{k=1,\ldots,K} \left( \| P_k(x_i - \theta_k) - (x_i - \theta_k) \|^2 + \text{tr}(P_k) \right) \right], \quad (2.5)$$

where $e_{ik}$ is the residual error for the $i$th sample given by the $k$-th subspace with the error defined by our likelihood model. The above loss function corresponds to computing for each sample the residual error to the closest subspace weighted by the dimension of the subspace. The penalty weighting the dimension of the subspace enforces a prior that puts more mass on subspaces of lower dimension. Given the likelihood or loss function we state the following Gibbs posterior

$$g(P_{1:[k]}, \theta_{1:[k]} \mid X) \propto \exp(-n\psi L(P_{1:[k]}, \theta_{1:[k]}, X)) \pi(P_{1:[k]}) \pi(\theta_{1:[k]}), \quad (2.6)$$

where $\psi$ is a chosen temperature parameter. A Gibbs posterior is simply a loss oriented alternative to the likelihood based posterior distribution. Traditionally it is used to account for model misspecification. Here the Gibbs posterior is used to avoid overfitting by arbitrarily increasing the dimension of the subspace and for computational efficiency in sampling.
2.2.1 Sampling $U_{[1:K]}$ and $\theta_{[1:K]}$ from the Gibbs posterior

In this subsection we outline our procedure for sampling from the model parameters $U_{[1:K]}$ and $\theta_{[1:K]}$ using a Metropolis–Hastings algorithm which is effectively a random walk on the sphere. We first state a few facts that we will use. First recall that there is a deterministic relation between $U_k$ and $P_k$, so given a $P_k$ we can compute $U_k$. Also recall that a point sampled from $S^m$ is not the pre-image of a subspace. Given a point $s_k^0 \in S^m$ we denote the subspace corresponding to this point as $P_k = \varphi^{-1}(s_k^0)$, this is the closest projection matrix to $s_k^0$ corresponding to a subspace. The procedure to compute $P_k$ from $s_k^0$ is given in Theorem 2. We obtain $U_k$ correspond to the top $d_k$ eigenvectors of $P_k$ where $d_k$ is the trace of $P_k$.

We now state two procedures. The first procedure initializes the parameters $U_{[1:K]}$ and $\theta_{[1:K]}$. The second procedure computes the $\ell$-th sample of the parameters.

The first procedure which we denote as Initialize($U_{[1:K]}, \theta_{[1:K]}$) proceeds as follows:

1. Draw $\sigma \sim \mathcal{G}_K$, the symmetric group of permutations on $K$ elements.

2. For $i = 1, \ldots, K$,

   (a) draw $z_{\sigma(i)}^0 \sim \mathcal{N}_{m(m+1)/2}(0, \tau I)$;
   
   (b) compute $s_{\sigma(i)}^0 = (\sqrt{m(m+1)/8})z_{\sigma(i)}^0/\|z_{\sigma(i)}^0\| + \varphi(I_m)$;
   
   (c) compute $P_{\sigma(i)}^0 = \varphi^{-1}(s_{\sigma(i)}^0)$;
   
   (d) compute $d_{\sigma(i)}^0 = \text{tr}(P_{\sigma(i)}^0)$;
   
   (e) compute $U_{\sigma(i)}^0$ as the top $d_{\sigma(i)}^0$ eigenvectors of $P_{\sigma(i)}^0$;
   
   (f) draw $\beta_{\sigma(i)}^0 \sim \mathcal{N}(0, I_m)$;
   
   (g) compute $\theta_{\sigma(i)}^0 = (I_m - P_{\sigma(i)}^0)\beta_{\sigma(i)}^0$.  

17
The first step permutes the order we initialize the $K$ components. Step (a) samples a point from a multivariate normal with the dimension of the sphere. In Step (b) we normalize the sampled point, recenter it, and embed it onto the sphere $S^{m(m+1)/2}$. In Step (c) we compute the projection matrix by computing the closest subspace to the embedded point computed in Step (b). Given the projection matrix we compute the dimension in Step (d) and the basis of the subspace in Step (e). Steps (e) and (f) we compute the $\theta$ parameters.

The second procedure which we denote as $\textbf{Update}(U^{(\ell)}_{[1:K]}, \theta^{(\ell)}_{[1:K]})$ computes the $\ell$-th sample as follows:

1. Draw $\sigma \sim \mathcal{S}_K$, the symmetric group of permutations on $K$ elements.

2. For $i = 1, \ldots, K$,
   
   (a) draw $z_{\sigma(i)} \sim \mathcal{N}_{m(m+1)/2}(z^{(\ell-1)}_{\sigma(i)}, \tau I)$;
   
   (b) compute $s_{\sigma(i)} = (\sqrt{m(m+1)/8})z_{\sigma(i)}/\|z_{\sigma(i)}\| + \varphi(I_m)$;
   
   (c) compute $P_{\sigma(i)} = \varphi^{-1}(s_{\sigma(i)})$;
   
   (d) compute $d_{\sigma(i)} = \text{tr}(P_{\sigma(i)})$;
   
   (e) compute $U_{\sigma(i)}$ as the top $d_{\sigma(i)}$ eigenvectors of $P_{\sigma(i)}$;
   
   (f) draw $u \sim \text{Unif}[0, 1]$;
   
   (g) set
   $$P_{[1:K]} = \left[ P^{(\ell-1)}_{[1:K]-\sigma(i)}, P_{\sigma(i)} \right];$$
   
   (h) set
   $$\theta_{[1:K]} = \left[ \theta^{(\ell-1)}_{[1:K]-\sigma(i)}, (I_m - U_{\sigma(i)}U_{\sigma(i)}^T)\theta^{(\ell-1)}_{\sigma(i)} \right];$$
   
   (i) compute the acceptance probability
   $$\alpha = \frac{\exp(-n\psi L(P_{[1:K]}, \theta_{[1:K]}; X))}{\exp(-n\psi L(P^{(\ell-1)}_{[1:K]}, \theta^{(\ell-1)}_{[1:K]}; X))};$$

18
(j) set
\[
(U^{(\ell)}_{\sigma(i)}; z^{(\ell)}_{\sigma(i)}) = \begin{cases} (U^{(\ell-1)}_{\sigma(i)}; z^{(\ell-1)}_{\sigma(i)}) & \text{if } \alpha > u, \\ (U^{(\ell)}_{\sigma(i)}; z^{(\ell)}_{\sigma(i)}) & \text{otherwise}; \end{cases}
\]

(k) draw $\beta_{\sigma(i)} \sim N_m(\beta^{(\ell-1)}_{\sigma(i)}, I_m)$;

(l) compute $\theta_{\sigma(i)} = (I_m - P^{(\ell-1)}_{\sigma(i)})\beta_{\sigma(i)}$;

(m) draw $u \sim \text{Unif}[0, 1]$;

(n) set
\[
\theta_{[1:K]} = [\theta^{(\ell-1)}_{[1:K]-\sigma(i)}, \theta^{(\ell)}_{\sigma(i)}];
\]

(o) compute the acceptance probability
\[
\alpha = \frac{\exp\left(-n\psi L(P^{(\ell-1)}_{[1:K]}, \theta^{(\ell)}_{[1:K]}, X)\right)}{\exp\left(-n\psi L(P^{(\ell-1)}_{[1:K]}, \theta^{(\ell-1)}_{[1:K]}, X)\right)};
\]

(p) set
\[
(\theta^{(\ell)}_{\sigma(i)}; \beta^{(\ell)}_{\sigma(i)}) = \begin{cases} (\theta^{(\ell-1)}_{\sigma(i)}; \beta^{(\ell-1)}_{\sigma(i)}) & \text{if } \alpha > u, \\ (\theta^{(\ell)}_{\sigma(i)}; \beta^{(\ell-1)}_{\sigma(i)}) & \text{otherwise}. \end{cases}
\]

Many steps of this procedure are the same as the first procedure with the following exceptions. In Steps (a) and (k) we are centering the random walk to the previous values of $z_{\sigma(i)}$ and $\beta_{\sigma(i)}$ respectively. Step (g) updates the set of $K$ projection matrices by replacing the $i$-th projection matrix in the set with the proposed new matrix. Step (h) is analogous to Step (g) but for the set of $\theta$ vectors. In Step (j) we update the subspace and in Step (p) we update the $\theta$ vector.

2.2.2 Sampling algorithm

We now state the algorithm we use to sample from the posterior. To simplify notation we work with precision matrices $J_k = \Sigma^{-1}_k$ instead of the inverse of covariance
matrices for each mixture component. Similarly, we work with the precision of the
$k$-th component $\gamma_k$ instead of the inverse of the variance, $\gamma_k = \sigma_k^{-2}$.

The following procedure provides posterior samples:

1. Draw $U^{(0)}_{[1,K]}, \theta^{(0)}_{[1,K]}, d^{(0)}_{[1,K]} \sim \text{Initialize}(U_{[1,K]}, \theta_{[1,K]})$.

2. Draw $J_{k(j_k)} \sim \text{Ga}(a, b)$ for $k = 1, \ldots, K$ and $j_k = 1, \ldots, d^{(0)}_k$.

3. For $t = 1, \ldots, T$,
   (a) for $i = 1, \ldots, n$ and $k = 1, \ldots, K$, compute
   \[ e_{ik} = \| P_k(t-1) (x_i - \theta_k^{(t-1)}) - (x_i - \theta_k^{(t-1)}) \|^2; \]
   (b) for $i = 1, \ldots, n$, set
   \[ w_i = \left( \frac{\exp(-\kappa r_{i1})}{\sum_{j'=1}^K \exp(-\kappa r_{ij'})}, \ldots, \frac{\exp(-\kappa r_{iK})}{\sum_{j'=1}^K \exp(-\kappa r_{ij'})} \right); \]
   (c) for $i = 1, \ldots, n$, draw $\delta_i \sim \text{Mult}(w_i)$;
   (d) update for $k = 1, \ldots, K$ each $\mu_k^{(t)} \sim \mathcal{N}(m_k^*, S_k^*)$ where
   \[ S_k^* = \left( n_k J_k^{(t-1)} + \lambda^{-1} I \right)^{-1}, \quad m_k^* = S_k^* \left( U_k^{(t-1)^T} J_k^{(t-1)} \sum_{\delta_i = k} x_i \right), \]
   and $n_k = \# \{ i : \delta_i = k \}$;
   (e) update for $k = 1, \ldots, K$, and each $\gamma_k^{(t)} \sim \text{Ga}(a_k^*, b_k^*)$,
   \[ a_k^* = n_k (m - d_k) + a, \]
   \[ b_k^* = b + \frac{n_k}{2} \left( \theta_k^{(t-1)^T} \theta_k^{(t-1)} \right) + \sum_{\delta_i = k} \left( \frac{1}{2} x_i^T x_i - x_i^T U_k^{(t-1)^T} x_i \right) - \theta_k^{(t-1)^T} \sum_{\delta_i = k} x_i; \]
   (f) update for $k = 1, \ldots, K$, and $j_k = 1, \ldots, d^{(t)}_k$,
   \[ J_{k(j_k)}^{(t)} \sim \text{Ga}\left( \frac{n_k}{2} + a, b + \frac{1}{2} \sum_{\delta_i = k} \left( U_k^{(t-1)^T} x_i - \mu_k \right)^2 \right), \]
   where $(u)_j$ denotes the $j$th element of the vector $u;$
(g) draw

\[ U_{[1:K]}^{(t)}; \theta_{[1:K]}^{(t)}, q_{[1:K]}^{(t)} \sim \text{Update}(U_{[1:K]}^{(t-1)}, \theta_{[1:K]}^{(t-1)}) \]

The update steps for \( \mu, \sigma^2, \Sigma \) are (d), (e), (f) respectively and are given by the conditional probabilities given all other variables. Steps (a), (b), and (c) assign the latent membership variable to each observation based on the distance to the \( K \) subspaces. We set the parameter \( \kappa \) very large which effective assigns membership of each \( x_i \) to the subspace with least residual error.

When drawing from the Gibb’s posterior distribution via a Metropolis–Hastings algorithm, the proposal distribution and temperature are adjusted through a burn-in period. In the first stage of burn-in, the proposal variance parameter \( \tau = 1 \) is fixed, while temperature is selected by a decreasing line search on a log-scale grid, from \( 10^{-20} \) to \( 10^{20} \) until the acceptance ratio reaches the 10%–90% range. With temperature fixed, the proposal variance \( \tau \) is adjusted until the acceptance ratio falls in the 25%–45% range during the burn-in period. Thinning was applied in that every third draw of the sampler was kept, this was determined from autocorrelation analysis.

2.3 Posterior consistency

In this section, an asymptotic analysis of our model is given which provides some theoretical guarantees for our estimation procedure. There is extensive literature on posterior consistency of Bayesian models (Ghoshal (2010)). However, extending standard results to our specific model is non-trivial. Let \( \mathcal{M} \) be the space of all the densities in \( \mathbb{R}^m \) and \( f_0 \) be the true data generating density. We first define some notion of distances and neighborhoods in \( \mathcal{M} \). A weak neighborhood of \( f_0 \) with radius
\( \epsilon \) is defined as

\[
W_\epsilon(f_0) = \left\{ f : \left| \int g f \, dx - \int g f_0 \, dx \right| \leq \epsilon \text{ for all } g \in C_0(\mathbb{R}^m) \right\},
\]

where \( C_0(\mathbb{R}^m) \) is the space of all continuous and bounded functions on \( \mathbb{R}^m \). The Hellinger distance \( d_H(f, f_0) \) is defined as

\[
d_H(f, f_0) = \left( \frac{1}{2} \int \left[ \sqrt{f(x)} - \sqrt{f_0(x)} \right]^2 \, dx \right)^{1/2}.
\]

Denote \( U_\epsilon(f_0) \) an \( \epsilon \)-Hellinger neighborhood around \( f_0 \) with respect to \( d_H \). The Kullback–Leibler (KL) divergence between \( f_0 \) and \( f \) is defined to be

\[
d_{KL}(f_0, f) = \int f_0(x) \log \frac{f_0(x)}{f(x)} \, dx,
\]

with \( K_\epsilon(f_0) \) denoting an \( \epsilon \)-KL neighborhood of \( f_0 \).

Let \( \Pi_s \) be a prior on the sphere \( S^{(m-1)(m+2)/2} \) which can be taken to be the uniform distribution or the von Mises–Fisher distribution. By projecting the samples from \( \Pi_s \) onto the cross-sections of the sphere, \( \Pi_s \) induces a prior distribution on the subspaces basis \( U \) which we denote by \( \Pi_U \).

Note that our model induces a prior \( \Pi \) on \( M \). Assume the true density \( f_0 \) follows the following the regularity conditions, i.e.,

(i) \( f_0(x) \) is bounded away from zero and bounded above by some constant \( M \) for all \( x \in \mathbb{R}^m \);

(ii) \( |\int \log(f_0(x)) f_0(x) \, dx| < \infty \);

(iii) for some \( \delta > 0, \int \left[ \log \left( \frac{f_0(x)}{f_\delta(x)} \right) \right] f_\delta(x) \, dx < \infty \), where \( f_\delta(x) = \inf_{y:|y-x|<\delta} f_0(y) \);

(iv) there exists \( \alpha > 0 \) such that \( \int |x|^{2(1+\alpha)m} f_0(x) \, dx < \infty \).
We will show that the posterior distribution $\Pi(\cdot \mid x_1, \ldots, x_n)$ concentrates around any true density $f_0$ as $n \to \infty$. The following theorem is on weak consistency.

**Theorem 3.** The posterior distribution $\Pi(\cdot \mid x_1, \ldots, x_n)$ is weakly consistent. That is, for all $\epsilon > 0$,

$$\Pi(W_\epsilon(f_0) \mid x_1, \ldots, x_n) \to 1 \quad P_{f_0}^\infty{\text{-almost surely as } n \to \infty},$$

where $W_\epsilon(f_0)$ is a weak neighborhood of $f_0$ with radius $\epsilon$ and $P_{f_0}^\infty$ represents the true probability measure for $(x_1, x_2, \ldots)$.

For a proof, see appendix A. In proving the following strong consistency theorem, we assume that the parameters $\sigma_i^2, \sigma^2$ and the diagonal elements of $\Sigma_i$ ($i = 1, \ldots, K$) follow i.i.d. truncated Gamma priors supported on some bounded interval $[0, M]$ for some large enough constant $M$.

**Theorem 4.** The posterior distribution $\Pi(\cdot \mid x_1, \ldots, x_n)$ is strongly consistent. That is, for all $\epsilon > 0$,

$$\Pi(U_\epsilon(f_0) \mid x_1, \ldots, x_n) \to 1 \quad P_{f_0}^\infty{\text{-almost surely as } n \to \infty},$$

where $U_\epsilon(f_0)$ is a neighborhood of $f_0$ with radius $\epsilon$ with respect to the Hellinger distance.

For a proof, see appendix A.

2.4 Results on real and simulated data

We use two data analysis examples to illustrate the utility of our model. In the first example, we generate synthetic data with simple geometric structure to contrast the performance of our method with two classic clustering algorithms, $k$-means clustering and a Gaussian mixture model. In the second example, we demonstrate superior
classification accuracy as compared to supervised classification models and a Gaussian mixture model on three data sets from the UCI machine learning repository (K. Bache (2013)).

2.4.1 Simulated data

A simple example of a stratified space that is a mixture of subspaces is a line puncturing a plane. We will use this geometric example to illustrate our model. We compare three methods the mixture of subspaces model, a mixture of normals, and \(K\)-means clustering.

The mixture model for a line intersecting a plane in \(\mathbb{R}^3\) comprises two components: subspace \(U_1\) corresponding to a line and subspace \(U_2\) corresponding to the plane. Although simple, this example can be a challenging situation to infer. To contrast the mixture of subspace model with a mixture of normals we consider the line and plane with different levels of thickening in the ambient space. The idea is that the performance of the mixture of Gaussians should catch up to the performance of the mixture of subspaces as the subspaces become thicker. The data is specified by the following distribution with the following five values for the precision parameter of the isotropic noise around the subspaces, \(\nu = [10, 5, 1, 0.5, 0.2]\):

\[
\text{LINE} \\
U_1 \sim \text{Unif}(V(1, 3)), \\
\mu_1 \sim \mathcal{N}(0, I), \\
\Sigma_1^{-1} \sim \text{TGa}(1, 1, \nu), \\
(I - U_1U_1^T)^{-1}\theta_1 \sim \mathcal{N}(0, I),
\]

\[
\text{PLANE} \\
U_2 \sim \text{Unif}(V(2, 3)), \\
\mu_2 \sim \mathcal{N}(0, I), \\
\text{diag}(\Sigma_2^{-1}) \overset{iid}{\sim} \text{TGa}(1, 1, \nu), \\
(I_3 - U_2U_2^T)^{-1}\theta_2 \sim \mathcal{N}(0, I),
\]

where \(\text{TGa}(1, 1, \nu)\) is a left truncated Gamma truncated at precision \(\nu\). Given these parameters for the two mixture components we specify the following two conditional
distributions

\[ x \mid \text{Line} \sim_{\text{id}} \mathcal{N}_3(U_1\mu_1 + \theta_1, U_1(\Sigma_1 - \sigma_1^2)U_1^T + \sigma_1^2 I), \]

\[ x \mid \text{Plane} \sim_{\text{id}} \mathcal{N}_3(U_2\mu_2 + \theta_2, U_2(\Sigma_2 - \sigma_2^2)U_2^T + \sigma_2^2 I). \]

We generated 500 observations from both the line and the plane, see Figure 2.2. For each of the five variance levels, ten data sets were generated, and a holdout set of 50 observations from the line and plane.

In Table 2.1 the range of assignment accuracy is reported for each method on the holdout set of the ten runs. Note that \( K \)-means performs poorly and as the precision parameter decreases the Gaussian mixture model catches up to the subspace model. For the subspace model we set the temperature parameter for the Gibbs posterior was set to \( 10^{-6} \), and acceptance rates between 38\% and 48\% were achieved for the subspace and affine mean parameters.

Table 2.1: Range of cluster assignment accuracy for the simulated data for the three algorithms: \( K \)-means, mixture of subspaces, and mixture of normals.

<table>
<thead>
<tr>
<th>Precision</th>
<th>( K )-means</th>
<th>Subspace</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>(0.66, 0.77)</td>
<td>(0.95, 0.99)</td>
<td>(0.89, 0.97)</td>
</tr>
<tr>
<td>0.5</td>
<td>(0.64, 0.82)</td>
<td>(0.90, 0.98)</td>
<td>(0.87, 0.98)</td>
</tr>
<tr>
<td>1</td>
<td>(0.57, 0.70)</td>
<td>(0.87, 0.98)</td>
<td>(0.85, 0.98)</td>
</tr>
<tr>
<td>2</td>
<td>(0.64, 0.72)</td>
<td>(0.87, 0.96)</td>
<td>(0.87, 0.95)</td>
</tr>
<tr>
<td>5</td>
<td>(0.59, 0.80)</td>
<td>(0.84, 0.97)</td>
<td>(0.84, 0.97)</td>
</tr>
</tbody>
</table>

2.4.2 Real data

The utility of the model on real data was examined on three data sets from the UCI Data Repository: the Statlog Vehicle Silhouettes data (Siebert (1987)), the Wisconsin Breast Cancer data (Mangasarian and Wolberg (1990)), and the Statlog Heart data (Detrano et al. (1989)). The statistical problem underlying all these data...
is classification, for example in the Wisconsin Breast Cancer data the objective is to classify a tumor as benign or malignant based on ten features of the biopsy taken from the tumor. We compared three classification procedures: logit and multinomial logit models, our mixture of subspaces, and a Gaussian mixture model. Our model and the Gaussian mixture model do not use the class labels in inferring a classifier. A point of interest is that our model does as well or better than the logit models which use both the response variable and the covariates to fit the model.

For our model we set the temperature parameter in the Gibbs posterior by tuning it during the burn-in period to obtain an acceptance ratio in the range of 20–40%. An alternative idea is to use cross-validation to set this parameter. This would however be computationally very costly. To compute predictive accuracy we use the MAP estimate from our MCMC to classify a new point.
Predicting on a 10% hold out of the data, the percentage of correctly predicted classifications is given in the following table. The Heart Data Set contains 270 observations of two classes with 13 covariates, the Vehicle Data Set contains 846 observations of four classes on 18 covariates, and the Breast Cancer Data Set contains 569 observations of two classes on 30 covariates. Each of the models predicted ten different hold out samples, and the intervals given are the range of those ten different predictions. See Table 2.2 for the results.

Table 2.2: Range of cluster assignment accuracy for the three data sets using the three classifiers on the left-out-data.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Logit</th>
<th>Model</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>(0.78, 0.86)</td>
<td>(0.89, 0.94)</td>
<td>(0.64, 0.70)</td>
</tr>
<tr>
<td>Heart</td>
<td>(0.72, 0.78)</td>
<td>(0.77, 0.81)</td>
<td>(0.56, 0.60)</td>
</tr>
<tr>
<td>Vehicle</td>
<td>(0.46, 0.59)</td>
<td>(0.77, 0.83)</td>
<td>(0.74, 0.79)</td>
</tr>
</tbody>
</table>

Another benefit of our model is that we can estimate the dimension of the linear subspace that each class is concentrated around. This is not something that can be done using either the mixture models proposed in either the Bayesian or penalized cost methods for mixture of subspaces (Lerman and Zhang (2010); Page et al. (2013)), since the classes are all assumed to have the same dimension. In Table 2.3 we state posterior probabilities for the dimension of each class in our mixture model for each of the three data sets.

2.5 Discussion

We present a method for learning or inferring mixtures of linear subspaces of different dimensions. The key idea in our procedure was using the observation that subspaces of different dimensions can be represented as points on a sphere is very useful for inference. The utility of this representation is that sampling from a sphere
Table 2.3: Posterior probabilities for the dimension of the subspace of clusters in each data set.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>0.45</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0.53</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.66</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.34</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.78</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.12</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.79</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.69</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.31</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.76</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.24</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.78</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.22</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.93</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.07</td>
</tr>
</tbody>
</table>

is straightforward, there exists a distance between subspaces of different dimensions that is differentiable and can be computed using principal angles; we avoid MCMC algorithms that jump between models of different dimensions. We suspect that this idea of embedding or representing models of different dimensions by embedding them into a common space with a distance metric that allows for ease of computation and sampling as well as nice analytic properties may also be of use in other settings besides subspaces.

Scaling our estimation procedure to higher dimensions and more samples will require greater computational efficiency and an EM-algorithm for this model holds promise. It is also of interest to examine if we can replace the Gibbs posterior with an efficient fully Bayesian procedure.
Although the main focus in statistics has been on data belonging to Euclidean spaces, it is common for data to have support on non-Euclidean geometric spaces. Perhaps the simplest example is to directional data, which lie on circles or spheres. Directional statistics dates back to R.A. Fisher’s seminal paper (Fisher, 1953) on analyzing the directions of the earth’s magnetic poles, with key later developments by Watson (1983), Mardia and Jupp (2000), Fisher et al. (1987) among others. Technological advances in science and engineering have lead to the routine collection of more complex geometric data. For example, diffusion tensor imaging (DTI) obtains local information on the directions of neural activity through $3 \times 3$ positive definite matrices at each voxel (Alexander et al., 2007). In machine vision, a digital image can be represented by a set of $k$-landmarks, the collection of which form landmark based shape spaces (Kendall, 1984). In engineering and machine learning, images are often preprocessed or reduced to a collection of subspaces, with each data point (an image) in the sample data represented by a subspace. One may also encounter data that are stored as orthonormal frames (Downs et al., 1971), surfaces, curves, and networks.
Statistical analysis of data sets whose basic elements are geometric objects requires a precise mathematical characterization of the underlying space and inference dependent on the geometry of the space. In many cases (e.g., space of positive definite matrices, spheres, shape spaces, etc), the underlying space corresponds to a manifold. Manifolds are general topological spaces equipped with a differentiable/smooth structure which induces a geometry that does not in general adhere to the usual Euclidean geometry. Therefore, new statistical theory and models have to be developed for statistical inference of manifold-valued data. There have been some developments on inferences based on i.i.d. (independent and identically distributed) observations on a known manifold. Such approaches are mainly based on obtaining statistical estimators for appropriate notions of location and spread on the manifold. For example, one could base inference on the center of a distribution on the Fréchet mean, with the asymptotic distribution of sample estimates obtained (Bhattacharya and Patrangenaru, 2003, 2005; Bhattacharya and Lin, 2013). There has also been some consideration of nonparametric density estimation on manifolds (Bhattacharya and Dunson, 2010; Lin et al., 2013; Pelletier, 2005). Bhattacharya and Bhattacharya (2012) provides a recent overview of such developments.

There has also been a growing interest in modeling the relationship between a manifold-valued response $Y$ and Euclidean predictors $X$. For example, many studies are devoted to investigating how brain shape changes with age, demographic factors, IQ and other variables. It is essential to take into account the underlying geometry of the manifold for proper inference. Approaches that ignore the geometry of the data can potentially lead to highly misleading predictions and inferences. Some geometric approaches have been developed in the literature. For example, Fletcher (2011) develops a geodesic regression model on Riemannian manifolds, which can be viewed as a counterpart of linear regression on manifolds, and subsequent work of Hinkle et al. (2012) generalizes polynomial regression model to the manifold. These para-
metric and semi-parametric models are elegant, but may lack sufficient flexibility in certain applications. Shi et al. (2009) proposes a semi-parametric intrinsic regression model on manifolds, and Davis et al. (2007) generalizes an intrinsic kernel regression method on the Riemannian manifold, considering applications in modeling changes in brain shape over time. Yuan et al. (2012) develops an intrinsic local polynomial model on the space of symmetric positive definite matrices, which has applications in diffusion tensor imaging. A drawback of intrinsic models is the heavy computational burden incurred by minimizing a complex objective function along geodesics, typically requiring evaluation of an expensive gradient in an iterated algorithm. The objective functions often have multiple modes, leading to large sensitivity to start points. Further, existence and uniqueness of the population regression function hold under relatively restrictive conditions. Therefore, the descent algorithms used in obtaining the estimate in general do not guarantee to converge to a global optima.

With the motivation of developing general purpose computationally efficient, theoretically sound and practically useful regression modeling frameworks for manifold-valued response data, we propose a nonparametric extrinsic regression model by first embedding the manifold where the response resides onto some higher-dimensional Euclidean spaces. We use equivariant embeddings, which preserve a great deal of geometry for the images. A local regression estimate (such as a local polynomial estimate) of the regression function is obtained after embedding, which is then projected back onto the image of the manifold. Outside the regression setting, both intrinsic and extrinsic approaches have been proposed for modeling of manifold-valued data and for mathematically studying the properties of manifolds. However, to our knowledge, our work is the first in taking an extrinsic approach in the regression modeling context. Our approach is general, has elegant asymptotic theory and ourperforms intrinsic models in terms of computation efficiency. In addition, there is essentially no difference in inference with the examples considered.
This chapter is organized as follows. Section 3.1 introduces the extrinsic regression model. In Section 3.2, we explore the full utilities of our models by applying our model to three examples in which the response resides on different manifolds. A simulation study is carried out for data on the sphere (example 1) applying both intrinsic and extrinsic models. The results indicate the overall superiority of our extrinsic method in terms of computational complexity and time compared to that of intrinsic methods. The extrinsic models are also applied to planar shape manifolds in example 2, with an application considered to modeling the brain shape of the Corpus Callosum from an ADHD (Attention Deficit/Hyperactivity Disorder) study. In example 3, our model is applied to data on the Grassmannian considering both simulated and real data. Section 3.3 is devoted to studying the asymptotic properties of our estimators in terms of asymptotic distribution and convergence rate.

3.1 Extrinsic local regression on manifolds

Let \( Y \in M \) be the response variable in a regression model where \((M, \rho)\) is a general metric space with distance metric \( \rho \). Let \( X \in \mathbb{R}^m \) be the covariate or predictor variable which can be random or fixed. Given data \((x_i, y_i) (i = 1, \ldots, m)\), the goal is to model a regression relationship between \( Y \) and \( X \). The typical regression framework with \( y_i = F(x_i) + \epsilon_i \) is not appropriate here as expressions like \( y_i - F(x_i) \) are not well-defined due to the fact that the space \( M \) (e.g., a manifold) where the response variable lies is in general not a vector space. Let \( P(x, y) \) be the joint distribution of \((X, Y)\) and \( P(x) \) be the marginal distribution of \( X \) with marginal density \( f_X(x) \). Denote \( P(y|x) \) as the conditional distribution of \( Y \) given \( X \) with conditional density \( p(y|x) \). One can define the population regression function or map \( F(x) \) (if it exists) as

\[
F(x) = \arg\min_{q \in M} \int_M \rho^2(q, y) P(dy|x),
\] (3.1)
where $\rho$ is the distance metric on $M$.

Let $M$ be a $d$-dimensional differentiable or smooth manifold. A manifold $M$ is a topological space that locally behaves like a Euclidean space. In order to equip $M$ with a metric space structure, one can employ a Riemannian structure, with $\rho$ taken to be the geodesic distance, which defines an intrinsic regression function. Alternatively, one can embed the manifold onto some higher dimensional Euclidean space via an embedding map $J$ and use the Euclidean distance $\| \cdot \|$ instead. The latter model is referred to as an extrinsic regression model. One of the potential hurdles for carrying out intrinsic analysis is that uniqueness of the population regression function in (3.1) (with $\rho$ taken to be the geodesic distance) can be hard to verify. Le and Barden (2014) establish several interesting and deep results for the regression framework and provide broader conditions for verifying the uniqueness of the population regression function. Intrinsic models can be computationally expensive, since minimizing their complex objective functions typically require a gradient descent type algorithm. In general, this requires fine tuning at each step, which results in an excessive computational burden. Further, these gradient descent algorithms are not always guaranteed to converge to a global minimum or only converge under very restrictive conditions. In contrast, the uniqueness of the population regression holds under very general conditions for extrinsic models. Extrinsic models are extremely easy to evaluate and are orders of magnitude faster than intrinsic models.

Let $J : M \to E^D$ be an embedding of $M$ onto some higher dimensional ($D \geq d$) Euclidean space $E^D$ and denote the image of the embedding as $\widetilde{M} = J(M)$. By the definition of embedding, the differential of $J$ is a map between the tangent space of $M$ at $q$ and the tangent space of $E^D$ at $J(q)$; that is, $d_q J : T_q M \to T_{J(q)} E^D$ is an injective map and $J$ is a homeomorphism of $M$ onto its image $\widetilde{M}$. Here $T_q M$ is the tangent space of $M$ at $q$ and $T_{J(q)} E^D$ is the tangent space of $E^D$ at $J(q)$. Let $\| \cdot \|$
be the Euclidean norm. In an extrinsic model, the true extrinsic regression function is defined as

\[ F(x) = \arg\min_{q \in M} \int_M \|J(q) - J(y)\|^2 P(dy|x) \]

\[ = \arg\min_{q \in M} \int_{\tilde{M}} \|J(q) - z\|^2 \tilde{P}(dz|x) \tag{3.2} \]

where \( \tilde{P}(\cdot \mid x) = P(\cdot \mid x) \circ J^{-1} \) is the conditional probability measure on \( J(M) \) given \( x \) induced by the conditional probability measure \( P(\cdot \mid x) \) via the embedding \( J \).

We now proceed to propose an estimator for \( F(x) \). Let \( K : \mathbb{R}^m \rightarrow \mathbb{R} \) be a multivariate kernel function such that \( \int_{\mathbb{R}^m} K(x)dx = 1 \) and \( \int_{\mathbb{R}^m} xK(x)dx = 0 \). One can take \( K \) to be a product of \( m \) one-dimensional kernel functions for example. Let \( H = \text{diag}(h_1, \ldots, h_m) \) with \( h_i > 0 \) \( (i = 1, \ldots, m) \) be the bandwidth vector and \( |H| = h_1 \ldots h_m \). Let \( K_H(x) = \frac{1}{|H|} K(H^{-1}x) \) and

\[ \hat{F}(x) = \arg\min_{y \in \mathcal{D}} \sum_{i=1}^n \frac{K_H(x_i - x) ||y - J(y_i)||^2}{\sum_{i=1}^n K_H(x_i - x)} = \sum_{i=1}^n \frac{J(y_i)K_H(x_i - x)}{\sum_{i=1}^n K_H(x_i - x)}, \tag{3.3} \]

which is basically a weighted average of points \( J(y_1), \ldots, J(y_n) \). We are now ready to define the extrinsic kernel estimate of the regression function \( F(x) \) as

\[ \hat{F}_E(x) = J^{-1}\left( \mathcal{P}(\hat{F}(x)) \right) = J^{-1}\left( \arg\min_{q \in \tilde{M}} \|q - \hat{F}(x)\| \right), \tag{3.4} \]

where \( \mathcal{P} \) denotes the projection map onto the image \( \tilde{M} \). Basically, our estimation procedure consists of two steps. In step one, it calculates a local regression estimate on the Euclidean space after embedding. In step two, the estimate obtained in step one is projected back onto the image of the manifold.

**Remark 1** The embedding \( J \) used in the extrinsic regression model is in general not unique. It is desirable to have an embedding that preserves as much geometry as
An equivariant embedding preserves a substantial amount of geometry. Let $G$ be some large Lie group acting on $M$. We say that $J$ is an equivariant embedding if we can find a group homomorphism $\phi : G \to GL(D, \mathbb{R})$ from $G$ to the general linear group $GL(D, \mathbb{R})$ of degree $D$ such that

$$J(gq) = \phi(g)J(q)$$

for any $g \in G$ and $q \in M$. The intuition behind equivariant embedding is that the image of $M$ under the group action of the Lie group $G$ is preserved by the group action of $\phi(G)$ on the image, thus preserving many geometric features. Note that the choice of embedding is not unique and in some cases constructing an equivariant embedding can be a non-trivial task, but in most of the cases a natural embedding would arise and such embeddings can often be verified as equivariant.

**Remark 2** Alternatively, we can obtain some robust estimator under our proposed framework. The regression estimate is taken as the projection of the following estimator onto the image $\widetilde{M}$ of $M$ after an embedding $J$. We can call it the extrinsic median regression model. Specifically, we define

$$\hat{F}(x) = \text{argmin}_{y \in E^D} \sum_{i=1}^{n} \frac{K_H(x_i - x)||y - J(y_i)||}{\sum_{i=1}^{n} K_H(x_i - x)}$$

and

$$\hat{F}_E(x) = J^{-1}\left(\text{argmin}_{q \in \tilde{M}} ||q - \hat{F}(x)||\right).$$

One can use the Weizfield formula (Weiszfeld, 1937) in calculating the weighted median of (3.5) (if it exists). Such estimates can be shown to be robust to outliers and contaminations.

**Remark 3** A kernel estimate is obtained first in (3.3) before projection. However, the framework can be easily generalized using higher order local polynomial regression estimates (of degree $p$)(Fan and Gijbels, 1996). For example, one can have the
local linear estimator for $\hat{F}(x)$ before projection. That is, for any $x$, let

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \|J(y_i) - \beta_0 - \beta_1'(x_i - x)\|^2 K_H(x_i - x).$$

(3.7)

Then, we have

$$\hat{F}(x) = \hat{\beta}_0(x),$$

(3.8)

$$\hat{F}_E(x) = J^{-1}\left(\mathcal{P}(\hat{F}(x))\right) = J^{-1}\left(\arg\min_{q\in\tilde{M}} \|q - \hat{F}(x)\|\right).$$

(3.9)

The properties of the estimator $\hat{F}_E(x)$ where $\hat{F}(x)$ is given by the general $p$th local polynomial estimator of $J(y_1), \ldots, J(y_n)$ are explored in Theorem 8.

3.2 Examples and applications

The proposed extrinsic regression framework is very general and has appealing asymptotic properties as will be shown in Section 4. To illustrate the wide applicability of our model and validate its finite sample performance, we carry out a study by applying our model to various examples with the response taking values in many well-known manifolds. For each of the examples considered, we provide details on the embeddings, verify such embeddings are equivariant, and give explicit expressions for the projections to obtain the final estimate in each case. In example 1, we simulate data from a 2-d sphere and compare the estimates from our extrinsic regression model with that of an intrinsic model. The result indicates that the extrinsic models clearly outperform the intrinsic models by orders of magnitude in terms of computational complexity and time. In example 2, we study a data example with response from a planar shape, in which the brain shape of the subjects are represented by landmarks on the boundary. Example 3 provides details of the estimator when the responses take values on a Stiefel or Grassmann manifold. The method is illustrated with a synthetic data set and small financial time series data set, both
of which have subspace responses of possibly mixed dimension and covariates, which are the corresponding time points.

**Example 1**

Statistical analysis on i.i.d data from the 2-dimensional sphere $S^2$, often called directional statistics, has a long history (Fisher, 1953; Watson, 1983; Mardia and Jupp, 2000; Fisher et al., 1987). In this example, we work out the details in an extrinsic regression model with the responses lying on a $d$-dimensional sphere $S^d$. The model is illustrated with data $\{(x_i, y_i), i = 1, \ldots, n\}$, where $y_i \in S^2$.

Note that $S^d$ is a submanifold of $\mathbb{R}^{d+1}$; therefore, the inclusion map $\iota$ serves as a natural embedding onto $\mathbb{R}^{d+1}$. It is easy to check that the embedding is equivariant with the Lie group $G = SO(d+1)$, the special orthogonal group of $(d+1)$ by $(d+1)$ matrices $A$ with $AA^T = 1$ and $|A| = 1$. Take the homomorphism map from $G$ to $GL(d + 1, \mathbb{R})$ to be the identity map. Then it is easy to see that $J(gp) = gp = \phi(g)J(p)$, where $g \in G$ and $p \in S^d$.

Given $J(y_1), \ldots, J(y_n)$, one first obtains $\widehat{F}(x)$ as given in (3.3). Its projection onto the image $\overline{M}$ is given by

$$\hat{F}_E(x) = \hat{F}(x)/||\hat{F}(x)||, \text{ when } \hat{F}(x) \neq 0. \quad (3.10)$$

There are many well defined parametric distributions on the sphere. A common and useful distribution is the von Mises-Fisher distribution (Fisher, 1953) on the unit sphere, which has the density with respect to the normalized volume measure on the sphere

$$p_{MF}(y; \mu, \kappa) = \exp(\kappa \mu^T y),$$

where $\kappa$ is a concentration parameter and $\mu$ is a location parameter and $E(y) = \mu$ holds. We simulate the data from the unit sphere by letting the mean function be covariate-dependent. That is, let

$$\mu = \frac{\beta \circ x}{|\beta \circ x|},$$
where \( \beta \circ x \) is the Hadamard product \((\beta_1 x^1, \ldots, \beta_m x^m)\).

For this example, we will use data generated by the following model
\[
\beta \sim N_3(0, I), \quad x_1^i \sim N(0, 1), \quad x_2^i \sim N(0, 1), \quad x_3^i = x_1^i \circ x_2^i, \tag{3.11}
\]
\[
y_i \sim MF(\mu_i, \kappa), \quad \mu_i = \frac{\beta \circ x_i}{|\beta \circ x_i|}, \quad i = 1, \ldots, n,
\]
\(\kappa\) some fixed known value.

As an example of what the data looks like, we generate one thousand \((n = 1000)\) observations from the above model with \(\kappa = 10\) so that realizations are near their expected value. Figure 3.1 shows this example in which 100 predictions from the extrinsic model are plotted against their true values using 900 training points. To select the bandwidth \(h\) we use 10-fold cross-validation with \(h\) ranging from \([.1, .2, \ldots, 1.9, 2]\) and choose the value that gives minimum average mean square error. Residuals for the mean square error are measured using the intrinsic distance, or greater circle distance, on the sphere.

![Figure 3.1](image)

**Figure 3.1: Left** The training values on the sphere. **Middle** The held out values to be predicted through extrinsic regression. **Right** The extrinsic predictions (blue) plotted against the true values (red).

To illustrate the utility and advantages of extrinsic regression models, we compare our method to an intrinsic kernel regression model that uses intrinsic distance of the sphere to minimize the objective function. Computations on the sphere are in general
not as intensive compared to more complicated manifolds such as shape spaces, etc, but it still requires an iterative algorithm, such as gradient descent, for the intrinsic model in order to obtain a kernel regression estimate. The following simulation results demonstrate extrinsic kernel regression gives at least as accurate estimates as intrinsic kernel regression but in much less computation time even for $S^2$.

**Comparison with an intrinsic kernel regression model:** The intrinsic kernel regression estimate minimizes the objective function $f(y) = \sum_{i=1}^{n} w_i d^2(y, y_i)$, where $y$ and $y_i$ are points on the sphere $S^2$, $w_i$ are determined by the Gaussian kernel function, and $d(\cdot, \cdot)$ in this case is the greater circle distance. Then the gradient of $f$ on the sphere is given by

$$\nabla f(y) = \sum_{i=1}^{n} w_i 2d(y, y_i) \frac{\log_y(y_i)}{d(y, y_i)} = \sum_{i=1}^{n} 2w_i \frac{\arccos(y^T y_i)}{\sqrt{1 - (y^T y_i)^2}} (y_i - (y^T y_i)y),$$

where $\log_y(y_i)$ is the log map or the inverse exponential map on the sphere. Estimates for $y$ can be obtained through a gradient descent algorithm with step size $\delta$ and error threshold $\epsilon$. We applied the intrinsic and extrinsic models to the same set of data using the Gaussian kernel function.

Twenty different data sets of 2000 observations were generated from the above sphere regression model with von-Mises Fisher concentration parameter

$$\kappa = \{1, 2, \ldots, 20\}.$$ 

Of the 2000 observations, 50 were used to check the accuracy of the extrinsic and intrinsic estimates. To see the effect of training sample size on the quality of the estimates, the estimates were also made on subsets of the 1950 training observations, starting with 2 observations and increasing to all 1950 observations. The same training observations were always used for both models. In both models, the bandwidth was chosen through 10-fold cross validation. The intrinsic kernel regression was fit with step size $\delta = .01$ and error threshold $\epsilon = .001$. The performance of the two
models are compared in terms of MSE and predictive MSE. The MSE is calculated using the greater circle distance between predicted values and the true expected value, while predictive MSE is calculated using the greater circle distance between the predicted values and the realized values. The performance results using 50 hold out observations can be seen in Figure 3.2.

Predictive MSE does not converge to 0 because the generating distribution has a high variance; however, as the concentration increases, the predictive MSE does approach 0. The extrinsic and intrinsic kernel regressions perform similarly with large sample sizes. The extrinsic kernel regression drops in predictive MSE faster than the intrinsic model, which may stem from only having the kernel bandwidth as
a tuning parameter which can be selected more easily than choosing the bandwidth, step-size, and error thresholds even through cross-validation.

A significant advantage of the extrinsic kernel regression is the speed of computation. Both methods were implemented in C++ using Rcpp (Eddelbuettel and François, 2011), and resulted in up to a 60× improvement in speed in making a single prediction using all of the training observations. For speed comparisons, a single prediction was made given the same number of test observations, and the time to produce the estimate was recorded. Each of these trials was done five times, and we compare the mean time to producing the estimate in Figure 3.3.

Note that the same kernel weights are computed in both algorithms, so the difference is attributable to the gradient descent versus extrinsic optimization procedures. Since the speed comparisons were done for computing a single prediction and the difference is due almost entirely to the gradient descent steps, making multiple predictions results in an even more favorable comparison for the extrinsic model. This experiment shows that the extrinsic kernel regression applied to sphere data performs at least as well on prediction and can be computed significantly faster.

**Example 2** We now consider an example with planar shape responses. Planar shapes are one of the most import classes of landmark based shapes spaces. Such spaces were defined by Kendall (1977) and Kendall (1984) with pioneering work by Bookstein (1978) motivated from applications on biological shapes. We now describe the geometry of the space which will be used in obtaining regression estimates for our model. Let \( z = (z_1, \ldots, z_k) \) with \( z_1, \ldots, z_k \in \mathbb{R}^2 \) be a set of \( k \) landmarks. Let \( \bar{z} = (\bar{z}, \ldots, \bar{z}) \) where \( \bar{z} = \sum_{i=1}^{k} z_i/k \). Denote \( u = \frac{z - \langle \bar{z} \rangle}{\|z - \langle \bar{z} \rangle\|} \) which can be viewed as an element on the sphere \( S^{2k-3} \), which is called the pre shape. The planar shape \( \Sigma_2^k \) can now be represented as the quotient of the pre shape under the group action by \( SO(2) \), the 2 by 2 special orthogonal group. That is, \( \Sigma_2^k = S^{2k-2-1}/SO(2) \).
Figure 3.3: Speed comparisons between the extrinsic and intrinsic kernel regressions as a function of the number of training observations. The average seconds to produce an estimate for a single test observation are plotted in red for the intrinsic model, and black for the extrinsic model. The multiple between the speed for the intrinsic and extrinsic estimates plotted are also plotted for reference.

\[ \Sigma^k \] can be shown to be equivalent to the complex projective space \( \mathbb{CP}^{k-2} \). Therefore, a point on the planar shape can be identified as the orbit or equivalent of \( z \) which we denote \( \sigma(z) \). Viewing \( z \) as elements in the complex plane, one can embed \( \Sigma^k \) onto the \( S(k, \mathbb{C}) \), the space of \( k \times k \) complex Hermitian matrices via the Veronese-Whitney embedding (see Bhattacharya and Patrangenaru (2005), Bhattacharya and...
Bhattacharya (2012):

\[ J(\sigma(z)) = uu^* = (u_i \bar{u}_j)_{1 \leq i, j \leq k}. \] (3.12)

One can verify the Veronese-Whitney embedding is equivariant (see Kendall (1984)) by taking the Lie group \( G \) to be special unitary group \( SU(k) \) with

\[ SU(k) = \{ A \in GL(k, \mathbb{C}), AA^* = I, det(A) = 1 \}. \]

The action is on the left,

\[ A\sigma(z) = \sigma(Az). \]

The homomorphism map \( \phi \) is taken to be

\[ \phi : S(k, \mathbb{C}) \rightarrow S(k, \mathbb{C}) : \phi(A)B = ABA^*. \]

Therefore, one has

\[ J(A\sigma(z)) = Auu^*A^* = \phi(A)J(\sigma(z)). \]

We now describe the projection after \( \hat{F}(x) \) is given by (3.3), where \( J(y_i) \) \( (i = 1, \ldots, n) \) are obtained using the equivalent embedding given in (3.12). Letting \( v^T \) be the eigenvector corresponding to largest eigenvalue of \( \hat{F}(x) \), by a careful calculation, one can show that the projection of \( \hat{F}(x) \) is given by

\[ P_{\lambda}(\hat{F}(x)) = v^T \hat{\nu}. \]

Therefore, the extrinsic kernel regression estimate is given by

\[ \hat{F}_E(x) = J^{-1}(v^T \hat{\nu}). \] (3.13)

**Corpus Callosum (CC) data set:** We study ADHD-200 dataset \(^1\) in which the shape contour of the brain Corpus Callosum are recorded for each subject along

\(^1\) http://fcon_1000.projects.nitrc.org/indi/adhd200/
with variables such as gender, age, and ADHD diagnosis. The subjects consist of patients who are diagnosed with ADHD. 50 landmarks were placed outlining the CC shape for 647 patients for the ADHD-200 dataset. The age of the patients range from 7 to 21 years old, with 404 typically developing children and 243 individuals diagnosed with some form of ADHD. The original data set differentiates between types of ADHD diagnoses, and we simplify the problem of choosing a kernel by using a binary response for an ADHD diagnosis.

According to the findings in Huang et al. (2015), there is not a significant effect of gender on the area of different segments of the CC; however diagnosis and the interaction between diagnosis and age were found to be statistically significant ($p < .01$). With knowledge of these results, we performed the extrinsic kernel regression method for the CC planar shape response using diagnosis, $x_1$, and the age, $x_2$, for covariates. The choice of kernel between two sets of covariates $x_1 = (x_1^1, x_1^2)$ and $x_2 = (x_2^1, x_2^2)$ is

$$k(x_1, x_2) = \begin{cases} 
\exp \left(-\frac{(x_2^1-x_2^2)^2}{h}\right) / h & \text{if } x_1^1 = x_2^1 \\
0 & \text{if } x_1^1 \neq x_2^1
\end{cases}$$

Although Huang et al. (2015) explores clustering the shape by specific diagnosis, we visualize how the CC shape develops over time by making predictions at different time points. We show predictions for ages 9, 12, 16, and 19 year old children of ADHD diagnosis or typical development. The results can be seen in Figure 3.4.

What we can observe from the two plots is that the CC shapes for the 8 year olds seem to be close, but by age 12 the shapes have diverged substantially, with shrinking of the CC being apparent in later years in development. This quality of the CC shapes between ADHD and normal development is consistent with results found in the literature (Huang et al. (2015)).

In previous studies, ADHD diagnoses were clustered using the shape information
Figure 3.4: Predicted CC shape for children ages 9, 12, 16, and 19. The black shape corresponds to typically developing children, while the red shape corresponds to children diagnosed with ADHD. Kernel regression allows us to visualize how CC shape changes through development. Here sections of CC appear smaller in ADHD diagnoses than in normal development.

to predict the diagnosis class, and the centroid of the cluster is the predicted shape for that class (Huang et al. (2015)). Our method adds to this analysis by taking the diagnosis and predicting the CC shape as a function of age. Our model also has the benefit of evaluating quickly, making selecting the bandwidth for the kernel through cross-validation feasible.

**Example 3** We now consider another two classes of important manifolds, Stiefel manifolds and Grassman manifolds (Grassmannians). The Stiefel manifold, $V_k(\mathbb{R}^m)$, is the collection of $k$ orthonormal frames in $\mathbb{R}^m$. That is, the Stiefel manifold consists
of the set of ordered \(k\)-tuples of orthonormal vectors in \(\mathbb{R}^m\), which can be represented as \(\{X \in S(m,k), XX^T = I_m\}\). The Stiefel manifold includes the \(m\) dimensional sphere \(S^m\) as a special case with \(k=1\) and \(O(m)\) the orthogonal group when \(k = m\). Examples of data on the Stiefel manifold includes the orbit of the comets and the vector cardiogram. Applications of Stiefel manifold are present in earth sciences, medicine, astronomy, meteorology and biology. The Stiefel manifold is a compact manifold of dimension \(km - k - k(k - 1)/2\) and it is a submanifold of \(\mathbb{R}^{km}\). The inclusion map can be further shown to be an equivalent embedding with the Lie group taken to the orthogonal group \(O(m)\).

Given \(\hat{F}(x)\) obtained by kernel regression after embedding the points \(y_1, \ldots, y_n\) on the Stiefel manifold to the Euclidean space \(\mathbb{R}^{km}\), the next step is to obtain the projection of \(\hat{F}(x)\) onto \(\widetilde{\mathcal{M}} = J(M)\). We first make an orthogonal decomposition of \(\hat{F}(x)\) by letting \(\hat{F}(x) = US\), where \(U \in V_{k,m}\), which can be viewed as the orientation of \(\hat{F}(x)\) and \(S\) is positive semi-definite, which has the same rank as \(\hat{F}(x)\). Then the projection of \(\hat{F}(x)\) (or projection set) is given by

\[
P_M(\hat{F}(x)) = \{U \in V_{k,m} : \hat{F}(x) = U(\hat{F}(x)^T \hat{F}(x))^{1/2}\}.
\]

See Theorem 10.2 in Bhattacharya and Bhattacharya (2012) for a proof of the results. Then the projection is unique, that is, the above set is a singleton if and only if \(\hat{F}(x)\) is of full rank.

The Grassmann manifold or the Grassmannian \(Gr_k(\mathbb{R}^m)\) is the space of all the subspaces of a fixed dimension \(k\) whose basis elements are vectors in \(\mathbb{R}^m\), which is closely related to the Stiefel manifold \(V_{k,m}\). Recall a subspace can be viewed as the span of an orthonormal basis. Let \(\mathbf{v} = \{v_1, \ldots, v_k\}\) be such an orthonormal basis for a subspace on the Grassmannian. Note that the order of the vector does not matter unlike in the case of Stiefel manifold. For any two elements on the Stiefel manifold whose span correspond to the same subspace, there exists an orthogonal
transformation (mapped by a orthogonal matrix in $O(k)$) between the two orthonormal frames. These two points will be identified as the same point on the Grassman manifold. Therefore, the Grassmannian can be viewed as the collection of the equivalent classes on the Stiefel manifold, i.e., a quotient space under the group action of $O(k)$, the $k$ by $k$ orthogonal group. Then one has $Gr_k(\mathbb{R}^m) = V_k(\mathbb{R}^m)/O(k)$. There are many applications of Grassmann manifolds, in which the subspaces are the basic element in signal processing, machine learning and so on.

The equivariant embedding for $Gr_k(\mathbb{R}^m)$ also exists (Chikuse, 2003). Let $X \in V_{k,m}$ be a representative element of the equivalent classes in $Gr_k(\mathbb{R}^m) = V_k(\mathbb{R}^m)/O(k)$. So an element in the quotient space can be represented by the orbit $\sigma(X) = XR$ where $R \in O(k)$. Then an embedding can be given by

$$J(\sigma(X)) = XX^T.$$ 

The collection of $XX^T$ forms a subspace of $\mathbb{R}^{m^2}$. We now verify that $J$ is an equivariant embedding under the group action of $G = O(m)$. Letting $g \in G = O(m)$, one has $J(gX) = gXX^Tg^T = \phi(g)J(X)$, where the map $\phi(g) = g$ acts on the image $J(X)$ by conjugation map. That is, $\phi(g)J(X) = gXX^Tg^T$.

Given the estimate $\hat{F}(x)$, the next step is to derive the projection of $\hat{F}(x)$ onto $\tilde{M} = J(M)$. Since all $XX^T$ form a subspace, one can use the following procedure to calculate the map from $\hat{F}(x)$ to the Grassmann manifold by finding an orthonormal basis for the image. This algorithm is a special case of the projection via Conway embedding (St. Thomas et al., 2014).

1. Find the eigendecomposition $\hat{F}(x) = QAQ^{-1}$

2. Take the $k$ eigenvectors corresponding to the top $k$ eigenvalues in $\Lambda$ as an orthonormal basis for $\hat{F}_E(x)$, $Q_{[1:k]}$.

We now consider two illustrative examples, one synthetic and one from a financial situation.
time series, for extrinsic kernel regression with subspace response variables. The technique is unique compared to other subspace regression techniques because the extrinsic distance offers a well defined and principled distance between responses of different dimension. This prevents having to constrain the responses to be a fixed dimension or hard coding a heuristic distance between subspaces of different dimension into the distance function.

We now consider a synthetic example in which the predictors are the time points and the responses are points on the Grassmann manifold. Since we represent subspaces with draws from the Stiefel manifold, we draw orthonormal bases from the Matrix von Mises-Fisher distribution as their representation. We generate $N$ draws from the following process with concentration parameter $\kappa$, in which the first $n_1$ draws are dimension 4 and the last $n_2$ draws are dimension 5,

\[
\begin{align*}
\text{for } 1 \leq t \leq N \text{ do} \\
\text{Draw } X &\sim MN(0, I_m, I_5) \\
\mu_{[1]} &:= t + X_{[1]} \\
\mu_{[2]} &:= t - X_{[2]} \\
\mu_{[3]} &:= t^2 + X_{[3]} \\
\mu_{[4]} &:= t X_{[4]} \\
\text{if } t > n_1 \text{ then} \\
\mu_{[5]} &:= t + t X_{[5]} \\
\text{end if} \\
Y_t &:= \text{vMF}(\kappa M)
\end{align*}
\]

Here the only covariate associated with $Y_t$ is $t$. With a concentration of $\kappa = 1$, and $n_1 = n_2 = 50$, we generate much noisier data than before, and are able to correctly predict the dimension of subspace at each time point. When examining the
pairwise distance between the realizations in Figure 3.5, it is clear that the extrinsic distance distinguishes between dimension and does not require any specification of the dimension. The predicted dimension at each time point and the residuals are plotted in figure 3.6.

**Figure 3.5:** Pairwise distance between two observations generated by the specified model indexed by $t$ measured by distance between points in the Conway embedding. This visualization of the extrinsic distance shows the cluster by dimension.

The key advantage with this method is not requiring any constraints on the dimension of the input or output subspaces. This is important in some examples such as high dimensional time series analysis with data such as high frequency trading where the analysis usually culminates in analyzing principal components, or eigenvectors of the large covariance matrix estimated between assets. Market events can change asset covariances which in turn changes the number of significant eigenvectors, so a method automatically interpolating time points must not depend on specifying the number of significant eigenvectors.
Figure 3.6: The estimated dimension and residual for the extrinsic kernel regression estimate at each time point $t$ from data generated from the specified model. The regression estimate is accurate on the dimension of the subspace and prediction residuals are consistent with a concentration parameter $\kappa = 1$.

We apply this method to the Istanbul Stock Exchange on UCI Machine Learning Repository Akbilgic et al. (2013), using the 5 index funds S&P 500 (SP), the Istanbul Stock Exchange (ISE), stock market return index of Japan (NIKKEI), MSCI European index (EU), and the stock market return index of Brazil (BOVESPA). The data contains 97 full weeks over 115 weeks of daily market closing values from January 5, 2009 to February 18, 2011. For each week, a covariance matrix is estimated between the assets, of which the eigenvectors with eigenvalues greater than $10^{-10}$ are retained as the orthonormal basis corresponding to the covariance matrix. As can be seen in Figure 3.7, these covariance matrices change significantly over time. These
orthonormal matrices are given to the model along with the corresponding week and each week is predicted.

![Covariance matrices estimated from daily closing prices from various stock market index funds over various weeks.](image)

**Figure 3.7**: The covariance matrices estimated from the daily closing prices from various stock market index funds over various weeks. Covariance between markets change substantially from week to week.

The residuals are shown in Figure 3.8 compared to the observed distance between subspaces between two consecutive observations. The method is predicting the subspaces within the variance of data, which means there is information or at least structure to how the covariance matrices are evolving over time – the relationships are not purely random week to week.

### 3.3 Asymptotic properties of the extrinsic regression model

In this section, we investigate the large sample properties of our extrinsic regression estimates. We assume the marginal density $f_X(x)$ is differentiable and the absolute value of any of the partial derivatives of $f_X(x)$ of order two are bounded by some
constant $C$. In our proof, we assume our kernel function $K$ takes a product form. That is, $K(x) = K_1(x^1) \cdots K_m(x^m)$ where $x = (x^1, \ldots, x^m)$ and $K_1, \ldots, K_m$ are one dimensional symmetric kernels such that $\int_\mathbb{R} K_i(u) du = 1$, $\int_\mathbb{R} u K_i(u) du = 0$ and $\int_\mathbb{R} u^2 K_i(u) du < \infty$ for $i = 1, \ldots, m$. The results can be generalized to kernels with arbitrary form and with $H$ given by a more general positive definite matrix instead of a diagonal matrix. Theorem 5 derives the asymptotic distribution of the extrinsic regression estimate $\hat{F}_E(x)$ for any $x$.

**Theorem 5.** Let $\mu(x) = E\left( \tilde{P}(dy|x) \right)$, which is the conditional mean regression function of $\tilde{P}$ and assume $\mu(x)$ is differentiable. Assume $n|H| \to \infty$. Denote $x = (x^1, \ldots, x^m)$. Let $\tilde{\mu}(x) = \mu(x) + \frac{Z(x)}{\int_X Z(x)}$, where the $i$th component $Z_i(x)$ ($i = 1, \ldots, D$)
of $Z(x)$ is given by

$$Z_i(x) = h_1^2 \left( \frac{\partial f}{\partial x^1} \frac{\partial \mu_i}{\partial x^1} + \frac{1}{2} f_x(x) \left( \frac{\partial^2 \mu_i}{\partial (x^1)^2} + \ldots + \frac{\partial^2 \mu_i}{\partial x^1 x^1} \right) \right) \int v_1^2 K_1(v_1) dv_1 + \ldots$$

$$+ h_m^2 \left( \frac{\partial f}{\partial x^m} \frac{\partial \mu_i}{\partial x^m} + \frac{1}{2} f_x(x) \left( \frac{\partial^2 \mu_i}{\partial (x^m)^2} + \ldots + \frac{\partial^2 \mu_i}{\partial (x^1)^2} \right) \right) \int v_m^2 K_m(v_m) dv_m$$

(3.14)

Assume the projection $P$ of $\tilde{\mu}(x)$ onto $\tilde{M} = J(M)$ is unique and $P$ is continuously differentiable in a neighborhood of $\tilde{\mu}(x)$. Then the following holds assuming $P(dy \mid x) \circ J^{-1}$ has finite second moments:

$$\sqrt{n} \|d_p(x)\mathcal{P} \left( \mathcal{F}(x) - \tilde{\mu}(x) \right) \xrightarrow{L} N(0, \tilde{\Sigma}(x)), \quad (3.15)$$

where $d_p(x)\mathcal{P}$ is the differential from $T_{\tilde{\mu}(x)}\mathbb{R}^D$ to $T_{\mathcal{P}(\tilde{\mu}(x))}\tilde{M}$ of the projection map $\mathcal{P}$ at $\tilde{\mu}(x) = \mu(x) + \frac{Z(x)}{f_x(x)}$. Here $\Sigma(x) = B'\tilde{\Sigma}(x)B$, where $B$ is the $D \times d$ matrix of the differential $d_p(x)\mathcal{P}$ with respect to given orthonormal bases of $T_{\tilde{\mu}(x)}\mathbb{R}^D$ and $T_{\mathcal{P}(\tilde{\mu}(x))}\tilde{M}$, and the $(j, k)$th entry of $\tilde{\Sigma}(x)$ is given by (B.13) with

$$\tilde{\Sigma}_{jk} = \frac{\sigma(J_j(y), J_k(y))}{f_x(x)} \int K(v)^2 dv,$$

(3.16)

where $\sigma(J_j(y), J_k(y)) = \text{Cov}(J_j, J_k)$, and $J_j$ is the $j$th element of $J(y)$. Here $\xrightarrow{L}$ indicates convergence in distribution.

Corollary 6 is on the mean integrated squared error of the estimates.

**Corollary 6.** Assuming the same conditions of Theorem 5 and the domain of the covariates is bounded, the mean integrated squared error of $\mathcal{F}_E(x)$ is of the order $O(n^{-4/(m+4)})$, with the choice of $h_i$’s $(i = 1, \ldots, m)$ to be of the same order, that is, of $O(n^{-1/(m+4)})$. 

53
Remark. Note that in nonparametric regression with both predictors \((m\text{-dimensional})\) and responses in the Euclidean space, the optimal order of the mean integrated squared error is \(O(n^{-4/(m+4)})\) under the assumption that the true regression function has bounded second derivative. Our model achieves the same rates. However, whether such rates are minimax in the context of manifold valued response is not known.

Theorem 7 shows some results on uniform convergence rates of the estimator.

**Theorem 7.** Assume the covariate space \(x \in \mathcal{X} \subset \mathbb{R}^m\) is compact and \(\mathcal{P}\) has continuous first derivative. Then

\[
\sup_{x \in \mathcal{X}} d_{\mu(x)} \mathcal{P} \left( \hat{F}(x) - E(\hat{F}(x)) \right) = O_p \left( \frac{\log^{1/2} n}{\sqrt{n} |H|} \right).
\]

(3.17)

As pointed out in Remark 3.1, it is ideal in many cases to fit a higher order (say \(p\)th order) local polynomial model in estimating \(\mu(x)\) before projecting back onto the image of the manifold. Such estimates are more appealing especially when \(F(x)\) is more curved over a neighborhood of \(x\). One can show that similar results as those of Theorem 5 hold, though with much more involved argument.

We now give details of such estimators and their asymptotic distributions are derived in Theorem 8. Recall \(F(x) = E(P(dy \mid x))\) and \(\mu(x) = E\left( \hat{P}(dy \mid x) \right)\) and \(J(y_1), \ldots, J(y_n)\) are the points on \(\widetilde{M} = J(M)\) after embedding \(J\). We first obtain an estimate \(\hat{F}(x)\) of \(\mu(x)\) using \(p\)th order local polynomials estimation. The intermediate estimate \(\hat{F}(x)\) is then projected back to \(\widetilde{M}\) serving as the ultimate estimate of \(F(x)\).
The general framework is given as follows:

\[ \{ \hat{\beta}_k^j(x) \}_{0 \leq |k| \leq p, 1 \leq j \leq D} \] (3.18)

\[
= \arg\min_{\{ \beta_k^j(x) \}_{0 \leq |k| \leq p, 1 \leq j \leq D}} \sum_{i=1}^{n} \left( \left| \left| \beta_k^j(x_i) - x \right| \right| - \sum_{1 \leq |k| \leq p} \beta_k^j(x_i - x)^{|k|} \right)^2
\]

\[
\sum_{1 \leq |k| \leq p} \beta_k^D(x_i - x)^{|k|} \right)^2
\]

\[
\times K_H(x_i - x) \right)
\] (3.19)

\[
\hat{K}_H(x) \right)
\] (3.20)

Some of the notation used in (3.18) are given as follows:

\[
k = (k_1, \ldots, k_m), \, |k| = \sum_{l=1}^{m} k_l, \, |k| \in \{0, \ldots, p\},
\]

\[
k! = k_1! \times \ldots \times k_m!, \, x^k = (x^1)^{k_1} \times \ldots \times (x^m)^{k_m}
\]

\[
\sum_{1 \leq |k| \leq p} = \sum_{j=1}^{p} \sum_{k_1=0}^{j} \cdots \sum_{k_m=0}^{j} \sum_{|k|=k_1 + \cdots + k_m = j}
\]

When \( k=0 \), \( \left( \hat{\beta}_0^1, \ldots, \hat{\beta}_0^D \right)^T \) corresponds to the kernel estimator, which is the same as the estimator given in (3.3). When \( p = 1 \), \( \left( \hat{\beta}_k^1 = 0, \ldots, \hat{\beta}_k^D = 0 \right)^T \) coincides with the estimator \( \beta_0 \) in (3.7).

Finally, we have

\[
\hat{F}(x) = \hat{\beta}_0(x) = \left( \hat{\beta}_k^1 = 0, \ldots, \hat{\beta}_k^D = 0 \right)^T
\] (3.21)

\[
\hat{F}(x) = J^{-1} \left( \mathcal{P} \left( \hat{F}(x) \right) \right) = J^{-1} \left( \arg\min_{q \in \mathbb{R}} ||q - \hat{F}(x)|| \right)
\] (3.22)

Theorem 8 derives the asymptotic distribution of \( \hat{F}(x) \), with \( \hat{F}(x) \) obtained using \( p \)th order polynomials local regression of \( J(y_1), \ldots, J(y_n) \) given in (3.21).

**Theorem 8.** Let \( \hat{F}(x) \) be given in (3.22). Assume the \((p+2)\)th moment of the kernel function \( K(x) \) exists and \( \mu(x) \) is \((p+2)\)th order differentiable in a neighborhood of
Assume the projection $\mathcal{P}$ of $\tilde{\mu}(x)$ onto $\tilde{M} = J(M)$ is unique and $\mathcal{P}$ is continuously differentiable in a neighborhood of $\tilde{\mu}(x)$, where $\tilde{\mu}(x) = \mu(x) + \text{Bias}(x)$, with $\text{Bias}(x)$ given in (B.28). If $P(dy \mid x) \circ J^{-1}$ has finite second moments, then we have:

$$\sqrt{nH}d_{\tilde{\mu}(x)}\mathcal{P} \left( \tilde{F}(x) - \tilde{\mu}(x) \right) \xrightarrow{L} N(0, \tilde{\Sigma}(x)), \quad (3.23)$$

where $d_{\tilde{\mu}(x)}\mathcal{P}$ is the differential from $T_{\tilde{\mu}(x)}\mathbb{R}^D$ to $T_{\mathcal{P}\tilde{\mu}(x)}\tilde{M}$ of the projection map $\mathcal{P}$ at $\tilde{\mu}(x)$. Here $\Sigma(x) = B'\tilde{\Sigma}(x)B$, where $B$ is the $D \times d$ matrix of the differential $d_{\tilde{\mu}(x)}\mathcal{P}$ with respect to given orthonormal basis of tangent space $T_{\tilde{\mu}(x)}\mathbb{R}^D$ and tangent space $T_{\mathcal{P}\tilde{\mu}(x)}\tilde{M}$ and the $jk$th entry of $\tilde{\Sigma}(x)$ is given by (B.31). Here $\xrightarrow{L}$ indicates convergence in distribution.

**Remark** Note that order of bias term $\text{Bias}(x)$ (given in (B.28)) differs when $p$ is even (see (B.25)) and when $p$ is odd (see (B.25)).

### 3.4 Conclusion

We have proposed an extrinsic regression framework for modeling data with manifold valued responses and showed desirable asymptotic properties of the resulting estimators. We applied this framework to a variety of applications, such as the responses restricted to the sphere, shape spaces, and linear subspaces. The principle motivating this framework is that kernel regression and Riemannian geometry both rely on locally Euclidean structures. This property allows us construct inexpensive estimators without loss of predictive accuracy as demonstrated by the asymptotic behavior of the mean integrated square error, and also the empirical results. Empirical results even suggest that the extrinsic estimators may perform better due to their reduced complexity and ease of optimizing tuning parameters such as kernel
bandwidth. Future work may also use this principle to guide sampling methodology when trying to sample parameters from a manifold or optimizing an EM-algorithm, where it may be computationally or mathematically difficult to restrict intermediate steps to the manifold.
Projection Approach to Multiple Monotone Regression

Shape-constrained (e.g. monotone constrained) statistical inference is applied in a variety of applications. In environmental toxicology, for instance, monotone constraints are imposed based on natural assumptions that the response of subjects exposed to certain chemical pollutants will not in general decrease with the increasing pollutant dose (Piegorsch et al. (2014)). Another common application can be found in disease screening, where the probability of disease is assumed non-decreasing with increasing measurements of a pertinent biomarker (McIntosh and Pepe (2002); Baker (2000)). Or in economics, the demand and supply curve is in general assumed to be monotone (Chehrazi and Weber (2010)). Motivated by this large panoply of applications, a large body of statistical methods have been developed for estimating monotone curves, i.e., one-dimensional monotone functions. Frequentist methods in general fall into three categories; the first involves kernel based approaches such as described by Mueller and Schmitt (1988), Mammen (1991) and Dette et al. (2005). The second class of methods model the regression function as a linear span of a spline basis such as in
Ramsay (1988) and Kong and Eubank (2006). The third class of methods are based on isotonic regression (Robertson et al. (1988); Barlow (1972)), recent developments of which can be found in Bhattacharya and Kong (2007); Bhattacharya and Lin (2010, 2011, 2013) and Lin et al. (2015). In addition, a few Bayesian approaches have been proposed in, for example, Bornkamp and Ickstadt (2009), Lin and Dunson (2014), Shively et al. (2009) and Shively et al. (2011). Although there is a large body of work on estimating monotone curves, shape-constrained problems with respect to multiple predictors is more challenging. Multiple predictor problems are important in drug interaction studies or in risk analyses involving the joint action of multiple pollutants. The challenge is due to the difficulty in incorporating the multivariate shape constraints. Along these lines, Saarela and Arjas (2011) proposes a Bayesian approach for multiple regression using marked point processes, while Lin and Dunson (2014) combines Gaussian processes with projections for estimating a multivariate monotone function. Chernozhukov et al. (2009) applies a monotone arrangement procedure (see Hardy et al. (1952)) to an initial unconstrained estimator.

Our motivation here is to develop a theoretically appealing, computationally feasible, and convenient-to-implement approach for estimating monotone constrained functions with multiple predictors. We propose use of a projection of some initial, unconstrained estimator of the regression function, i.e., finding the monotone function closest to this initial estimator in some distance norm. Such estimates are intuitive, and can result in reduction of the estimation error compared to that of a naïve initial estimator. Note that a general projection framework for constrained functional parameters, in particular for constrained functions forming a closed convex cone in a Hilbert space, is proposed in Fils-Villetard et al. (2005). Our approach falls into this general framework; however, our projection algorithm makes use of the fact that the convex cone of multivariate monotone function is an intersection of a collection of convex cones of univariate monotone function. We then derive an expression for the
projection estimate, making use of unidimensional projections. It can be shown that
the projection of an initial estimator onto the space of monotone functions with mul-
tiple predictors can be decomposed into sequential projections of an adjusted initial
estimator along each univariate direction. This simplifies the operation substantially,
and allows us to suggest computation algorithms for approximating such functions.

In the next section, we study in detail this monotone projection framework and
the properties of our projection estimates. In section 4.2, we describe a bootstrap
methodology for constructing confidence intervals. In section 4.3, we carry out a
simulation study to explore their operating characteristics, and we apply our methods
to a two-dose, joint-action data set from environmental toxicology.

4.1 A projection framework for monotone regression with multiple
predictors

Let \( x \in \mathbb{R}^p \) be a \( p \)-dimensional predictor and \( y \) be a response variable. The response
variable can be, e.g., binary or continuous, depending on the application of interest.
We define the regression function \( F(x) \) in a general framework as

\[
F(x) = E(y \mid x).
\]  

(4.1)

For instance, if \( y \) is binary, taking values 0 or 1, then take \( F(x) \) as the response
probability \( F(x) = P(y = 1 \mid x) \).

Denote the data as \( (x_i, y_i), i = 1, \ldots, n \). Without loss of generality, we assume
the predictors or covariates \( x_i = (x_{i1}, \ldots, x_{ip}) \in [0, 1]^p \subseteq \mathbb{R}^p \). The regression function
\( F(x) \) is assumed to be monotone with respect to a natural partial ordering on \( \mathbb{R}^p \),
that is, for, \( x_1 = (x_{11}, \ldots, x_{1p}), x_2 = (x_{21}, \ldots, x_{2p}) \) and \( x_{1j} \leq x_{2j} \) (for \( j = 1, \ldots, p \)),
one has \( F(x_1) \leq F(x_2) \). We are interested in conducting inference on \( F(x) \) under
the monotonicity constraint.

Denote \( \mathcal{M} \) as the space of monotone functions on \( \mathcal{X} = [0, 1]^p \). It can be seen that
\( \mathcal{M} \) is a closed convex cone. Our approach relies on projecting an initial estimator of
For instance, the initial estimator could be a local polynomial estimator such as the popular kernel estimator or a local linear estimator. Alternatively, one could employ expansions of spline bases such as a $B$-spline basis. In any case, we will show the resulting projection estimates exhibit desirable theoretical properties as well as good finite-sample performance. Efficient computational algorithms may also be developed for implementing our approach.

To illustrate, we first consider a local polynomial regression estimator for a one-dimensional curve. Let $K(x)$ be a kernel function with $\int K(x)dx = 1$, $\int xK(x)dx = 0$ and $\int x^2K(x)dx < \infty$. Denote $K_h(x) = h^{-1}K(x/h)$. Let $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ be a $p+1$-dimensional coefficient vector, and take

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} K_h(x - x_i) \left( y_i - \sum_{j=0}^{p} \beta_j(x - x_i)^j \right)^2.$$ 

An initial local polynomial estimator of $F(x)$ can be simply

$$\hat{F}(x) = \hat{\beta}_0,$$

which can be fitted very easily via weighted least squares.

Another popular class of methods for nonparametric regression involves spline models (Zhang (2012)). More precisely, one can construct a class of initial estimators by modeling the regression function as a linear span of a spline basis, the most of popular of which is the $B$-spline basis (Eilers and Marx (2010)). Given the a knot sequence $\tau_1, \tau_2, \ldots, \tau_N$, the cubic B-spline basis $\{B_{j,l}\}_{k=1}^{4}$ is defined recursively as follows:

$$B_{j,1}(x) = \begin{cases} 1 & \tau_j \leq x \leq \tau_{j+1} \\ 0 & \text{otherwise}, \end{cases}$$

$$B_{j,l}(x) = \frac{x - \tau_j}{\tau_{j+l-1} - \tau_j} B_{j,l-1}(x) + \frac{\tau_{j+l} - x}{\tau_{j+l} - \tau_{j+1}} B_{j+1,l-1}(x) \quad (l = 2, 3, 4).$$

With this, one elects to obtain an initial estimator of $F(x)$ as a linear combination
of the spline basis. Using the $B$-spline basis above, this sets

$$F(x) = \sum_{j=1}^{N} \beta_j B_j(x), \quad (4.2)$$

where $(\beta_1, \ldots, \beta_m)$ is the vector of coefficients. The estimate is

$$\hat{F}(x) = \sum_{i=1}^{m} \hat{\beta}_i B_i(x)$$

with

$$(\hat{\beta}_1, \ldots, \hat{\beta}_m) = \arg \min_{\beta \in \mathbb{R}^m} \left( \lambda \sum_{i=1}^{n} (y_i - F(x_i))^2 + (1 - \lambda) \int (F''(x))^2 \, dx \right).$$

Here $\lambda$ is a smoothing parameter that controls the tradeoff between tighter fit to the data and smoothness of the estimates. We refer to Wahba (1990) for a general reference on smoothing splines.

For the multivariate case, one can easily obtain a kernel based initial estimator by employing a multivariate kernel $K$. Or one can obtain an initial estimator using tensor product B-splines (Eilers and Marx (2010)). For example, take $p = 2$ with the two-dimensional function $F(x_1, x_2)$. Let $\{B_{i1}(x_1)\} i = 1, \ldots, N_1$, be a B-spline basis along the $x_1$ direction, and $\{B_{j2}(x_2)\} j = 1, \ldots, N_2$, be a spline basis along the $x_2$ direction. A tensor product spline basis is given by $\{B_{i1}B_{j2}\}, i = 1, \ldots, N_1, j = 1, \ldots, N_2$. The multivariate function is modeled as a linear span of the tensor product B-spline basis. An initial estimator $\hat{F}(x_1, x_2)$ can be obtained by minimizing the objective function

$$\lambda \sum_{i=1}^{n} (y_i - F(x_{i1}, x_{i2}))^2 + (1 - \lambda) \int \int \left( \left( \frac{\partial^2 F}{\partial x_1^2} \right)^2 + 2 \left( \frac{\partial^2 F}{\partial x_1 \partial x_2} \right)^2 + \left( \frac{\partial^2 F}{\partial x_2^2} \right)^2 \right) \, dx_1 \, dx_2.$$

With any initial estimator $\hat{F}(x_1, x_2)$, one then projects $\hat{F}(x_1, x_2)$ onto the monotone space $\mathcal{M}$, which produces our ultimate estimator $F(x_1, x_2)$. We now proceed to give a rigorous definition of this projection and characterize its properties.
Let $w$ be a function on $\mathcal{X} = [0,1]^p$. We define the projection of $w$ onto the constrained space $\mathcal{M}$ as

$$P_w = \arg\min_{G \in \mathcal{M}} \int_{\mathcal{X}} \{w(t) - G(t)\}^2 dt. \quad (4.3)$$

That is, the projection estimate is defined to be the element in $\mathcal{M}$ that is closest to the initial (pre-projected function) estimator in $L^2$ distance.

Recall $\mathcal{M} = \mathcal{M}[0,1]^p$, which is the space of monotone functions on $[0,1]^p$. Focusing initially on the $p = 1$ case, (4.3) has the following closed form solution (see Lin and Dunson (2014) and Anevski and Soulier (2011))

$$P_w(x) = \inf_{v \geq x} \sup_{u \leq x} \frac{1}{v-u} \int_u^v w(t) dt, \quad x \in [0,1]. \quad (4.4)$$

The existence and uniqueness of the projection follow from Theorem 1 in Rychlik (2001).

Letting $h$ be any function on $[0,1]$, the greatest convex minorant of $h$ is defined by

$$T(h) = \arg\max \{z : z \leq h, \ z \text{ convex}\} \quad (4.5)$$

The solution (4.4) is the slope of the greatest convex minorant of $\pi(t) = \int_0^t w(s) ds$.

**Remark** The projection in (4.4) can be well approximated using the pooled adjacent violators algorithm (Barlow (1972)).

One can easily generalize the above unidimensional projection algorithm to multiple dimensions ($p > 1$). Take $p = 2$ for example: the following algorithm (Lin and Dunson (2014)) converges to the two-dimensional projection

$$P_w = \arg\min_{G \in \mathcal{M}} \int_0^1 \int_0^1 (w(s,t) - G(s,t))^2 dsdt.$$
Algorithm 1. For any fixed $t$, $w(s, t)$ is a function of $s$ and we use the projection (4.4) to obtain a monotone function in $s$. We perform this projection for all values of $t$, and denote the resulting surface by $\hat{w}^{(1)}(s, t)$. Letting $S^{(1)} = \hat{w}^{(1)} - w$, for any fixed $s$, project $w + S^{(1)}$ as a function of $t$ onto $M[0, 1]$ using (4.4). Perform this projection for all values of $s$ and denote the resulting surface by $\hat{w}^{(1)}(s, t)$. Letting $T^{(1)} = \hat{w}^{(1)} - (w + S^{(1)})$. Letting $i = 2, \ldots, k$, in the $i$th step we obtain $\hat{w}^{(i)}$ by projecting $w + T^{(i-1)}$ along the $s$ direction for every fixed $t$ value in $[0, 1]$ and $\hat{w}^{(i)}$ as the projection of $w + S^{(i)}$ along the $t$ direction for every fixed $s$ value in $[0, 1]$. The algorithm terminates when $\hat{w}^{(i)}$ or $\hat{w}^{(i)}$ is monotone with respect to both $s$ and $t$ for some step $i$.

Via an induction argument, one can show that projecting a $p$-dimensional function onto $M[0, 1]^p$ with $p > 2$ can be characterized similarly to Algorithm 1, above, by introducing $p$ residual sequences.

Given an initial estimate $\hat{F}(x)$, we denote the projection estimate of $F(x)$ under the shape constraints as $\bar{F}(x)$ with

$$\bar{F}(x) = P_{\bar{F}(x)}(x).$$

(4.6)

$\bar{F}(x)$ is the ultimate estimate used for inference.

Now, let

$$\|\bar{F}(x) - F(x)\|_2 = \left( \int_{[0, 1]^p} (\bar{F}(x) - F(x))^2 dx \right)^{1/2}$$

$$= \left( \int_0^1 \cdots \int_0^1 (\bar{F}(x) - F(x))^2 dx_1 \cdots dx_p \right)^{1/2},$$

which is the $L^2$ norm for a $p$-dimensional function $F(x)$.

The following propositions show that $\bar{F}(x)$ is ‘closer’ to the true regression monotone function $F(x)$, compared to the initial estimator $\hat{F}(x)$. As a consequence, $\bar{F}(x)$ produces smaller error in the $L^2$ norm compared to that of the initial estimator $\hat{F}(x)$. 

64
Proposition 9. Let $x \in \mathbb{R}^p$ ($p \geq 1$). Let $\hat{F}(x)$ be an initial estimator of $F(x)$ and $\hat{F}(x) = P_{\hat{F}(x)}$. Then the following holds:
\[
\|\hat{F}(x) - F(x)\|_2 \leq \|\hat{F}(x) - F(x)\|_2
\] (4.7)

For a proof, see the Appendix.

Corollary 10 follows immediately from Proposition 9.

Corollary 10.
\[
\|\hat{F}(x) - F(x)\|_2 = O(\lambda_n).
\]

where $\lambda_n \to 0$ as $n \to \infty$. Then
\[
\|\hat{F}(x) - F(x)\|_2 = O(\lambda_n).
\]

In the 1-dimensional case, we can derive more general results on the reduction in estimation error for the projection estimator.

Theorem 11. Assume $x \in \mathbb{R}$. Let $\Phi$ be any convex function and $\hat{F}(x)$ be an initial estimator of $F(x)$, such as a kernel estimator or a spline estimator. Let $\hat{F}(x) = P_{\hat{F}(x)}(x)$. One has
\[
\int_0^1 \Phi \left( \hat{F}(x) - F(x) \right) dx \leq \int_0^1 \Phi \left( \hat{F}(x) - F(x) \right) dx.
\] (4.8)

For a proof, see the Appendix.

Corollary 12. Assume $x \in \mathbb{R}$. Let $\hat{F}(x)$ be an initial estimator of $F(x)$ such as a kernel estimator or a spline estimator. Let $\hat{F}(x) = P_{\hat{F}(x)}(x)$. One has
\[
\sup_x |\hat{F}(x) - F(x)| \leq \sup_x |\hat{F}(x) - F(x)|,
\] (4.9)

and
\[
\int_0^1 |\hat{F}(x) - F(x)|^q dx \leq \int_0^1 |\hat{F}(x) - F(x)|^q dx,
\] (4.10)

where $q \in [1, \infty)$.
Proof. Taking $\phi = \|x\|^q$ in Theorem 11, (4.10) follows. As $q \to \infty$, (4.9) holds. Note that a different proof for (4.9) is given in Lin and Dunson (2014).

The following proposition concerns the asymptotic distribution of the projection, which follows from the general results of Theorem 3.4 in Fils-Villetard et al. (2005).

**Proposition 13.** Let the tangent cone of $\mathcal{M}$ be $T_{\mathcal{M}}(F)$, which is the the closure of $\{\lambda(G - F), G \in \mathcal{M}, \lambda > 0\}$. If

$$\frac{\hat{F}(x) - F(x)}{t_n} \xrightarrow{\mathcal{L}} U(x)$$

for some $t_n \to 0$ as $n \to \infty$, then

$$\frac{\tilde{F}(x) - F(x)}{t_n} \xrightarrow{\mathcal{L}} \tilde{P}_U(x),$$

(4.11)

where $\tilde{P}_U(x)$ is the projection of $U(x)$ onto the tangent cone $T_{\mathcal{M}}(F)$ and $\xrightarrow{\mathcal{L}}$ indicates convergence in distribution.

**Remark** Let $\mathcal{M}_1, \ldots, \mathcal{M}_p$ be $p$ convex cones of functions such that $\mathcal{M}_k$ is the convex cone of a function which is monotone with respect to the $k$th direction of $x$. That is, for any $F \in \mathcal{M}_k$, $F(x_1, \ldots, x_p)$ is monotone with respect to $x_k$ at any fixed value of other coordinates. It is not difficult to see that $\mathcal{M}$ is the intersection of the $p$ convex cones $\mathcal{M}_1, \ldots, \mathcal{M}_p$, i.e., $\mathcal{M} = \cap_{k=1}^p \mathcal{M}_k$. The projection algorithm (Algorithm 1) can be viewed as a sequential projection of a $p$-variate function onto each convex cone $\mathcal{M}_i$, while at each step the projected function is adjusted by adding residual sequences from the previous step. Note that similar algorithms consisting of a projection step and an adjustment step hold for any space that can be written as the intersection of a collection of convex cones (see Bauschke and Borwein (1994)).
4.2 Bootstrap confidence intervals

In this section, we appeal to the bootstrap for constructing confidence interval from our estimators. Consider the following procedure for generating a nonparametric estimate, along with confidence intervals, for a monotone function $F(x)$. The data are $(x_i, y_i)$ pairs, $i = 1, \ldots, n$.

1. Fit the curve to obtain an initial estimator $\hat{F}$ which need not satisfy the monotonicity constraint(s).

2. Project the estimate onto $\mathcal{M}$, the space of monotone functions. Call this projected estimate $\tilde{F}$, and let $e_i = y_i - \tilde{F}(x_i)$ be the residual of the $i$th point from the monotone estimate.

3. Repeat the following procedure at least $B = 2000$ times.
   
   (a) Resample the residuals with replacement, yielding $e_i^*, \ldots, e_n^*$.
   
   (b) Set $y_i^* = \tilde{F}(x_i) + e_i^*$.
   
   (c) Apply steps 1 and 2 to the bootstrapped points $(x_i, y_i^*)$ to yield $\tilde{F}^*$.

4. Collect the $G$ different bootstrapped estimates $\tilde{F}^*$. One can adopt the percentile based methods for constructing point-wise confidence intervals. For example, the lower and upper 95% confidence bounds of $F(x)$ are given by the 2.5 percentile and 97.5 percentile of the ranked bootstrapped estimates respectively.

The bootstrap procedure is applied in constructing 95% pointwise confidence intervals for the monotone regression function in Section 4.3 in a simulation study and a real data example. In Figures 4.4 and 4.5, one can see that the confidence
limits includes the true monotone regression function (in both the one- and two-dimensional case) for almost all of the \( x \) values. A empirical coverage study is carried out in a simulation study.

4.3 Simulation studies and data analysis

4.3.1 MSE study

We first carry out a simulation study on our projection estimates. We simulate data from some true monotone functions in a regression model and we consider monotone functions with one, two, and three predictors. The average root squared error across 50 trials of experiment with varying standard error are calculated and recorded in the following table. Data of size \( n = 100 \) are simulated from a normal error model with varying standard deviation. The true mean functions for the 1-d cases, which are proposed by Holmes and Heard (2003) and Neelon and Dunson (2004) are

(a) flat function, \( F_1(x) = 3, x \in (0, 10] \);

(b) sinusoidal function, \( F_2(x) = 0.32\{x + \sin(x)\}, x \in (0, 10] \);

(c) step function, \( F_3(x) = 3 \) if \( x \in (0, 8] \) and \( F_3(x) = 6 \) if \( x \in (8, 10] \);

(d) linear function, \( F_4(x) = 0.3x, x \in (0, 10] \);

(e) exponential function, \( F_5(x) = 0.15 \exp(0.6x - 3), x \in (0, 10] \);

(f) logistic function, \( F_6(x) = 3/\{1 + \exp(-2x + 10)\}, x \in (0, 10] \);

(g) half-normal function, \( F_7(x) = 3 \exp(-0.5(0.02)^2(0.1x - 1)^2), x \in (0, 10] \).

(h) mixture function, \( F_8(x) = 6F(0.1x), \) where \( F(\cdot) \) is the mixture of \( N(0.25, 0.004^2) \) and \( N(0.75, 0.04^2) \).
The results on the square root of mean squared error (RMSE) of the above models with $\sigma = 1$ and $\sigma = 0.5$, which are averaged across 50 different trials, are reported in Table 4.1 with a spline estimator or kernel estimator as an initial estimator. Overall the projection estimates with a kernel estimator as the initial estimator yield smaller root mean squared error compared to that of the projection estimates with the spline estimator as an initial estimator. The same set of models are also used in a compared study in Shively et al. (2009) with $n = 100$ and $\sigma = 1$. Our projection estimate of an initial kernel estimator produces smaller RMSE values for 5 out of 8 models. Although the to samples of data in our model are not the same as the ones used in their study, but the RMSE are relatively stable after the averaging of 50 trials. The results for with data generated from other noise levels are plotted in Figure 4.1. Example fits are plotted in Figure 4.4.

Table 4.1: [1-d functions] Root mean square error for simulated data with $n = 100$ and the results averaged across 50 simulation replicates in each case.

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>flat</th>
<th>sinusoidal</th>
<th>step</th>
<th>linear</th>
<th>exponential</th>
<th>logistic</th>
<th>half-normal</th>
<th>mixture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kernel</td>
<td>0.0485</td>
<td>0.1228</td>
<td>0.3144</td>
<td>0.1129</td>
<td>0.1155</td>
<td>0.1475</td>
<td>0.0470</td>
<td>0.3045</td>
</tr>
<tr>
<td>Spline</td>
<td>0.0573</td>
<td>0.1255</td>
<td>0.3209</td>
<td>0.0813</td>
<td>0.1060</td>
<td>0.1369</td>
<td>0.0570</td>
<td>0.3180</td>
</tr>
<tr>
<td>$\sigma = 1$</td>
<td>flat</td>
<td>sinusoidal</td>
<td>step</td>
<td>linear</td>
<td>exponential</td>
<td>logistic</td>
<td>half-normal</td>
<td>mixture</td>
</tr>
<tr>
<td>Kernel</td>
<td>0.0858</td>
<td>0.2217</td>
<td>0.3875</td>
<td>0.2009</td>
<td>0.1895</td>
<td>0.2438</td>
<td>0.0950</td>
<td>0.3893</td>
</tr>
<tr>
<td>Spline</td>
<td>0.1102</td>
<td>0.2307</td>
<td>0.4048</td>
<td>0.1631</td>
<td>0.2179</td>
<td>0.2430</td>
<td>0.1166</td>
<td>0.4092</td>
</tr>
</tbody>
</table>
The two dimensional true monotone functions for simulation, which were first considered in Saarela and Arjas (2011), are given in the following. Data of size 100 are generated with varying noise levels, and $(x_1, x_2) \in [0, 1] \times [0, 1]$. Example Bootstrap confidence interval estimates are plotted in Figure 4.5.

(a) $F_1(x_1, x_2) = \sqrt{x_1}$;

(b) $F_2(x_1, x_2) = 0.5x_1 + 0.5x_2$;

(c) $F_3(x_1, x_2) = \min(x_1, x_2)$;

(d) $F_4(x_1, x_2) = 0.25x_1 + 0.25x_2 + 0.5 \times 1_{\{x_1+x_2>1\}}$;

(e) $F_5(x_1, x_2) = 0.25x_1 + 0.25x_2 + 0.5 \times 1_{\{\min(x_1,x_2)>0.5\}}$;

(f) $F_6(x_1, x_2) = 1_{\{(x_1-1)^2+(x_2-1)^2<1\}} \times \sqrt{1 - (x_1 - 1)^2 - (x_2 - 1)^2}$.

We also simulated data from models whose true regression functions are three-dimensional monotone functions, where $x_1, x_2, x_3 \in [0, 10]$, and the RMSE results for $\sigma = 0.5$ and $\sigma = 1$ are included in Table 4.3. The RMSE estimates more noise values are plotted in Figure 4.3.

(a) $F_1(x_1, x_2, x_3) = 0.15(x_1 + x_2 + x_3)$;

(b) $F_2(x_1, x_2, x_3) = 0.5x_1x_2x_3$;

(c) $F_3(x_1, x_2, x_3) = \min(x_1, x_2, x_3)$;

(d) $F_4(x_1, x_2, x_3) = 1/(1 + \exp(-(x_1 + x_2 + x_3)))$;

(e) $F_5(x_1, x_2, x_3) = \exp(0.01x_1 + 0.1\sqrt{x_2}) + \sin(x_3/5)$;

The RMSE results for the estimates in models generated with other noise levels are plotted in Figure 4.2.
Table 4.2: [2-d functions] Root mean square error for simulated data with $n = 100$ and the results averaged across 50 simulation replicates in each case.

<table>
<thead>
<tr>
<th>$\sigma = 0.5$</th>
<th>$F_1$</th>
<th>$F_2$</th>
<th>$F_3$</th>
<th>$F_4$</th>
<th>$F_5$</th>
<th>$F_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kernel</td>
<td>0.1125</td>
<td>0.1274</td>
<td>0.1356</td>
<td>0.1783</td>
<td>0.1530</td>
<td>0.1687</td>
</tr>
<tr>
<td>Spline</td>
<td>0.1153</td>
<td>0.1439</td>
<td>0.1360</td>
<td>0.1929</td>
<td>0.1768</td>
<td>0.1715</td>
</tr>
<tr>
<td>$\sigma = 1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kernel</td>
<td>0.1923</td>
<td>0.1874</td>
<td>0.2175</td>
<td>0.2551</td>
<td>0.2601</td>
<td>0.2491</td>
</tr>
<tr>
<td>Spline</td>
<td>0.2383</td>
<td>0.2484</td>
<td>0.2514</td>
<td>0.3036</td>
<td>0.2861</td>
<td>0.2849</td>
</tr>
</tbody>
</table>

**Figure 4.2:** RMSE for 2-dimensional projection estimates

### 4.3.2 Bootstrap confidence intervals

Following the procedures described in Section 4.2, we construct confidence limits for in one- and two-dimensional examples. Here are some plots of the 95% bootstrap confidence intervals for some simulated data for the 1-d and 2-d functions. See Figure 4.4 and Figure 4.5. The 95% bootstrap confidence intervals almost always contain the true monotone regression function.

**Figure 4.3:** RMSE for 3-dimensional projection estimates
Figure 4.4: Bootstrap confidence interval estimates for sinusoidal ($F_2$) and logistic ($F_6$) functions. The black line is the estimate, while the blue lines give the confidence intervals. The true function is given in green, while the training points are given in black.

Figure 4.5: Bootstrap confidence interval estimates for $F_1$ and $F_6$ functions. The estimate is given as the black plane, while the confidence intervals are given by the blue planes. The true function is given in green, while the training data is given in red.
Table 4.3: [3-d functions] Root mean square error for simulated data with $n = 100$ and the results averaged across 50 simulation replicates in each case.

<table>
<thead>
<tr>
<th>$\sigma = 0.5$</th>
<th>$F_1$</th>
<th>$F_2$</th>
<th>$F_3$</th>
<th>$F_4$</th>
<th>$F_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kernel</td>
<td>0.0841</td>
<td>0.1677</td>
<td>0.1616</td>
<td>0.1590</td>
<td>0.2249</td>
</tr>
<tr>
<td>Spline</td>
<td>0.2078</td>
<td>0.2216</td>
<td>0.2138</td>
<td>0.2600</td>
<td>0.2857</td>
</tr>
<tr>
<td>$\sigma = 1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kernel</td>
<td>0.1846</td>
<td>0.2586</td>
<td>0.2714</td>
<td>0.2511</td>
<td>0.3687</td>
</tr>
<tr>
<td>Spline</td>
<td>0.3449</td>
<td>0.4232</td>
<td>0.4108</td>
<td>0.5119</td>
<td>0.5454</td>
</tr>
</tbody>
</table>

An empirical coverage study of the bootstrap procedure is carried out. In the study, 2000 samples of data are first simulated from some models. Bootstrap confidence intervals are constructed for each of the sample data. At any fixed point, among these 2000 confidence intervals constructed, we record the percentage of the confidence intervals that contains the true regression function evaluated at that point. The results are recorded in Table 4.4 for two one-dimensional functions. We studied the empirical coverage for 10 points of the one-dimensional regression function.

Table 4.4: Empirical coverage of the bootstrap procedure for $F_1$ at evaluated at 10 covariate levels for two one-dimensional functions.

<table>
<thead>
<tr>
<th></th>
<th>0.5</th>
<th>1.5</th>
<th>2.5</th>
<th>3.5</th>
<th>4.5</th>
<th>5.5</th>
<th>6.5</th>
<th>7.5</th>
<th>8.5</th>
<th>9.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1\sigma=1.0$</td>
<td>82.4</td>
<td>97.0</td>
<td>97.6</td>
<td>95.0</td>
<td>85.9</td>
<td>98.1</td>
<td>94.2</td>
<td>85.2</td>
<td>78.3</td>
<td>96.2</td>
</tr>
<tr>
<td>$F_1\sigma=0.5$</td>
<td>70.8</td>
<td>95.5</td>
<td>88.7</td>
<td>95.2</td>
<td>83.2</td>
<td>83.4</td>
<td>88.1</td>
<td>81.8</td>
<td>81.2</td>
<td>99.1</td>
</tr>
<tr>
<td>$F_2\sigma=1.0$</td>
<td>98.1</td>
<td>98.5</td>
<td>90.6</td>
<td>79.8</td>
<td>77.4</td>
<td>75.4</td>
<td>84.7</td>
<td>86.8</td>
<td>98.1</td>
<td>99.7</td>
</tr>
<tr>
<td>$F_2\sigma=0.5$</td>
<td>94.4</td>
<td>97.3</td>
<td>96.8</td>
<td>69.5</td>
<td>68.1</td>
<td>76.0</td>
<td>75.1</td>
<td>87.2</td>
<td>97.5</td>
<td>97.3</td>
</tr>
</tbody>
</table>

4.3.3 Real data analysis

An applied example of monotonic regression occurs in quantitative risk assessment for exposure to hazardous chemical or biological agents. For our example we use a data set of the rates of cellular damage after being exposed to dichlorodiphenyl-trichloroethane (DDT) and titanium dioxide nanoparticles found in Deutsch and Piegorsch (2012). Since increasing exposure to these hazardous agents are expected to cause increasing cellular damage, monotonic regression is a valid method.
Applying the kernel estimate of monotonic regression and bootstrapping we can obtain an estimate of cellular damage as a function of DDT and TiO\(_2\). With 2,000 bootstrap samples, the 95% bootstrap confidence interval and point estimate for the function is plotted in Figure 4.6. The estimate shows that the effect of TiO\(_2\) is slight, but being exposed to a concentration of DDT greater than 0.05 µmol/L results in a large jump in cellular damage, such that it appears to be essentially a step function.

4.4 Discussion

We have proposed a multivariate monotone constrained regression algorithm based on kernel and spline initial estimates. Monotone estimates are inferred via repeated projections of the initial estimates onto one dimensional monotone functions. We also outline a procedure for finding point-wise confidence intervals. The accuracy of the estimates and coverage of the confidence intervals are empirically demonstrated.
through simulation study. The method is then briefly applied to a small toxicology data set.

A potential area for future work is to establish functional confidence intervals as opposed to point-wise confidence intervals. An obvious problem with point-wise confidence intervals is that the 95th percentile line may look quite different than any of the functions that appeared in the bootstrap sample.
An ubiquitous application in “Big Data” is the analysis of a large body of text, or corpus. Topic models are a diverse set of tools used in the analysis of a corpus that are united by an admixture of latent features called “topics”. Topic models can be used by sociologists to study vocabulary expression as it varies by region (Yang et al. (2013)), or as it varies over time (Blei and Lafferty (2006)). It’s generally useful for finding clusters of documents and interpreting themes as a powerful exploratory tool (Griffiths and Steyvers (2004)). In this chapter, first we focus on Latent Dirichlet Allocation (LDA), which models text and documents as a set of multinomial counts or “bag of words”. We also look at how to create a statistical experiment using Natural Language (or Zipfian) priors, the performance of ensemble models in this arena, and apply the Conway Embedding to create a powerful non-parameteric extension of Spherical Topic Models.

LDA is widely employed across the topic modeling literature and in practice. As
an admixture model, it can be stated as
\[
\begin{align*}
\phi_k &\sim \text{Dir}_V(\beta) \quad k \in \{1, \ldots, K\} \\
\theta_d &\sim \text{Dir}_K(\alpha) \quad d \in \{1, \ldots, D\} \\
z_w &\sim \text{Multinom}_K(\theta_d) \quad d \in \{1, \ldots, D\}, w \in \{1, \ldots, N_d\} \\
v_w &\sim \text{Multinom}_V(\phi_z) \quad d \in \{1, \ldots, D\}, w \in \{1, \ldots, N_d\}
\end{align*}
\]

where \( K \) is the number of latent factors or “topics”, \( V \) is the number terms in the model, and \( N_d \) is the number of terms in the \( d^{th} \) document. A geometric interpretation of the model is learning \( K \) points on the \( V-1 \)-simplex parameterizing the convex hull of the observed documents, or the *corpus*. Many variants of LDA have developed over the last decade providing a flexible family of topic models. Some include frequently available metadata about documents, such as the time of the publication in Dynamic Topic Models (Blei and Lafferty (2006)), or the author and location of the publication (Blei (2012)).

The Dirichlet (and Dirichlet Process) priors lead to two problems we will address. The first is in simulating data from the priors for simulation studies; the second is posterior inconsistency in the number of topics (Womack et al. (2013)). In this chapter we propose a topic model based on the hierarchical Bayesian structure of LDA, but incorporate priors that are reflective of natural language distributions of terms. Our goal is to establish realistic synthetic term frequencies to help evaluate how standard topic models perform on corpora with different numbers of documents, and documents of varying length.

To compare different topic model methodologies there are several common evaluations for goodness of fit for topic models (Chang et al. (2009), Inouye et al. (2014)). These either rely on human judgement of topic coherence, or comparing word frequencies to documents internal or external to the corpus. These metrics are useful
for assessing fit in a predictive setting, but little is understood about how well admixture models perform inference. Additionally, they fail to address common statistical modeling questions including understanding the effect of sample size on bias. In this chapter we propose a method for generating corpora that have realistic term frequency distributions.

We also present a method for improving LDA estimates in a straight-forward and intuitive manner by using an ensemble of LDA model estimates produced from collapsed Gibbs sampling. This allows us to fit a more flexible prior that reflects a natural language law known as Zipf’s law, and approximates $K$ without a high computational cost. We show that our model ensures topic coherence if they can be learned from the LDA framework.

Finally, we show another example of the utility of the Conway Embedding, an embedding of subspaces of different dimension into Euclidean space, for model construction. The Spherical Admixture Model (SAM) is briefly introduced and a prior on the Stiefel manifold is discussed. With a prior on the Stiefel manifold established, we use the Conway Embedding to quickly specify a joint prior on the topics and the number of topics, effectively creating a non-parametric version of SAM.

5.1 Limitations of LDA

The natural language properties of real word data make studying the properties of LDA difficult. In practice it is impossible to have a data set where the topics are known beforehand. The inability to have data where the “right” answers are known a priori limits researchers’ ability to ascertain accuracy of inference. Instead LDA is measured by its predictive performance, however this is not informative of LDA’s properties as a statistical model. For example, performance in terms of parameter MSE is a standard statistical metric.

There are several methods for evaluating the goodness of fit of LDA models
(Chang et al. (2009), Inouye et al. (2014)), which rely on human judgment for the coherence of topics or comparing word frequencies to other documents, internal or external to the corpus. One popular method uses a metric called perplexity, which prioritizes the recovery of corpus wide term frequencies (Asuncion et al. (2009)). To compute perplexity, a document is held out of the fitting process. Half of the words in the document are used to evaluate the most likely set of topic proportions, and the distribution of the second half of the words are compared to the expected distribution based on the model estimate of topics. This approach ignores whether the topics are coherent, so loses its usefulness in social science applications and LDA’s use as an exploratory tool.

Topic modelers interested in interpretability alternatively use the “Intruder Test” (Chang et al. (2009)). Here a human is shown four high probability words from a topic and one low probability word from the same topic, and is tasked with finding the word that does not belong on the list. Similarly, four documents associated with the topic with high probability are shown, and one document with low probability are shown. The human is tasked with locating the “intruder” document. This has been a gold standard for assessing topic coherence, but is not scalable and has unknown asymptotic behavior as a performance metric. Another set of methods that require human intervention involves tagging documents manually. The tags are taken to be a proxy for the latent topics. Metrics are constructed for how closely the resulting model’s topics are to the topical tags of the documents. These metrics are useful for assessing fit on a particular data set, however since the true parameters are never known in these cases, it is difficult to assess the risk of new topic models to guide the development of topic modeling methodology. Mean Squared Error studies would be useful in these instances.

We propose attempting to generate corpora that are realistic as possible to understand how changes in the properties of parameter inputs effect model fit. This
process itself helps understand what appropriate prior parameters might be. The key principle is that synthetic data from Monte Carlo simulation allows researchers to know and control for factors in a way that is impossible with real data.

5.2 LDA and Zipf’s Law

Unfortunately, drawing from the Dirichlet distributions commonly used as priors in LDA do not produce corpora following Zipf’s law, a core empirical property of natural language. Zipf’s law states that the term frequency distribution of any corpus of language in any language or context, follows a power law distribution. The conventional wisdom is that this property holds only for documents of sufficient length and that the lowest-frequency words in the corpus do not follow the power law. To the latter point, however, in a 2002 paper, Ha et al. show that the inclusion of compound words holds Zipf’s law through the lowest frequencies in the corpus. From Goldwater et al. (2011), “[Zipf’s law] has been largely ignored in modern machine learning and computational linguistics.” Yet they point out that “it is important for models used in unsupervised learning to be able to describe the gross statistical properties of the data that they are intended to learn from. Otherwise, these properties may distort inferences about the parameters of the model.”

Consider the two examples below. In each case, the Dirichlet distributions used as priors from two well-known LDA papers are used to generate corpora in a process described in the previous section. In both cases, the simulated term frequencies do not follow a power law distribution.

Griffiths and Steyvers wrote a highly-cited paper in 2007 that introduced Gibbs sampling as a method of estimating an LDA model. They chose $\alpha = 50/K$ and $\beta = 0.1$ as the symmetric parameters of the Dirichlet distributions for topics in documents and words in topics. This is a point mass prior on each of the concentration parameters which results in flat topics, so it is not surprising when the corpus wide
term frequencies do not fit a power law curve.

In 2009, Mimno et al. propose an asymmetric $\alpha$ and symmetric $\beta$, “as a new standard for topic modeling.” Since they optimize $\alpha$ to their data, they do not specify the exact type of asymmetry. Using a common Gamma prior on the Dirichlet hyper parameters such that we retrieve the same mean of $50/K$, but with shape and rate parameters of shape $a = 50/K$ and rate $b = 1$. The key difference is that now the concentration parameters come from a power law like distribution instead of being fixed points as seen in Figure 5.1.

![Gamma hyperprior Density](image1)

![Symmetric Prior Density](image2)

**Figure 5.1:** The prior densities to produce concentration parameters for standard LDA models. The Gamma shape parameter is $50/300$ and the rate parameter is $1$, to center the topic distribution on the proposed symmetric Dirichlet prior by Steyvers and Griffiths (2007), but still provide a power-law like shape in values.

The hope would be that using Mimno et al.’s approach that the generated corpus would reflect observed corpus term frequencies accurately, however the observed performance is not significantly different from that of using the point mass prior as seen in Figure 5.2.

Neither of these examples generates corpora with Zipfian term frequency distributions. While such corpora can still be used to estimate LDA models, they are of limited utility. When the distribution of the input data is significantly different than what is observed in theory the behavior of the estimator may be suboptimal.
Fortunately, it is possible to generate simulated corpora with term frequencies that follow Zipf’s law to help evaluate these properties.

5.3 Comparison of Synthetic Data to NSF Grant Award Abstracts

To demonstrate using the principals of Zipf’s law to generate counts similar to real data, we compare the counts sampled from an LDA models with power laws to the counts found in a data set of NSF grant award abstracts. All NSF data can be found online from the NSF website http://www.nsf.gov/funding/.

We use 13,092 award abstracts from 2010. The data was tokenized using the Mallet package with extracting bigrams. Across the abstracts there are 1,160,115 terms, with an average document length of 721. The grant awards are mostly related to cell biology and environmental sciences, with awards from a wide variety of other fields such as computer science, statistics, and materials science as well. From the word frequency distribution, we estimate a power law of approximately 1.1.

To generate data comparable to the NSF grant awards, we use the information
we gain exploring the data set to set the parameters. To generate the documents, we use the following model, with $V = 1,160,115$, $D = 13,092$, and we speculate that $K = 40$.

Our Monte Carlo-generated corpus is specified as follows:

\[
\begin{align*}
\beta_v & \sim \text{Pareto}(1, 1.1) & v \in \{1, \ldots, V\} \\
N_d & \sim \text{Pois}(721) & d \in \{1, \ldots, D\} \\
\phi_k & \sim \text{Dir}_V(\beta/||\beta||_1) & k \in \{1, \ldots, K\} \\
\theta_d & \sim \text{Dir}_K(1/40_K) & d \in \{1, \ldots, D\} \\
z_{d,w} & \sim \text{Multinomial}(\theta_d) & d \in \{1, \ldots, D\}, w \in \{1, \ldots, N_d\} \\
v_{d,w} & \sim \text{Multinomial}(\phi_{z_w}) & d \in \{1, \ldots, D\}, w \in \{1, \ldots, N_d\}
\end{align*}
\]

We also generate a data set using symmetric Dirichlet distributions, the only difference being using $1/V$ for the concentration parameter on the Dirichlet distribution generating the topics. We compare the two synthetic term frequencies to the observed frequencies in the NSF award abstract data in Figure 5.3.

The synthetic data generated from the power law generates the term frequencies significantly better than the symmetric Dirichlet generated data on the corpus wide level. Although we cannot duplicate individual documents directly, we can compare the shape of relative term frequencies in individual documents. These comparisons are illustrated in Figure 5.4. We observe that the shape still holds, although there appears to be less variability in the distribution shape for synthetic documents.

The implication of being able to generate data similar to real data, is we can validate statistical methodology on synthetic data by seeing how well true parameters can be recovered. In addition to recovering the topics and topic proportions well, we are also interested in estimating the number of topics.
5.4 Inference of Topic Models Using Zipfian Priors and Ensemble Estimation

We also attempt to use power law priors to improve topic estimation. A problem that arises immediately is in prior-specification, since any power law prior implicitly places a higher a priori frequency on some terms than others, when no such prior knowledge typically exists (or can be provided in an automatic way). For example, a naive ordering of the terms may be indexing them in alphabetical order. Placing a power law prior, which is a function of the index, would suggest a prior belief that terms that start with “a” will occur drastically more frequently than terms that start with “z”. Even if the terms are ordered by their frequency in the document, so that in aggregate the power law holds, the true topics will contain only a subset of the terms, while the rest of the top terms in aggregate have extremely low frequency.

A strong solution to this problem is to integrate the model posterior over the set of all power law priors. This is prohibitively computationally expensive. We instead
Figure 5.4: Document term frequencies ranked and plotted by the log the sorting index and the log of the term frequency. The black is NSF data while red is power law distributed data. There is notably less variability in the synthetic data, however the shapes of the distributions are similar throughout.
propose a weak solution to this problem. We approximate the integral by averaging over many models with different power law priors.

We use an ensemble model for an approximate estimate for the true number of topics. We also observed that our ensemble estimates are useful in selecting coherent and informative topics for the corpus analyzed. This provides an easy to implement and intuitive method to select interesting topics and does not require choosing the number of topics.

The benefits of techniques such as model averaging have been well documented in both machine learning and statistics literatures. The model average may perform better than individual models because each model captures a different part of a complex problem, or because the model fit is approximate and therefore the model average is an estimate more robust to the fitting algorithm. In topic modeling, model averaging has been shown to be superior in unsupervised LDA models in terms of perplexity compared to stand alone LDA models (Nguyen et al. (2014)).

There have been a variety of approaches to approximate an LDA model, including Gibbs sampling (Kuzmenko (2011)), variational Bayesian inference, and collapsed Gibbs sampling (Blei and Lafferty (2009)). In collapsed Gibbs sampling, the topics and topic proportions are integrated out, and the topic assignments of each token in the corpus is sampled.

For our ensemble model, each model is specified in the following way. Given a power law parameter $\rho$, and $D$ documents of length $N_d$ respectively,
\[
\beta_i = i^\nu \\
\phi_i = \text{Permute}(\beta_i) \\
\theta_d \sim \text{Dir}_K(\alpha) \\
z_w \sim \text{Multinomial}(\theta_d) \\
v_w \sim \text{Multinomial}(\phi_w) \\
\]

Notice that with no prior knowledge, a permutation over the vocabulary essentially provides us a way of searching a broader part of the topic parameter space, even if we do not permute the power law into the “true” topics.

Each model \( M \) in the ensemble \( \mathbb{M} \) is estimated through collapsed Gibbs sampling. Since collapsed Gibbs sampling relies on allocating each observed token in the corpus to a topic, but does not require storing any other model parameters, this was the method that performed best with speed and quality of estimates given our data size. In larger applications, Gibbs sampling may yield faster speed since a table of counts and model parameters becomes more memory efficient at larger corpus sizes. Since the model will be estimated repeatedly, the algorithm choice is essential in time sensitive applications. After a burn-in period, the topics and topic proportions are taken from the last iteration in the chain for each model.

For some index \( c \), we take the models \( M_c \in \mathbb{M} \). The \( k \)th topic and the \( d \)th topic proportions in model \( c \) are denoted \( \phi_k^{(c)} \) and \( \theta_d^{(c)} \) respectively. During model fitting, it is not required that \( \phi_k^{(x)} \) and \( \phi_k^{(y)} \) are related. Therefore in order to make an ensemble estimate, it is required to determine what topics in which models are “about” the same things. To find these pairings, we use a clustering algorithm.

We perform a \( K \)-means clustering algorithm, where \( K \) is the ensemble size \( |\mathbb{M}| \). The motivating idea behind fitting a \( K \)-means algorithm here is that if the true
number of distinct topics $\kappa$ is the same as the number of topics fit in the model, and there is enough information in the documents to estimate these topics, then the $K$-means clustering should return $\kappa$ clusters, with each cluster containing $|M|$ members. An alternative method is to explore hierarchical clustering methods.

Once we have clusters of topics, there are a few approaches to getting topics for the ensemble estimate. One option is to use all the clusters from the $K$-means clustering algorithm, however in practice some of the clusters contain low quality topics. For example, Figure 5.5 illustrates a typical output of the clustering algorithm. We choose to only make topic estimates from clusters that have a number of members within a given threshold of $|M|$. We chose clusters with $|M| \pm 0.2|M|$ members and discard other clusters. Clusters that exceed the upper threshold contain uninformative terms, essentially corpus specific stop words. Clusters that fall short of the lower threshold are overly specific to one document, or a product of a model with a particularly malformed prior generated through the permutation step.

The selection of topic clusters is an important step in providing only high quality and meaningful topics in contexts where interpreting topics is a goal of the user. This procedure does not optimize perplexity measures or making predictions for text data, since the smaller and larger clusters still contain information about the corpus. It instead favors topic coherence and accurate inference on the identified topics.

Given the set of topic clusters, there are multiple ways to combine the topics and topic proportions to retrieve the ensemble estimate. A simple way is to take the centroid of each retained cluster for the topic. Instead, we took a weighted average of the topics in the cluster, where the weights were proportional to a performance metric of the model. The performance metric used was the Jensen-Shannon divergence between the observed and expected corpus wide term frequencies given each model.

There are two methods for aggregating topic proportions across models. Topic proportions can be estimated relatively quickly by solving the matrix problem $D\hat{\Phi}^{-1} = \ldots$
\( \hat{\Theta} \) where \( \mathcal{D} \) is the \( D \times V \) document term matrix, \( \hat{\Phi}^{-1} \) is the \( V \times k \) generalized inverse for the topic estimate, and \( \hat{\Theta} \) is the desired topic proportions estimate. Topic proportions can also be aggregated to get an ensemble estimate in a manner similar to the topic estimates by averaging the topics proportions over the documents. Since some of the topics are discarded, using the latter method requires that the topic proportions be rescaled such that the topic proportions sum to one. We choose the latter method for aggregation.

5.5 Ensemble Estimation on NSF Grant Award Abstracts

We applied our ensemble method to a set of 13,092 NSF grant award abstracts from 2010. We fit our model with an ensemble size \( |\mathcal{M}| = 100 \), each with a number of topics \( k = 50 \). The topics were clustered using a K-means clustering algorithm, returning 50 clusters. The discarded clusters were those containing more than 120 members or fewer than 80 members. That threshold resulted in 41 of the clusters being retained to give topic estimates. The clustering of the topics can be seen in Figure 5.5.

The topics that are discarded tend to be of low quality and not informative for the corpus wide word distributions. Upon visual inspection, the larger clusters are centered around topics with high probability on statistically independent words. Example word clouds are illustrated in Figure 5.6.

Many of the discarded topics offer little information about the corpus as a whole, which is reflected in their topic proportions. These low quality, low proportion topics may cause overfitting problems when an individual model is used to make out of sample predictions about word frequencies and therefore hurts performance by perplexity or similar measures. Discarding these topics provides a sparser and more robust estimate. The corpus wide topic proportions of both the discarded and retained topics are compared in Figure 5.7.
The resulting cluster of topics from an ensemble of 100 models. Each model contains 50 topics, so 5000 topics are clustered using $K$-means clustering, with the number of clusters set to 50. Clusters with a number of members far away from 100 members indicate that the member topics are of low quality since they tend to be too generic, or overly specific to a small number of documents.

The quality topics we retained can be inspected to show that they are more related to a theme in the set of grant awards. Example word clouds are shown in Figure 5.8. An important application that this ensemble selection of topics to retain is that the topics tend to be the more coherent topics. This is valuable in settings where a small group of analysts must work with a large number of documents, and the topics are used to discover themes of interests.

5.6 Synthetic Data Comparison

Although it is known that Latent Dirichlet Allocation and its non-parametric relative, the Hierarchical Dirichlet Process, are inconsistent in the number of topics (Womack et al. (2013)), there are few scalable alternatives to estimate the “true” number of topics that generated data. Although our model averaging technique does
Figure 5.6: Word clouds of large (more than 120 members) clusters. These topics put high probability mass on words that are independent. The prevalence of these words in every document makes the presence of any one word uninformative of the others, i.e. $P(\text{science}|\text{education}) = P(\text{science})$ in a corpus of NSF grant awards.

Figure 5.7: The corpus wide topic proportions of the retained and discarded topic. Of the topics, 68% were retained and 32% were discarded as junk topics.
not provide any theoretical basis for arguing that we have a consistent estimator for the number of topics, we believe it can provide a useful approximation. To see if we could get close to the true number of topics, we need to generate data with a known number of topics. However, since LDA with symmetric priors does not generate realistic text data, we invoke Zipf’s law.

To generate different synthetic data sets to test the properties of the ensemble method, we used the principles of Zipf’s law above to generate data that was comparable to the word frequencies seen in the NSF grant award data by setting the Dirichlet parameter to a power law with parameter 1.4 to match the NSF grant award data. We generated 14,000 documents of an average length of 700 terms from 40 topics \((k = 40)\) with 1.2 million terms. Using that Dirichlet parameter on topics and a symmetric Dirichlet on topic proportions, we generated synthetic documents with a true number of topics \(k = 40\). The number of topics estimated from fitting an ensemble on 15 different data sets is given below. Each model had symmetric priors.
on both topics and topic proportions. Five were fit with \( k = 50 \), five with \( k = 60 \), and five with \( k = 70 \). The results are in Table 5.1.

Table 5.1: The number of estimated topics from the proposed ensemble method on 15 different synthetic corpora with a true number of topics \( k = 40 \). The clustering method reliably reduces the number of topics to approximate the true number.

<table>
<thead>
<tr>
<th>Estimated Number of Topics</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>2</td>
</tr>
<tr>
<td>38</td>
<td>2</td>
</tr>
<tr>
<td>39</td>
<td>6</td>
</tr>
<tr>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>41</td>
<td>3</td>
</tr>
</tbody>
</table>

Based on the synthetic results, it seems that this model works well in approximating \( k \) for text data, or data that follow power law distributions. The intuition driving the idea is that the true topics generate a strong signal in the data, but a small group of words dominate the corpus. When the model is learning an “extra” topic, it fits the small group of words that dominate the corpus – words towards the top of the power law distribution. This causes the large cluster sizes that can be discarded. When the data is symmetric, there is not a small group of words dominating the corpus, and so the model is more likely to learn the “extra” topic as any single one of the 40 true topics, so in the clustering phase almost no clusters are discarded.

5.7 Spherical Admixture Model with Orthogonality Constraint

Now we discuss another application in topic modeling concerning estimating the true number of topics. We first recast a topic model with topic priors over the unit sphere, originally proposed in Reisinger et al. (2010), onto the Stiefel manifold therefore inducing an orthogonality constraint between topics. Next, we use the tools developed in Chapter 2 to give a model specification with a joint prior on the topics and the number of topics.
5.7.1 Generative Model on the Stiefel

The Spherical Admixture Model (SAM), instead of modeling the documents as multinomial counts, models the documents as unit vectors. This allows for more flexible processing of the vocabulary (e.g. using tf-idf weighting), and also allows modeling the data with the von Mises-Fisher (vMF) distribution. The vMF can be thought of as a directional distribution based on the cosine distance, and a key advantage of the model is that documents which are equiprobable under the multinomial model have different probabilities under a vMF distribution. For example, if the probability of each word occurring is $1/n$, all documents of length $n$ are equiprobable under a multinomial model like LDA. However under the vMF distribution a document with one word repeated $n$ times will have a different density than a document with $n$ words with one occurrence.

For a vocabulary of size $V$ over $D$ documents with $K$ topics, SAM is a Bayesian admixture model of documents on $S^{V-1}$, and uses a weighted directional average of the topic matrix $\phi$, whose columns are topics $\phi_t$, to compose the documents. Documents are generated as $l_2$-normalized unit vectors $v_d$ through the following generative process

$$
\mu | \kappa_0 \sim \text{vMF}_V(m, \kappa_0) \quad \text{(corpus average)}
$$

$$
\phi_k | \mu, \xi \sim \text{vMF}_V(\mu, \xi), \quad k \in K \quad \text{(topics)}
$$

$$
\theta_d | \alpha \sim \text{Dir}_K(\alpha), \quad d \in D \quad \text{(topic proportions)}
$$

$$
\bar{\phi}_d | \phi, \theta_d = \text{Avg}(\phi, \theta_d), \quad d \in D \quad \text{(spherical average)}
$$

$$
v_d | \bar{\phi}_d, \kappa \sim \text{vMF}_V(\bar{\phi}_d, \tau) \quad d \in D \quad \text{(generate document)}
$$

Here, the spherical average is computed $\text{Avg}(\phi, \theta_d) = \frac{\phi \theta_d}{||\phi \theta_d||}$, which is not the Buss-Fillmore spherical average but has the advantage of not being computed iteratively.
In most Bayesian topic models, including SAM, topics are drawn independently of each other from some distribution. In principle information is shared through the hierarchical structure to differentiate the topics, however in practice there are many common terms that can overwhelm the likelihood causing topics to be redundant. Strategies to eliminate this problem is to reference a list of stop words from the corpus such as words like “such”, “as”, and “and.” However there are often corpus specific stop words that are not included in general reference lists. For example in a corpus of documents about topic modeling, words like “topic” is likely to be uninformative and worth removing.

Introducing an orthogonality constraint on the topics can enforce the prior knowledge that they should be interpretable as distinct. This can be introduced by making an adjustment to the SAM generative process. Instead of being generated independently from a vector vMF distribution, the topics are generated from a matrix vMF distribution on the Stiefel manifold ($V_{V,K}$, the set of all $V \times K$ matrices with orthonormal columns). This generative process can be given by

$$
\mu|\kappa_0 \sim \text{vMF}_V(m, \kappa_0) \quad \text{(corpus average)}
$$

$$
\phi|\mu, \xi \sim \text{vMF}_{V \times K}(\mu, \xi) \quad \text{(orthogonal topics)}
$$

$$
\theta_d|\alpha \sim \text{Dir}_K(\alpha), \quad d \in D \quad \text{(topic proportions)}
$$

$$
\tilde{\phi}_d|\phi, \theta_d = \text{Avg}(\phi, \theta_d), \quad d \in D \quad \text{(spherical average)}
$$

$$
v_d|\tilde{\phi}_d, \kappa \sim \text{vMF}_V(\tilde{\phi}_d, \tau) \quad d \in D \quad \text{(generate document)}
$$

By specifying only a vector as the corpus average, we imply that the orthogonal topics prior parameter is a matrix where each column is the corpus average. Since the only difference then between the two models is the orthogonality constraint, the sampling scheme for the posterior distribution does not change from SAM except for the sampling scheme of the topics. When drawing the topics from the SAM, the
draws are constrained to the null space of the rest of the topics, a strategy outlined by Hoff (2009). The procedure for estimating SAM with the orthogonality constraint is that instead of updating the full set of topics simultaneously, the update rule becomes the following

1. For each $k \in \{1, \ldots, K\}$ in random order:
   
   (a) Construct $N_r$ as an orthonormal basis for the null space of $\phi_{[1, -r]}$

   (b) Update $\zeta$ with topic mean parameter $N^T_r \bar{\mu}_{[r]}$ as in Reisinger et al. (2010)

   (c) Set $\phi_{[r]} = N_r \zeta$.

Note that because in these topic models that we always have $V - K > 0$ because of the naturally large vocabulary sizes and a priori small number of topics that the null space of $\phi_{[1, -r]}$ will never be one dimensional, which would force the update rule for $\phi_r$ to just be sign changes.

While this model does satisfy the conditions for convergence outlined in both Hoff (2009) and Reisinger et al. (2010), we have not overcome the local mode issues observed with vanilla LDA above. Per the previous observations, these occur due to imperfect vocabulary filtering and misspecification of the number of topics. While working in the spherical framework allows us to improve our vocabulary expression in terms of a selective metric (e.g. tf-idf), we have still not addressed a misspecified number of topics. In SAM with orthogonal constraints, although we no longer see topics that are almost identical, there are instead low coherence, low proportion topics that hurt the overall coherence and efficiency of the model.

To introduce the ability to learn the number of topics in the model, we will use the flexibility of the Conway Embedding to put a joint prior on the topics and the number of topics. We use $S^C$ to refer to $S^{(V-1)(V+2)/2}$ the Conway Sphere, and $\varphi(\cdot)$ as the embedding function, we propose the following model:
\[\mu|\kappa_0 \sim \text{vMF}_{SC}(m, \kappa_0)\] (corpus average)

\[\eta|\mu, \xi \sim \text{vMF}_{SC}(\mu, \xi)\] (embedded orthogonal topics)

\[(\phi, K) = \varphi^{-1}(\eta)\] (orthogonal topics)

\[\theta_d|\alpha \sim \text{Dir}_K(\alpha), \quad d \in D\] (topic proportions)

\[\bar{\phi}_d|\phi, \theta_d = \text{Avg}(\phi, \theta_d), \quad d \in D\] (spherical average)

\[v_d|\bar{\phi}_d, \kappa \sim \text{vMF}_V(\bar{\phi}_d, \tau) \quad d \in D\] (generate document)

One note is that for distributions vMF_{SC} the values are assumed to lay on the Conway Sphere instead of the unit sphere. The Conway Sphere in this case is extremely high dimensional, so using strict selection criteria is critical to reduce the vocabulary size (with which the method scales). The priors on the concentration parameters are specified for the distributions as if they are over the unit sphere.

Without some sort of penalty on the number of topics, the larger number of topics will fit the data better and any estimation method will tend towards \(\frac{V}{2}\) topics. Thus to introduce a penalization on the number of topics, we use a Metropolis-Hastings scheme to sample from the Gibbs posterior, similar to Section 2.2. Instead we replace the loss function with following form

\[
L(\phi, K|D, v_{deD}, \tau) = \frac{2|D|K}{V} - \sum_{deD} (\tau \bar{\phi}_d^T v_d) \tag{5.1}
\]

Which gives the Gibbs posterior

\[
g(\phi, K|D, v_{deD}, \tau) \propto \exp (-|D|\psi L(\phi, K|D, v_{deD}, \tau)) \pi(\phi) \tag{5.2}
\]

Notice that the maximum penalty is \(|D|\) and would counterbalance a perfect fit to each document for a loss value of zero. In order to compute the Gibbs posterior, we do need to compute an approximation to \(\bar{\phi}\), which requires a set of topic proportions.
Since this procedure does not jointly sample the topic proportions, we compute \( \bar{\phi} \) as a temporary dummy variable. The topic proportions can be approximated similarly to how they are estimated in the ensemble method. If the dummy topic proportions are \( \hat{\alpha}_d \) for document \( d \in D \) and \( v \) is the document term matrix \( (|D| \times V) \), then 
\[
\hat{\alpha} = v(\phi^T \phi)^{-1} \phi^T.
\]
The rows of \( \hat{\alpha} \) give the dummy proportions. To get the Gibbs temperature, a random search on a log10 scale from \(-10\) to \(10\) is performed (to range from \(10^{-10}\) to \(10^{10}\)).

Because of our use of the Gibbs posterior, we do not need to estimate the corpus average parameter \( \mu \) since the remaining parameters and \( \mu \) are conditionally independent given the topics. The topics themselves are linked to \( \mu \) via the likelihood, however since we use a loss function instead of the likelihood \( \mu \) does not play a role in the inference.

The remaining parameters are estimated the same as they are in SAM once the topics and number of topics are sampled. However we see that this framework of using the extrinsic measure with a Gibbs posterior lead to a novel model. The large dimension of this problem can lead to slow mixing of the topics, so an EM or Hamiltonian Monte Carlo scheme may be more successful, or at least require less tuning of proposal distributions.

The main disadvantage of this is that because of movements up the sphere, the topics as they get added into the model may not be coherent with each other. Therefore taking a posterior mean of these draws may not, and frequently does not, yield a coherent topic although individual draws are coherent.

\subsection{Analysis of NSF Award Abstracts}

The NSF award abstracts are hosted publicly at \url{http://www.nsf.gov/awardsearch/download.jsp}. We look at 13,092 abstracts from the 2010 awards. The vocabulary is constructed using the tokenizer from the Mallet package with bi-grams extraction.
The vocabulary was then reduced by taking terms only with the top 10% score in terms of $tf-idf$ that also occurred in at least 5 documents. This left a vocabulary size of 78,343 terms. The average length of the documents after trimming the vocabulary is 379 words.

We applied both LDA and SAM (with orthogonal constraint) to the data. Since the models can’t be directly compared (an issue noted by Reisinger et al. in 2010), we inspect the topics for their coherence by looking at the top words. The models were both fit with 20 topics. Example word clouds are included in Figures 5.9 and 5.10.

One thing to notice about the output of the spherical topic models is that when investigating the topics, the positive and the negative parts of the topic vectors tend to be thematically coherent (also noted by Reisinger et al. (2010)). This is interesting because in a sense it allows a denser representation of the topics.

One interesting result of SAM with orthogonality constraints on this particular data set is that the positive and negative components of the topics tend to relate to broader impact terms on one value with field or discipline terms on the other. It may not be surprising that writers of grants from different disciplines use different
Figure 5.10: Topics for SAM with orthogonal constraints. The size of the terms correspond to the absolute value of the terms in the topics. Since there are both negative and positive values, and since the positive and negative values yield coherent themes, both are shown.

goals in broader impacts, however the SAM topic structure gives us a tool for making thematic connections that would not be obvious from simply looking at the top terms of a topic.

As can be seen in Figure 5.11, a novel aspect of SAM on the Conway Sphere is that it gives a posterior distribution for the number of topics. The results imply that there are in the neighborhood of 30 topics, compared to the number of topics estimated by the ensemble topic model which was nearer to 40 topics. This suggests that SAM on the Conway Sphere may be a more conservative estimate of the number of topics, however it is also significantly more computationally expensive.
Figure 5.11: Distribution on the number of topics as estimated from SAM on the Conway Sphere. The uncertainty in the number of topics seems to be more than in the mixtures of subspaces application of the Conway Sphere, in which the dimensions of each subspace tended to have clear posterior modes.

5.8 Conclusion

In this chapter we have addressed a core issue with topic modeling, in that the commonly used generative stochastic process for text data fails to generate text data well. Previous strategies such as picking low concentration parameters for Dirichlet distributions did not generate realistic term frequencies as judged by adherence to Zipf’s law. We resolve this issue by setting the prior on term frequencies directly proportional to a power law, and observe more realistic count data. These synthetic data sets are useful for assessing topic model methodology by measuring a new model or technique’s ability to recover true parameters.

We demonstrate this utility by implementing a straightforward and parallelizable ensemble of LDA models. Using data generated to conform to Zipf’s law, we show the technique is able to approximate the true number of topics that generated the
data. We apply this data to a corpus of NSF grant awards and showed that the topic selection step in aggregating the ensemble tends to result in high quality topics.

We also demonstrate geometrically motivated variations on the Spherical Admixture Model to ultimately give a version that does not require the number of topics as an input. This method can also characterize the uncertainty in the number of topics. Unlike the ensemble model however, this version of SAM has scalability concerns and requires strict filtering of the vocabulary. Other sampling methods such as HMC or incorporating the topics into a variational inference scheme are promising solutions to this issue.
In this dissertation we have presented novel approaches to statistical problems from a geometric perspective. Considering the extrinsic structure of common statistical objects allowed us to construct new models and computational strategies.

Using key extrinsic properties of linear subspaces, we have proposed a novel approach for jointly estimating the subspaces and the dimension of those subspaces in a mixture. Also using general extrinsic distances, we specified a method for performing kernel regression that is capable of handling both manifold and stratified-space valued responses. We then followed this line of research into less general problems, giving novel approaches in multivariate monotone regression and topic modeling.

Modeling stratified-spaces is a difficult problem, and may remain one for quite some time due to the computational and theoretical complications that the singularities in these structures induce. Hopefully the progress of this research opens that door a little further.
Appendix A

Appendix to Chapter 2

Proof. A result due to Schwartz (1965) states that if $\Pi$ assumes positive mass to any Kullback–Leibler neighborhood of $f_0$, then the resulting posterior is weakly consistency. Therefore, one needs to show for all $\epsilon > 0$,

$$\Pi(\{f : d_{KL}(f_0, f) \leq \epsilon\}) > 0. \quad (A.1)$$

Note that $\Pi(d_i = m) > 0$ for any $i = 1, \ldots, K$. Then one has

$$\Pi(K_\epsilon(f_0)) \geq \int_{O(m) \times \ldots \times O(m)} \Pi(K_\epsilon(f_0) \mid U_1, \ldots, U_K, d_1 = \ldots = d_K = m) d\Pi(U_1, \ldots, U_K \mid d_1 = \ldots = d_K = m). \quad (A.2)$$

Therefore it suffices to show that $\Pi(K_\epsilon(f_0) \mid U_1, \ldots, U_K) > 0$ where $U_1, \ldots, U_K$ are the bases of the respective $m$-dimensional subspaces $U_1, \ldots, U_K$.

We will show that there exists $K$ large enough such that given $m$-dimensional subspaces $U_1, \ldots, U_K$, the following mixture model assigns positive mass to any KL
neighborhood of $f_0$,

$$f(x, U, \Sigma) = \sum_{j=1}^{K} w_j \mathcal{N}(\phi(\mu_j), \tilde{\Sigma}_j),$$  \hspace{1cm} (A.3)

with $\phi(\mu_j) = U_j \mu_j + \theta_j$ and $\tilde{\Sigma}_j = U_j (\Sigma_j - \sigma_2^2 I_{d_j}) U_j^T + \sigma_2^2 I_m$.

If $U_1, \ldots, U_K$ have the same dimension $m$ and $U_1, \ldots, U_K$ are a choice of orthonormal bases on the respective subspaces, then an infinite-dimensional version of our model can be given by

$$X \sim g(x, \Gamma) = \int_{\mathbb{R}^m} \mathcal{N}(x; \phi(\mu), \Sigma) P(\mu) d\mu,$$

$$\phi(\mu) = U \mu + \theta,$$

$$\Sigma = U(\Sigma_0 - \sigma_2^2 I_m) U^T + \sigma^2 I_m$$

with parameters $\Gamma = (U, \theta, \Sigma_0, \sigma, P)$. The prior for $P$ can be given by a Dirichlet process prior whose base measure has full support in $\mathbb{R}^m$ while the priors of the rest parameters $(U, \theta, \Sigma_0, \sigma)$ can be given the same as those of our model. By Theorem 3.1 of Page et al. (2013) or Theorem 2 of Wu and Ghosal (2010), there exists an open subset $P$ of the space of all the probability measures on $\mathbb{R}^m$ such that for for all $\epsilon > 0$, any $P \in \mathcal{P}$ such that

$$\int_{\mathbb{R}^m} f_0(x) \log \frac{f_0(x)}{g(x, \Gamma)} dx \leq \epsilon. \hspace{1cm} (A.4)$$

We will first show that for any $\epsilon' > 0$ there exists $K$ large enough and $w_1, \ldots, w_K$, $\theta_1, \ldots, \theta_K$, $\mu_1, \ldots, \mu_K$, $U_1, \ldots, U_K$, $\Sigma_1, \ldots, \Sigma_K$, $\sigma_1, \ldots, \sigma_K$, and $\sigma^2$ such that

$$\left| g(x, \Gamma) - \sum_{j=1}^{K} w_j \mathcal{N}(\phi(\mu_j), \tilde{\Sigma}_j) \right| \leq \epsilon',$$

for any $P \in \mathcal{P}$.

Let $L$ be some large enough number. We first partition $\mathbb{R}^m$ into $L^m + 1$ sets. Let

$$A_{i_1, \ldots, i_m} = \prod_{j=1}^{m} \left( -L + (i_j - 1) \frac{\log L}{L}, -L + i_j \frac{\log L}{L} \right]$$

105
and

\[ A_0^L \cup \left( \bigcup_{i_1, \ldots, i_m = 1}^L A_{i_1, \ldots, i_m} \right) = \mathbb{R}^m. \]

Pick \( \mu_j \in A_{i_1, \ldots, i_m} \) where \( j = 1, \ldots, L^m \) and let \( w_j = P(A_{i_1, \ldots, i_m}) \). Approximating the integral by the finite sum over the cubes in \( \mathbb{R}^m \), for all \( \epsilon'' > 0 \), there exists \( L \) large enough such that

\[
\left| g(x, \Gamma) - \sum_{j=1}^{L^m} w_j \mathcal{N}(\mu_j U + \theta_j, \Sigma) \right| \leq \epsilon''.
\]

Let \( K = L^m \). Now let \( U_1, \ldots, U_K \) be \( K \) points in the \( \epsilon'' \) neighborhood of \( U, \theta_1, \ldots, \theta_K \) in the \( \epsilon'' \) neighborhood of \( \theta, \Sigma_1, \ldots, \Sigma_K \) be in the \( \epsilon'' \) neighborhood of \( \Sigma_0 \), and \( \sigma_1, \ldots, \sigma_K \) in the \( \epsilon'' \) neighborhood of \( \sigma \), then by the continuity of \( \sum_{j=1}^K w_j \mathcal{N}(\mu_j U + \theta, \Sigma) \), one can show that

\[
\left| g(x, \Gamma) - \sum_{j=1}^K w_j \mathcal{N}(\mu_j U + \theta_j, \Sigma_j) \right| \leq \left| g(x, \Gamma) - \sum_{j=1}^K w_j \mathcal{N}(\mu_j U + \theta, \Sigma) \right| \\
+ \left| \sum_{j=1}^K w_j \mathcal{N}(\mu_j U + \theta, \Sigma) - \sum_{j=1}^K w_j \mathcal{N}(\mu_j U_j + \theta_j, \Sigma_j) \right|
\]

\[
\leq \epsilon'' + \delta(\epsilon''),
\]

where \( \delta(\epsilon'') \to 0 \) when \( \epsilon'' \to 0 \). For the above choices of \( w_1, \ldots, w_K, \mu_1, \ldots, \mu_K, \theta_1, \ldots, \theta_K, U_1, \ldots, U_K, \Sigma_1, \ldots, \Sigma_K, \sigma_1, \ldots, \sigma_K \) corresponding to any \( P \in \mathcal{P} \), one looks at

\[
\int_{\mathbb{R}^m} f_0(x) \log \frac{f_0(x)}{\sum_{j=1}^K w_j \mathcal{N}(\mu_j U_j + \theta_j, \Sigma_j)} \, dx = \int_{\mathbb{R}^m} f_0(x) \log \frac{f_0(x)}{g(x, \Gamma)} \, dx \\
+ \int_{\mathbb{R}^m} f_0(x) \log \frac{g(x, \Gamma)}{\sum_{j=1}^K w_j \mathcal{N}(\mu_j U_j + \theta_j, \Sigma_j)} \, dx.
\]

Take \( \epsilon'' = \epsilon \). By the continuity of the log function, one has

\[
\int_{\mathbb{R}^m} f_0(x) \log \frac{f_0(x)}{\sum_{j=1}^K w_j \mathcal{N}(\mu_j U_j + \theta_j, \Sigma_j)} \, dx \leq \epsilon + \delta'(\epsilon) \quad (A.5)
\]
where $\delta'(\epsilon) \to 0$ when $\epsilon \to 0$. Note that our prior assigns positive mass to arbitrary neighborhood of $w_1, \ldots, w_K, \theta_1, \ldots, \theta_K, \mu_1, \ldots, \mu_K, U_1, \ldots, U_K, \Sigma_1, \ldots, \Sigma_K, \sigma_1, \ldots, \sigma_K, \sigma^2$. By (A.4), (A.5) and the continuity of the model, our assertion follows.

In proving the following strong consistency theorem, we assume that the parameters $\sigma_i^2, \sigma^2$ and the diagonal elements of $\Sigma_i$ $(i = 1, \ldots, K)$ follow i.i.d. truncated Gamma priors supported on some bounded interval $[0, M]$ for some large enough constant $M$.

**Theorem 14.** The posterior distribution $\Pi(\cdot \mid x_1, \ldots, x_n)$ is strongly consistent. That is, for all $\epsilon > 0$,

$$\Pi(U_\epsilon(f_0) \mid x_1, \ldots, x_n) \to 1 \quad P f_0^\infty - \text{almost surely as } n \to \infty,$$

where $U_\epsilon(f_0)$ is a neighborhood of $f_0$ with radius $\epsilon$ with respect to the Hellinger distance.

**Proof.** By Theorem 3, the true density $f_0$ is in the weak support of our model. Then by a result due to Barron (1988) (also see Theorem 2 in Ghoshal et al. (1999)), for all $\epsilon > 0$ if we can construct sieves $D_{\delta,n} \subseteq \mathcal{M}$ with $\delta < \epsilon$ such that the metric entropy

$$\log N(\delta, D_{\delta,n}) \leq n\beta$$

for some $\beta < \epsilon^2/2$ and $\Pi(D_{\epsilon,n}^c) \leq C_1 \exp(-n\beta_1)$ with some constants $C_1$ and $\beta_1$. Then the posterior distribution $\Pi(\cdot \mid x_1, \ldots, x_n)$ is strongly consistent at $f_0$.

Denote

$$\Theta = (K, \theta_1, \ldots, \theta_K, w_1, \ldots, w_K, \mu_1, \ldots, \mu_K, U_1, \ldots, U_K, \Sigma_1, \ldots, \Sigma_K, \sigma_1, \ldots, \sigma_K, \sigma)$$

and

$$\sigma = (\sigma_{11}, \ldots, \sigma_{1d_1}, \ldots, \sigma_{K1}, \ldots, \sigma_{Kd_K}, \sigma, \sigma_1, \ldots, \sigma_K)$$
where \( \sigma^2_{ij} \) is the \( j \)th diagonal element of \( \Sigma_i \) which we assume to be diagonal in our model. Let

\[
\Theta_n = \{ \Theta : K = K_n, \| \theta_i \| \leq \bar{\theta}_n, \| \mu_i \| \leq \bar{\mu}_n, \max(\sigma) \leq M, \min(\sigma) \geq h_n, \ i = 1, \ldots, K \}
\]

where \( K_n, \bar{\theta}_n, \bar{\mu}_n, M \) and \( h_n \) are some sequences depending on \( n \).

Let \( D_{\delta,n} = \{ f(x, \Theta) : \Theta \in \Theta_n \} \) where \( f(x, \Theta) \) is given by (A.3) for any \( \Theta \). We need to verify the metric entropy and the prior mass of \( D_{\delta,n} \). Let

\[
\Theta^i_{K_n} = (d^i_1, \ldots, d^i_K, \theta^i_1, \ldots, \theta^i_K, w^i_1, \ldots, w^i_K, \mu^i_1, \ldots, \mu^i_K, U^i_1, \ldots, U^i_K, \sigma^i), \ i = 1, 2.
\]

Note that posterior consistency with respect to the Hellinger distance is equivalent to posterior consistency with respect to the \( L^1 \)-distance due to the equivalence of the two distances. For \( i = 1, 2 \), let \( \phi^i(\mu_j) = U^i_j \mu_j^i + \theta^i_j \) and \( \tilde{\Sigma}^i_j = U^i_j (\Sigma^i_j - (\sigma^i)^2 I_d^j) (U^i_j)^T + (\sigma^i)^2 I_m \). One has

\[
\int_{\mathbb{R}^m} |f(x, \Theta^1_{K_n}) - f(x, \Theta^2_{K_n})| \, dx
\]

\[
= \int_{\mathbb{R}^m} \left| \sum_{j=1}^{K_n} w^1_j N_m(\phi^1(\mu_j), \tilde{\Sigma}^1_j) - \sum_{j=1}^{K_n} w^2_j N_m(\phi^2(\mu_j), \tilde{\Sigma}^2_j) \right| \, dx
\]

\[
= \int_{\mathbb{R}^m} \left| \sum_{j=1}^{K_n} w^1_j N_m(\phi^1(\mu_j), \tilde{\Sigma}^1_j) - \sum_{j=1}^{K_n} w^2_j N_m(\phi^2(\mu_j), \tilde{\Sigma}^2_j) \right| \, dx
\]

\[
+ \sum_{j=1}^{K_n} w^1_j N_m(\phi^2(\mu_j), \tilde{\Sigma}^2_j) - \sum_{j=1}^{K_n} w^2_j N_m(\phi^2(\mu_j), \tilde{\Sigma}^2_j) \right| \, dx
\]

\[
\leq \int_{\mathbb{R}^m} \left| \sum_{j=1}^{K_n} N_m(\phi(\mu_j), \tilde{\Sigma}_j) - N_m(\phi^2(\mu_j), \tilde{\Sigma}^2_j) \right| \, dx
\]

\[
+ \int_{\mathbb{R}^m} \left| \sum_{j=1}^{K_n} (w^1_j - w^2_j) N_m(\phi^2(\mu_j), \tilde{\Sigma}^2_j) \right| \, dx
\]

\[
\leq \int_{\mathbb{R}^m} \left| \sum_{j=1}^{K_n} w^1_j N_m(\phi(\mu_j), \tilde{\Sigma}_j) - N_m(\phi^2(\mu_j), \tilde{\Sigma}^2_j) \right| \, dx + \sum_{j=1}^{K_n} |w^1_j - w^2_j|\]

\[
\leq \sum_{j=1}^{K_n} w^1_j \int_{\mathbb{R}^m} \left| N_m(\phi(\mu_j), \tilde{\Sigma}_j) - N_m(\phi^2(\mu_j), \tilde{\Sigma}^2_j) \right| \, dx + \sum_{j=1}^{K_n} |w^1_j - w^2_j|.
\]
Note that
\[
\int_{\mathbb{R}^m} \left| \mathcal{N}_m(\phi^1(\mu_j), \tilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \tilde{\Sigma}_j^2) \right| \, dx
\]
\[
= \int_{\mathbb{R}^m} \left| \mathcal{N}_m(\phi^1(\mu_j), \tilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \tilde{\Sigma}_j^1) + \mathcal{N}_m(\phi^2(\mu_j), \tilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \tilde{\Sigma}_j^2) \right| \, dx
\]
\[
\leq \int_{\mathbb{R}^m} \left| \mathcal{N}_m(\phi^1(\mu_j), \tilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \tilde{\Sigma}_j^1) \right| \, dx
\]
\[
+ \int_{\mathbb{R}^m} \left| \mathcal{N}_m(\phi^2(\mu_j), \tilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \tilde{\Sigma}_j^2) \right| \, dx.
\]

By the proof of Lemma 5 of Wu and Ghosal (2010), one has for the first term of the above expression
\[
\int_{\mathbb{R}^m} \left| \mathcal{N}_m(\phi^1(\mu_j), \tilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \tilde{\Sigma}_j^1) \right| \, dx \leq \sqrt{\frac{2}{\pi}} \frac{\|\phi^1(\mu_j) - \phi^2(\mu_j)\|}{\lambda_1(\tilde{\Sigma}_j^1)^{1/2}}
\]
\[
= \sqrt{\frac{2}{\pi}} \frac{\|(U_j^1 \mu_j + \theta_j^1) - (U_j^2 \mu_j + \theta_j^2)\|}{\lambda_1(\tilde{\Sigma}_j^1)^{1/2}}
\]
\[
\leq \sqrt{\frac{2}{\pi \lambda_1(\tilde{\Sigma}_j^1)}} \left( \|\theta_j^1 - \theta_j^2\| + \|U_j^1 \mu_j - U_j^2 \mu_j\| \right),
\]
where \(\lambda_1(\tilde{\Sigma}_j^1)\) is the smallest eigenvalue of \(\tilde{\Sigma}_j^1\). Therefore combining all the terms above
\[
\int_{\mathbb{R}^m} |f(x, \Theta_{K_n}^1) - f(x, \Theta_{K_n}^2)| \, dx
\]
\[
\leq \sum_{j=1}^{K_n} w_j^1 - w_j^2 + \max_{j=1, \ldots, K_n} \left\{ \sqrt{\frac{2}{\pi \lambda_1(\tilde{\Sigma}_j^1)}} \left( \|\theta_j^1 - \theta_j^2\| + \|U_j^1 \mu_j - U_j^2 \mu_j\| \right) \right\}
\]
\[
+ \sum_{j=1}^{K_n} w_j \int_{\mathbb{R}^m} \left| \mathcal{N}_m(\phi^2(\mu_j), \tilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \tilde{\Sigma}_j^2) \right| \, dx.
\]
Without loss of generality, assume $\det(\tilde{\Sigma}_j^2) \geq \det(\tilde{\Sigma}_j^1)$. One has

$$
\int_{\mathbb{R}^m} \left| N_m(\phi^2(\mu_j), \tilde{\Sigma}_j^1) - N_m(\phi^2(\mu_j), \tilde{\Sigma}_j^2) \right| \, dx
$$

$$
= \frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} \left| \frac{1}{\det(\tilde{\Sigma}_j^1)^{1/2}} \exp\left[ -\frac{1}{2} (x - \phi^2(\mu_j))^T (\tilde{\Sigma}_j^1)^{-1} (x - \phi^2(\mu_j)) \right] \right| dx
$$

$$
- \frac{1}{\det(\tilde{\Sigma}_j^2)^{1/2}} \exp\left[ -\frac{1}{2} (x - \phi^2(\mu_j))^T (\tilde{\Sigma}_j^2)^{-1} (x - \phi^2(\mu_j)) \right] \left| dx \right|
$$

$$
\leq \frac{2^{m+1}}{(2\pi)^{m/2}} \int_{[0,\infty)^m} \max\left\{ 0, \frac{1}{\det(\tilde{\Sigma}_j^1)^{1/2}} \exp\left[ -\frac{1}{2} (x - \phi^2(\mu_j))^T (\tilde{\Sigma}_j^1)^{-1} (x - \phi^2(\mu_j)) \right] \right\} \, dx.
$$

(A.7)

We take $\delta = \epsilon/2$. We first partition $[0, 1]^{K_n}$ into $N_w$ grid points. With a choice of the grid points given in Lemma 1 in Ghoshal et al. (1999), the number grid points needed for $\sum_{j=1}^{K_n} |w^1_j - w^2_j| \leq \epsilon/8$ is bounded by

$$
\log N_w \leq K_n \left[ 1 + \log \left( \frac{1 + \epsilon/8}{\epsilon/8} \right) \right].
$$

Note that the number of balls of radius $\delta R$ used to cover a Euclidean ball centered at the origin of radius $R$ in $\mathbb{R}^m$ is bounded by $(3/\delta)^m$. Therefore, the number of balls of radius $\epsilon [2\pi \lambda_1(\tilde{\Sigma}_1)]^{1/2}/32$ needed to cover $\|\theta\| \leq \tilde{\theta}_n$ is bounded by $(96\tilde{\theta}_n/\epsilon [2\pi \lambda_1(\tilde{\Sigma}_1)]^{1/2})^m$. Then one can always find $\theta_1^i, \ldots, \theta_{K_n}^i$, $i = 1, 2$, such that

$$
\max_{j=1, \ldots, K_n} \sqrt{\frac{2}{\pi \lambda_1(\tilde{\Sigma}_j^1)}} \|\theta_1^j - \theta_2^j\| \leq \frac{\epsilon}{8}.
$$

Note that $\mu_j^1$ and $\mu_j^2$ do not necessarily have the same dimension. We view them as elements in $\mathbb{R}^m$ by filling the last $m - d_j^i$, $i = 1, 2$, elements with 0. Then one can show that $\|U^1_j \mu_j^1 - U^2_j \mu_j^2\| \leq \|\mu_j^1 - \mu_j^2\|$. Then the covering number for $\mu$ is bounded
by \((96\bar{\mu}_n/\epsilon[2\pi\lambda_1(\tilde{\Sigma}_1)]^{1/2})^m\) with which one can find \(\mu_1', \ldots, \mu_{K_n}'\), \(i = 1, 2\), such that

\[
\max_{j=1,\ldots,K_n} \sqrt{\frac{2}{\pi \lambda_1(\tilde{\Sigma}_j)}} \|U_j^1 \mu_j^1 - U_j^2 \mu_j^2\| \leq \frac{\epsilon}{8}.
\]

Note that for \(\Theta \in \Theta_n\), all the eigenvalues of \(\tilde{\Sigma}\) lie in the interval \([h_n^2, M^2]\). We divide the \(m\)-dimensional cube \([h_n^2, M^2]^m\) into smaller cubes such that there are \(N_\lambda\) grid points for the \(m\) eigenvalues \(\lambda = (\lambda_1, \ldots, \lambda_m)\) of \(\tilde{\Sigma}\). Note that for any \(\tilde{\Sigma}_1\) whose eigenvalues fall into one of the cubes in the grid, one can always find \(\tilde{\Sigma}_2\) such that \(\tilde{\Sigma}_1^{-1} - \tilde{\Sigma}_2^{-1}\) is positive definite. For example, one may take \(\tilde{\Sigma}_2^{-1} = \tilde{\Sigma}_1^{-1} - \tilde{\epsilon}I_m\) where \(\tilde{\epsilon}\) is small enough and \(0 < \tilde{\epsilon} < \min\{1/\lambda_1(\tilde{\Sigma}_1), \ldots, 1/\lambda_m(\tilde{\Sigma}_1)\}\). Then

\[
\exp\left[-\frac{1}{2}(x - \phi^2(\mu_j)) \trans (\tilde{\Sigma}_j^1)^{-1}(x - \phi^2(\mu_j))\right] \leq \exp\left[-\frac{1}{2}(x - \phi^2(\mu_j)) \trans (\tilde{\Sigma}_j^2)^{-1}(x - \phi^2(\mu_j))\right].
\]

Then from the above inequality and (A.7), one has

\[
\int_{\mathbb{R}^m} \left| N_m(\phi^2(\mu_j), \tilde{\Sigma}_j^1) - N_m(\phi^2(\mu_j), \tilde{\Sigma}_j^2) \right| \, dx \\
\leq 2^{m+1} \int_{[0,\epsilon]^m} \left( \frac{1}{\det(\tilde{\Sigma}_j^1)^{1/2}} - \frac{1}{\det(\tilde{\Sigma}_j^2)^{1/2}} \right) \\
\times \exp\left[-\frac{1}{2}(x - \phi^2(\mu_j)) \trans (\tilde{\Sigma}_j^2)^{-1}(x - \phi^2(\mu_j))\right] \, dx \\
\leq 2^m \frac{\det(\tilde{\Sigma}_j^2)^{1/2} - \det(\tilde{\Sigma}_j^1)^{1/2}}{\det(\tilde{\Sigma}_j^1)^{1/2}}.
\]

We divide the range of each of the eigenvalues \([h_n^2, M^2]\) into \(L\) equidistant intervals and let \(\lambda_{jl} = h_n^2(1 + \epsilon/2^{m+3})^{2l_j/m}\) where \(j = 1, \ldots, m\) and \(1 \leq l_j \leq L\). We pick \(L\) to be the smallest integer which satisfies \(h_n^2(1 + \epsilon/2^{m+3})^{2l_j/m} \geq M^2\). We pick the \(j\)th eigenvalue of \(\tilde{\Sigma}_j^1\) and \(\tilde{\Sigma}_j^2\) to be in some interval \([\lambda_{j(l-1)}, \lambda_{jl}]\) and satisfying the
ordering on the eigenvalues. Then one has

\[
2^m \frac{\det(\Sigma_j^2)^{1/2} - \det(\Sigma_j^1)^{1/2}}{\det(\Sigma_j^1)^{1/2}}
\]

\[
\leq 2^m \prod_{j=1}^m h_n^2 (1 + \epsilon/2^{m+3})^{2j/m} - \prod_{j=1}^m h_n^2 (1 + \epsilon/2^{m+3})^{2(j-1)/m}
\]

\[
\prod_{j=1}^m h_n^2 (1 + \epsilon/2^{m+3})^{2(j-1)/m}
\]

\[
\leq \frac{\epsilon}{8},
\]

and the metric entropy of \(\lambda\) in \([h_n^2, M^2]^m\) is bounded by

\[
\log N_\lambda \leq m \left[ \frac{\log(M^2/h_n^2)}{\log(1 + \epsilon/2^{m+3})} + 1 \right].
\]

Letting \(\bar{\theta}_n = \bar{\mu}_n\) and combining all the terms above on the entropy numbers, the metric entropy number of \(D_{n,\epsilon}\) is bounded by

\[
\log N(\delta, D_{\delta,n}) \leq K_n \left[ 1 + \log \left( \frac{1 + \epsilon/8}{\epsilon/8} \right) \right] + 2m \log \left[ \frac{96\bar{\theta}_n}{\epsilon(2\pi\lambda_1(\Sigma_1))^{1/2}} \right]
\]

\[
+ m \left[ \frac{2 \log(M/h_n)}{\log(1 + \epsilon/2^{m+3})} + 1 \right].
\]

Let \(K_n = c_1\sqrt{n}, \bar{\theta}_n = c_2\sqrt{n}\), and \(h_n = c_3n^{-1/b}\) for some constants \(b > 0\) and \(c_1, c_2\)
and \(c_3\) small enough. Then one has

\[
\log N(\delta, D_{\delta,n}) \leq n\beta.
\]

It remains to verify the condition that the prior mass outside \(D_{\delta,n}\) is exponentially small given our priors on the parameters and the above choice of \(K_n, \bar{\theta}_n, h_n\) and \(M\). We assume multivariate normal priors (with diagonal covariance matrices) for \(\theta\) and \(\mu\), thus with a choice of \(\bar{r}_n, \bar{\mu}_n\), one can show that using changing of variables and Mill’s inequality, the tail \(P(||\theta|| \geq \bar{\theta}_n)\) decays exponentially. The elements \(\sigma_1^2, \ldots, \sigma_K^2, \sigma^2\) and the variance (i.e., diagonal) terms of \(\Sigma_j, j = 1, \ldots, m\), are assumed to follow the i.i.d. Gamma priors with density \(b^a/\Gamma(a)x^{-a-1}\exp(-b/x)\) and
hyperparameters $a$ and $b$. Then by direct calculation of the integrals,

$$P(\min(\sigma) \leq h_n) \leq c_4 n \exp(-\tilde{C}_4 h_n^{-b}) \leq c_4 \exp(-\tilde{c}_4 n + \log n),$$

which decays exponentially fast. \qed
Appendix B

Appendix to Chapter 3

Proof of Theorem 5. Recall

\[ \hat{F}(x) = \frac{\frac{1}{n} \sum_{i=1}^{n} J(y_i) K_H(x_i - x)}{\frac{1}{n} \sum_{i=1}^{n} K_H(x_i - x)}. \]

Denote the denominator of \( \hat{F}(x) \) as

\[ \hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_H(x_i - x) = \frac{1}{n} \left| H \right| \sum_{i=1}^{n} K(x_i - x). \]

It is standard to show

\[ \hat{f}(x) \xrightarrow{p} f_X(x) \quad (B.1) \]
where $\to^p$ indicates convergence in probability. For the numerator term of $\hat{F}(x)$, one has
\[
E \left( \frac{1}{n} \sum_{i=1}^{n} J(y_i) K_H(x_i - x) \right) = \frac{1}{n} \sum_{i=1}^{n} E \left( J(y_i) K_H(x_i - x) \right) \\
= \frac{1}{n} \sum_{i=1}^{n} \int E \left( J(y_i) K_H(x_i - x) \right) | x_i \rangle f_X(x_i) dx_i \\
= \frac{1}{n} \sum_{i=1}^{n} \int \mu(x_i) K_H(x_i - x) f_X(x_i) dx_i \\
= \int \mu(\bar{x}) K_H(\bar{x} - x) f_X(\bar{x}) d\bar{x}.
\]
Noting that $\mu(x) = (\mu_1(x), \ldots, \mu_D(x))' \in \mathbb{R}^D$, we slightly abuse the integral notation above meaning that the $j$th entry of $E \left( n^{-1} \sum_{i=1}^{n} J(y_i) K_H(x_i - x) \right)$ is given by
\[
\int \mu_j(\bar{x}) K_H(\bar{x} - x) f_X(\bar{x}) d\bar{x}.
\]
Letting $v = H^{-1}(\bar{x} - x)$ by changing of variables, the above equations become
\[
E \left( \frac{1}{n} \sum_{i=1}^{n} J(y_i) K_H(x_i - x) \right) = \int \mu(x + Hv) K(v) f_X(x + Hv) dv.
\]
By the multivariate Taylor expansion,
\[
f_X(x + Hv) = f_X(x) + (\nabla f) \cdot ( Hv ) + R,
\] (B.2)
where $\nabla f$ is the gradient of $f$ and $R$ is the remainder term of the expansion. The remainder $R$ can be shown to be bounded above by
\[
R \leq \frac{C}{2} \| Hv \|^2, \quad \| Hv \| = |h_1 v_1| + \ldots |h_m v_m|.
\]
Note that $\mu(x + Hv)$ is a multivariate map valued in $\mathbb{R}^D$. We can make second order multivariate Taylor expansions for $\mu(x + Hv) = (\mu_1(x + Hv), \ldots, \mu_D(x + Hv))'$ at each of its entries $\mu_i$ for $i = 1, \ldots, D$. We have
\[
\mu(x + Hv) = \mu(x) + A(Hv) + V + R,
\] (B.3)
where $A$ is a $D$ by $m$ matrix whose $i$th row is given by the gradient of $\mu_i$ evaluated at $x$. $V$ is a $D$ dimensional vector, whose $i$th term is given by $\frac{1}{2}( Hv )^T T_i( Hv )$, where $T_i$ is the Hessian matrix of $\mu_i(x)$ and $R$ is the remainder vector. Thus,

$$E \left( \frac{1}{n} \sum_{i=1}^{n} J(y_i) K_H(x_i - x) \right)$$

$$\approx \int \left( (f_X(x) + (\nabla f) \cdot (Hv)) K(v)(\mu(x) + A(Hv) + V) \right) dv$$

$$= f_X(x) \mu(x) + f_X(x) \int K(v)A(Hv)dv + f_X(x) \int K(v)Vdv$$

$$+ \mu(x) \int (\nabla f) \cdot (Hv)K(v)dv + \int (\nabla f) \cdot (Hv)K(v)A(Hv)dv$$

$$+ \int (\nabla f) \cdot (Hv)K(v)Vdv. \tag{B.6}$$

By the property of the kernel function, we have $\int K(u)udu = 0$; therefore the second term of equation (B.5) is zero by simple algebra. To evaluate the third term of equation (B.5), we first calculate for $\int K(v)Vdv$. From here onward until the end of the proof, we denote $x = (x^1, \ldots, x^m)$ where $x^i$ is the $i$th coordinate of $x.$ Note that the $i$th term of $V$ ($i = 1, \ldots, D$) is given by $\frac{1}{2}( Hv )^T T_i( Hv )$, where $T_i$ is the Hessian matrix of $\mu_i$, which is precisely

$$\frac{1}{2} h_i^2 v_i^2 \left( \frac{\partial^2 \mu_i}{\partial(x^1)^2} + \ldots + \frac{\partial^2 \mu_i}{\partial x^m x^1} \right) + \ldots + \frac{1}{2} h_m^2 v_m^2 \left( \frac{\partial^2 \mu_i}{\partial x^1 x^m} + \ldots + \frac{\partial^2 \mu_i}{\partial(x^m)^2} \right).$$

Therefore, the $i$th entry of the third term of equation (B.5) is given by

$$U_i = \frac{1}{2} f_X(x) \left( h_1^2 \left( \frac{\partial^2 \mu_i}{\partial(x^1)^2} + \ldots + \frac{\partial^2 \mu_i}{\partial x^m x^1} \right) \right) \int v_1^2 K_1(v_1)dv_1 + \ldots$$

$$+ h_m^2 \left( \frac{\partial^2 \mu_i}{\partial x^1 x^m} + \ldots + \frac{\partial^2 \mu_i}{\partial(x^m)^2} \right) \int v_m^2 K_m(v_m)dv_m. \tag{B.7}$$
The first term of equation (B.6) is given by
\[
\mu(x) \int (\nabla f) \cdot (Hv) K(v) dv = \int \left( h_1 v_1 \frac{\partial f}{\partial x_1} + \ldots + h_m v_m \frac{\partial f}{\partial x_m} \right) K(v) dv = 0.
\]

The \(i\)th entry of the second term of equation (B.6) is given by
\[
h_1^2 \frac{\partial f}{\partial x_1} \frac{\partial \mu_i}{\partial x_1} \int v_1^2 K_1(v_1) dv_1 + \ldots + h_m^2 \frac{\partial f}{\partial x_m} \frac{\partial \mu_i}{\partial x_m} \int v_m^2 K_m(v_m) dv_m. \quad \text{(B.8)}
\]

The third term of equation (B.6) can be shown to be zero, since odd moments of symmetric kernels are 0. Therefore, we have
\[
E \left( \frac{1}{n} \sum_{i=1}^{n} J(y_i) K_H(x_i - x) \right) \approx f_X(x) \mu(x) + Z, \quad \text{(B.9)}
\]

where the \(i\)th coordinate of \(Z\) is
\[
Z_i = h_1^2 \left\{ \frac{\partial f}{\partial x_1} \frac{\partial \mu_i}{\partial x_1} + \frac{1}{2} f_X(x) \left( \frac{\partial^2 \mu_i}{\partial (x_1)^2} + \ldots + \frac{\partial^2 \mu_i}{\partial x_m^2} \right) \right\} \int v_1^2 K_1(v_1) dv_1 \\
+ \ldots \\
+ h_m^2 \left\{ \frac{\partial f}{\partial x_m} \frac{\partial \mu_i}{\partial x_m} + \frac{1}{2} f_X(x) \left( \frac{\partial^2 \mu_i}{\partial x_1 x_m} + \ldots + \frac{\partial^2 \mu_i}{\partial (x_m)^2} \right) \right\} \int v_m^2 K_m(v_m) dv_m \quad \text{(B.10)}
\]

combining equation (B.7) and equation (B.8). The reminder term of (B.2) is of order \(o(\max\{h_1, \ldots, h_m\})\) and each entry of the remainder vector in (B.3) is of order \(o(\max\{h_1^2, \ldots, h_m^2\})\).

We now look at the covariance matrix of \(n^{-1} \sum_{i=1}^{n} J(y_i) K_H(x_i - x)\), which we denote by \(\Sigma(x)\). Denote the \(j\)th entry \((j = 1, \ldots, D)\) of \(J(y_i)\) as \(J_j(y_i)\). Denote \(\sigma(y^j, y^k)\) as the conditional covariance between the \(i\)th entry and \(j\)th entry of \(y\). We
have

\[ \Sigma_{jk} = E \left[ \left( \frac{1}{n} \sum_{i=1}^{n} J_j(y_i)K_H(x_i - x) \right) - E \left( \frac{1}{n} \sum_{i=1}^{n} J_j(y_i)K_H(x_i - x) \right) \right] \]

\[ \left( \frac{1}{n} \sum_{i=1}^{n} J_k(y_i)K_H(x_i - x) \right) - E \left( \frac{1}{n} \sum_{i=1}^{n} J_k(y_i)K_H(x_i - x) \right) \] \]

\[ = E \left[ \left( \frac{1}{n} \sum_{i=1}^{n} \left( J_j(y_i)K_H(x_i - x) - \int \mu_j(\bar{x})K_H(\bar{x} - x)f_X(\bar{x})d\bar{x} \right) \right. \]

\[ \left. \left( \frac{1}{n} \sum_{i=1}^{n} \left( J_k(y_i)K_H(x_i - x) - \int \mu_k(\bar{x})K_H(\bar{x} - x)f_X(\bar{x})d\bar{x} \right) \right) \right] \]

\[ = \frac{1}{n} \int E \left[ \left( J_j(y_1)K_H(x_1 - x) - \int \mu_j(\bar{x})K_H(\bar{x} - x)f_X(\bar{x})d\bar{x} \right) \right. \]

\[ \left. \left( J_k(y_1)K_H(x_1 - x) - \int \mu_k(\bar{x})K_H(\bar{x} - x)f_X(\bar{x})d\bar{x} \right) \mid x_1 \right] f_X(x_1)dx_1 \]

\[ = \frac{1}{n} \int \sigma(J_j(y_1)K_H(x_1 - x), J_k(y_1)K_H(x_1 - x))f_X(x_1)dx_1 \]

\[ = \frac{1}{n} \int K_H(x_1 - x))^2 \sigma(J_j(y_1), J_k(y_1))f_X(x_1)dx_1. \]

By the change of variable \( v = H^{-1}(x_1 - x) \), the above equation becomes

\[ \Sigma_{jk} = \frac{1}{n|H|} \int K(v)^2 \sigma(J_j(y_v), J_k(y_v))f_X(Hv + x)dv \]

\[ = \frac{1}{n|H|} \int K(v)^2 \sigma(J_j(y_v), J_k(y_v)) \left( f_X(x) + \nabla f \cdot (Hv) + o(\max\{h_1, \ldots, h_m\}) \right) dv \]

\[ = \frac{1}{n|H|} \int K(v)^2 \sigma(J_j(y_v), J_k(y_v))f_X(x)dv + o \left( \frac{1}{n|H|} \right). \quad \text{(B.11)} \]

By (B.1), (B.9) and (B.29), and applying central limit theorem and Slustky’s theorem, one has

\[ \sqrt{n|H|} \left( \hat{F}(x) - \hat{\mu}(x) \right) \xrightarrow{L} N(0, \Sigma(x)), \quad \text{(B.12)} \]
where \( \tilde{\mu}(x) = \mu(x) + \frac{Z}{f_X(x)} \) and the \( i \)th entry \( (i = 1, \ldots, D) \) of \( Z \) is given by \( (B.10) \)

\[
\Sigma_{jk} = \frac{\sigma(J_j(y_v), J_k(y_v)) \int K(v)^2 dv}{f_X(x)}. \tag{B.13}
\]

One can show

\[
\sqrt{n|H|} \left( \tilde{F}_E(x) - P (\tilde{\mu}(x)) \right) = \sqrt{n|H|} \tilde{d}_{\tilde{\mu}(x)} P \left( \tilde{F}(x) - \tilde{\mu}(x) \right) + o_P(1).
\]

Therefore, one has

\[
\sqrt{n|H|} \tilde{d}_{\tilde{\mu}(x)} P \left( \tilde{F}(x) - \tilde{\mu}(x) \right) \xrightarrow{L} N(0, \tilde{\Sigma}(x)). \tag{B.14}
\]

Here \( \tilde{\Sigma}(x) = B^t \Sigma(x) B \), where \( B \) is the \( D \times d \) matrix of the differentiable \( d_{\tilde{\mu}(x)} P \) with respect to given orthonormal bases of \( T_{\tilde{\mu}(x)} \mathbb{R}^D \) and \( T_{P_{\tilde{\mu}(x)}} \mathcal{M} \).

\[\Box\]

\textbf{Proof of Corollary 6.} In choosing the optimal order of bandwidth, one can consider choosing \( (h_1, \ldots, h_m) \) such that the mean integrated squared error is minimized.

Note that

\[
\tilde{F}_E(x) - F(x) = \text{Jacob}(P)_{\mu(x)} \left( \hat{F}(x) - \mu(x) \right) + o_P(1). \tag{B.15}
\]

Here \( \text{Jacob}P \) is the Jacobian matrix of the projection map \( P \). One has

\[
\text{MISE}(\tilde{F}_E(x)) = \int \left\| \tilde{F}_E(x) - F(x) \right\|^2 dx
\]

\[
= \int \left\| \text{Jacob}(P)_{\mu(x)} \left( \hat{F}(x) - \mu(x) \right) + o_P(1) \right\|^2 dx
\]

\[
= \int \left( \sum_{i=1}^D \left( \sum_{j=1}^D \mathcal{P}_{ij} \left( \hat{F}_j(x) - \mu_j(x) \right) \right)^2 + o_P(1) \right) dx
\]

\[
= O(1/n|H|) + \ldots + O(1/n|H|) + O(h_1^4) + \ldots + O(h_m^4).
\]

The last terms follow from Fatou’s lemma, and that the Jacobian map is differentiable at \( \mu(x) \) for every \( x \). Therefore, if \( h_i \)’s \( (i = 1, \ldots, m) \) are taken to be of the same
order, that is, of $O(n^{-1/(m+4)})$, then one can obtain $\text{MISE}(\hat{F}_E(x))$ with an order of $O(n^{-4/(m+4)})$.

Proof of Theorem 7. Note that for all $x$,
\[ \hat{F}_E(x) - E\left(\hat{F}_E(x)\right) = \text{Jacob}_x(\mathcal{P}) \left( \hat{F}(x) - E\left(\hat{F}(x)\right) \right) + o_p(1). \] (B.16)

Then one has
\[ \|\hat{F}_E(x) - E\left(\hat{F}_E(x)\right)\| = \left[ \sum_{i=1}^{D} \left( \sum_{j=1}^{D} \mathcal{P}_{ij} \left( \hat{F}_j(x) - E\left(\hat{F}_j(x)\right) \right) \right) \right]^2 + g(o_p(1)), \] (B.17)

where \{\mathcal{P}_{ij}\} is the $ij$th entry of the Jacobian matrix of $\mathcal{P}$ at $x$, and $g(o_p(1))$ is a function of the term $o_p(1)$ in (B.16) which goes to zero as $n \to \infty$ with probability 1, for every $x$. Since the covariate space $\mathcal{X}$ is compact, then by Dini’s theorem (Rudin, 1976), $g(o_p(1))$ goes to zero uniformly with probability 1. Therefore, one has
\[ \sup_{x \in \mathcal{X}} \|\hat{F}_E(x) - E\left(\hat{F}_E(x)\right)\| = \sup_{x} \|\text{Jacob}(\mathcal{P}) \left( \hat{F}(x) - E\left(\hat{F}(x)\right) \right)\| + o_p(1) \] (B.18)

where $o_p(1)$ is uniform over $\mathcal{X}$. Now one looks at
\[ \sup_{x} \|\text{Jacob}(\mathcal{P}) \left( \hat{F}(x) - E\left(\hat{F}(x)\right) \right)\| = \sup_{x} \left[ \sum_{i=1}^{D} \left( \sum_{j=1}^{D} \mathcal{P}_{ij} \left( \hat{F}_j(x) - E\left(\hat{F}_j(x)\right) \right) \right) \right]^2. \] (B.19)

Note that the projection map is differentiable around the neighborhood of $\mu(x)$ and $\mathcal{X}$ is compact, so $\mathcal{P}_{ij}(x)$ are bounded. Let $C_{ij} = \sup_{x \in \mathcal{X}} \mathcal{P}_{ij}^2(x)$ and $C = \max C_{ij}$. For each term note that, by Cauchy-Schwarz,
\[ \sup_{x \in \mathcal{X}} \left( \sum_{j=1}^{D} \mathcal{P}_{ij} \left( \hat{F}_j(x) - E\left(\hat{F}_j(x)\right) \right) \right)^2 \leq \sup_{x} \sum_{j=1}^{D} \mathcal{P}_{ij}^2 \left( \hat{F}_j(x) - E\left(\hat{F}_j(x)\right) \right)^2 \] (B.20)
\[ \leq C \sum_{j=1}^{D} \sup_{x \in \mathcal{X}} \left( \hat{F}_j(x) - E\left(\hat{F}_j(x)\right) \right)^2. \] (B.21)
By Theorem 2 in Hansen (2008), one can see that
\[ \sup_{x \in \mathcal{X}} \left| \left( \hat{F}_j(x) - E(\hat{F}_j(x)) \right) \right| = O(r_n), \tag{B.22} \]
where \( r_n = \log^{1/2} n/\sqrt{n|H|}. \) Then one has
\[ \sup_{x \in \mathcal{X}} \left( \sum_{j=1}^{D} \left( \sum_{j=1}^{D} \left( \mathcal{P}_{ij} \left( \hat{F}_j(x) - E(\hat{F}_j(x)) \right) \right) \right)^2 = O(r_n^2). \tag{B.23} \]

Now let \( \nu_n(x) = \sqrt{\sum_{i=1}^{D} \left( \sum_{j=1}^{D} \mathcal{P}_{ij} \left( \hat{F}_j(x) - E(\hat{F}_j(x)) \right) \right)^2}, \) which is a sequence of functions indexed by \( n, \) then for every \( x, \) one has
\[ \nu_n(x) = O(r_n). \]
Since \( \mathcal{X} \) is compact, then by Dini’s theorem,
\[ \sup_{x \in \mathcal{X}} \nu_n(x) = O(r_n). \tag{B.24} \]

Then by (B.18) and (B.24), one has
\[ \sup_{x \in \mathcal{X}} \| \hat{F}_E(x) - E(\hat{F}_E(x)) \| = O(r_n) = O\left( \log^{1/2} n/\sqrt{n|H|} \right). \]

Proof of Theorem 8. Given the higher order smoothness assumption on \( \mu(x), \) one can make higher order approximations and using a local polynomials regression estimate would result in the reduction of bias term in estimating \( \mu(x). \) The asymptotic distribution for multivariate local regression estimator for Euclidean responses has been derived (Gu et al., 2014; Ruppert and Wand, 1994; Masry, 1996), and and we leverage on some of their results in our proof.

Note that
\[ \hat{F}(x) = \left( \hat{F}_1(x), \ldots, \hat{F}_D(x) \right) \in \mathbb{R}^D. \]
\[ E(\hat{F}(x)) = \left( E(\hat{F}_1(x)), \ldots, E(\hat{F}_D(x)) \right)^T \]

and the expectation taken in each component is with respect to the marginal distribution of \( \hat{P}(dy|x) \). Then by Theorem 1 of Gu et al. (2014), the following holds:

1. If \( p \) is odd, then for \( j = 1, \ldots, D \)
   \[
   \text{Bias}_j(\hat{F}(x)) = E(\hat{F}_j(x)) - \mu_j(x) 
   = (\mathcal{M}_p^{-1} \mathcal{B}_{p+1} \mathbf{H}^{(p+1)} \mathbf{m}_{p+1}^j(x))_1 ,
   \] (B.25)
   which is of order \( O(\|h\|^{p+1}) \). Here \((\cdot)_1\) represents the first entry of the vector inside the parenthesis;

2. If \( p \) is even, then for \( j = 1, \ldots, D \)
   \[
   \text{Bias}_j(\hat{F}(x)) = E(\hat{F}_j(x)) - \mu_j(x) 
   = \left( \sum_{l=1}^{m} h_l \frac{f_l(x)}{f_x(x)} (\mathcal{M}_p^{-1} \mathcal{B}_{p+1}^l - \mathcal{M}_p^{-1} \mathcal{M}_p^{l} \mathcal{M}_p^{-1} \mathcal{B}_{p+1}) \mathbf{H}^{(p+1)} \mathbf{m}_{p+1}^j(x) \right. 
   \]
   \[
   \left. + \mathcal{M}_p^{-1} \mathcal{B}_{p+2} \mathbf{H}^{(p+2)} \mathbf{m}_{p+2}^j(x) \right)_1 ,
   \] (B.27)
   which is of order \( O(\|h\|^{p+2}) \).

For any \( k \in \{0,1, \ldots, p\} \). Let \( N_k = \frac{(k+m-1)}{m-1} \) and \( N_p = \sum_{k=0}^{p} N_k \). Here \( \mathcal{M}_p \) is a \( N_p \times N_p \) matrix whose \((i,j)\)th block \( (0 \leq i, j \leq p) \) is given by \( \int_{\mathbb{R}^m} u^{i+j} K(u) du \) and \( \mathcal{M}_p^{l} \ (l = 1, \ldots, m) \) is a \( N_p \times N_p \) matrix whose \((i,j)\)th block \( (0 \leq i, j \leq p) \) is given by \( \int_{\mathbb{R}^m} u_l u^{i+j} K(u) du \). \( \mathcal{B}_{p+1} \) is a \( N_p \times N_{p+1} \) matrix whose \((i,p+1)\)th \((i = 1, \ldots, p)\) block is given by \( \int_{\mathbb{R}^m} u_l u^{i+p+1} K(u) du \) and \( \mathcal{B}_{p+1}^{l} \ (l = 1, \ldots, m) \) is a \( N_p \times N_{p+1} \) matrix whose \((i,p+1)\)th \((i = 1, \ldots, p)\) block is given by \( \int_{\mathbb{R}^m} u_l u^{i+p+1} K(u) du \). We have

\[
\mathbf{H}^{(p+1)} = \text{diag}\{h_1^{p+1}, \ldots, h_m^{p+1}\}, \quad f_l(x) = \frac{\partial f_x(x)}{\partial x^l} \quad \text{and} \quad \mathbf{m}_{p+1}^j(x) \ (j = 1, \ldots, D) \quad \text{is the vector of all the } p+1 \text{ order partial derivative of } \mu_j(x), \quad \text{that is, } m_{p+1}^j(x) = \begin{pmatrix}
\frac{\partial \mu_j^{p+1}(x)}{\partial (x^1)^{p+1}}, & \frac{\partial \mu_j^{p+1}(x)}{\partial (x^1)^{p} \partial (x^2)}, & \cdots, & \frac{\partial \mu_j^{p+1}(x)}{\partial (x^m)^{p+1}}
\end{pmatrix}.
\]
With \( \text{Bias}_j(\hat{F}(x)) \) \((j = 1, \ldots, D)\) given above, one has
\[
\text{Bias}(x) = E(\hat{F}(x))) - \mu(x) = \left( \text{Bias}_1(\hat{F}(x)), \ldots, \text{Bias}_D(\hat{F}(x)) \right)^T. \tag{B.28}
\]

Although higher order polynomial regression results in the reduction in the order of bias with the higher order smoothness assumptions on \( \mu(x) \), the order and expression of the covariance remains the same. That is,
\[
\Sigma_{jk} = \text{Cov}(\hat{F}_j(x), \hat{F}_k(x))
= \frac{1}{n|H|} f_X(x)^{-1} \int K(v)^2 \sigma(J_j(y_v), J_k(y_v)) dv + o \left( \frac{1}{n|H|} \right), \tag{B.29}
\]
where \( \sigma(J_j(y_v), J_k(y_v)) \) is the covariance between \( J_j(y_v) \) and \( J_k(y_v) \).

Applying the central limit theorem, one has
\[
\sqrt{n|H|} \left( \hat{F}(x) - \mu(x) - \text{Bias}(x) \right) \xrightarrow{L} N(0, \Sigma(x)) \tag{B.30}
\]
where the \( j \)th \((j = 1, \ldots, D)\) entry of \( \text{Bias}(x) \) is given in (B.25) or (B.26) depending on \( p \) is odd or even, and
\[
\Sigma_{jk} = \frac{\sigma(J_j(y_v), J_k(y_v)) \int K(v)^2 dv}{f_X(x)}. \tag{B.31}
\]

Letting \( \tilde{\mu}(x) = \mu(x) + \text{Bias}(x) \), one has
\[
\sqrt{n|H|} \left( \hat{F}_E(x) - \mathcal{P}(\tilde{\mu}(x)) \right) = \sqrt{n|H|d_{\tilde{\mu}(x)}P} \left( \hat{F}(x) - \tilde{\mu}(x) \right) + o_P(1).
\]
Therefore by applying Slutsky’s theorem, one has
\[
\sqrt{n|H|d_{\tilde{\mu}(x)}P} \left( \hat{F}(x) - \tilde{\mu}(x) \right) \xrightarrow{L} N(0, \tilde{\Sigma}(x)). \tag{B.32}
\]
Here \( \tilde{\Sigma}(x) = B^{-1}\Sigma(x)B \) where \( B \) is the \( D \times d \) matrix of the differentiable \( d_{\tilde{\mu}(x)}P \) with respect to given orthonormal bases of the tangent space \( T_{\tilde{\mu}(x)}\mathbb{R}^D \) and tangent space \( T_{\tilde{\mu}(x)}\hat{M} \).
Appendix C

Appendix to Chapter 4

Proof of Proposition 9. By Algorithm 1, the multivariate projection is obtained by a collection of sequential 1-dimensional projections. Let $w = \hat{F}(x)$ be the pre-projected estimate and $P_w = \hat{F}(x)$ be the projection of $w$. We first prove Proposition 9 holds for $p = 2$, the 2-dimensional projection. Therefore,

$$P_w = \arg\min_{G \in \mathcal{M}} \int_0^1 \int_0^1 \{w(s, t) - G(s, t)\}^2 dsdt. \quad (C.1)$$

Note that $P_w$ is the limit of $\hat{w}^{(k)}$ and $\tilde{w}^{(k)}$, where $\hat{w}^{(k)}$ is the 1-dimensional projection of $w + T^{(k-1)}$ along the $s$ direction for any $t$, and $\tilde{w}^{(k)}$ is the 1-dimensional projection of $w + S^{(k)}$ along the $t$ direction for any $s$.

Define the norm of any two-dimensional function $f(s, t)$ as $||f|| = \langle f, f \rangle^{1/2} = \left[\int \{f^2(s, t)\} dsdt\right]^{1/2}$ with $\langle \cdot, \cdot \rangle$ denoting the inner product. Then by the property of projection, one has

$$\langle w - P_w, P_w \rangle = 0, \quad (C.2)$$
and
\[ \langle w - P_w, h \rangle \leq 0 \text{ for any } h \in \mathcal{M}. \]  
(C.3)

In the following, we proceed to show that for any \( k \),
\[ ||\hat{w}^{(k)}|| \geq ||\hat{w}^{(k)}|| \geq ||\hat{w}^{(k+1)}||, \]  
(C.4)
i.e., that the norm of the sequence \( \hat{w}^{(k)} \) and \( \hat{w}^{(k)} \) is not increasing. In order to do so, we first introduce the notion of cones and dual cones of functions. Let \( C_s \) be the cone of the continuous functions \( f(s, t) \) which are monotone with respect to \( s \) for any \( t \), and \( C_t \) be the cone of continuous functions which are monotone with respect to \( t \) for any \( s \). Define their dual cones \( C^*_s \) and \( C^*_t \) as
\[
C^*_s = \left\{ g(s, t) \in C[0, 1]^2 : \int f(s, t)g(s, t)ds \leq 0, \text{ for all } t \text{ and } f \in C_s \right\},
\]
and
\[
C^*_t = \left\{ g(s, t) \in C[0, 1]^2 : \int f(s, t)g(s, t)dt \leq 0, \text{ for all } s \text{ and } f \in C_t \right\}.
\]
Denote \( P(w \mid C_s) \) as the projection of \( w \) over \( C_s \) found by minimizing \( \int (w - f)^2 ds \) for all \( f \in C_s \) and any fixed \( t \). Denote \( P(w \mid C_t) \) as the projection of \( w \) over \( C_t \) found by minimizing \( \int (w - f)^2 dt \) for all \( f \in C_t \) and any fixed \( s \). By Lemma A1 of Lin and Dunson (2014), the following holds:
\[
P(w \mid C^*_s) = w - P(w \mid C_s) \text{ and } P(w \mid C^*_t) = w - P(w \mid C_t). \]  
(C.5)
Note that \( -S^{(k+1)} = (w + T^{(k)}) - \hat{w}^{(k)} = (w + T^{(k)}) - P(w + T^{(k)} \mid C_s) = P(w + T^{(k)} \mid C^*_s) \)
where the last equality follows from (C.5). Here \( P(w + T^{(k)} \mid C_s) \) denotes the projection of \( w + T^{(k)} \) onto \( C_s \) and \( P(w + T^{(k)} \mid C^*_s) \) is the projection onto \( C^*_s \). Therefore \( -S^{(k+1)} \) minimizes \( ||(w + T^{(k)}) - f|| \) for all \( f \in C^*_s \) and \( -T^{(k)} \) minimizes \( ||(w + S^{(k)}) - f|| \) for all \( f \in C^*_t \). Then, one concludes that
\[
||\hat{w}^{(k)}|| = ||w + S^{(k)} - (-T^{(k-1)})|| \geq ||w + S^{(k)} - (-T^{(k)})|| \geq ||w + T^{(k)} - (-S^{(k+1)})||
\]
for all \( k \). Therefore, one has \( ||\hat{\mu}^{(k)}|| \geq ||\hat{\mu}^{(k+1)}|| \).

Now, for sufficiently large \( k \), one has
\[
\|P_w\| \leq \|\hat{\mu}^{(k+1)}\| \leq \|\hat{\mu}^{(k)}\| \leq \|\hat{\mu}^{(k)}\| \leq \ldots \leq \|w\|. 
\] (C.6)

By the above equation and (C.3),
\[
\|P_w - F\|^2 = \|P_w\|^2 + \|F\|^2 - 2\langle F, P_w \rangle \\
\leq \|w\|^2 + \|F\|^2 - 2\langle F, w \rangle \\
= \|w - F\|^2,
\]
proving our contention.

\[ \square \]

**Lemma 15.** Denote \( \mathcal{C} \) as the convex cone of a function on some domain set \( X = [0,1]^p \), which includes \( \mathcal{M} \), the convex cone of monotone functions on \( X \). Let \( g \) be any function on \( X \) and \( g^* \in \mathcal{C} \) such that
\[
g^* = \arg \min_{f \in \mathcal{C}} \int_X (g(x) - f(x))^2 \, dx.
\] (C.7)

Then, one has for every \( f \in \mathcal{C} \),
\[
\int_{x \in X} (g(x) - g^*(x))(g^*(x) - f(x)) \, dx \geq 0,
\] (C.8)
so that
\[
\int_{x \in X} (g(x) - g^*(x))g^*(x) \, dx = 0,
\] (C.9)
and
\[
\int_{x \in X} (g(x) - g^*(x))f(x) \, dx \leq 0.
\] (C.10)

**Proof.** By the definition of a convex cone, for any \( \alpha \in [0,1] \) and any \( f \in \mathcal{C} \), \( (1 - \alpha)g^* + \alpha f \in \mathcal{C} \). Then
\[
\int_X (g(x) - ((1 - \alpha)g^* + \alpha f))^2 \, dx
\]

achieves its minimum at $\alpha = 0$. Now, take the derivative of the above objective function to find

$$2 \int_X (g(x) - ((1 - \alpha)g^* + \alpha f))(g^*(x) - f(x))dx,$$

which is non-negative at $\alpha = 0$. Therefore,

$$\int_{x\in X} (g(x) - g^*(x))(g^*(x) - f(x))dx \geq 0.$$

Now let $f(x) = cg^*(x)$, so that

$$\int_{x\in X} (g(x) - g^*(x))(1 - c)g^*(x)dx \geq 0.$$

By letting $0 < c \leq 1$, then $c \geq 1$, one can see that

$$\int_{x\in X} (g(x) - g^*(x))g^*(x)dx = 0,$$

which further implies,

$$\int_{x\in X} (g(x) - g^*(x))f(x)dx \leq 0.$$

\[ \square \]

**Proof of Theorem 11.** Let $\phi(u)$ be the derivative of the convex function $\Phi(u)$. By the property of convex functions, one has

$$\Phi(v) - \Phi(u) \geq (v - u)\phi(u).$$

Let $u = \tilde{F}(x) - F(x)$ and $v = \hat{F}(x) - F(x)$. Then

$$\Phi(\tilde{F}(x) - F(x)) \geq \Phi \left( \tilde{F}(x) - F(x) \right) + \left( \tilde{F}(x) - \hat{F}(x) \right) \phi \left( \tilde{F}(x) - F(x) \right).$$

Thus,

$$\int_0^1 \Phi \left( \tilde{F}(x) - F(x) \right) dx \geq \int_0^1 \Phi \left( \tilde{F}(x) - F(x) \right) dx$$

$$+ \int_0^1 \left( \tilde{F}(x) - \hat{F}(x) \right) \phi \left( \tilde{F}(x) - F(x) \right) dx.$$
It suffices to show that
\[ \int_0^1 \left( \hat{F}(x) - \bar{F}(x) \right) \phi \left( \bar{F}(x) - F(x) \right) dx \geq 0. \]

Note that \( \hat{F}(x) \) is the slope of \( T(\hat{F}(x)) \), the greatest convex minorant of \( \hat{F}(x) = \int_0^x \hat{F}(s) ds \). One can write the unit interval \([0,1]\) as the union of the sets \( \{ x : T(\hat{F}(x)) = \hat{F}(x) \} \), over which the function \( \hat{F}(x) \) is monotone, and the disjoint open sets \( \{ x : T(\hat{F}(x)) < \hat{F}(x) \} \). One can see that over each of the disjoint sets \( \{ x : T(\hat{F}(x)) < \hat{F}(x) \} \), \( T(\hat{F}(x)) \) is a linear function. (Otherwise, one can always construct a convex function above it, leading to a contradiction.) Therefore, \( \hat{F}(x) \), which is the derivative of \( T(\hat{F}(x)) \), is a constant function over each of these sets. Let \( \{ x : T(\hat{F}(x)) < \hat{F}(x) \} = \cup U_i \) where \( U_i \) and \( U_j \) are disjoint intervals for \( i \neq j \). Let \( \bar{F}(x) = c_i \) over the set \( U_i \). One has \( c_i = \frac{\int_{U_i} \hat{F}(x) dx}{|U_i|} \) (see Lemma 2 of Groeneboom and Jongbloed (2010)), where \( |U_i| \) denote the length of the intervals, which can be viewed as the projection of \( \hat{F}(x) \) restricted to the set \( U_i \). Then
\[ \int_0^1 \left( \hat{F}(x) - \bar{F}(x) \right) \phi \left( \bar{F}(x) - F(x) \right) dx = \sum_i \int_{U_i} \left( \hat{F}(x) - \bar{F}(x) \right) \phi \left( \bar{F}(x) - F(x) \right) dx. \]

Since \( F(x) \) is a monotone increasing function, \( -F \) is decreasing, thus
\[ \phi \left( \hat{F}(x) - F(x) \right) \]
is decreasing over \( U_i \). Then by equation (C.10) of Lemma 15, one has
\[ \int_{U_i} \left( \hat{F}(x) - \bar{F}(x) \right) \phi \left( \bar{F}(x) - F(x) \right) dx \geq 0. \]

Then
\[ \int_0^1 \left( \hat{F}(x) - \bar{F}(x) \right) \phi \left( \bar{F}(x) - F(x) \right) dx \geq 0, \]
which implies that

\[ \int_0^1 \Phi \left( \tilde{F}(x) - F(x) \right) dx \leq \int_0^1 \Phi \left( \hat{F}(x) - F(x) \right) dx. \]
Bibliography


Biography

Brian Stephen St. Thomas was born December 15, 1988 in Wellesley, Massachusetts. In 2011 he graduated *summa cum laude* from Bates College. He will earn his Ph.D. from Duke University in 2015. In the summer of 2015 he will join Digitalsmiths, a TiVo subsidiary, as a Data Scientist.