5-2023

Geometric Inference in Machine Learning: Applications of Fisher Information for Model Selection and Other Statistical Applications

Trevor Herntier

Geometric Inference in Machine Learning: Applications of Fisher Information for Model Selection and Other Statistical Applications

by

Trevor Herntier

A dissertation
submitted to the College of Engineering and Science of Florida Institute of Technology
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in
Systems Engineering

Melbourne, Florida
May, 2023

We the undersigned committee
hereby approve the attached dissertation

Geometric Inference in Machine Learning: Applications of Fisher Information for Model Selection and Other Statistical Applications by Trevor Herntier

---

Adrian M. Peter, Ph.D.
Associate Professor
Computer Engineering and Sciences
Major Advisor

Georgios Anagnostopoulos, Ph.D.
Associate Professor
Computer Engineering and Sciences

Luis Daniel Otero, Ph.D.
Associate Professor
Computer Engineering and Sciences

William Arrasmith, Ph.D..
Professor
Computer Engineering and Sciences

Philip J. Bernhard, Ph.D.
Associate Professor and Department Head
Computer Engineering and Sciences

ABSTRACT

Geometric Inference in Machine Learning: Applications of Fisher Information for Model Selection and Other Statistical Applications

Trevor Herntier

We consider the problem of model selection using the Minimum Description Length (MDL) criterion for distributions with parameters on the hypersphere. Model selection algorithms aim to find a compromise between goodness of fit and model complexity. Variables often considered for complexity penalties involve number of parameters, sample size and shape of the parameter space, with the penalty term often referred to as stochastic complexity. Because Laplace approximation techniques yield inaccurate results for curved spaces, existing criteria incorrectly penalize complexity. We demonstrate how the use of a constrained Laplace approximation on the hypersphere yields a novel complexity measure that more accurately reflects the geometry of these spherical parameters spaces. We refer to this modified model selection criterion as *spherical MDL*. As proof of concept, spherical MDL is used for bin selection in histogram density estimation, performing favorably against other model selection criteria.

Furthermore, we consider the problem of identifying the most similar distribution from a constrained set to a given distribution. We measure similarity using a symmetric distance on the manifold governed by the Fisher information metric, with a smaller distance on the manifold indicating distributions that are more similar. For the most part, research into the geodesic problem on manifolds is limited to the paths between two known distribution. Allowing one or both of the endpoint distributions to belong to a constrained surface on the manifold requires the introduction of *transversality conditions*

and the techniques from variational calculus. We show the efficacy of this approach by applying it to different manifolds and constraint surfaces, including Gaussian manifolds with the isotropic constraint surface.

Table of Contents

Abstract iii

List of Figures viii

List of Tables x

Acknowledgments xi

Dedication xii

1 Introduction 1

1.1 Motivation 1

1.2 Main Contributions 2

1.3 Introduction to Model Selection 4

1.4 Introduction to Geodesics on Manifolds 7

2 Related Works 9

3 Spherical Minimum Description Length 15

3.1 Bayesian Approach to Model Selection 15

3.1.1 Comparing Models 15

3.1.2 An Inappropriate Prior 16
3.1.3 Geometry of Probabilistic Models 17
3.1.4 Fisher Information 18
3.1.5 Observed vs. Expected Fisher Information 19
3.1.6 An Appropriate Prior 23
3.2 Asymptotic MDL in $\mathbb{R}^K$ 25
3.3 Spherical MDL 27
3.3.1 Derivation of the Spherical MDL Criterion 27
3.3.2 Riemannian Volume of a Hypersphere 34
3.4 Case Study: Spherical MDL for Histograms 37
3.4.1 Theoretical Development 38
3.4.2 Experimental Results 43

4 Geodesics and Transversality Conditions 45

4.1 Background 45
4.2 Euler-Lagrange Equation 48
4.3 Transversality Conditions 53
4.4 Euler-Lagrange for univariate Gaussian 62
4.4.1 Euler-Lagrange Equation for Gaussian 62
4.4.2 Example for Gaussian Variable Endpoint 66
4.5 Euler-Lagrange Equation for Multivariate Gaussian Distributions 71
4.5.1 Euler-LaGrange equations for Bivariate Gaussian Distributions . 73
4.6 Results 75
4.6.1 Isotropic Terminal Distribution 75
4.6.1.1 Constant Mean Vector 76

4.6.1.2 Constant Mean Vector with Initial Covariance 77
4.6.1.3 Starting on the Constraint 79
4.6.2 Variable Initial and Final Conditions 79
4.7 Normal Approximation to the Poisson Distribution 85

5 Conclusions and Further Study **92**
5.1 Conclusions 92
5.2 Further Study 93
5.2.1 Future in Model Selection 93
5.2.2 Geodesics 94

6 Publication **96**

A Proof **104**
A.1 Fisher Information for Gaussian 104
A.2 Fisher Information of a 2-Dimensional Gaussian 108
A.3 Euler-Lagrange on 2-Sphere 112
A.3.1 Transversality Conditions on the 2-Sphere 114

List of Figures

3.1 Bernoulli Fisher Information. 20

3.2 Fisher information for Bernoulli random variable $\theta = 0.75$ 22

3.3 Riemannian volume of hyperspheres. 35

3.4 Bimodal, skewed, trimodal and claw densities. 44

4.1 Optimal path and neighboring path with perturbing function. 49

4.2 General transversality condition. 54

4.3 Area under neighboring function as a result of transversality condition. . 56

4.4 Approximation of additional area using MVT for integrals. 57

4.5 Contributors to δy as a result of variable endpoint. 58

4.6 Fisher information plot in parameter space of univariate Gaussian. . . . 64

4.7 Visual comparison of normal distributions with different variances. . . . 65

4.8 Shortest path from $N(0, 0.5)$ to $\sigma = \mu^2$. 68

4.9 Shortest path from $N(0, 0.5)$ to $\sigma = \mu^2$ as terminal surface. 68

4.10 Evolution of Gaussian from N(0,0.5) to final distribution. 69

4.11 Shortest path to $S(\mu) = \mu - 7 = 0$. 70

4.12 Evolution of distributions to $\mu = 7$. 70

4.13 Isotropic example 1. 77

4.14 Isotropic example 2. 78

4.15 Starting on transversality constraint. 80
4.16 Error ellipses showing movement of distribution along constraint. . . . 81
4.17 Starting on constraint, example 2. 82
4.18 Error ellipses showing movement of distribution from Quadrant III to Quadrant I. 83
4.19 Initial and final variable endpoint. 84
4.20 Error ellipses form initial and final variable endpoints. 85
4.21 Three Poisson distributions. 86
4.22 Normal approximation to the Poisson distribution. 88
4.23 Errors in the normal approximation to the Poisson distributions. 90

A.1 Example of transversality on 2-sphere. 116

List of Tables

3.1 Fisher information for Bernoulli random variable $\theta = 0.5$ 21

3.2 Fisher information for Bernoulli random variable $\theta = 0.75$. 22

3.3 Vertex distance of hypercubes and dimension. 37

3.4 Comparison of different model selection criteria for histogram density estimation. 44

4.1 Errors in normal approximation for $\lambda = 30$. 89

4.2 Errors in normal approximation for $\lambda = 5$. 91

Acknowledgements

Thanks to Dr. Adrian Peter for the constant patience, pushing and perspective.

Dedication

To Matt and Eliza, you were there from the beginning , and to Sam, you almost made it to the end. Finally to my 2022 Multivariable Calculus class for always asking about the "nonsense" on which I spend my free time.

Chapter 1

Introduction

1.1 Motivation

At the heart of all machine learning applications is understanding the distribution of the data. After all, every machine learning algorithm is at the mercy of the data it is provided. The probability distributions that generate data can be considered points on a statistical manifold, the study of which is known as information geometry. Information geometry allows us to exploit aspects of our prior knowledge by understanding how the distributions that describe that data interact with other distributions in the model family with rules governed by the geometry of the manifold they are on. Additionally, it opens up geometric vocabulary allowing researchers to reason intuitively about statistical problems.

The geometry of these manifolds precisely links the data to the parameters, giving the parameters context and the adding clarity to the task at hand. These manifolds are Riemannian manifolds and, accordingly, are endowed with a natural geometry that allows for the definition of common geometrical concepts. Unlocking this geometry on the manifold requires an appropriate *metric tensor* or *metric*. This metric tensor provides a notion of distance on the manifold, which allows the measurement of concepts such as surface areas on the manifold (sometimes referred to as Riemannian volumes) both locally and globally and arc lengths between two distributions In [40], it is shown that the the appropriate tensor is the Fisher Information matrix since it is invariant to reparameterisation of the distributions occupying the manifold.

Most importantly for this research is that the geometry of these statistical manifolds allows for a comparison between the distributions using the properties of the manifold. These comparisons will allow for the identification of optimal distributions given sampled data as well as quantify the magnitude of dissimilarities between distributions, which are sometimes measured by divergence formulae, like Kullback-Leibler [61] or cross-entropy [92]. As effective as these divergence formulae are, perhaps a more in-

stinctual comparison of distributions would be the symmetric distance between their locations on the manifold, something that divergence formulae do not provide, in general.

Armed with a data set and this idea of distance on a manifold, information geometry is well suited to aid the choice of which model on a manifold best fits the distribution suggested by the data. Model Selection research has uncovered many criterion to determine the optimal distribution, each with its own unique definition of optimal. In this definition, most criteria ignore the geometry of the manifolds at worst and, at best, inappropriately penalize distributions on curved manifolds. Developing a suitable model selection criterion that properly accounts for this geometry is a key motivation of this research

Furthermore, the study of distances on these manifolds leads directly to the properties of the shortest path between two points, or geodesics. Geodesics are interesting in the study of geometry because they provide insight into the curvature of a surface. The behavior of geodesics on a surface can be used to study the global structure of the surface.and to identify which distributions on the manifold are closest to a chosen model. In this research,we will make use of techniques from variational calculus to develop a novel use of geodesics on Riemannian manifolds. As will be seen with this research, working with the geometry of these manifolds can further applications in machine learning and model selection.

Typically, research regarding geodesics involves the distance between two points on a manifold, almost entirely focused on closed form expressions for the that distance. Absent from current research is relaxing the boundary conditions for geodesics by allowing one of the endpoints of the geodesics some degrees of freedom to move along the manifold. Here,we relax one or both boundary conditions, allowing the geodesic to start or end on a subsurface on the manifold. This shift in thinking expands the questions that can be answered. Instead of simply asking what is the distance between two distributions, the question becomes what is the closest isotropic distribution, for example. Having a variable boundary condition opens up countless new research questions and applications.

1.2 Main Contributions

The proposed research makes several unique contributions to advancing the exploration of machine learning and model selection using information geometry to compare statistical distributions. Most significantly:

- A correct Laplace approximation on hyperspherical manifolds;
- A novel model selection criteria for distributions with parameters residing on hyperspheres called spherical Minimum Description Length;
- Incorporating a logical complexity penalty with a geometric interpretation for distributions on hyperspheres

- Illustrate the efficacy of spherical Minimum Description Length by choosing appropriate bin numbers for histogram density estimators;
- Uses of the Fisher Information matrix to find the shortest path between a known distribution on a variety of statistical manifolds and a subset of distributions on the manifold defined by a constraint surface;
- Derive general transversaility conditions for the Multivariate Gaussian distrubtion
- Employ transversality conditions on the manifold of Guassian distributions to satisfy isotroptic boundary conditions;
- Improving the normal approximation to the Poisson Distribution by selecting an appropriate Gaussian distribution from the submanifold of univariate Gaussian distributions with equal variance and mean that is closest to a distribution realized from data.

More specifically and with regards to model selection, we will use differential geometry to develop a novel model selection criteria for distributions with hyperspherical parameters, called spherical Minimum Description length with a pleasing geometric interpretation of the penalty term. Separating this criterion for ones already existing in the literature will be correcting the Laplace approximation for curved parameter spaces. With regards to the study of geodesics,we will expand on the current research which typically restricts itself to finding the shortest path between two well defined points by introducing variable endpoints, or transversality condition. This shift in thinking allows asking more pertinent questions by focusing on finding the closest distribution to a given distribution instead of limiting questions to length of the shortest path.

The remainder of this document is organized as follows: In Section 1.3, we provide a brief introduciton to the key principles of model selection in which we present pioneering concepts in the field, to be expanded later in the document. In Section 1.4 we introduce the concept of geodesics on manifolds and their role in measuring a distance between statistical distributions, something that is foreign without the ideas of information geometry.

Chapter 2 covers the most relevant works to the present development, including some historical reflection on various model selection criteria. Also in this section will provide a summary of current relevant works regarding techniques of measuring distances along curved parameter spaces using the Fisher Information matrix as the metric.

In Chapter 3, a detailed analysis of model selection is found. With an ambitious goal of making the paper more self-contained,we briefly recap the background of model selection for a Bayesian perspective in Section 3.1, paying specific attention to geometric motivations behind the use of the Fisher information matrix. Section 3.2 details the geometric derivation of MDL in $\mathbb{R}^K$. An approach not as familiar as the original information theoretic formulation, yet enabling the analogous development of MDL on the

hypersphere $\mathbb{S}^{K-1}$ is detailed in Section 3.3. By comparing the developments in $\mathbb{R}^K$ and $\mathbb{S}^{K-1}$, one can readily see where the modifications must be made for the constrained parameter space. Next, in Section 3.4,we consider a practical application of spherical MDL for selecting the bin width of the ubiquitous histogram. Our experimental results validate the utility of spherical MDL in comparison to other state-of-the-art model selection criteria.

The paper then makes a transition to exploring techniques of calculus of variation on statistical manifolds in Chapter 4 with the ultimate goal of finding the shortest path between a given distribution on the manifold and a final distribution somewhere on a surface residing on the manifold. This will require a brief introduction to the techniques of variational calculus, which is provided in Section 4.1. This outline is far from exhaustive and is just to provide minimal foundation for the current topic. With that in mind, the Euler-Lagrange equation and transversality conditions are paramount for the current research, so a reasonably detailed explanation and proof are given in Section 4.2 and Section 4.3, respectively. In Section A.3, the Euler-Lagrange equation with transversality conditions are applied on the 2-sphere, to show proof of concept with the goal of applying it to more general probability distributions.

Once the proficiency of transversality conditions to select the closest model on the constraint surface is established, we then apply them by first examining geodesics on the univariate Gaussian manifold between a distribution and constraint surface, as shown in Section 4.4.

Next, in Section 4.5 we study the Fisher information of the multivariate Gaussian use the results to find geodesics on the bivariate Guassian distribution on the isotropic constraint surface. Additionally, in order to explore the limits of transversality conditions, we use them to find geodesics with both variable initial and terminal boundaries as well as solving for a geodesic when the intitial distribution already satisfies the contsraint surface we show how transversality conditions on the univariate Gaussian can better the normal approximation to the Poisson distribution given sampled data, a technique that receives far less relative attention compared to the binomial distribution, especially when considering the usefulness of the Poisson distribution.

Finally, in Chapter 5, we provide some further direction to logical extensions of this current research.

1.3 Introduction to Model Selection

The premise of model selection is to objectively choose, from a set of competing models, one that most parsimoniously obtains a good fit to the observed data. The difficulty arises from the fact that goodness of fit and parsimony are inherently conflicting properties. A more philosophical view is sufficiently captured by Ockham's razor: "Pluralities are never to be put forward without necessity." The widely established measure of good-

ness of fit is the likelihood of the observed data. With this issue settled for the most part, research has focused on how to penalize models that overfit the data. Almost all popular model selection criteria differ primarily on the method of this penalizing factor. Simple penalties can depend on only the number of parameters of the model and perhaps the sample size, while more complex criteria take into account the geometric complexity of the parameter manifold. In this paper,we revisit the geometry associated with the complexity measure for the Minimum Description Length (MDL) criterion [87, 84]. We note that almost all of this previous development is restricted to unconstrained parameter spaces. In this paper,we are mainly interested in model selection criteria when parameters are implicitly constrained.

A simple way to accommodate constraints in parametric models is the explicit removal of the constrained set leaving behind a reduced set of unconstrained parameters. Unfortunately, this is difficult to analytically perform when the constraints are nonlinear. In the present paper, we show that a constrained MDL-like criteria can be derived in such situations, referring to this new model selection criterion as *spherical MDL*. We also show that there is no requirement to explicitly reduce the set of parameters to a smaller unconstrained set. Instead, we work with the constraints implicitly, extending MDL naturally to such situations. We argue that this opens up MDL to more interesting and constrained parametric models than hitherto seen in the literature. Before introducing spherical MDL and the general methodology behind constrained parametric spaces, we present a simplified version of the current model complexity landscape.

Paramount to every criterion is the value it places on parametric complexity. When making a decision, however, this value is not the greatest concern. It would be natural to think that if models with few parameters are chosen consistently by a a certain criteria, it must be placing a large complexity penalty on models with many parameters. While this may be true, what is actually happening is that, according to this criterion, when models get more complex the *increase* in penalty is larger, making it more undesirable to choose the next most complicated model. In other words, a K parameter model may be considered extremely complex, but if the $K+1$ parameter model isn't exceedingly more complex, there is little harm in choosing the $K+1$ parameter model. Every model selection criterion compares this increase in complexity from one model to the next to the improvement in fit and makes its choice accordingly.

Arguably, the three most widely used selection criteria are Akaike's information criterion (AIC) and its incarnations [3, 2, 52], Bayesian information criterion (BIC) [91] and Minimum Description Length (MDL). The AIC criterion is given by

$$AIC = -2\log\ f(X;\hat{\theta}) + 2K \tag{1.1}$$

and BIC

$$BIC = -2\log\ f(X;\hat{\theta}) + K\log(N), \tag{1.2}$$

where $X = \left\{x_i\right\}_{i=1}^{N}$ is the observed data, K is the cardinality of the parameters in the

candidate model, N is the sample size and $\log\ f(X;\hat{\theta})$ is the log-likelihood of the model evaluated at the maximum likelihood estimate (MLE) $\hat{\theta}$. In the original forms of Equations (1.1) and (1.2), the parameters in the set θ specific to the density f are assumed to lie in an Euclidean space, i.e., $\theta \in \mathbb{R}^K$. The candidate model which minimizes the above in each case will be the appropriate model for the data according to the respective criteria. Both criteria use the negative log-likelihood as the measure for goodness of fit and employ similar complexity penalties that reward paucity of parameters. However, BIC includes the sample size in its penalty term and will tend to choose less complex models as more data is collected. Interestingly, in Equation (1.1), the penalty term can be derived principally from a bias correction between the unknown true model and approximation by the selected model family (and, for more details, see [74]).

The criticisms of the complexity penalties for AIC and BIC are tied to the failure of either to consider how the parameters interact within the model. This shortcoming was addressed in [87, 88, 85] with the introduction of the Minimum Description Length principle

$$MDL = -\log\ f(X;\hat{\theta}) + \frac{K}{2}\log\left(\frac{N}{2\pi}\right) + \log\int \sqrt{\det I(\theta)}d\theta, \tag{1.3}$$

where $I(\theta)$ is the Fisher information matrix. Even though the predecessors to MDL acted as inspirations, Rissanen approached model selection from a unique perspective, that of information theory as opposed to probability theory. Both schools of thought use data to select an appropriate model that can be used to explain the data. However, where probability models aim at searching for the true underlying distribution that generated the data, MDL merely looks at compressing the data. In fact, Rissanen argues [47] that it is entirely inappropriate to look for this "true" distribution since the existence of it is questionable and, as such, the task of trying to estimate it is impracticable. This leaves MDL with the central idea of finding regularities in data and to use these to compress the data such that the data can be described using less symbols. Data is compressed by means of a code and models that offer shorter code lengths are considered to describe the data better. Even though MDL doesn't concern itself with finding the "true" model, the search for regularities in the data often results in identifying the distribution which generated the data [47].

In this work, and, as mentioned above,we propose a novel MDL-like criterion specifically designed for models with spherical parameter spaces, i.e., $\theta \in \mathbb{S}^{K-1}$. We derive the new criterion by revisiting the geometric derivation of MDL—as opposed to its original code-length inspired formulation—and show how when dealing with spherical parametric spaces one can constrain the Laplace approximation to respect this geometry.

The geometric derivation of MDL [9] is predicated on carving up parametric manifolds into disjoint regions within which parametric models are indistinguishable. While this approach *prima facie* looks quite different from the standard MDL code length approach, it is shown that the geometric derivation is entirely equivalent to standard MDL.

We begin with this geometric approach in the present work since the carving up of parametric manifolds into disjoint regions can be readily extended to constrained parameter spaces. As we show, for the case of spherical MDL, this results in a new complexity term that penalizes based on the normalizing constant of the Fisher–Bingham distribution [59] generalized appropriately to higher dimensions. The MDL criterion, as it is presently formulated, assumes that parameters lie in a Euclidean $\mathbb{R}^K$ space and are otherwise unconstrained. Asymptotic analyses based on this model are prone to inaccuracies for spherical models. In the remainder of this work,we detail the theoretical connections with the original MDL criterion and more importantly offer insight into the interpretation of spherical MDL in the context of distinguishable distributions in the model space.

1.4 Introduction to Geodesics on Manifolds

The importance of a metric to measure the similarity between two distributions has weaved itself into a plethora of applications. Fields concerning statistical inference [58, 4, 37], model selection [2, 87, 51] and machine learning have found it necessary to quantify the likeness of two distributions. A common approach to measure this similarity is to define a divergence between distributions using the tenets of information geometry, e.g., the Fisher–Rao distance or the f-divergence [81], respectively.

With information geometry, it is possible to define a distance between two statistical distributions. In the context of information geometry, statistical distributions are represented by points on a statistical manifold. Of the many paths that connect two points on a manifold, the minimizing path is known as a geodesic. The first rigorous mathematical treatment of geodesics and the geodesic problem is usually attributed to the German mathematician Carl Friedrich Gauss, who published a paper on the topic in 1828 [43]. It was in this paper that Gauss defined the geodesic as the shortest path between two points on a surface and derived the geodesic equation which, in the context of his research, described the path of a particle moving under the influence of gravity on a curved surface. Since Gauss's work, the geodesic problem has been studied extensively by mathematicians and physicists, and has applications in a wide range of fields, including geometry, topology, physics, engineering, and computer science. Because of its varied use, studying the behavior of geodesics has become an attractive topic for researchers, most of the time finding a closed form equation for the length of a geodesic. On some manifolds, this has been a fruitful endeavour, with satisfying equations defining this distance. However, on many statistical manifolds, closed form solutions for the length of a geodesic remains unsolved.

To the best of our knowledge, research and results in information geometry have predominantly focused on establishing similarities between two given distributions. Here, we consider an important class of problems where one or both endpoint distributions are not fixed, but instead, constrained to live on subset of the parameter manifold.

When one relaxes the fixed endpoint requirements, the development of finding the shortest path between a given distribution and constraint surface (not single distribution) must be reconsidered using transversality conditions [45, 44] for the standard length-minimizing functional. This is precisely the focus of the present work, where we derive the transversality conditions for working in the Riemannian space of multivariate Gaussian distributions. This approach opens new avenues for research and application areas where one no longer needs to provide the ending distribution but rather a description of the constraint set where the most similar model must be discovered.

Chapter 2

Related Works

In this chapter, we provide a history and comparison of the effectiveness of popular model selection criterion and how geometry aided in the evolution an interpretations of the field. Rissanen's Minimum Description Length is looked at in more detail, considering it was the first to employ the geometric structures of statistical manifolds in its complexity penalty. Furthermore, the Minimum Description length has matured through its development, applications and interpretations, which is germane to my research.

Next, relevant current and historical research are presented that motivated this work on the applications of the geometry of statistical manifolds, beginning with a history of divergence measures on manifolds and differential geometry. Then, using divergence measures, main results of shortest paths between two distributions on manifolds are discussed. Additionally, some important applications of the calculus of variations are outlined, showing its efficacy on a wide range of research areas.

Statistical manifolds are a mathematical framework used to study the geometry of probability distributions. Rao introduced the concept in [80], where he established that each unique set of parameters of a probability distribution can reside on a statistical manifold. He showed that these manifolds can be endowed with a geometric structure, paving the way for further studies in information geometry. Rao based this geometric structure on the works of Fisher [42], where the basis of the Fisher information was first introduced

Decades later, the seminal work of Rao was expanded on by Amari in [6, 4, 5]. His approached focused on the Riemannian geometry of the manifold and proposed invariant divergence measures that allow for calculating the distance between distributions on the manifold. His work with the Fisher metric also provided insight into the sensitivity of the probability distribution to changes in the underlying parameters, which reveals information about the local geometry around a particular distribution.

These works laid the foundation for the applications of information geometry in a variety of statistical fields. Kass explored how the distance between distributions has ap-

plications to inference [58]. Here, Kass showed how ideas in statistical inference merge nicely with ideas from information geometry. Recognizing the unconventional use of geometric ideas in While doing so, he was a strong advocate for using geometrical concepts into Bayesian statistical ideas.

As outlined above, some of the pioneering ideas of model selection criteria are credited to Akaike [2] with his development of Aikaike's Information Criterion, or AIC. Akaike's original idea generated numerous variants of AIC involving bootstrapping [29] and small sample corrections [28]. Even with these corrections, the AIC is still considered to under-penalize complex models. With the Bayesian Information Criterion, Schwarz [91] tried to address some of these issues. However, neither of these popular approaches consider the geometry of the manifold. The functional relationship between the parameters and the distribution wasn't explored until Rissanen's Minimum Desciption Length, yet still not with a geometric interpretaion.

Rissanen's first offering was an early two part code version of MDL. Originally, the MDL criterion was given by

$$MDL = -\log\ f(X;\theta) + \frac{K}{2}\log\left(\frac{N}{2\pi}\right) \tag{2.1}$$

and later evolved to become the three part code seen in Equation (1.3). Similar to AIC and BIC, this two part code fails to penalize for the geometry of the parameter manifold. The third term in Equation (1.3) penalizes a model for geometric complexity by incorporating the Riemannian volume [19] of the parameter manifold. MDL deviates from AIC [3, 2] and BIC [91] in that its objective is not to search for the underlying true model, but to encode regularities in the data. The difficulty with this is that the optimal distribution in the family is required to describe the data properly, but also requires too much information to be optimal. This motivated the idea of identifying a universally represented distribution from a model family, one that compresses every data set almost as well as the best model for every single unique data set. Rissanen coined the term *stochastic complexity* to describe the code length associated with this universal distribution.

In [11], the normalized maximum likelihood (NML) was shown to be the universal distribution of every model family. Specifically, the probability distribution associated with the NML distribution is

$$p(X) = \frac{f(X|\hat{\theta}_X)}{\int f(Y|\hat{\theta}_Y)dY}, \tag{2.2}$$

where X denotes the collected data, Y represents any potential data set that could be observed by the experiment and $\hat{\theta}_X$ denotes their respective maximum likelihood estimate for X with a similar notation used for Y. The normalizing constant, $\int f(Y|\hat{\theta}_Y)dY$, for the distribution can be thought of as the sum of all maximum likelihood estimates from all possible data sets the experiment could generate. The code length associated with

this distribution is found by taking the negative logarithm of Equation (2.2)

$$SC = -\log\ f(X|\hat{\theta}_X) + \log \int f(Y|\hat{\theta}_Y)dY \tag{2.3}$$

and provides us with the mathematical definition of stochastic complexity. Like all other model selection criteria, the first term is a goodness of fit term and the last term is a penalty for complexity, which is sometimes referred to as the parametric complexity and is independent of the data in the sample set. The model that minimizes the stochastic complexity is the one which MDL would choose as optimal.

As elegant as the NML definition of stochastic complexity is, it is not without its flaws. Mainly, the normalizing integral is usually computationally costly to compute making the NML distribution elusive in general and as such it is difficult to compare the stochastic complexity of competing models. In fact, this normalizing integral may not even be finite, a problem which has been named the *infinity problem* [47]. Several solutions have been proposed to fix the infinity problem [86, 101], but only in specific cases. Without a satisfactory solution to the problem in general, the NML definition of stochastic complexity is limited in its practical applicability.

Recognizing these issues, an asymptotic formula for the stochastic complexity was derived for larger sample sizes by Balasubramanian in [9]. A brief proof of this formula will be provided in Section 3.2. The penalty for complexity in this asymptotic formula can be understood in terms of the geometry of the statistical manifold on which the parameters reside. Briefly, instead of trying to compress data using regularities within it, Balasubramanian defines stochastic complexity as the ratio of the volume of an ellipsoid near the MLE to the volume of the entire manifold. An undesirable model would be one in which this ellipsoid is very small when compared to the volume of the entire manifold. We leverage similar geometric arguments when developing spherical MDL in Section 3.3.

Prior to Rissanen's MDL, the Minimum Message Length (MML) was introduced in [23]. Rissanen's MDL is, at its foundation, similar to MML in the sense that both selection criteria aim at finding the model that minimizes the code length that is used to describe the data. However, MDL and MML differ in two important facets [22]. First, MML assumes a prior distribution over the parameters, whereas MDL does not. Intuitively, this prior distribution requires a code length so the code length terms in MDL are inherently shorter. Secondly, the goal of MML is to find the best specific model for the given data. In fact, MML is almost unconcerned as to which model family the selected model belongs. In contrast, MDL searches just for a model class that minimizes the code length needed to explain the data. Further analysis is required to find which specific model within the class best fits the data.

My model selection criterion is specifically suited for distributions that have parameters residing on a spherical manifold. In [64, 98], it was shown that the parameters for histogram density estimation can appropriately be placed on the hypersphere. In [78,

79], while showing that MLE theory can be used to estimate the coefficients for wavelet density estimation, it was shown that the coefficients of any square-root density estimator expanded in an orthogonal series resides on a unit hypersphere. In [17], the normalizing constant for the density function for spherical data (not parameters) was studied in detail. Here, Bingham showed that normalizing distributions on the sphere requires a confluent hypergeometric function of matrix argument. Furthermore, in [75], it was suggested that the Laplace approximation employed in the derivation of the asymptotic version of MDL is erroneous when applied on curved manifolds. Here, we show that the spherical MDL integral is instead equal to the normalizing constant of the Fisher–Bingham distribution when the parameters (not data) are constrained to lie on a hypersphere. Even though it can be difficult to calculate, reference [62] offers efficient numerical ways to estimate the value of this normalizing constant.

As anecdotal empirical evidence of our theoretical development of spherical MDL, Section 3.4 evaluates its use for histogram optimal bin width selection. The authors in [49] detailed the first use of MDL to find the optimal number of bins for histogram estimation. In this case, stochastic complexity for histograms was developed using the notion of code lengths, which is aligned with Rissanen's original formation of the MDL. Along with the criteria obtained from the code length, two asymptotic versions of the criteria were developed. These three variants of MDL proved to give results that are comparable with other methods of histogram density estimation. The capabilities of the use of NML in MDL has been explored in [60] where the author applies MDL to histogram density estimators with unequal bin widths. Here, histograms vary based on the location and quantity of cut points within the range of the data. Stochastic complexity is found using the normalized maximum likelihood distribution. In [38], the performance of 11 different bin selection criteria were analyzed, among them variants of AIC, BIC and MDL. Here, all the criteria were used to calculate the optimal number of bins for 19 different density shapes and real data. The densities were chosen to analyze the efficacy of each criterion when recognizing varying characteristics of densities, like skewness, kurtosis and multimodality. The performance of each criterion was measured with two different metrics: Peak Identification Loss and the Hellinger risk. Among these results, it was shown that AIC performs relatively poorly when considering either metric, while BIC and MDL were better performers with MDL performing well with both metrics.

Most efforts towards measuring distances on statistical manifolds build on the foundation started by Fisher in [42], in which he introduces the idea of the information matrix. In [61] Kullback and Leibler published a pioneering effort to describe this distance. Works such as [16, 53], endowed statistical distributions with geometrical properties. However, it was Rao [80] that expanded on the ideas of Fisher that defined a metric for statistical models based on the information matrix. Here, Rao showed that the information matrix satisfies the condition of a metric on a Riemannian statistical manifold, and is widely used because of its invariance [30]. This connection between distance and distributions encouraged others to explore the distance between specific families of

distributions [54]. Among these families include special cases of the multivariate normal model [95], the negative binomial distribution [66], the gamma distribution [82, 7], Poisson distribution [73], among others.

In [25], the authors offer a detailed exploration of geodesics on a multivariate Gaussian manifold. They show that there exists a geodesic connecting any two distributions on a Gaussian manifold. Furthermore, they find these distances for specific instances of distributions on the manifold, but a closed-form solution for the most general case remains an open problem.

In [21] and expanded on in [35], the authors offer a very detailed discussion, focusing primarily on the univariate normal distribution for which a closed-form solution for the Fisher–Rao distance is known. Here, the authors focus on a geometrical approach, abandoning the "proposition-proof" format offered in previous research. With this geometric approach, closed-form solutions to various special cases are derived: univariate Gaussian distributions, isotropic Gaussian distributions, and Gaussian distributions with diagonal covariance matrix.

Another novel application of geodesics on a Gaussian statistical manifold is explored in [78], where the authors use information geometry for shape analysis. Shapes were represented using a K-component Gaussian Mixture Model, with the number of components being the same for each shape. With this, each shape occupied a unique point on a common statistical manifold. Upon mapping two shapes to their points on this manifold, the authors use an iterative approach to calculate the geodesic between these two points, with the length of the geodesic offering a measure of similarity of the shapes. Furthermore, because of the iterative approach to solving for the geodesic, all intermediate points along path are revealed. These points can be mapped to their own unique shapes, essentially showing the evolution from one shape to another. This shape deformation exhibits the benefit of analyzing more than just the distance between points on a manifold and that "walk" along the path has real substance.

In [24], the authors explore the complexity of Gaussian geodesic paths, with the ultimate goal of relating the complexity of a geodesic path on a manifold to the correlation of the variables labeling its macroscopic state. Specifically, the authors show that, if there is a dependence between the variables, the complexity of the geodesic path decreases. Complexity, for these purposes is defined as the volume of the manifold traversed by the geodesic connecting a known initial state to a future state, which is well defined. It is shown that this volume decays by a power law at a rate that is determined by the correlation between the variables on the Gaussian manifold.

In [33], the authors use the geometry of statistical manifolds to study how the quantum characteristics of a system are affected by its statistical properties. Similar to our work, the authors prescribe an initial distribution on the manifold of Gaussians and examine the geodesics emanating from it, without dictating a specific terminating distribution. The authors show that these paths tend to terminate at distributions that minimize Shannon entropy. However, unlike our work, these paths are free to roam on the manifold

and are not required to terminate on a specific surface on the manifold. Furthermore, the most relevant part of the author's work considers only univariate Gaussians with a two-dimensional parameter manifold, without ever considering higher dimensions.

Though we have chosen to work with Riemannian geometry, it is worth mentioning that information geometry often employs dualistic geometries that can be established using divergence measures. In [103], the authors detail the use of divergence measures to obtain the dual coordinates for the space of multivariate Gaussians. However, they point out that the choice of divergence measure is not unique and resulting geometries lack the same interpretative power of the natural parameterization.

Though these previous works operate in the space of multivariate Gaussians and deriving geodesics therein, they all require defining the initial and terminal distributions on the manifold. In this work, we address a novel problem of finding the geodesics when the terminal conditions are hypersurface constraints rather than a single point. Technically, these transversality conditions are variable boundary conditions placed on the initial and final distributions requiring them to reside on a parametric surface typically defined by constraining the coordinates. The usefulness of these variable boundary conditions has emerged in many areas including physics [20] in which the author studied wetting phenomenon on rough surfaces and in [46], where the authors studied the elasticity of materials. Additionally, in [48, 69, 57], transversality conditions were employed in economic optimal control problems with a free-time terminal condition. However, as practical as transversality conditions have been in the above fields, their application in information geometry literature is deficient.

Here, we address the deficiency by adopting the techniques of variational calculus. Calculus of variation has its history rooted in the works of Euler, Lagrange and Laplace. The simplest problems employing the calculus of variations are boundary value problems involving a particular set of differential equations called the Euler-Lagrange equations. Like much of calculus it searches for stationary values, but instead of finding the stationary points, it searches for stationary curves of functionals . Historically, common problems which find variational calculus useful are the brachistochrone problem, minimal surfaces and shortest paths, the lasts of which will be the focus of this current research [32, 70].

Chapter 3

Spherical Minimum Description Length

In this chapter, a brief introduction to a Bayesian approach to model selection criteria is provided, followed by a detailed proof of an asymptotic form of the Minimum Description Length principle, focusing on the geometric aspects of the criterion. It will be shown how these interpretations fall short when dealing with parameters that fall on spherical manifolds, providing motivation for the bulk of this chapter; the development of spherical Minimum Description Length model selection criteria. All model selection criteria share the early ideas of Akaike with AIC and Schwarz with BIC. Spherical Minimum Description Length does not compromise what is at the core of these model selection criteria. As always, spherical Minimum Description Length looks for a balance between the complexity of a model and its goodness of fit to the sampled data. It does this in part by comparing the probabilities of each candidate model being responsible for generating the sampled data which, thanks to differential geometry, has a pleasing geometric interpretation rooted in comparing volumes on a statistical manifold. While its foundation was built on simple model selection criteria, spherical Minimum Description Length more closely resembles a consolidation of Rissanen's MDL and Bayesian statistics.

3.1 Bayesian Approach to Model Selection

3.1.1 Comparing Models

Suppose we have the parameter space Θ^K, such that for all $\theta \in \Theta$ we have $\theta : \theta^T\theta = 1$. This places all distributions in this space on the $(K-1)$-dimensional hypersphere. We assume data $X = \left\{x_i\right\}_{i=1}^{N}$ are a sample realization from the density function $f(x;\theta)$

where $\theta \in \Theta$ and the corresponding likelihood function is given by

$$l(\theta; X) = \prod_{i=1}^{N} f(x_i; \theta). \tag{3.1}$$

As in [9], we begin with a Bayesian viewpoint of model selection. Taking the simplest case, suppose we have two candidate models A and B with the goal of choosing one to represent our data. Let θ_A and θ_B be the parameters for each model, most likely of unequal dimensions. For the moment, assume that the parameter spaces are unconstrained. Later, in Section 3.4, we enforce constraints on them—specifically hypersphere constraints—and will perform model selection in this space.

We wish to examine the posterior of both models and choose the most likely of the candidate models. By Bayes' rule, the posterior probability of model A is

$$\Pr(A|X) = \frac{\Pr(A)}{\Pr(X)} \int_{\mathbb{S}^{K-1}} l(\theta; X)\pi(\theta)d\theta. \tag{3.2}$$

Here, $\Pr(A)$ is the prior probability of model A, $\pi(\theta)$ is a prior density over the model parameters and $\Pr(X)$ is a prior density function of the data. Candidate model B has a similar expression for its posterior. Henceforth, $\Pr(X)$ is ignored since it is a common factor. In addition, we take the prior probabilities of each candidate model to be equal and therefore disregarded. The comparison between two posteriors $\Pr(A|X)$ and $\Pr(B|X)$ therefore devolves into the comparison of two integrals, one with model parameters θ_A and the other with θ_B with the posterior probability being larger for the larger integral. Thus, our goal is the evaluation and maximization of the integral

$$\mathcal{I}(X) = \int_{\mathbb{S}^{K-1}} l(\theta; X)\pi(\theta)d\theta \tag{3.3}$$

over all valid models.

3.1.2 An Inappropriate Prior

Before evaluating Equation (3.3), we need to define a prior probability in the parameter space. While a uniform prior seems to be a logical choice [68]—following Laplace's principle of insufficient reason [99]—it is not reparametrization invariant. That is, choosing a uniform distribution as the prior for a specific parametrization does not guarantee that the prior for all parametrizations will be uniform. Let a model be defined by parameter θ with $\theta \in [0, 1]$ for the sake of convenience. Assume a uniform prior probability density function given by

$$p(\theta) = 1, \ \theta \in [0, 1]. \tag{3.4}$$

Now assume a second parametrization of the parameters, ψ along with a monotonic

transformation $\theta \rightarrow \psi$, i.e., $\psi = r(\theta)$. Of course, under this new parametrization, we would want the prior distribution to be uniform as well. The prior probability density function over ψ, $p(\psi)$ from Equation (3.4) is expressed as

$$p(\psi) = p(\theta) \left| \frac{\partial r^{-1}(\psi)}{\partial \psi} \right| \neq 1 \tag{3.5}$$

in general. Clearly, this is undesirable. In fact, $p(\psi) = 1$ is only guaranteed to be true if the transformation, $\psi = r(\theta)$ is a translation, which is a very limited reparametrization. That is, an unbiased prior for an arbitrary parametrization fails to give equal weight to the values of the parameters in other parametrizations. A more appropriate prior would be one with a structure that remains the same regardless of the parametrization used. The motivation for a more appropriate prior can be explained using the geometry of hypothesis testing. Below, after a brief discussion on parameter space geometries and their Fisher information, we will revisit this issue of developing a reparameterization invariant prior and its connection to the MDL criterion.

3.1.3 Geometry of Probabilistic Models

Applying geometrical constructs to statistical models is not a new idea. Rao [80] and Jeffreys [55] pioneered the idea of a measure of the distance between two distributions on a parameter manifold. The usefulness of differential geometry in exploring statistical inference is discussed in even greater detail in [58, 10]. Here, geometry is tasked with the challenge of finding a metric to measure distances on a statistical manifold. Distributions that are similar to one another reside closer together on the parametric manifold, as measured by the chosen metric. As such, deciding on the appropriate metric opens up geometrical representations for statistical tests. Even though many metrics can be defined, the Fisher information matrix is a natural metric on a parametric manifold due to its invariance property [5, 40].

The manifold associated with a family of models is populated with many distributions. Let a sample set $X = \left\{x_i\right\}_{i=1}^{N}$ be drawn from one of the distributions. A logical statistical question would be, if someone were just given the data, what the probability is with which they would choose the distribution on the manifold which produced the data. The problem of model selection is to pick the best model given a finite sample. Where one distribution can be mistaken for another, we consider the two distributions to be indistinguishable. Distinguishable distributions then can be defined as two distributions that are sufficiently far enough—as measured by the chosen metric—that the probability of mistaking one distribution for another is reasonably small.

Given two probability distributions f and g defined on the same manifold, relative

entropies between f and g can be defined as [36]:

$$D(f\|g) = \int f(x) \ln\left(\frac{f(x)}{g(x)}\right) dx. \tag{3.6}$$

The parameter vectors associated with each distribution are θ_f and θ_g (i.e., $f(x) = p(x;\theta_f)$ and $g(x) = p(x;\theta_g)$. Employing Stein's lemma [9] in Equation (3.6) results in

$$D(f\|g) \approx \frac{1}{2}\Delta\theta^T I \Delta\theta, \tag{3.7}$$

where $\Delta\theta = \theta_f - \theta_g$ and I is the Fisher information matrix (with details below). This strongly suggests that the Fisher information matrix acts as the natural metric on the parameter manifold.

The above discussion shifts attention from unique sets of parameters to counting the number of distinguishable distributions. For an in-depth discussion of distinguishable distributions, please see [9]. For completeness, we include a brief discussion as follows. While it is true that every single distribution is indexed by a unique parameter vector, there is a region around any individual distribution such that distributions in that region are statistically indistinguishable from one another. That is, there is a reasonable probability of mistaking one of the distributions for a neighboring distribution. The size of this elliptical region depends on the natural metric of the manifold, which is the Fisher information, as well as the sample size, since distributions can be more consistently differentiated with a larger sample size.

3.1.4 Fisher Information

The Fisher information matrix is a measure of how much information about the parameter of interest is available from the data collected. Traditionally, the Fisher information matrix is given by

$$I_{i,j}(\theta) = \int f(x;\theta)\frac{\partial}{\partial\theta_i}\log f(x;\theta)\frac{\partial}{\partial\theta_j}\log f(x;\theta)dx, \tag{3.8}$$

for continuous distributions and

$$I_{i,j} = f(x;\theta)\sum\left(\frac{\partial}{\partial\theta_i}\log(f(x;\theta)\frac{\partial}{\partial\theta_j}\log(f(x;\theta))\right) \tag{3.9}$$

for discrete distributions, where the index (i, j) represents the appropriate parameter pair of the multivariate parameter vector θ. In this form, the Fisher information matrix is the expectation of the variance of the score vector for the multi parameter distribution $f(x;\theta)$.

There are two alternate forms of the Fisher information, if certain regularity conditions are satisfied. Firstly, we can compute the Fisher information matrix from the expectation of the Hessian of the log likelihood. Specifically,

$$I_{i,j}(\theta) = -\mathcal{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(x;\theta)\right] = -\mathcal{E}\left[H\right], \tag{3.10}$$

where H is the Hessian matrix of the log-likelihood.

Alternatively, the Fisher information can be calculated from the variance of the score function

$$I(\theta) = \mathrm{Var}(S_f(x;\theta)), \tag{3.11}$$

where

$$S_f(x;\theta) = \nabla \log f(x;\theta). \tag{3.12}$$

Here, we use it for two closely related ideas. The Fisher information is the foundation for developing Jeffreys prior, a non-informative prior that is reparametrization invariant, which solves the issues raised in Section 3.1.2. In addition, it provides a natural Riemannian metric for a statistical manifold, which will allow us to find volumes of entire closed manifolds as well as the local volume of distinguishability around a single value of the parameter. These volumes will help to interpret the complexity parameter in the spherical MDL criterion proposed in this work. Additionally, this imparted geometric structure will allow me to measure the distances between to distributions on a manifold, which is paramount to finding geodesic paths between distributions.

Adjacent to the current research is the use of the Fisher information for parameter estimation. The Cramer-Rao lower bound (CRLB) provides a lower bound on the variance of the MLE for an unknown parameter. In terms of the Fisher information, we can find this lower bound with

$$CRLB = I^{-1}(\theta_{i,j}) \tag{3.13}$$

This lower bound is a measure of the uncertainty of the MLE. The CRLB also provides a benchmark against which the proficiency of different parameters can be compared. It provides the standard of efficiency and validates the preference of estimates that meet this standard. Using this lower bound to measure the efficiency of a statistic has been found useful for evaluating optimality of machine learning algorithms [56, 65]

3.1.5 Observed vs. Expected Fisher Information

In this current work, the Fisher information will be found using Equation (3.10). For some manifolds, this is difficult to calculate, since the expectations involved are not always attainable. However, if you can calculate the log-likelihood of a function, you can calculate the *observed* Fisher information. To clarify, what we refer to as the Fisher information, I, is sometimes called the *expected* Fisher information. Whether to use

the expected information or the observed information is at best situational and at worst subject to personal choice and at different times, the literature has made cases for the preference of both [96, 41, 97, 26]. Here, we provide a brief explanation of both the observed and expected information via an example using the Bernoulli distribution. Also, we validate the choice of the expected Fisher information for the current topic.

The observed Fisher information is defined as

$$\boldsymbol{J}_{i,j}(\theta) = \frac{\partial}{\partial \theta_i} f(x;\theta) \frac{\partial}{\partial \theta_j} f(x;\theta) \tag{3.14}$$

and is often used when trying to estimate the value of a parameter. The literature often misses what differentiates the observed information from the expected information: observations. Until a sample is observed, it is difficult to calculate the information in that sample. As an example to call attention to their differences , we examine a Bernoulli random variable, taking samples of varying sizes and calculating the observed Fisher information at different values of the parameter. The density function for the Bernoulli distribution is given by

$$f(x;\theta) = \theta^x (1-\theta)^{1-x} \tag{3.15}$$

with Fisher information, $I(\theta)$

$$I(\theta) = \frac{1}{\theta(1-\theta)} \tag{3.16}$$

plotted in Figure 3.1.

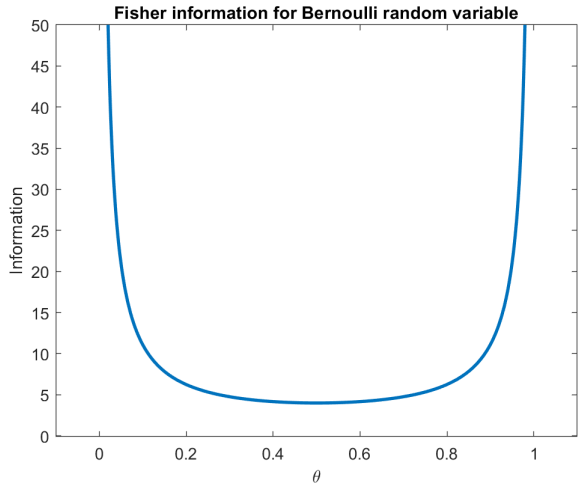


Figure 3.1: Bernoulli Fisher Information.
The Fisher information vs. parameter value for the Bernoulli distribution.

The graph of the Fisher information confirms some intuitive insights about the metric.

The definition of the Fisher information is how much information about the parameter is learned by a single sample. If that sample comes from a distribution with $\theta = 0$ or $\theta = 1$, we get all the information about the parameter we could possibly need from this single observation. As such, the information obtained from the single sample is large. Accordingly, with such overwhelming evidence regarding the parameter, the variance in future samples is small. In fact, the variance in these extreme cases would be 0.

However, if the sample comes from a population with $\theta = 0.5$, many samples would be required in order to obtain a reasonable estimate of the parameter, hence the minimum value of the Fisher information at $\theta = 0.5$ in Figure 3.1. The variance of future samples is largest at $\theta = 0.5$ as expected by the CRLB, since future observation are unpredictable.

In Table 3.1, the observed Fisher information is displayed for a distribution with $\theta = 0.5$ at various sample sizes. These results are averaged over 1000 trials. For this value of the parameter, the expected information is $I(0.5) = 4$, as shown at the last entry of the table. Considering the MLE of the Bernoulli distribution is an unbiased estimator of the parameter, one would expect the observed information to the expected information as the sample size gets larger which is confirmed by the data. The graph of the observed information versus sample size is shown in Figure 3.2.

$\theta = 0.5$	
n	J
5	4.87
10	4.57
20	4.23
50	4.09
100	4.04
$I(\theta)$	4

Table 3.1: Fisher information for Bernoulli random variable $\theta = 0.5$
Table showing the observed Fisher information for differing sample sizes for a Bernoulli distribution with $\theta = 0.5$. The observed Fisher information converges to the expected Fisher information, $I(\theta) = 4$ as n gets large.

To observe the effects that the value of the parameter has on the rate of convergence, Table 3.2 repeats the experiment with $\theta = 0.75$. As you can see, as the sample size increases, the observed Fisher information approaches the expected Fisher information, but at a slower rate than parameters closer to $\theta = 0.5$.

The small value for the observed Fisher information at $n = 5$ is curious. It appears to be because the sample space of sample parameters is limited with such a small sample size. As such, the expected Fisher information does not fit the trend seen with sample sizes that are more reasonable. Additionally, as the value of the parameter approaches $\theta = 1$, the probability that the sample yields an MLE of $\hat{\theta} = 1$ increases, causing the

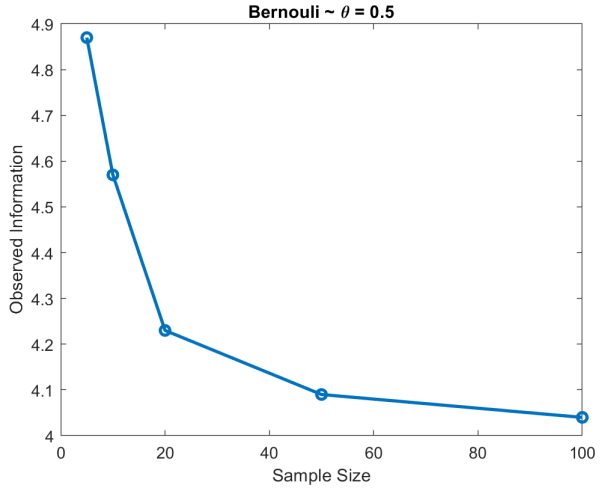


Figure 3.2: Fisher information for Bernoulli random variable $\theta = 0.75$
Value of the observed Fisher information for a Bernoulli distribution with $\theta = 0.5$. at various sample sizes.

information to diverge. Other than throwing off the average value in our table, this divergence is a clear disadvantage to using the observed Fisher information for small sample sizes.

$\theta = 0.75$	
n	J
5	5.27
10	6.32
20	6.21
50	5.6
100	5.46
$\mathbf{I}(\theta)$	5.33

Table 3.2: Fisher information for Bernoulli random variable $\theta = 0.75$.
Table showing the observed Fisher information for differing sample sizes for a Bernoulli distribution with $\theta = 0.75$. The observed Fisher information converges to the expected Fisher information, $I(\theta) = 5.33$, as n gets large, but more slowly than $\theta = 0.5$.

In Section 3.2 we'll show some key ideas of Balasubramanian's asymptotic MDL. Here, he refers to both the observed and expected Fisher information in his model selection criteria. He explains that, considering his is an asymptotic selection criteria, the observed Fisher information approaches the expected Fisher information, a claim corroborated by the data above. However, he adds the disclaimer that the data is drawn from a

candidate model that is on the manifold of the family of distributions being considered. if the candidate model resides on a different manifold with its own unique geometry, a correction factor needs to be included in the criteria to ensure that comparison of volumes has geometric meaning and consistency. This correction factor involves the observed Fisher information, which introduces the geometry local to the candidate model into the model selection criteria.

For all of the research we present here, we concern ourselves with the interaction between distributions on the same manifold. When we use geometric vocabulary, it will always apply to the entire manifold, whose geometry is defined by the expected Fisher information, thus making it the proper choice over the observed information, for our purposes.

3.1.6 An Appropriate Prior

Armed with this geometric definition of the Fisher information, an appropriate non-informative prior can be chosen again starting with two K-dimensional parametrizations θ and ψ. Furthermore, we assume there is a transformation $\psi = r(\theta)$ with both r and its inverse r^{-1} being differentiable, i.e., r is a diffeomorphic map. Consider a density function $f(x;\theta)$ with score vector

$$S_f(x;\theta) = \left[\frac{\partial}{\partial\theta_1}\log f(x;\theta), \quad \cdots, \quad \frac{\partial}{\partial\theta_K}\log f(x;\theta)\right]^T.$$

Define the Jacobian transformation matrix for both r and r^{-1} such that

$$J_r(\theta)J_{r^{-1}}(\psi) = \mathtt{I}_K, \tag{3.17}$$

where $\mathtt{I}_K$ is the $K \times K$ identity matrix.

The non-informative prior, which is tantamount to Jeffreys prior [55] is $\pi(\theta) \propto \sqrt{\det I(\theta)}$. To show that Jeffreys prior is invariant to reparametrization, we consider the transformation proposed above. This reparametrization yields a new density function

$$g(x;\psi) = f(x;r^{-1}(\psi))\left|\det(J_{r^{-1}}(\psi))\right| \tag{3.18}$$

and a new score function

$$S_g(x;\psi) = (J_{r^{-1}}(\psi))^T S_f(x;r^{-1}(\psi)). \tag{3.19}$$

The new Fisher information, $\tilde{I}(\psi)$, for the distribution with respect to the new parametriza-

tion is

$$\begin{aligned}\tilde{I}(\psi) &= \text{cov}(S_g(x;\psi)) \\ &= \text{cov}\left(J_{r^{-1}}(\psi)^T S_f(x; r^{-1}(\psi))\right) \\ &= (J_{r^{-1}}(\psi))^T I(r^{-1}(\psi)) J_{r^{-1}}(\psi).\end{aligned} \tag{3.20}$$

The Jeffreys prior for the ψ parametrization is

$$\begin{aligned}\tilde{\pi}(\psi) &\propto \sqrt{\det(\tilde{I}(\psi))} \\ &= \sqrt{\det\left[(J_{r^{-1}}(\psi))^T I(r^{-1}(\psi)) J_{r^{-1}}(\psi)\right]} \\ &= \det(J_{r^{-1}}(\psi)) \det\sqrt{I_\theta(r^{-1}(\psi))} \\ &= \det(J_{r^{-1}}(\psi)) \pi(r^{-1}(\psi)).\end{aligned} \tag{3.21}$$

Since the infinitesimals $d\theta$ and $d\psi$ also transform using the same Jacobian with $d\psi = \det(J_r(\theta))d\theta$, we get

$$\tilde{\pi}(\psi)d\psi = \pi(\theta)d\theta. \tag{3.22}$$

Therefore, the Jeffreys prior remains unchanged under the reparametrization $\psi = r(\theta)$ and the value of Equation (3.3) is indifferent to different representations of the parameter. With this, we can see that the Fisher information directs us to an appropriate prior to use in the evaluation of Equation (3.3). Specifically, Jeffreys prior is

$$\pi(\theta) = \frac{\sqrt{\det(I(\theta))}}{\int \sqrt{\det(I(\theta))}d\theta}, \tag{3.23}$$

where $\int \sqrt{\det(I(\theta))}d\theta$ is necessary in order to normalize the prior. In fact, this normalizing constant can sometimes be the largest shortcoming of Jeffreys prior because, in some cases, the integral may not converge, making the prior improper. If the interval diverges, it is possible to place artificial bounds on the limits of integration in order to make the integral converge. However, this won't be an issue in spherical MDL since, in many applications, the integral will be in closed form and convergent. Selecting a non-informative prior to evaluate Equation (3.3) is mathematically preferred, making Jeffreys prior the most suitable choice. However, differential geometry gives an interesting interpretation of the Jeffreys prior which will help explain the complexity penalty in spherical MDL. In Riemannian geometry, Equation (3.7) yields the squared distance element between two nearby points on a parameter manifold, implying again that the Fisher information is a natural metric for the manifold. This metric tensor can be used to calculate volumes of the manifold. Firstly,

$$V_{\mathcal{M}} = \int \sqrt{\det\left(I(\theta)\right)}d\theta \tag{3.24}$$

measures the Riemannian volume of an entire manifold. This integral is evaluated across all possible values of the parameter and, as such, only depends on the model family. As mentioned, this integral is known in closed form for the hypersphere.

Secondly, we are interested in partitioning the volume of the entire manifold into smaller local volumes encompassing indistinguishable distributions. The number of distributions in each volume need not be the same, but every volume is given an equal prior probability. Essentially then, Jeffreys prior provides a uniform prior with regard to these volumes and not individual parameters. With this, the numerator of Equation (3.23) represents these volumes on the parameter space. More specifically,

$$V_d = \sqrt{\det\left(I(\theta)\right)}d\theta \tag{3.25}$$

can be thought to represent the *infinitesimal* Riemannian volume local to each distinguishable distribution. The complexity parameter for spherical MDL will be interpreted geometrically with these volumes. The reader is pointed to [58] for a more detailed discussion on the appropriateness of Jeffreys prior for spherical distributions.

3.2 Asymptotic MDL in $\mathbb{R}^K$

In [9], the author develops an alternative derivation of MDL. Instead of attempting to find shortest code lengths, stochastic complexity is approached from a geometric perspective, which is more aligned with our development of spherical MDL. The author begins with a Bayesian approach to model selection and evaluates Equation (3.3). Again, given a set of data $X = \left\{x_i\right\}_{i=1}^{N}$, the likelihood of any given model with density $f(x;\theta)$ and a K-dimensional parameter vector is given in Equation (3.1) and its average negative log-likelihood is given by

$$L(\theta) = -\frac{1}{N}\log(l(\theta;X)). \tag{3.26}$$

Our goal is still to evaluate Equation (3.3):

$$\mathcal{I}(X) = \int \exp\left\{-N\,L(\theta)\right\}\pi(\theta)d\theta. \tag{3.27}$$

To evaluate the integral in Equation (3.27), we employ standard Laplace approximation techniques [63]. In order to do so, we first expand the integrand around the maximum likelihood estimate of the parameters $\hat{\theta}$ using a Taylor series approximation. The first order term of the expansion of the likelihood *vanishes* at the MLE, resulting in the integral being modified to

$$\mathcal{I}(X) \approx \exp\left\{-N\,L(\hat{\theta})\right\}\pi(\hat{\theta})\int \exp\left\{-\frac{N}{2}\left(\theta-\hat{\theta}\right)^T H_I\left(\theta-\hat{\theta}\right)\right\}d\theta, \tag{3.28}$$

where H_I is the Hessian of the *unconstrained* negative log likelihood. (We later distinguish this Hessian from that of the constrained log-likelihood.) Recognizing that the quadratic integral will result in a Gaussian integral (for unconstrained parameters), the final evaluation yields an expression for what Balasubramanian called the razor of the model,

$$RZR = \exp\left\{-N\,L(\hat{\theta})\right\}\pi(\hat{\theta})\left(\frac{\left(\frac{2\pi}{N}\right)^{K}}{\det(H_I)}\right)^{\frac{1}{2}}, \tag{3.29}$$

where K is the cardinality of the parameter set. Please note that the standard Laplace approximation has assumed that our parameter space is $\mathbb{R}^K$. With the aforementioned substitution, and evaluating the prior from Equation (3.23) at the maximum likelihood estimate, the final form of MDL is found by taking the negative log of the razor and is given by

$$MDL = -\log(RZR) = -\log f(X;\hat{\theta}) + \frac{K}{2}\log\left(\frac{N}{2\pi}\right) + \tag{3.30}$$

$$\log\int\sqrt{\det\left(I(\theta)\right)}d\theta. \tag{3.31}$$

The first term in Equation (3.31) addresses how well the model fits the data. The second and third terms concern the complexity of the model, which has three facets: the number of dimensions in the model, K, the form of the model as given by $I(\theta)$ and the domain of the parameter set as implied by the limits of integration on the third term.

During the development of Equation (3.31), the standard Laplace approximation was employed. The Laplace approximation is widely used to evaluate integrals with a unique global maximum over $\mathbb{R}^K$. However, the authors in [75] suggested that this approximation needs modification in order to be used on curved spaces. Thus, if Equation (3.31) is to be used on parameters that lie on a hypersphere, the penalty for overfitting will not be accurate, unless the tails of the integrand are ignored. This is the basis of spherical MDL—an extension of the razor approach to MDL to hyperspherical parameter spaces.

To summarize, spherical MDL addresses certain issues that arise in standard MDL. First, the Fisher information integral must exist and be finite. Since this integral represents the Riemannian volume of the model space, and the volumes of unit hyperspheres are available in closed form, this is usually not an (algebraic) concern for spherical MDL. In addition, if the value of the maximum likelihood estimate resides close to the edge of the parameter space, it becomes difficult to find the volume of the parameter space in the immediate vicinity of the MLE. Of course, if the MLE lies on a symmetric space like a hypersphere, then every parameter lies sufficiently in the interior of the model space, so this is not an issue either. Finally, spherical MDL does not ignore parameter constraints (such as restriction to a hypersphere) thereby resulting in a more accurate but

still efficiently computable model complexity.

3.3 Spherical MDL

3.3.1 Derivation of the Spherical MDL Criterion

Geometrically, the concept of penalizing a model for complexity can be interpreted as comparing the volume of the manifold in the vicinity of the model corresponding to the MLE to the volume of the entire parameter manifold. If the candidate model occupies very little space on a manifold, it is considered undesirable. This line of development led to the need to evaluate the integral in Equation (3.3) while constraining it to a $(K-1)$-dimensional hypersphere.

First, we use standard constrained optimization to enforce the unit length of the coordinate vectors for the parameters. Let

$$M(\theta, \lambda) = L(\theta) + \frac{\lambda}{N}(\theta^T\theta - 1) \tag{3.32}$$

be the Lagrangian corresponding to the constrained optimization problem. Next, the Lagrange parameter λ is set during the process of obtaining the optimal maximum likelihood estimate $\hat{\theta}$. The bulk of the development below attempts to convince the reader that we can rewrite Equation (3.3) as

$$\mathcal{I}(X) = \int_{\mathbb{S}^{K-1}} \exp\left\{-N\,M(\theta, \hat{\lambda})\right\} \pi(\theta) d\theta \tag{3.33}$$

with the domain of the integral restricted to coordinate vectors on the unit hypersphere.

To evaluate Equation (3.33), we employ the Laplace approximation methodology but now with the hypersphere constraint enforced. At the outset, this involves finding $\hat{\theta}$, the maximum likelihood estimate of the parameter vector θ at which M is minimized. At this minimum value, M is stationary, i.e., $\nabla_\theta M = 0$. We then expand $M(\theta, \hat{\lambda})$ around this minimum (with the Lagrange parameter set to a fixed value $\hat{\lambda}$). The resulting expansion is

$$M(\theta) = M(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T H (\theta - \hat{\theta}) + \mathcal{O}(\|\theta - \hat{\theta}\|_2^3), \tag{3.34}$$

where H is the Hessian of M (with the Lagrange parameter λ set to its optimum value $\hat{\lambda}$). We now show that this is a principled approach.

If the maximum likelihood problem had been unconstrained, we could have set θ to its MLE value $\hat{\theta}$, expanded the objective function around $\hat{\theta}$ and then employed Laplace's approximation to obtain the value of the integral in Equation (3.3). Since the ML parameters θ are constrained to a hypersphere, this route is closed to us. However, we show below that we can begin with a set of independent coordinates (defining a hypersphere) and then prove that the second order Taylor series expansion in Equation (3.34) is entirely *equivalent* to the corresponding expansion using independent coordinates. That is,

we begin with independent coordinates θ_R and a dependent coordinate θ_K and relate the quadratic form emanating from the Taylor series expansion using a carefully constructed "Hessian" of M to a corresponding quadratic form driven by the independent Hessian. The derivation closely follows the more general derivation in [15].

When we use independent coordinates to describe a $(K-1)$-dimensional hypersphere, we get

$$O(\theta_R) \equiv L(\theta_R, \theta_K(\theta_R)), \tag{3.35}$$

where the Kth parameter θ_K has been explicitly written out as a function of the remaining parameters $\theta_R \equiv \left\{\theta_1,\ \theta_2,\ \ldots,\ \theta_{K-1}\right\}$. The new objective function $O(\theta_R)$ corresponds to substituting $\theta_K(\theta_R)$ into the negative log-likelihood objective function $L(\theta_R,\ \theta_K)$. The partial derivatives of $O(\theta_R)$ can then be related to the corresponding ones from $L(\theta_R,\ \theta_K(\theta_R))$. Taking partial derivatives, we obtain

$$\frac{\partial O}{\partial \theta_k} = \frac{\partial L}{\partial \theta_k} + \frac{\partial \theta_K}{\partial \theta_k}\frac{\partial L}{\partial \theta_K}, \tag{3.36}$$

where the explicit dependence of θ_K on θ_k has been included. The second partials are tedious but straightforward to evaluate:

$$\frac{\partial^2 O}{\partial \theta_k \partial \theta_l} = \frac{\partial^2 L}{\partial \theta_k \partial \theta_l} + \frac{\partial \theta_K}{\partial \theta_l}\frac{\partial^2 L}{\partial \theta_k \partial \theta_K} + \frac{\partial \theta_K}{\partial \theta_k}\frac{\partial^2 L}{\partial \theta_l \partial \theta_K} + \frac{\partial \theta_K}{\partial \theta_k}\frac{\partial \theta_K}{\partial \theta_l}\frac{\partial^2 L}{\partial \theta_K^2} + \frac{\partial^2 \theta_K}{\partial \theta_k \partial \theta_l}\frac{\partial L}{\partial \theta_K}. \tag{3.37}$$

The quadratic form corresponding to the independent coordinates θ_R can in turn (after some simplification) be written as

$$\begin{aligned}\sum_{kl} u_k u_l \frac{\partial^2 O}{\partial \theta_k \partial \theta_l} = \sum_{kl} u_k u_l \frac{\partial^2 L}{\partial \theta_k \partial \theta_l} + 2\left(\sum_{k=1}^{K-1} u_k \frac{\partial \theta_K}{\partial \theta_k}\right)\left(\sum_{l=1}^{K-1} u_l \frac{\partial^2 L}{\partial \theta_l \partial \theta_K}\right) \\ + \left(\sum_{k=1}^{K-1} u_k \frac{\partial \theta_K}{\partial \theta_k}\right)^2 \frac{\partial^2 L}{\partial \theta_K^2} + \frac{\partial L}{\partial \theta_K}\sum_{kl} u_k u_l \frac{\partial^2 \theta_K}{\partial \theta_k \partial \theta_l}.\end{aligned} \tag{3.38}$$

Here, $u = \left[u_1,\ u_2,\ \ldots,\ u_{(K-1)}\right]^T$ and the double summation indices in $\sum_{kl}$ each range from 1 to $(K-1)$. So far, we have made no contact with our constrained optimization problem. From the Lagrangian

$$M(\theta, \hat{\lambda}) = L(\theta) + \frac{\hat{\lambda}}{N}\left(\sum_{k=1}^{K} \theta_k^2 - 1\right), \tag{3.39}$$

where $\hat{\lambda}$ is the optimal value of the Lagrange parameter, we see that the MLE of θ satisfies

the relation

$$\frac{\partial M}{\partial \theta_k} = \frac{\partial L}{\partial \theta_k} + 2\theta_k \frac{\hat{\lambda}}{N} = 0, \tag{3.40}$$

with the optimal value of the Lagrange parameter

$$\hat{\lambda} = -\frac{N}{2} \sum_{k=1}^{K} \hat{\theta}_k \frac{\partial L}{\partial \theta_k}\bigg|_{\theta=\hat{\theta}} \tag{3.41}$$

obtained by multiplying Equation (3.40) by θ_k, summing over all $k \in \{1, \ldots, K\}$ and enforcing the constraint $\theta^T\theta = 1$. Furthermore, Equation (3.40) gives us a relation connecting $\frac{\partial L}{\partial \theta_K}$ and $\hat{\lambda}$. We can also obtain a relation connecting $\frac{\partial \theta_K}{\partial \theta_k}$ and (θ_R, θ_K) by differentiating the constraint equation $\sum_{k=1}^{K-1} \theta_k^2 + \theta_K^2 = 1$ once to get

$$2\theta_k + 2\theta_K \frac{\partial \theta_K}{\partial \theta_k} = 0. \tag{3.42}$$

This relation holds for all θ on the hypersphere unlike Equation (3.40) which is valid only at the MLE. Taking second derivatives, we obtain

$$2\delta_{kl} + 2\frac{\partial \theta_K}{\partial \theta_k}\frac{\partial \theta_K}{\partial \theta_l} + 2\theta_K \frac{\partial^2 \theta_K}{\partial \theta_k \partial \theta_l} = 0. \tag{3.43}$$

We now have all the ingredients necessary to evaluate Equation (3.38) for the constrained problem. Substituting Equations (3.40), (3.42) and (3.43) into Equation (3.38), we get

$$\begin{aligned}\sum_{kl} u_k u_l \frac{\partial^2 O}{\partial \theta_k \partial \theta_l} = \sum_{kl} u_k u_l \frac{\partial^2 L}{\partial \theta_k \partial \theta_l} - \frac{2}{\theta_K}\left(\sum_{k=1}^{K-1} u_k \theta_k\right)\left(\sum_{l=1}^{K-1} u_l \frac{\partial^2 L}{\partial \theta_l \partial \theta_K}\right) \\ + \frac{1}{\theta_K^2}\left(\sum_{k=1}^{K-1} u_k \theta_k\right)^2 \frac{\partial^2 L}{\partial \theta_K^2} + \frac{2\hat{\lambda}}{N}\sum_{kl} u_k u_l \left(\delta_{kl} + \frac{\theta_k \theta_l}{\theta_K^2}\right).\end{aligned} \tag{3.44}$$

This can be reorganized with a view toward our goal of obtaining a quadratic form corresponding to a "Hessian" derived from the Lagrangian in Equation (3.39). We get

$$\begin{aligned}\sum_{kl} u_k u_l \frac{\partial^2 O}{\partial\theta_k \partial\theta_l} &= \sum_{kl} u_k u_l \left(\frac{\partial^2 L}{\partial\theta_k \partial\theta_l} + \frac{2\hat{\lambda}}{N}\delta_{kl}\right) - \frac{2}{\theta_K}\left(\sum_{k=1}^{K-1} u_k\theta_k\right)\left(\sum_{l=1}^{K-1} u_l \frac{\partial^2 L}{\partial\theta_l \partial\theta_K}\right) \\ &+ \frac{1}{\theta_K^2}\left(\sum_{k=1}^{K-1} u_k\theta_k\right)^2 \left(\frac{\partial^2 L}{\partial\theta_K^2} + \frac{2\hat{\lambda}}{N}\right) \\ &= \sum_{kl} u_k u_l \frac{\partial^2 M}{\partial\theta_k \partial\theta_l} - \frac{2}{\theta_K}\left(\sum_{k=1}^{K-1} u_k\theta_k\right)\left(\sum_{l=1}^{K-1} u_l \frac{\partial^2 M}{\partial\theta_l \partial\theta_K}\right) \\ &+ \frac{1}{\theta_K^2}\left(\sum_{k=1}^{K-1} u_k\theta_k\right)^2 \frac{\partial^2 M}{\partial\theta_K^2},\end{aligned} \tag{3.45}$$

where we have taken care to set $\hat{\lambda}$ to its optimum MLE value (while not treating it as a function of θ). Consequently, the second partials of M *do not* include the dependence of $\hat{\lambda}$ on $\hat{\theta}$. To further simplify this expression, we now define

$$v \equiv \left[u_1,\ u_2,\ \ldots,\ u_{(K-1)}, -\frac{1}{\theta_K}\sum_{k=1}^{K-1} u_k\theta_k\right]^T. \tag{3.46}$$

Note that v satisfies the constraint $\sum_{k=1}^{K} v_k\theta_k = 0$ implying that v is orthogonal to θ. This will be important later on in the specification of the constrained quadratic form. Using the definition of the Lagrangian in Equation (3.39), we get

$$\sum_{kl} u_k u_l \frac{\partial^2 O}{\partial\theta_k \partial\theta_l}\bigg|_{\theta=\hat{\theta}} = \sum_{k=1}^{K}\sum_{l=1}^{K} v_k v_l \frac{\partial^2 M}{\partial\theta_k \partial\theta_l}\bigg|_{\theta=\hat{\theta}}, \tag{3.47}$$

which implies the equality of the independent and constrained quadratic forms. Note that the constraint $\sum_{k=1}^{K}\theta_k^2 = 1$ implies that

$$\sum_{k=1}^{K}\theta_k d\theta_k = 0, \tag{3.48}$$

where $d\theta_k$ is an infinitesimal quantity. Assuming this remains valid for a small (but not infinitesimal) vector $\Delta\theta$ (up to second order correction factors), this in turn implies that the increment vector $\left[\Delta\theta_1,\ \Delta\theta_2,\ \ldots,\ \Delta\theta_K\right]^T$ is orthogonal to the gradient of the constraints, equal to $\left[2\theta_1,\ 2\theta_2,\ \ldots,\ 2\theta_K\right]^T$. Therefore, the quadratic form obtained from the Lagrangian M is only valid in the subspace spanned by increment vectors $\left\{v|\ \sum_{k=1}^{K} v_k\theta_k = 0\right\}$. This further implies that this quadratic form is equivalent to the

independent quadratic form in Equation (3.38) provided the increments are confined to the correct subspace.

Given the above analysis, the second order Taylor series expansion of M around the MLE estimate $\hat{\theta}$ in Equation (3.34), where the (k, l) element of the Hessian is

$$H_{kl} = \left.\frac{\partial^2 M}{\partial\theta_k \partial\theta_l}\right|_{\theta=\hat{\theta}} = \left.\frac{\partial^2 L}{\partial\theta_k \partial\theta_l}\right|_{\theta=\hat{\theta}} + \frac{2\hat{\lambda}}{N}\delta_{kl}, \tag{3.49}$$

emerges as the quantity most closely connected to the expansion of the independent objective O using coordinates θ_R. When the increments $\theta - \hat{\theta}$ are confined to the subspace orthogonal to the gradient vector $\left[2\hat{\theta}_1,\ 2\hat{\theta}_2,\ \ldots,\ 2\hat{\theta}_K\right]^T$, i.e.,

$$\sum_{k=1}^{K} 2(\theta_k - \hat{\theta}_k)\hat{\theta}_k = 0, \tag{3.50}$$

then the quadratic form $(\theta - \hat{\theta})^T H(\theta - \hat{\theta})$ is equivalent to the independent one as shown above in Equation (3.47). In the subsequent calculations, we set $\hat{\theta}$ to the constrained maximum likelihood solution (wherein $\hat{\theta}$ is constrained to lie on the surface of a unit hypersphere) and allow θ to vary over just the surface of the same unit hypersphere. For values of θ close to $\hat{\theta}$, $\theta - \hat{\theta}$ will approximately satisfy Equation (3.50), thereby validating our choice of "Hessian" for the hyperspherically constrained Laplace approximation.

A question may arise at this juncture as to why we could not have directly worked with the independent coordinates in the first place. Insofar as parameter constraints remain implicit (and hypersphere constraints fall into this category), it is much easier to work with constrained and implicit parameterizations than explicit ones (since the latter are typically harder to come by). Provided the manifold integrals can be carried out without defaulting to Gaussian integrals—and we make this case throughout the present work—implicit parameterizations should be preferred, especially given the correspondence worked out above between the constrained and independent quadratic forms.

With the Hessian defined in this manner (and related to the Lagrangian M), an asymptotic solution to Equation (3.33) can be found. Since Equation (3.32) represents the Lagrangian as a Taylor expansion around the MLE, it will be useful to redefine the entire integrand as a Taylor expansion. As such, the prior, $\pi(\theta)$, needs to be expanded as well. Expanding the prior around the MLE, we get

$$\pi(\theta) = \pi(\hat{\theta}) + (\theta - \hat{\theta})^T \nabla\pi(\hat{\theta}) + \mathcal{O}(\|\theta - \hat{\theta}\|_2^2). \tag{3.51}$$

We now rewrite Equation (3.33) as a product of Equations (3.34) and (3.51) to get

$$\begin{aligned}\mathcal{I}(X) &= \int_{\mathbb{S}^{K-1}} \exp\left\{-N\,M(\theta,\hat{\lambda}\right\}\pi(\theta)d\theta \\ &\approx \int_{\mathbb{S}^{K-1}} \exp\left[-N\,M(\hat{\theta}) - \frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right]\left[\pi(\hat{\theta}) + (\theta-\hat{\theta})^T\nabla\pi(\hat{\theta}) + \dots\right]d\theta \\ &\approx \exp\left\{-N(M(\hat{\theta})\right\}\pi(\hat{\theta})\int_{\mathbb{S}^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\}d\theta.\end{aligned} \tag{3.52}$$

Here, we have used the fact that $\theta \to \hat{\theta}$ as $N \to \infty$ on the order of $N^{-\frac{1}{2}}$ [14] which makes

$$\int_{\mathbb{S}^{K-1}} (\theta-\hat{\theta})^T\nabla\pi(\hat{\theta})d\theta = 0. \tag{3.53}$$

The evaluation of the integral of the quadratic term in Equation (3.52), when constrained to the $(K-1)$-dimensional hypersphere, is where Rissanen's MDL *inaccurately* penalizes the stochastic complexity of spherical parameter spaces. Instead of resulting in a Gaussian integral, there is no closed form solution in general. However, for distributions whose individual parameters contribute equally to the the Fisher information matrix, as is the case in the histogram, we can efficiently evaluate this integral. This assertion will be expanded upon in Section 3.4.

We continue solving Equation (3.52), and using the Jeffreys prior as the appropriate prior, we get

$$\mathcal{I}(X) = \exp\left\{-N\,M(\hat{\theta})\right\}\int_{\mathbb{S}^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\}d\theta\frac{\sqrt{\det(I(\hat{\theta}))}}{\int\sqrt{\det(I(\theta))}d\theta}. \tag{3.54}$$

As is customary with most model selection criteria, the optimal model according to

spherical MDL will be the one which minimizes the $-\log$ of Equation (3.54). Hence,

$$\begin{aligned} MDL_{\mathbb{S}^{K-1}} = \; & NM(\hat{\theta}) - \log\sqrt{\det(I(\hat{\theta}))} + \log\int \sqrt{\det(I(\theta))}d\theta \\ & - \log\int_{\mathbb{S}^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\} d\theta \\ = \; & N\left[L + \hat{\lambda}(\hat{\theta}^T\hat{\theta}-1)\right] - \log\sqrt{\det(I(\hat{\theta}))} + \log\int \sqrt{\det(I(\theta))}d\theta \\ & - \log\int_{\mathbb{S}^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\} d\theta \\ = \; & -\log l(\hat{\theta}) - \log\sqrt{\det(I(\hat{\theta}))} + \log\int \sqrt{\det(I(\theta))}d\theta \\ & - \log\int_{\mathbb{S}^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\} d\theta. \end{aligned} \tag{3.55}$$

The first term in Equation (3.55) is the log-likelihood and rewards a model for goodness of fit. The last three terms represent the parametric complexity penalty in spherical MDL:

$$\begin{aligned} C = & -\log\sqrt{\det(I(\hat{\theta}))} + \log\int \sqrt{\det(I(\theta))}d\theta - \\ & \log\int_{\mathbb{S}^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\} d\theta. \end{aligned} \tag{3.56}$$

The complexity penalty reflects the proportion of the volume of the total parameter space that lies close to the one model that best describes the data. The second term in Equation (3.56) is independent of the data and the candidate model, and therefore must reflect only the complexity in the inherent chosen model family. Specifically, this term represents the volume of the parameter manifold which is known in closed form. The first term represents the local volume around the model corresponding to the MLE, as measured by the natural measure of the parameter manifold. The final term is dependent upon the intrinsic properties of the model family, attributes of the data and on the candidate distribution. Essentially, it measures the volume of an ellipsoid around the parameter with respect to a local metric determined by the data. The essence of spherical MDL is within this integral. During the course of the evaluation of this integral, the small ellipse around the MLE is constrained to lie on the surface of the sphere.

Alternatively, the complexity term in Equation (3.56) can be represented as a ratio

of two terms

$$C = -\log\left[\frac{\sqrt{\det(I(\hat{\theta}))}\int_{\mathbb{S}^{K-1}}\exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\}d\theta}{\int\sqrt{\det(I(\theta))}d\theta}\right]. \tag{3.57}$$

Here, the denominator is the volume of the entire parameter manifold. The numerator is the volume of a small ellipsoid on the surface of the sphere around the MLE. If the volume around the MLE is small compared to the volume of the entire manifold, the model is considered complex and this term grows accordingly.

In contrast, the complexity penalty for the asymptotic version of Rissanen's MDL in $\mathbb{R}^K$ is

$$C_{\mathbb{R}^K} = \frac{K}{2}\log\left(\frac{N}{2\pi}\right) + \log\int\sqrt{\det\left(I(\theta)\right)}d\theta. \tag{3.58}$$

3.3.2 Riemannian Volume of a Hypersphere

The asymptotic version of MDL requires that the entire Riemannian volume of the manifold be finite. That is, $\int\sqrt{\det(I(\theta))}d\theta$ must converge. In complicated cases, this can be very impractical. If the manifold is unbounded, compromises such as artificially bounding the parameter space are required in order for the integral $\int\sqrt{\det(I(\theta))}d\theta$ to converge. Approximations using Monte Carlo integration are also utilized [89]. As difficult as this mathematical hurdle can be to overcome, it ends up being a big advantage for spherical MDL. In many cases, spherical MDL concerns itself with hyperspherical manifolds with Riemannian volumes equivalent to the surface area of a $(K-1)$-dimensional hypersphere, which is known in closed form.

The equation for the volume of a hypersphere is

$$V_{\mathcal{M}} = \begin{cases} K\pi^{K/2}, & K \text{ even}, \\ 2^K\pi^{\frac{K-1}{2}}\frac{\left(\frac{K-1}{2}\right)!}{(K-1)!}, & K \text{ odd}. \end{cases} \tag{3.59}$$

Figure 3.3 shows the volume of the unit hypersphere (technically the surface area) as a function of dimension. Curiously, this volume reaches a maximum at seven parameters after which the volume rapidly decreases approaching 0. Having the volume of the entire manifold decreasing to 0 can be troublesome since on the manifold, there must be room for a local volume around distinguishable distributions. In fact, in some cases, the local volume around a parameter can exceed the volume of the entire parameter space resulting in misspecified models. However, high dimensional models only make sense when dealing with large sample sizes. In this scenario, the dissimilarity between two neighboring distributions becomes more noticeable. This allows the volume around each detectable distribution to shrink. Thus, as the number of parameters increases, the

volume of the entire manifold decreases. However, it only makes sense to use these models with sample sizes that are large enough to force the volume around the MLE to be smaller than the overall manifold's volume [50]. Since spherical MDL represents an asymptotic approximation of NML, it is more suitable when applied to large sample data. We refer the reader to [72, 77] for more details regarding the issue of misspecification for high-dimensional models.

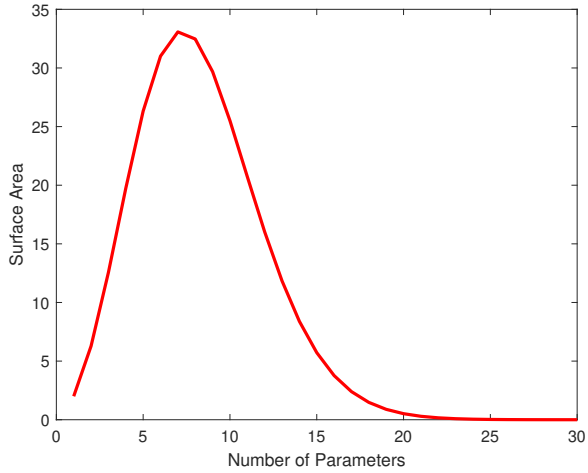


Figure 3.3: Riemannian volume of hyperspheres.

Riemannian Volume of hypersphere. The surface area of a hyperspherical manifold plotted against the cardinality of the parameters. Interestingly, the surface area grows to a maximum at seven dimensions and then monotonically decreases. Accordingly, a seven-dimensional model family requires a relatively large ellipsoid around the MLE in order to avoid excessive penalties for complexity.

This rapidly decreasing in surface area of higher dimensional unit hyperspheres has an interesting geometrical oddity. It's easy to imagine a unit circle (1-sphere) being placed inside a unit square (a square with side length of 2). In the context of Riemmanian geometery, the surface area of the 1-sphere is simply the perimeter of this circle and we have placed this circle into the smallest box in which it can fit. This 1-sphere touches the face of each sides of the unit box at its exact center. Similarly, we can place a unit 2-sphere into a unit cube. Once again, this 2-sphere is packed as tightly as possible into this cube. There is not bigger 2-sphere that could fit into the 3-dimensional box. Like before, this 2-sphere touches each face of the cube at its center. Here, the surface area of this hypersphere is $SA = 4\pi r^2$ and is clearly larger than the perimeter of a unit circle, as shown in Figure 3.3.

As we allow the thought experiment to grow in dimension, some facts are held true. First of all, this $(K-1)$-dimensional hypersphere touches this $(K-1)$-dimensional hypercube at the center of each face. Second of all, the $(K-1)$-dimensional unit hypersphere is the largest possible hypersphere that will fit into the $(K-1)$-dimensional hypercube. It is packed as tightly as possible. The contradiction is, if this is the largest possible hy-

persphere that will fit in this box, but the surface area of this hypersphere approaches 0 when the number of dimensions increases, where is this hypersphere. After all, the room inside the box is not getting smaller, considering the volume of the hypercube 2^K. It must be the case that these high dimension hyperspheres exist almost entirely at the center of each face of the hypercube. As K grows, almost all of the surface area congregates at a single point attached to the center of each face of the hypercube.

Adding support to this argument is a quick examination of what occupies the corners of each of these hypercubes. As we increase the number of dimensions, the amount of empty space in the corners grows. Essentially, these hyperspheres quickly abandon the corners in favor of the center point on the face of each hypercube. This tendency must increase with number of dimensions, leaving the corners of these boxes desolate.

Quantifying this is easy. Working only in the positive orthant starting at the origin, every point on the hypersphere is 1 unit away. If we were to draw a line from the origin towards the corner of the hypercube, this line would intersect the hypersphere 1 unit away and every dimensions component would have the same magnitude. That is, the K-dimensional vector,,$< a_1, a_2, \ldots, a_i, \ldots, a_K >$, pointing directly towards the corner of the hypercube with magnitude one has every component with magnitude. That is all a_i are equal. This magnitude depends on the number of dimensions and is given by

$$a_i = \frac{\sqrt{K}}{K}. \tag{3.60}$$

Also, the points at the vertices of the $(K-1)$-dimensional hypercube have K unit coordinates. The distance from the origin to the corner also only depends on the dimension. This distance, D is given by

$$D = \sqrt{K}. \tag{3.61}$$

These equations show that the distance from the origin to the corner of a hypercube grows on the order of $K^{0.5}$ while the distance along the same path to the intersection with the hypersphere grows on the order of $K^{-0.5}$. This quickly leaves the corners empty. A table showing the perpendicular distance between the vertex of the hypercube and the hypersphere is shown in Table 3.3.

How exactly the geometry appears in higher dimensions is difficult to picture. For certain, the behavior must be similar to what is described above. One additional thought experiment would be to imagine the disappointment of a 20 dimensional boy receiving a 20 dimensional ball in a 20 dimensional box for Christmas when he opens up his gift to see almost nothing but emptiness.

Dimension	Distance
2	0.707
3	1.155
4	1.5
5	1.789
6	2.041
7	2.268
8	2.475
9	2.667
10	2.846

Table 3.3: Vertex distance of hypercubes and dimension.
The above table summarizes the distances between the corners of hypercubes and the the intersection point on the hypersphere and the line connecting the origin with the corner. This distance monotonically increases with dimension.

3.4 Case Study: Spherical MDL for Histograms

The regular histogram is one of the most popular nonparametric density estimators. It is the go-to method for data scientists for quickly visualizing the regularities of their data. With relatively few parameters, a histogram can approximately model a variety of density functions without explicit knowledge of any underlying structure. Despite their simplicity, histograms can display very complicated characteristics of density functions like kurtosis and multimodality, which are often tied to the construction method.

Histogram construction always begs the question: what are the appropriate number of bins? While, in most cases, since the ultimate purpose of the histogram is to highlight features in the data, a subjective choice of number of bins would be one that best shows the features you wish to highlight. However, it is possible to use model selection to remove some of the subjectivity from this decision. Here, as a simple proof of concept, we show how spherical MDL can be used to select the optimal number of bins for a fixed-bin-width histogram. We detail the full scope of applying spherical MDL to this model, starting from the log likelihood and then deriving the relevant equations from Section 3.3 to reach the final criterion given in Equation (3.55).

Probably because of its prominence, approaches to bin selection for histograms are very popular, with many of the schemes deeply rooted in model selection theory [100, 38]. Here, we consider histograms with equal bin width, also known as regular histograms. When doing so, we can use Equation (3.55) to optimize the number of bins once a simple algebraic transformation produces the required hypersphere geometry. The geometric interpretation of spherical MDL also allows for a satisfying solution to the question of how to penalize empty bins, something ignored in much of the current research or addressed by allowing for unequal bin width.

3.4.1 Theoretical Development

The histogram can be realized by estimating an unknown density function via deploying piecewise constant functions and then using the maximum likelihood estimator, which results in the histogram. The height of each bin is proportional to the number of data points falling in its interval, i.e.,

$$f(x) = \begin{cases} c_i, & \text{if } x \text{ is in interval } i, \\ 0, & \text{otherwise.} \end{cases} \tag{3.62}$$

Given data $X = \{x_1,\ x_2,\ ...,\ x_N\}$, the likelihood function is given by

$$l(c) = \prod_{i=1}^{K} c_i^{v_i}, \tag{3.63}$$

where v_i is the number of data points in the i-th interval and K is the number of bins.This makes the average negative log-likelihood

$$L(c) = -\frac{1}{N} \sum_{i=1}^{K} v_i \log c_i. \tag{3.64}$$

As in [98], we choose to map the parameters of the histogram to the hypersphere. We begin by making the variable substitution $u_i^2 = c_i$ after which the average negative log-likelihood becomes (with a mild abuse of notation)

$$L(u) = -\frac{1}{N} \sum_{i=1}^{K} 2v_i \log u_i. \tag{3.65}$$

We now restrict the parameters to lie on a $(K-1)$-dimensional hypersphere by setting

$$\sum_{i=1}^{K} u_i^2 = h^{-1}, \tag{3.66}$$

where h is the regular bin width of the histogram. This ensures the volume under the density to be one. To emphasize the dependence of the complexity on the number of parameters K, we make the substitution $h = \frac{R}{K}$, where R is the range of the data. The

constrained average negative log-likelihood is then

$$\begin{aligned} M(u,\lambda) &= L(u) + \frac{\lambda}{N}(\sum_{i=1}^{K} u_i^2 - \frac{K}{R}) \\ &= -\frac{1}{N}\left[\sum_{i=1}^{K} 2v_i \log u_i - \lambda(\sum_{i=1}^{K} u_i^2 - \frac{K}{R})\right]. \end{aligned} \tag{3.67}$$

Minimizing $M(u,\lambda)$ w.r.t. u yields $\hat{u}_k = \sqrt{\frac{v_k K}{NR}}$ with the optimal value of the Lagrange parameter being $\hat{\lambda} = N\frac{R}{K}$.

We wish to solve Equation (3.49) for the histogram density log-likelihood function Equation (3.67). Starting with Equation (3.65),

$$L(u) = -\frac{1}{N}\sum_{i=1}^{K} 2v_i \log u_i, \tag{3.68}$$

the resulting gradient is

$$\frac{\partial L}{\partial u_k} = -\frac{1}{N}\frac{2v_k}{u_k} \tag{3.69}$$

and the Hessian

$$\frac{\partial^2 L}{\partial u_k^2} = \frac{1}{N}\frac{2v_k}{u_k^2} \tag{3.70}$$

with all other mixed partials equal to zero.

Next, we evaluate the Hessian of the average negative log likelihood at the MLE. Using Equation (3.49),

$$\begin{aligned} H_{kk} &= \frac{2v_k}{\frac{Kv_k}{R}} + 2\frac{R}{K} \\ &= 4\frac{R}{K}. \end{aligned} \tag{3.71}$$

Hence, the Hessian is a diagonal matrix with all positive entries, ensuring that it is positive definite as required by the Taylor expansion in Equation (3.49). To evaluate the Fisher information, we take the expectation of Equation (3.71) to get

$$\begin{aligned} I_{kk}(\theta) &= \int H_{kk} f(x) dx \\ &= H_{kk} \int f(x) dx \\ &= 4\frac{R}{K}, \end{aligned} \tag{3.72}$$

which does not depend on the histogram model parameters.

The parameters of the histogram density function do not lie on a unit hypersphere but, rather, they reside on a hypersphere with radius $\left(\frac{K}{R}\right)^{\frac{1}{2}}$. Additionally, this volume requires a scale factor of $\sqrt{\det(I(\hat{\theta}))} = \left(4\frac{R}{K}\right)^{\frac{K}{2}}$ based on the Fisher information from Equation (3.72) which defines a metric tensor on our manifold. Considering both of these influences, the volume of our sphere must be adjusted by a factor of $\left(\frac{K}{R}\right)^{\frac{K}{2}}\left(4\frac{R}{K}\right)^{\frac{K}{2}} = 2^K$, making the volume of the entire manifold for each family

$$V_H = 2^K V_{\mathcal{M}}, \tag{3.73}$$

where $V_{\mathcal{M}}$ is the hypersphere volume given in Equation (3.59). It may seem necessary to restrict the volume of the manifold even further, considering that the parameters of the histogram necessarily reside only in the positive hyperorthant of the hypersphere and the volume in Equation (3.73) accounts for the entire hypersphere. However, the same logic that would apply to restricting the Riemannian volume would also apply to the integral of the exponential of the quadratic form in Equation (3.55). With both restrictions having the same opposite effects on spherical MDL, restriction to the positive hyperorthant becomes unnecessary.

According to spherical MDL, the optimal number of bins for a histogram is the one which minimizes

$$\begin{aligned} MDL_{sphere} &= -\sum_{i=1}^{K} 2v_i \log u_i - \log\sqrt{\det(I(\hat{\theta}))} + \log V_H \\ &\quad - \log\int_{S^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\} d\theta \\ &= -\sum_{i=1}^{K} 2v_i \log u_i - \frac{K}{2}\log\left(4\frac{R}{K}\right) + \log V_H \\ &\quad - \log\int_{S^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\} d\theta. \end{aligned} \tag{3.74}$$

The third term, which is independent of the data, penalizes solely based on the number of parameters in the model family. If the model family being assessed has K parameters, of which l are empty, the model family is penalized as a K parameter family and *not* as a $K-l$ parameter family. This particular distribution with l empty bins simply is one which resides on the l axes of the hypersphere.

The final term can be elusive to find in general, but when the Hessian of the log-likelihood consists of identical elements as it does with the histogram, the integral now represents the normalizing constant of the von Mises distribution whose solution is

known in closed form. Focusing just on this integral, we first expand the quadratic form, recalling that every diagonal element of the Hessian is $4h$ and can be amalgamated with $\frac{N}{2}$ to get

$$\begin{aligned} Q(\hat{\theta}) &= \int_{S^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\} d\theta \\ &= \int_{S^{K-1}} \exp\left\{-2Nh(\theta-\hat{\theta})^T(\theta-\hat{\theta})\right\} d\theta \\ &= \int_{S^{K-1}} \exp\left\{-2Nh(\theta^T\theta - 2\theta^T\hat{\theta} + \hat{\theta}^T\hat{\theta})\right\} d\theta. \end{aligned} \tag{3.75}$$

Now, in the expanded quadratic, we have two quadratic terms that are subject to our constraint $\theta^T\theta = h^{-1}$. We can further simplify the integral to be

$$\begin{aligned} Q(\hat{\theta}) &= \int_{S^{K-1}} \exp\left\{-2Nh(h^{-1} - 2\theta^T\hat{\theta} + h^{-1})\right\} d\theta \\ &= \int_{S^{K-1}} \exp\left\{-4N + 4Nh\theta^T\hat{\theta})\right\} d\theta \\ &= \exp\{-4N\} \int_{S^{K-1}} \exp\left\{4Nh\theta^T\hat{\theta})\right\} d\theta. \end{aligned} \tag{3.76}$$

In order to satisfy the definition of the von Mises distribution, we will need to put this on the unit hypersphere. We do this by making the following substitutions:

$$\frac{x_i}{\sqrt{h}} = \theta_i, \frac{\hat{x_i}}{\sqrt{h}} = \hat{\theta}_i \text{ and } d\theta = \frac{dx}{\sqrt{h}}. \tag{3.77}$$

Once again, to more clearly show that complexity increases as the number of parameters increases, we make the substitution $h = \frac{R}{K}$. Equation (3.76) becomes (after a minor abuse of notation)

$$\begin{aligned} Q(\hat{x}) &= \exp(-4N) \int_{S^{K-1}} \exp\left\{ 4Nh \frac{x^T}{\sqrt{h}} \frac{\hat{x}^T}{\sqrt{h}} \right\} dx \frac{1}{\sqrt{h}^K} \\ &= \left(\frac{K}{R}\right)^{\frac{K}{2}} \exp(-4N) \int_{S^{K-1}} \exp\left\{4Nx^T\hat{x}\right\} dx. \end{aligned} \tag{3.78}$$

The integral in Equation (3.78) is now in the form of a von Mises distribution. In general, the von Mises distribution is

$$f_K(x, u, \kappa) = \frac{\exp(\kappa u^T x)}{C_K(\kappa)}, \tag{3.79}$$

where $\kappa \geq 0$, $\|u\| = 1$ and x is random unit vector. The distribution in Equation (3.79) must integrate to one, so

$$\int_{S^{K-1}} \exp(\kappa u^T x) dx = C_K(\kappa), \tag{3.80}$$

where

$$C_K(\kappa) = \left(\frac{2\pi}{\kappa}\right)^{\frac{K}{2}} \kappa \; I_{\frac{K}{2}-1}(\kappa) \tag{3.81}$$

and $I_\zeta(\kappa)$ is the modified Bessel function of order ζ [1]. The right side of Equation (3.81) will be used to determine the value of the integral in Equation (3.78).

By comparing the integral in Equation (3.78) to Equation (3.80), we can see that $\kappa = 4N$. With this, Equation (3.78) then becomes

$$Q(\hat{x}) = \left(\frac{K}{R}\right)^{\frac{K}{2}} \exp(-4N) \left(\frac{\pi}{2N}\right)^{\frac{K}{2}} 4N \; I_{\frac{K}{2}-1}(4N), \tag{3.82}$$

which is independent of $\hat{x}$. Substituting this into Equation (3.74), we obtain that spherical MDL will choose a histogram with the number of bins that minimizes

$$\begin{aligned} MDL_{sphere} &= -\sum_{i=1}^{K} 2v_i \log u_i - \frac{K}{2} \log\left(4\frac{R}{K}\right) + \log V_H \\ &\quad - \log \int_{S^{K-1}} \exp\left\{-\frac{N}{2}(\theta - \hat{\theta})^T H (\theta - \hat{\theta})\right\} d\theta \\ &= -\sum_{i=1}^{K} 2v_i \log u_i - \frac{K}{2} \log\left(4\frac{R}{K}\right) + \log V_H \\ &\quad - \log \left[\left(\frac{K}{R}\right)^{\frac{K}{2}} \exp(-4N) \left(\frac{\pi}{2N}\right)^{\frac{K}{2}} 4N \; I_{\frac{K}{2}-1}(4N)\right], \end{aligned} \tag{3.83}$$

where V_H is defined in Equation (3.73). We note in passing that, even though terms that only depend on the sample size will contribute to the complexity of the model, they don't contribute to the selection process since they are identical to every model. With

this, Equation (3.83) simplifies to

$$\begin{aligned} MDL_{sphere} &= -\sum_{i=1}^{K} 2v_i \log u_i + \frac{K}{2}\log\left(\frac{K}{4R}\right) + \log V_H \\ &\quad + \frac{K}{2}\log\left(\frac{R}{K}\right) + \frac{K}{2}\log\left(\frac{2N}{\pi}\right) - \log\left(\, I_{\frac{K}{2}-1}(4N)\right) \\ &= -\sum_{i=1}^{K} 2v_i \log u_i + \log(V_H) + \frac{K}{2}\log\left(\frac{N}{2\pi}\right) - \log\left(I_{\frac{K}{2}-1}(4N)\right). \end{aligned} \tag{3.84}$$

Spherical MDL closely tracks ordinary MDL when it comes to asymptotics. The modified Bessel function in Equation (3.84) can be considerably simplified as $N \to \infty$:

$$I_{\frac{(K-1)}{2}}(4N) \approx \frac{\exp\{4N\}}{\sqrt{2\pi}}\left[\frac{1}{(4N)^{\frac{1}{2}}} + \frac{(4K-3-K^2)}{8(4N)^{\frac{3}{2}}} + \mathcal{O}(\frac{1}{(4N)^{\frac{5}{2}}})\right]. \tag{3.85}$$

Since the leading term in Equation (3.85) is independent of K, we obtain that spherical MDL and ordinary MDL converge to the same complexity (after ignoring terms independent of K) as $N \to \infty$.

3.4.2 Experimental Results

Every model selection criterion uniquely penalizes parametric complexity. All penalties have mathematical foundations that validate their individual appropriateness. In the case of choosing a model for a distribution whose parameters lie on the hypersphere, as is the case for the histogram, criteria that ignore the geometry of the manifold or improperly apply asymptotic approximations are inherently less appropriate than a criterion that considers these characteristics.

Experiments were conducted generating results of optimal bin counts for histograms of differently shaped distributions. A variety of sampling distributions were created from mixtures of one-dimensional Gaussian distributions as in [67, 102]. The densities chosen represent many characteristics of real densities such as multimodality, skewness and spatial variability. The densities estimated were: Bimodal, Skewed Unimodal, Trimodal and Claw as shown in Figure 3.4. In addition, 2500 trials of sample size 60 were taken from each distribution. The optimal number of bins for AIC, BIC, two part MDL from Equation (2.1), Balasubramanian's asymptotic MDL from Equation (3.31) and spherical MDL was calculated for each trial. The frequency with which the decisions made by AIC, BIC, two part MDL and asymptotic MDL deviated from the decision made by spherical MDL are summarized in Table 3.4.

The results show that AIC and two part MDL penalize complex models the least with AIC most frequently making incorrect decisions. This is true to the reputation of AIC,

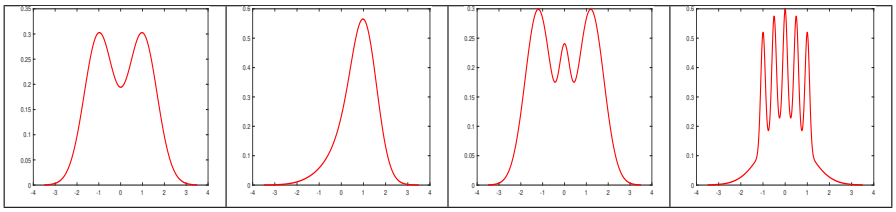


Figure 3.4: Bimodal, skewed, trimodal and claw densities.
Four different densities selected for varying characteristics. Bimodal (left), skewed (center left), trimodal (center right) and claw (right).

	AIC	**BIC**	**MDL2**	**MDL**
Bimodal	1407	221	1372	4
Skew	1441	200	1349	9
Trimodal	1478	197	1323	3
Claw	1569	257	1471	6
Total	5895	875	5515	22

Table 3.4: Comparison of different model selection criteria for histogram density estimation.
Frequency of deviation of 2500 trials of the choice made by Akaike's information criterion (AIC), Bayesian information criterion (BIC), two part Minimum Description Length (MDL) (MDL2) and asymptotic MDL from the choice of spherical MDL for a sample size of 60 drawn from different distributions. We found that BIC consistently penalizes complexity the most while AIC and MDL2 are consistently forgiving of complex models. Spherical MDL and ordinary MDL offer a compromise between goodness of fit and complexity, with spherical MDL always choosing a less complex model, showing that ordinary MDL underpenalizes the complexity of curved parameter spaces.

at reasonable sample sizes. This is expected considering that the number of parameters alone are used to penalize models, with sample size not considered. BIC always chooses models that are the least complex, showing the importance it places on sample size. Balasubramanian's asymptotic MDL and spherical MDL always choose models that have less extreme number of bins. When compared to MDL, spherical MDL tended to prefer less complex models, indicating that MDL underpenalizes the complexity of curved parameter spaces. While these results are somewhat anecdotal, they serve to demonstrate the importance of incorporating the histogram's hypersphere geometry into model selection. Furthermore, in general, we advocate for the modification of model selection criteria to respect their parameter space geometries.

Chapter 4

Geodesics and Transversality Conditions

One common problem geodesics are used to solve is to find the shortest path between two defined points. The goal of this chapter is to find the path with the shortest distance between a distribution on a manifold and a surface on the manifold making one or both of the endpoints of the geodesic variable. Furthermore, we wish to identify the distribution on the surface with which this path intersects. This will involve proofs and applications of the Euler-Lagrange equation and a variable endpoint condition called the transversality condition.

As an aside, the literature in model selection criteria designates the observed Fisher information as $J(\theta)$ and the expected Fisher information as $I(\theta)$, as have we in the previous chapter. A more common way to denote a generic metric tensor is to use g_{ij}, where the subscript represents the location inside the information matrix. Going forward, we adopt this more common approach when we refer to the Fisher information matrix.

4.1 Background

Among differential calculus's many applications are problems regarding finding the maxima and minima of functions. Typically, this involves finding the critical points of a function by setting its first derivative to 0 then using any number of tests, the second derivative test of concavity for example, to classify the critical points as maxima or a minima. After this classification, the desired value of the independent and dependent is determined based on the question asked. Simply put, in the single variable case, differential calculus searches for the values of the independent variable, x, that returns the maximum or minimum value of the function, $f(x)$. In the multi-variable case, the search turns to finding the vector $(x_1, x_2, ..., x_n)$, that returns the maximum value for the function, $f(x_1, x_2, ..., x_n)$.

Analogously, techniques of calculus of variation operate on *functionals*, which are mappings from a space of functions to its underlying field of scalars. That is, a functional is a function that, instead of taking real numbers as inputs, it has functions as inputs and returns a scalar. For example, a simple functional is the definite integral of a function. In this case, the functional takes a function and returns the scalar value of the area underneath the curve between the limits of integration. Clearly, for fixed limits of integration, the value for the area only depends on the function which you are integrating.

Considering the functional $I[y]$, the typical formulation of a calculation of variations problem is

$$\begin{aligned} \min \quad & I[y] = \int_{x_1}^{x_2} F(x, y, \dot{y}) dx \\ & y(x_1) = y_1 \qquad y(x_2) = y_2 \end{aligned} \tag{4.1}$$

where initial and terminal values are defined as (x_1, y_1) and (x_2, y_2) respectively, $\dot{y}$ is Newton's dot notation for the derivative, often adopted in differential geometry, and $F(x, y, \dot{y})$ is the function argument for the functional. In some cases, when the dependencies are obvious or to clarify expressions in proofs, the dependencies of F may be dropped. In general y can be a vector of functions dependent on x, a vector of independent variables. The theory behind finding the extremum to problems like this are analogous to single variable calculus, in which a vanishing first derivative is used to locate critical points. Here, we locate the extremal functions using functional derivatives, leading to solving the Euler-Lagrange equations outlined in Section 4.2.

One classic problem which can be solved by the calculus of variation is to find the shortest path between two points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$. In introductory calculus courses, it is often correctly explained that the total arc length of a curve connecting two points is the sum of infinite, infinitesimally small distances.

$$L = \int \sqrt{1 + \dot{y}^2(x)} dx. \tag{4.2}$$

In n-dimensional Euclidean space, the square of one of those small distances is given by

$$ds^2 = dq_1^2 + dq_2^2 + \cdots + dq_n^2 \tag{4.3}$$

where dq_i is a small distance along the n^{th} dimension.

Considering distributions as points on a statistical manifold, an intuitive measure of similarity is the shortest distance along the manifold between the two distributions. Early works [16, 53, 41] endowed statistical distributions with geometrical properties. However, it was Rao [80] that expanded on the ideas of Fisher [42] that defined a metric for statistical models based on the Fisher information matrix. This connection between distance and distributions, encouraged others to explore the distance between specific families of distributions. Among these families include special cases of the multivariate normal model [95], the negative binomial distribution [66], the gamma distribution [82],

Poisson distribution [73], among others. In differential geometry, these shortest paths between points on a manifold are known as geodesics.

Formulating a geodesic problem as a calculus of variations problem is rather elementary, though its solution is involved. In Euclidean geometry, this path is a straight line, and this distance is easily found. However, moving these ideas onto statistical manifolds complicates both the geometry and the calculus of this seemingly elementary problem. However, just like the elementary case in Equation (4.3), solving for the shortest path L on a manifold, involves the summation of many infinitely small arc lengths, ds.

$$ds^2 = \dot{\Theta}^T g(\Theta)\dot{\Theta} \tag{4.4}$$

where Θ is a parameter vector, $g(\Theta)$ is a metric tensor dependent on the parameter vector and $(\cdot)^T$ represents the transpose of a vector. The metric tensor for Euclidean space is the identity matrix but on the multivariate Gaussian manifold, this metric tensor is the Fisher Information matrix, discussed later in Section 4.2.

This makes the functional we wish to minimize

$$L = \int \sqrt{\dot{\Theta}^T g(\Theta)\dot{\Theta}}dx. \tag{4.5}$$

or, because the square root is a monotonically increasing function, we can conveniently use

$$K = \int \dot{\Theta}^T g(\Theta)\dot{\Theta}dx. \tag{4.6}$$

With this, the calculus of variation problem that solves for the minimum distance on a manifold is

$$\begin{aligned} \min \quad & K = \int \dot{\Theta}^T g(\Theta)\dot{\Theta}dx \\ & \Theta_0 = [\theta_{01}, \theta_{02}, \dots, \theta_{0n}] \qquad \Theta_1 = [\theta_{11}, \theta_{12}, \dots, \theta_{1n}] \end{aligned} \tag{4.7}$$

The metric tensor $g(\Theta) = g(\mu, \Sigma)$ on Riemannian manifolds is the Fisher information matrix, discussed earlier in section 3.1.

The solution to problems like this involve employing the Euler-Lagrange equations, which are a system of second order differential equations useful for finding extremals of functionals. Looking at Equation (4.1), the Euler-Lagrange equation is

$$K_y - \frac{d}{dx}K_{\dot{y}} = 0 \tag{4.8}$$

A detailed proof of Equation (4.8) is found in Section 4.2.

Historically and recently, research regarding geodesics on manifolds has focused on developing closed form equations for the distance between two distributions on the manifold. However, for many purposes, being able to choose the most acceptably similar

distribution outweighs the gravity of knowing the exact degree of likeness. In essence, given a single distribution, a more appropriate question would be which other distribution is most similar, with little bother given to the exact measure of similarity. Regardless, a very intuitive metric of similarity could utilize some concept of distance between distributions.

For example,the goal of all model selection criteria [87, 88, 2, 51] is to choose a model which is most similar to a given distribution. During the choosing method, model selection purposefully clouds the similarity of models with a penalty parameter, resulting in a metric that measures the relative goodness of models. Using this metric, the criteria can select the most optimal model, without relying on exactly how optimal the choice is. Essentially, when model selection chooses a best model, since it truly is better than all other models available, little concern is given to how good the choice really is. When employing model selection, the result is always a distribution, not a measure of similarity between distributions.

The current body of research concerning geodesics on statistical manifold focuses almost entirely on finding the shortest path between two distributions on the manifold and finding a closed form equation for the length of this path. Essentially, limiting research to the narrow focus concerning just the exact distance between two distribution ignores important questions regarding the relationship a single distribution has with all other possible distributions on the manifold. Essentially, questions like the latter focus on *both* the journey *and* the destination instead of narrowly posed questions about just the journey, like the former. Absent from the current research in this field are answers to questions concerning which distribution, from a subset on the manifold is most similar to a given distributions. Here, we examine in detail how to closest distribution to a given distribution given a constraint that the final distribution must satisfy. To our knowledge, this is the first time that geodesics on a Gaussian manifold is studied have been used t. This transfer of focus opens up possible applications to domain adaptation, model selection among other fields that differential geometry has been proven useful.

4.2 Euler-Lagrange Equation

When solving calculus of variation problems the Euler-Lagrange equations is used to find the extremal function. The Euler-Lagrange equations are a system of second-order differential equations that must be satisfied by the stationary function of the functional in question. In attempt to make this work self contained, what follows is a proof of the Euler-Lagrange equation.

First, let's denote the maximizing or minimizing function of Equation (4.1) to be $y^*(x)$. This function is unknown, but not variable. That is, $y^*(x)$ is known to exist, is not subject to change and it will be the function that minimizes (or maximizes) our functional. Because it is the extremal, it will yield a more optimal result than any other

function in the vicinity. We will designate all other functions, sometimes referred to as *admissible curves* or *admissible functions* to be $y(x)$ where $y(x)$ varies slightly from $y^*(x)$ at every x value in between the limits of integration except at the endpoints, for which we will force $y(x) = y^*(x)$. As will be seen shortly, we will capture these slight variations from y^* by simply adding a small value to the optimal function at every point, buy simply adding some different arbitrary to it at every x in its domain.

In Figure 4.1, we can see that we have an optimal function in red/solid. The blue/-dashed function sharing endpoints with the optimal function is an admissible function that possibly differs from the optimal function at every value of x, except (for now) at the end points. In other words, the difference between $y^*(x)$ and an admissible curve is a function of x and this difference is captured in a perturbing function, $\phi(x)$ shown in green/dashed touching the x axis.. By multiplying $\phi(x)$ by a scalar,it is possible to generate all possible differences from our optimal function, resulting in all possible admissible functions. Since $\phi(x)$ is arbitrary, all admissible functions need not be a member of the same family of functions. We can now define all admissible curves to be

$$y(x) = y^*(x) + \epsilon\phi(x), \tag{4.9}$$

where $\phi(x_1) = \phi(x_2) = 0$.

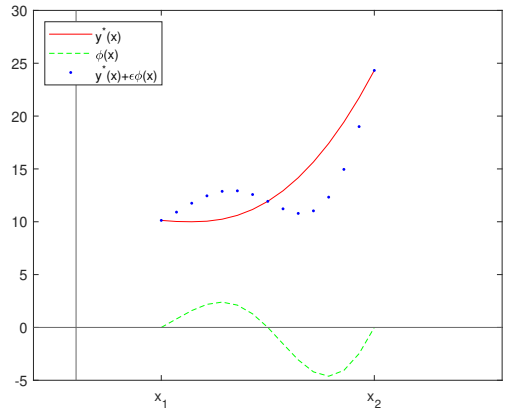


Figure 4.1: Optimal path and neighboring path with perturbing function.
The optimal function, $y^*(x)$ in red with the perturbing function, $\phi(x)$ in green. A generic function $y(x) = y^*(x) + \epsilon\phi(x)$ in the neighborhood of $y^*(x)$ is shown.

The vertical distance between the optimal function and the candidate function is denoted as δy and is called the *variation*. The *total variation* is measure of how sub-optimal a neighboring function is when compared to the optimal function. The total variation is given by

$$\begin{aligned}\Delta I &= I[y] - I[y^*] \\ &= \int_{x_1}^{x_2} \left(F[x, y, \dot{y}] - F[x, y^*, \dot{y^*}] \right) dx.\end{aligned} \tag{4.10}$$

In Section 4.3, we expand upon this proof to allow the optimal function, and therefor all candidate functions a variable endpoint by introducing a transversality condition. The concepts of the proof diverge from the usual Euler-Lagrange equation at this point. Essentially, when utilizing transversality conditions, we need to find both the optimal path, and the optimal boundary conditions of the functions. While this added complexity to the solution comes with added complexity to the mathematics, the proof branches off because of one simple change. In the fixed end point problem, all candidate functions are required to pass through the problem defined initial point and terminal point. This is precisely why the perturbing function, $\phi(x)$ must be 0 at the endpoints. During the proof of the transversality conditions, this restriction is relaxed at one or both of the endpoints. As shown later, this relaxation results in an orthogonality requirement between the optimal path and the terminating surface defined in the transversality condition. Using Equation (4.9), we can rewrite Equation (4.1) evaluated at an arbitrary curve as

$$I = \int_{x_1}^{x_2} F\left(x, y^*(x) + \epsilon\phi(x), \dot{y^*}(x) + \epsilon\dot{\phi}(x)\right) dx. \tag{4.11}$$

Since $\phi(x)$ could be any function and $y^*(x)$ is not variable (it is optimal), the value of Equation (4.11) depends entirely on ϵ, since x is integrated out of the expression. Because of this, and because we know that the optimal value of I happens at $y^*(x)$, we know that the first derivative evaluated at $\epsilon = 0$ (the condition that sets $y(x) = y^*(x)$) must be 0. That is,

$$\frac{dI}{d\epsilon}\Big|_{\epsilon=0} = 0. \tag{4.12}$$

To clarify future parts of the proof, we will redefine $y(x)$ as

$$\Phi(\epsilon) = y^*(x) + \epsilon\phi(x) = y(x), \tag{4.13}$$

reinforcing the sole dependence of the admissible curves on the scalar value of ϵ. Using Equations (4.11), (4.12) and (4.13), the condition which our optimal function, $y^*(x)$ must satisfy are

$$\frac{dI}{d\epsilon} = \int_{x_1}^{x_2} \left(\frac{\partial F}{\partial \Phi}\frac{d\Phi}{d\epsilon} + \frac{\partial F}{\partial \dot{\Phi}}\frac{d\dot{\Phi}}{d\epsilon}\right) dx = 0. \tag{4.14}$$

Since Φ is only a function of ϵ, $\frac{d\Phi}{d\epsilon} = \phi(x)$ and $\frac{d\dot{\Phi}}{d\epsilon} = \dot{\phi}(x)$. Then, (4.14) becomes

$$\frac{dI}{d\epsilon} = \int_{x_1}^{x_2} (\frac{\partial F}{\partial \Phi}\phi(x) + \frac{\partial F}{\partial \dot{\Phi}}\dot{\phi}(x))dx. \tag{4.15}$$

Noticing that the second term of Equation (4.15) has the structure of an integration by parts problem from introductory calculus

$$\int u(x)dv = u(x)v(x) - \int v(x)du \tag{4.16}$$

with $u = \frac{\partial F}{\partial \dot{\Phi}}$ and $dv = \dot{\phi}(x)$ By inspection, the second term in (4.15) becomes

$$\int_{x_1}^{x_2} \frac{\partial F}{\partial \dot{\Phi}}\dot{\phi}(x)dx = \phi(x)\frac{\partial F}{\partial \dot{\Phi}}\Big]_{x_1}^{x_2} - \int_{x_1}^{x_2} \frac{d}{dx}\frac{\partial F}{\partial \dot{\Phi}}\phi(x)dx. \tag{4.17}$$

Furthermore, $\phi(x_1) = \phi(x_2) = 0$ so the first term of Equation (4.17) vanishes. With these substitutions, (4.15) becomes

$$\begin{aligned} \frac{dI}{d\epsilon} &= \int_{x_1}^{x_2} (\frac{\partial F}{\partial \Phi}\phi(x) - \frac{d}{dx}\frac{\partial F}{\partial \dot{\Phi}}\phi(x))dx \\ &= \int_{x_1}^{x_2} (\frac{\partial F}{\partial \Phi} - \frac{d}{dx}\frac{\partial F}{\partial \dot{\Phi}})\phi(x)dx \\ &= 0. \end{aligned} \tag{4.18}$$

which we evaluate at $\epsilon = 0$,the condition to find our extremum. Equation (4.18) can be 0 if $\phi(x) = 0$ and/or $\frac{\partial F}{\partial \Phi} - \frac{d}{dx}\frac{\partial F}{\partial \dot{\Phi}} = 0$. The choice of $\phi(x)$ is only restricted by $\phi(x_1) = 0$ and $\phi(x_2) = 0$. While this doesn't exclude $\phi(x) = 0$ for every $x_1 < x < x_2$, letting the perturbing function be 0 everywhere makes little sense since it will cause no variation from our optimal function. This fact, along with Equation (4.18) having to be satisfied at all allowable $\phi(x)$, requires that the solution must satisfy.

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\frac{\partial F}{\partial \dot{y}} = 0, \tag{4.19}$$

which is known as the Euler-Lagrange equation, the solution to which will give use the optimal function, $y^*(x)$, Considering that solving (4.19) will require solving a second-order differential equation (not necessarily linear) with boundary conditions $y(x_1) = y_1$ and $y(x_2) = y_2$., the Euler-Lagrange equation can be mathematically difficult to solve.

Using typical notation for partial derivatives, the Euler-Lagrange equation will be written as

$$F_y - \frac{d}{dx}F_{\dot{y}} = 0. \tag{4.20}$$

Ending this section is an alternate approach to proving the Euler-Lagrange equation in Equation (4.20), using second order Taylor approximation of a functional. As will be seen, this approach can offers more generality and introduces the terms *first variation* and *second variation*. In short, this will involve writing the total variation given in Equation (4.10) as a second order Taylor polynomial and inspecting the each term in the polynomial.

For an arbitrary admissible curve y in the neighborhood of the maximizing function y^*, the second order Taylor approximation for the functional $F(x, y, \dot{y})$ is

$$\begin{aligned} F(x, y, \dot{y}) \approx\ & F(x, y^*, \dot{y^*}) + F_{y^*}(y - y^*) + F_{\dot{y^*}}(y - y^*) + \\ & \frac{1}{2}\left[F_{y^*y^*}(y - y^*)^2 + 2F_{y^*\dot{y^*}}(y - y^*)(\dot{y} - \dot{y^*}) + F_{\dot{y^*}\dot{y^*}}(\dot{y} - \dot{y^*})^2\right]. \end{aligned} \tag{4.21}$$

Using previously defined notation for arbitrary admissible curves, $y = y^* + \epsilon\phi(x)$ and subtracting the first term of Equation (4.21) to the left side, we have an equation that looks useful, considering how variation was defined in Equation (4.10).

$$\begin{aligned} F(x, y, \dot{y}) - F(x, y^*, \dot{y^*}) \approx\ & F_{y^*}\epsilon\phi(x) + F_{\dot{y^*}}\epsilon\dot{\phi}(x) \\ & + \frac{1}{2}\left[F_{y^*y^*}\epsilon^2\phi^2(x) + 2F_{y^*\dot{y^*}}\epsilon^2\phi(x)\dot{\phi}(x) + F_{\dot{y^*}\dot{y^*}}\epsilon^2\phi^2(x)\right]. \end{aligned} \tag{4.22}$$

This is the integrand of Equation (4.10). Substituting this approximation, the total variation is now.

$$\begin{aligned} \Delta I \approx\ & \epsilon\int_{x_1}^{x_2}\left[F_y^*\phi(x) + F_{\dot{y^*}}\dot{\phi}(x)\right]dx \\ & + \frac{\epsilon^2}{2}\int_{x_1}^{x_2}\left[F_{y^*y^*}\phi^2(x) + 2F_{y^*\dot{y^*}}\phi(x)\dot{\phi}(x) + F_{\dot{y^*}\dot{y^*}}\phi^2(x)\right]dx. \end{aligned} \tag{4.23}$$

The first integral in Equation (4.23) is called the *first variation* and is given the symbol δI. the second integral is called the *second variation* and is given the symbol $\delta^2 I$.

The first variation will result in the Euler-Lagrange equation with a proof that mimics the proof following Equation (4.15). The second variation offers insight into the nature of the extremum, similar to a second derivative test in single variable calculus. Those familiar with maximizing functions from single variable calculus should find the first approach to proving the Euler-Lagrange equation appealing. However, employing Taylor expansions to the funcitonal, then defining the total variation based on this expansion is more intuitive when we introduce variable endpoints.

4.3 Transversality Conditions

Problems which find only Equation (4.20) useful require that the endpoint conditions be fixed points. That is, when working on a statistical manifold and solving for the shortest path, Equation (4.20) are the only conditions that need to be satisfied, in addition to the prescribed endpoint distributions.

In many instances, however, it is desirable to allow the optimal function, $y^*(x)$, to start or end anywhere on a known subsurface of the manifold. In this way, the endpoints are not known, but are constrained by a user defined restriction. This extra freedom requires further conditions on the optimal path, called tranversality conditions. These new conditions are

$$[F + (\dot{\sigma} - \dot{y})F_{\dot{y}}]\Big|_{x=x_2} = 0 \tag{4.24}$$

and

$$[F + (\dot{\rho} - \dot{y})F_{\dot{y}}]\Big|_{x=x_1} = 0. \tag{4.25}$$

Here, $\sigma(x)$ is the surface at which the optimal function terminates, $\rho(x)$ is the surface at which the optimal function originates. Of course, the optimal function still needs to satisfy the original Euler-Lagrange equation in addition to these two new conditions.

What follows is a proof for a variable endpoint transversality condition in equation Equation (4.24), knowing a variable initial point in Equation (4.25) follows a very similar proof. Also, the proof will be done for a single variable path, but we will extend the results to multi-variable paths when appropriate. As a reminder, it was noted in Equation (4.18), that $\phi(x_2) = 0$, was the requirement for the known endpoint distribution, and that this requirement helped simplify the integration involved in the proof of the Euler-Lagrange equation. This is no longer true, and is the key difference which will motivate the proof that follows.

Figure 4.2 graphically represents the problem. The optimal path in red, labelled y^* and an arbitrary neighboring path in yellow, labelled y are shown. For clarity, not shown is a perturbing function, $\phi(x)$, but this function still exists, the effects of which result in the vertical difference of these paths. As in the proof of the Euler-Lagrange equation, this perturbing function is required to be 0 at the initial point, forcing the known initial distribution. However, and key to the transversality condition proof, is that $\phi(x_2) \neq 0$, resulting in an unknown final distribution for the geodesic. This perturbing function is exaggerated in magnitude in order to produce a clear difference between the optimal path and the candidate path. However, the candidate paths are required to be in the neighborhood of the optimal path, a fact that we will leverage in many parts of the proof.

Briefly, here is what the proof entails. We first formulate an expression for the total variation between the candidate function y and the optimal function y^*. This involves in part and most importantly, the difference in the values of the functional as a result of

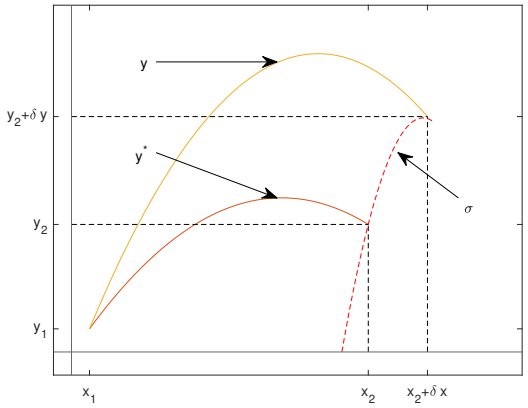


Figure 4.2: General transversality condition.
Showing the relationship between the optimal path y^*, an arbitrary neighboring candidate path y and the terminal transversality surface σ. For clarity purposes, the neighborhood of paths around the optimal path is generously extended to highlight the differences at the variable endpoint. These differences would be smaller, in order to validate the use of first order approximations for the behavior of the the paths at the endpoint

the paths intersecting the boundary surface at different distributions. Then, using first order approximations for the behavior of these paths, we redefine the first variation, δI from Equation (4.23). Like before, we define this first variation as just a function of ϵ, the scalar multiple of the arbitrary perturbing function. With this, we can use single variable calculus to come up with the additional conditions that satisfy the transversality constraint. The approximations involved to most of the heavy lifting for the proof. These approximations will rely on all candidate functions being in the neighborhood of the optimal path. This makes any deviations small on the order 1.

During the proof of the Euler-Lagrange equation, the variation of these paths were capture by a δy at every point in the path. Here, considering the concern is the endpoint and not the path itself, in an effort to simplify notation, the terms δy and δx will be used to signify the deviation from the optimal curve at the points at which they intersect $\sigma(x)$, rather than at any point within the interval. Noteworthy is the fact that generally, because of the variable endpoint surface, the arbitrary candidate function y and the optimal path y^* do not share a common value of the parameter at the endpoint, the idea of which is captured in δx. Because of this, δy is a vertical distance, but not exactly a distance between two paths at x_2. This δy can be written in terms of the paths or the terminal surface. Specifically, δy is given by the following

$$\begin{aligned} \delta y &= y(x_2 + \delta x) - y^*(x_2) \\ &= \sigma(x_2 + \delta x) - \sigma(x_2). \end{aligned} \tag{4.26}$$

As in the proof of the Euler-Lagrange equation, we are looking for an optimal path defined by the following minimization problem:

$$\text{minimize } I[y] = \int_{x_1}^{x_2} F[x, y, \dot{y}]dx \tag{4.27}$$

where x_1 is known, but now x_2 is unknown and restricted to the constraint surface.

Using the definition of total variation from Equation (4.10), we mimic the proof from the Euler-Lagrange equation. However, considering that the neighboring path is defined over the interval $(x_1, x_2 + \delta x)$, the integrals limits must change to reflect this.

$$\begin{aligned} \Delta I &= I[y] - I[y^*] \\ &= \int_{x_1}^{x_2+\delta x} F(x, y, \dot{y})dx - \int_{x_1}^{x_2} F(x, y^*, \dot{y^*})dx. \end{aligned} \tag{4.28}$$

Here, x_1 is defined, but neither x_2 or (of course) $x_2 + \delta x$ is defined. According to Figure 4.2, δx takes on a positive value, but this is not true in general.

The integrals in Equation (4.28) do not share common limits. However, the first integral's limits can be split, resulting in

$$\begin{aligned} \Delta I &= \int_{x_1}^{x_2} F(x, y, \dot{y})dx + \int_{x_2}^{x_2+\delta x} F(x, y, \dot{y})dx - \int_{x_1}^{x_2} F(x, y^*, \dot{y^*})dx \\ &= \int_{x_1}^{x_2} [F(x, y, \dot{y}) - F(x, y^*, \dot{y^*})]dx + \int_{x_2}^{x_2+\delta x} F(x, y, \dot{y})dx. \\ &\qquad\qquad A \qquad\qquad\qquad\qquad\qquad\qquad B \end{aligned} \tag{4.29}$$

Equation (4.29) has a fulfilling geometric interpretation when examined in conjunction with the behaviors of these paths, outlined in Figure 4.2. The first integral is the area between $F(x, y, \dot{y})$ and $F(x, y^*, \dot{y^*})$ between x_1 and x_2, which we have already considered in Equation (4.23) during the proof of the Euler-Lagrange equations. The second integral is the additional area under $F(x, y, \dot{y})$ resulting from the candidate path extending past x_2, to its intersection with the terminal surface, shown in Figure 4.3. The sum of these areas represents a geometric interpretation of the total variation.

Using the results from Equation (4.23), and shortening the notation for clarity, to a first order approximation, Equation (4.29) can be written as

$$\Delta I = \epsilon \int_{x_1}^{x_2} \left[F_y \phi(x) + F_{\dot{y}^*} \dot{\phi}(x)\right] dx + \int_{x_2}^{x_2+\delta x} F(x, y, \dot{y})dx. \tag{4.30}$$
$$B$$

Now, in order to address the second integral in Equation (4.30), we must extend the optimal path past its intersection with the constraint surface, $x = x_2 + \delta x$. The nature of this extension will be discussed later. However, considering that the candidate function is in the neighborhood of the optimal solution, the areas under each of these paths in the

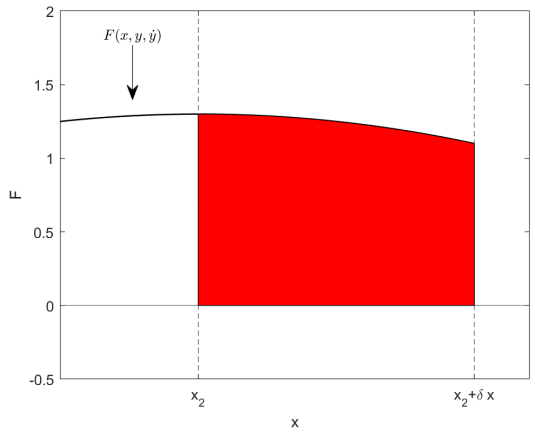


Figure 4.3: Area under neighboring function as a result of transversality condition. The candidate path y intersects the terminal surface at $x_2 + \delta x$. This part of the path adds to the total variation, mathematically represented by the second integral in Equation (4.29) and graphically by the red area above.

interval $(x_2, x_2 + \delta x)$ are almost identical. This is the first time we are leveraging the fact that the candidate path is in the neighborhood of the optimal path. Extending the optimal path past the constraint surface and writing integral $\boldsymbol{B}$ in Equation (4.30) with y^* instead of y allows us to take advantage of the calculus of optimal solutions later in the proof. As such, Equation (4.30) can be rewritten using the optimal path in the integrand of the second integral.

$$\Delta I \approx \epsilon \int_{x_1}^{x_2} \left[F_y \phi(x) + F_{\dot{y}^*} \dot{\phi}(x) \right] dx + \underbrace{\int_{x_2}^{x_2+\delta x} F(x, y^*, \dot{y}^*) dx}_{\boldsymbol{B}}. \tag{4.31}$$

In an attempt to simplify Equation (4.31) to a more useful form, some additional approximations must be made in order to have the upper limits of the integrals be identical. To do so, we are going to find it necessary to define the nature of the extension of the optimal path in Figure 4.2 through the constraint surface, and ending at $x_2 + \delta x$. As stated, he optimal path, y^*, is not defined over this part of the domain, so the nature of the extension is chosen at our convenience. Furthermore, considering that δx is small, the additional area underneath $F(x, y^*, \dot{y}^*)$ as a result of this extension, mathematically represented by the underset $\boldsymbol{B}$ integral is Equation (4.31), is relatively unaffected by the nature of the extension. Because of this, we choose to extend y^* such that $F(x, y^*, \dot{y}^*)$ remains constant in (F, x) space throughout this interval. We can take comfort knowing that the mean value theorem of integrals allows us to approximate the area as a result of any extension of the path using a simple rectangle. This last integral of equation (4.31) is then approximated by the area of a rectangle shown in Figure 4.4. Essentially, the green

rectangle in Figure 4.4 is equivalent to the red rectangle in Figure 4.3. Using the mean value theorem of integrals, we make the following assertion

$$\int_{x_2}^{x_2+\delta x} F(x, y^*, \dot{y}^*)dx \approx (\delta x)F(x, y^*, \dot{y}^*)\Big|_{x=x_2} \tag{4.32}$$

and substitute this result into Equation (4.31). Doing so results in

$$\Delta I \approx \epsilon \int_{x_1}^{x_2} \left[F_y\phi(x) + F_{\dot{y}^*}\dot{\phi}(x)\right] dx + (\delta x)F(x, y^*, \dot{y}^*)\Big|_{x=x_2} . \tag{4.33}$$
$$B$$

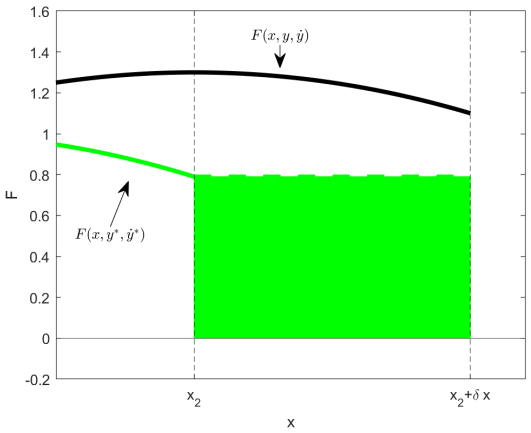


Figure 4.4: Approximation of additional area using MVT for integrals.
The above shows the horizontal extension of the $F(x, y^*.\dot{y}^*)$ into the interval $(x_2, x_2+\delta x)$. This horizontal extension, and its area in green, will allow the second integral of Equation (4.31) to be evaluated using simple geometry. This area is used as an approximate of the area in the same region under the $F(x, y, \dot{y})$, whose path is shown in black.

In addition to this substitution, we invoke the results of Equation (4.15), and rewrite the Equation (4.33) as

$$\Delta I \approx \epsilon \int_{x_1}^{x_2} \left[F_y^* - \frac{d}{dx}F_{\dot{y}^*}\right] \phi(x)dx + \epsilon\phi(x_2)F_{\dot{y}}\Big|_{x=x_2} + (\delta x)F(x, y^*, \dot{y}^*)\Big|_{x=x_2} . \tag{4.34}$$

At this point, we focus on what is contributing to δy. As we can see in Figure 4.5, it is the result of two independent vertical distances. First of all, at x_2, y and y^* differ because of the perturbing function by an amount of $\epsilon\phi(x_2)$. However, because in our example, the candidate function keeps increasing through the interval $(x_2, x_2 + \delta x)$ the perturbing function does not capture the entire δy. There is an additional contribution resulting from

the two paths terminating at different values of x. If we extend y^* tangentially so that it also terminates at $x_2 + \delta x$ as shown in Figure 4.5, we can graphically see where this additional contribution to δy originates. Mathematically, it is the sum of two terms; one from the perturbing function evaluated at x_2 and the other from the tangential extension of the optimal path, both of which are captured by

$$\delta y = \epsilon\phi(x_2) + \dot{y^*}(x_2)\delta x. \tag{4.35}$$

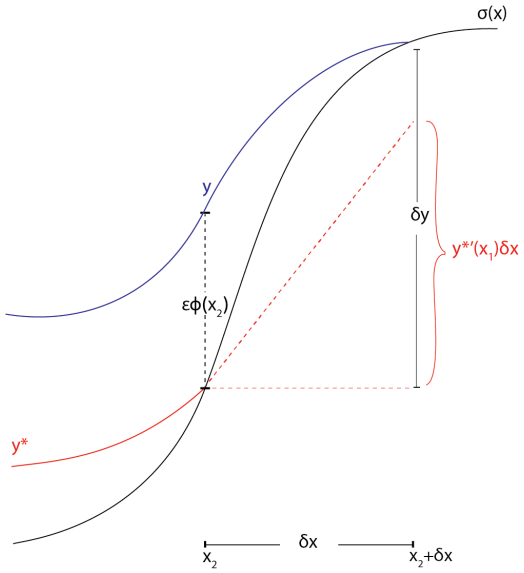


Figure 4.5: Contributors to δy as a result of variable endpoint.
The above shows relationships of the deviations of the neighboring curve from the optimal curve. The value for δy. is the sum of two different vertical difference: one from the perturbing function at x_2, $\epsilon\phi(x_2)$ in dashed black and one from the tangential extension of y^* to $x_2 + \delta x$, red closed brace.

We will use this relationship to substitute for $\epsilon\phi(x_2)$ in Equation (4.34). Since Equation (4.35) is a first order approximation of the extension of of y^*, this expression is called the first variation of I or δI. The motivation for this substitution is to define the first variation of I as a function of only ϵ and not the candidate path or the perturbing function, both of which are arbitrary. Later in the proof, we will take the derivative of

δI with respect to ϵ, leveraging the fact that $\epsilon = 0$ at the optimal function. Substituting Equation (4.35) into Equation (4.34), we get

$$\begin{aligned}\delta I \approx & \epsilon \int_{x_1}^{x_2} \left[F_y^* - \frac{d}{dx} F_{\dot{y}^*}\right] \phi(x) dx + (\delta y - \dot{y^*} \delta x) F_{\dot{y}}\Big|_{x=x_2} \\ & + (\delta x) F(x, y^*, \dot{y^*})\Big|_{x=x_2}.\end{aligned} \tag{4.36}$$

Collecting all terms with δx, Equation (4.36) becomes

$$\begin{aligned}\delta I \approx & \epsilon \int_{x_1}^{x_2} \left[F_y^* - \frac{d}{dx} F_{\dot{y}^*}\right] \phi(x) dx + F_{\dot{y}}\Big|_{x=x_2} (\delta y) \\ & + \left[F\Big|_{x=x_2} - \dot{y^*} F_{\dot{y}}\Big|_{x=x_2}\right] (\delta x).\end{aligned} \tag{4.37}$$

Previously, the terms involving δy and δx were eliminated from the fixed endpoint proof, since these are the deviations of the paths as a result of the transversality condition. That is, there is no deviation from at the endpoint when we define the final distribution exactly. During the fixed endpoint proof, the first variation, δI only had the integral term from Equation (4.37). At this point, it is worth assessing what we have. The ultimate goal is to minimize I, our functional. In doing so, we define the total variation as $I[y] - I[y^*]$. Since y^* is the minimum path, all neighboring paths are larger when evaluated in our functional, so the difference above is negative. That is

$$\Delta I \approx \delta I = I[y] - I[y^*] \leq 0. \tag{4.38}$$

With Equation (4.37), we have a first order approximation for this variation. Because ϵ is arbitrary in both magnitude and sign, Equation (4.38) is true if $\Delta I \approx \delta I$=0, since changing the sign of ϵ moves the candidate function across the optimal path. Of course, this does not require that every individual term in Equation (4.37) be 0. However, we recognize the first term involving the integral from our proof of the Euler-Lagrange equation and, identical to the fixed end point problem, this integral must be 0. Even though we do not yet know which point on the constraint surface the optimal path will intersect, that is x_2 is still unknown, we do know that the optimal path must be the best path that intersects it at that point, otherwise it would not be optimal. So, the integral term in Equation (4.37) must be 0 and thus a solution to the Euler-Lagrange equation. The rest of Equation (4.37) must also be 0

$$F_{\dot{y}^*}\Big|_{x=x_2} (\delta y) + \left[F\Big|_{x=x_2} - \dot{y^*} F_{\dot{y}^*}\Big|_{x=x_2}\right] (\delta x) = 0 \tag{4.39}$$

or

$$F_{\dot{y}^*}\Big|_{x=x_2} (\delta y) + (F - F_{\dot{y}^*} \dot{y^*})\Big|_{x=x_2} \delta x = 0. \tag{4.40}$$

Absent from Equation (4.39) is any explicit dependence on the constraint surface. This

surface is a crucial aspect to the transversality condition and must play a role in choosing the optimal path. To introduce the surface, we look at its behavior at the intersection point with the optimal path. In Equation (4.35), we have an expression for δy that relies on the perturbing function and the tangential extension of the optimal path. Alternatively, we could define δy as the tangential extension of the the constraint surface $\sigma(x)$ in the interval $(x_2, x_2 + \delta x)$, providing that δx is small enough to make a first order approximation of $\sigma(x)$ accurate, which it is. With this idea, we redefine δy as $\delta y \approx \dot{\sigma}(x_2)\delta x$.

Using this in conjunction with Equation (4.39)

$$\begin{aligned} & F_{\dot{y}^*}\Big|_{x=x_2}(\sigma'(x_2))(\delta x) + \left[F_{y^*}\Big|_{x=x_2} - \dot{y^*}F_{\dot{y}^*}\Big|_{x=x_2}\right](\delta x) = 0 \\ &= \left[F_{\dot{y}^*}\Big|_{x=x_2}(\sigma'(x_2)) + \left[F_{y^*}\Big|_{x=x_2} - \dot{y^*}F_{\dot{y}^*}\Big|_{x=x_2}\right]\right](\delta x) = 0 \\ &= \left[F + (\dot{\sigma} - \dot{y})F_{\dot{y}}\right]\Big|_{x=x_2}(\delta x) = 0. \end{aligned} \tag{4.41}$$

Considering that $\delta x \neq 0$ in general, for Equation (4.41) to be true

$$\left[F + (\dot{\sigma} - \dot{y})F_{\dot{y}}\right]\Big|_{x=x_2} = 0. \tag{4.42}$$

which is the transversality condition for a single parameter manifold, previewed in Equation (4.24)

With little effort, these results can be expanded to accommodate for multi-parameter distributions with higher dimensional transversality constraints. This is going to involve extending the single value parameter to an n-dimensional vector, and redefining the constraint as a multi-variable function. Fundamentally, the transversality requirement is captured in Equation (4.40). In order to assist the introduction of higher dimension notation, this equation can be rewritten as an inner product of two vectors.

$$\begin{bmatrix} (F - F_{\dot{y}^*}) \\ F_{y^*} \end{bmatrix} \cdot \begin{bmatrix} \delta x \\ \delta y \end{bmatrix} = 0 \tag{4.43}$$

evaluated at $x = x_2$.

Without changing direction, we change the magnitude of the second vector in the inner product in Equation (4.43) by a factor of (δx^{-1}).

$$\frac{1}{\partial x}\begin{bmatrix} \delta x \\ \delta y \end{bmatrix} = \begin{bmatrix} 1 \\ \partial y/\partial x \end{bmatrix} = \begin{bmatrix} 1 \\ \dot{\sigma}(x_2) \end{bmatrix}. \tag{4.44}$$

In Equation (4.44), we have a vector that is clearly tangent to the transversality surface at $x = x_2$. This equation reveals an orthogonality relationship between the tangent vector to the surface at $x = x_2$ and a vector involving the functional.

In order to take a closer look of the geometry of the constraint surface and to facilitate its extension to multi-parameter manifolds, we redefine it as a level curve of a higher

dimensional surface. Up until this point, the surface had been defined as $y = \sigma(x)$. Alternatively, we can define a three dimensional surface $S(x, y) = y - \sigma(x)$ and redefine the terminal surface as $S(x, y) = 0$ allowing us to use some concepts from multi-variable calculus to extend Equation (4.44) to higher dimensions.

With the transversality surface redefined in this way, we can find and utilise properties of the gradient of this surface. The gradient is typically defined as

$$\nabla S = \begin{bmatrix} \frac{\partial S}{\partial x} \\ \frac{\partial S}{\partial y} \end{bmatrix} = \begin{bmatrix} S_x \\ S_y \end{bmatrix}. \tag{4.45}$$

It is well known that the direction of gradient to a surface defines the orthogonal vector to the surface at the point which it is evaluated, and is a property easily extended to higher dimensional space. Being orthogonal to the surface, the gradient must also be orthogonal to any tangent vector to the surface, including $\begin{bmatrix} 1 \\ \dot{\sigma}(x_2) \end{bmatrix}$. In Equation (4.28), we already established an orthogonal vector to this. These mutually orthogonal vectors must be parallel to each other, which will yield a new expression for the transversality condition.

$$\begin{bmatrix} (F - F_{\dot{y}^*}) \\ F_{y^*} \end{bmatrix} = \alpha \begin{bmatrix} S_x \\ S_y \end{bmatrix}. \tag{4.46}$$

Equation (4.46) is, perhaps, a complicated notation for conditions that were more simply stated in Equation (4.24). However, written in this way, we can logically extend the conditions to multi-parameter distributions. First of all our only concern is that these vectors be parallel, so while the the scalar multiple α has a logical part in the proof, we can let $\alpha = 1$ without changing the conditions. Secondly, and more importantly is the extension of these ideas into higher dimensions. In that case, the functional F becomes dependent on a vector of parameters $\mathbf{y} = [y_1, y_2, \dots, y_n]$ and the surface becomes a function of these parameters $S(x, y_1, y_2, \dots, y_n) = 0$. The transversality conditions can be written as

$$\begin{bmatrix} F - [F_{\dot{\mathbf{y}}}] \cdot [\dot{\mathbf{y}}] \\ F_{\dot{y}_1{}^*} \\ F_{\dot{y}_2{}^*} \\ \vdots \\ F_{\dot{y}_n{}^*} \end{bmatrix} = \begin{bmatrix} S_x \\ S_{y_1} \\ S_{y_2} \\ \vdots \\ S_{y_n} \end{bmatrix}. \tag{4.47}$$

Obtaining closed form solutions to the Euler-Lagrange equation is mathematically difficult on complicated manifolds. Introducing transversality conditions adds orders of magnitude to the difficulty of the task. In this work, the first novel application of both concepts was done on the 2-sphere. Using this surface as an introduction it both offered a

natural transition from spherical MDL to geodesics as well as providing a relatively simple surface on which to apply transversality conditions. However, considering that the remaining on this dissertation focuses on the paramteer space for Gaussian distributions, the results of the working on the sphere are provided in the appendices.

4.4 Euler-Lagrange for univariate Gaussian

Here, we continue or exploration of transversality conditions as they apply to univariate Gaussian distributions. We show the derivation of the Fisher information matrix, followed by the Euler-Lagrange equations. Both of these are used to develop the transversality conditions and applied to surfaces on the manifold of all normal distributions.

4.4.1 Euler-Lagrange Equation for Gaussian

Let a sample X be realized from a univariate normal distribution. That is $X_i = \left\{x_1, x_2, \ldots, x_n\right\} \sim N(\mu, \sigma^2)$. The probability density function for X_i is

$$f(x_i; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(x_i - \mu\right)^2\right\} \tag{4.48}$$

and the log-likelihood is

$$l(\mu, \theta, X) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left\{\sum (x_i - \mu)^2\right\}. \tag{4.49}$$

Using Equation (3.10) and the above probability density function, we can find the entries g_{ij} to the Fisher Information matrix for the univariate Normal distribution

$$\begin{aligned} g_{11} &= -\mathcal{E}\left[\frac{\partial^2}{\partial\mu^2}\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left\{\sum (x_i - \mu)^2\right\}\right)\right] \\ &= -\mathcal{E}\left[-\frac{1}{2\sigma^2}\left\{\frac{\partial}{\partial\mu}\sum\left(-2(x_i - \mu)\right)\right\}\right] \\ &= -\mathcal{E}\left[-\frac{1}{\sigma^2}\right] \\ &= \frac{1}{\sigma^2}, \end{aligned} \tag{4.50}$$

$$
\begin{aligned}
g_{12} = g_{21} &= -\mathcal{E}\left[\frac{\partial^2}{\partial\mu\partial\sigma}\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left\{\sum(x_i-\mu)^2\right\}\right)\right] \\
&= -\mathcal{E}\left[-\frac{1}{2\sigma^2}\left\{\frac{\partial}{\partial\sigma}\sum\left(-2(x_i-\mu)\right)\right\}\right] \\
&= 0,
\end{aligned}
\tag{4.51}
$$

$$
\begin{aligned}
g_{22} &= -\mathcal{E}\left[\frac{\partial^2}{\partial\sigma^2}\left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left\{\sum(x_i-\mu)^2\right\}\right)\right] \\
&= -\mathcal{E}\left[\frac{\partial}{\partial\sigma}\left\{-\frac{1}{\sigma} + \frac{1}{\sigma^3}\sum\left((x_i-\mu)^2\right)\right\}\right] \\
&= -\mathcal{E}\left[\frac{1}{\sigma^2} - \frac{3}{\sigma^4}\sum\left((x_i-\mu)^2\right)\right] \\
&= \frac{2}{\sigma^2}.
\end{aligned}
\tag{4.52}
$$

Displaying these as a matrix, the Fisher information matrix for the univariate Gaussian is

$$
g = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}. \tag{4.53}
$$

This metric tensor has some interesting properties. First of all, it is entirely independent of μ, so the concept of distance does not change as we move along *isovariance* lines. Secondly, the metric rewards, in the sense of arc length, the paths that choose to traverse in the space with larger standard deviations. In Figure 4.6, contour lines are shown to show the relative distances on the manifold and different values for the standard deviation. Traversing along the manifold at a standard deviation of 1.5 is quintuple as far as traversing the between the same two means at a standard deviation of 3.4. Because of the behavior of the Fisher information, geodesics will favor traveling through parts of the manifold occupied by distributions with large standard deviations.

Understanding that the distances between distributions on a manifold is a measure of similarity, intuition confirms the idea that smaller standard deviations are less favorable than large standard deviations if one is concerned with minimizing distances. In Figure 4.7, four distributions are plotted in two pairs. In each pair, one distribution has $\mu = -5$ and the other has $\mu = 5$, making the distribution ten "μ" units from each other, a seemingly equal difference, However, in Figure 4.7a, each distribution has $\sigma = 1$ whereas in Figure 4.7b each distribution has $\sigma = 50$. It's clear that the distributions with large standard deviations are inherently more similar and should probably be closer together on the manifold. If data were drawn from one of the distributions in Figure 4.7a,

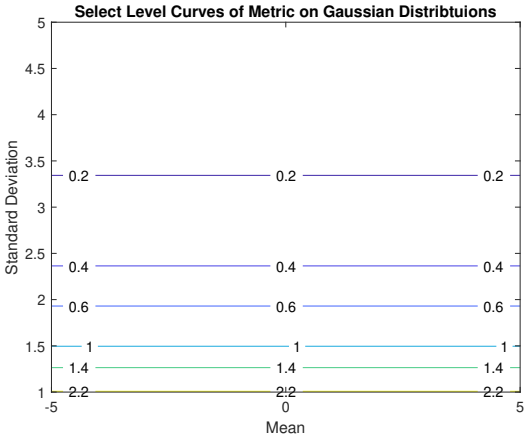


Figure 4.6: Fisher information plot in parameter space of univariate Gaussian. Contour lines displaying affects that the metric tensor has on paths traversed along different standard deviations. Paths traveling along higher standard deviations are shorter than equivalent Euclidean paths traversing along shorter standard deviations.

it would be easy to discern from which of the two distributions it originated. On the other hand, data drawn from a distribution in Figure 4.7b could easily be confused as coming from its counterpart. These ideas reinforce the definitions of the Fisher information and distinguishable distributions, outlined earlier in this work.

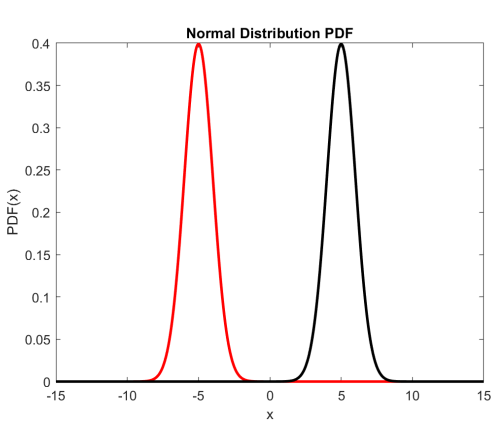
With the metric Equation (4.53), the argument for the arc length functional that we want to minimize is

$$F = \frac{1}{2}\left[\frac{\dot{\mu}^2}{\sigma^2} + \frac{2\dot{\sigma}^2}{\sigma^2}\right]. \tag{4.54}$$

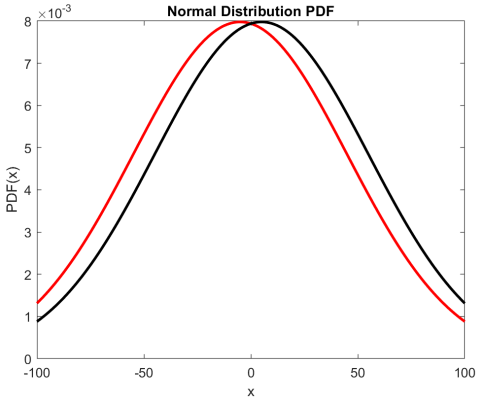
Here, we once again have adopted the common practice of ignoring the square root sign associated with Pythagorean distances, since the square root function is monotonically increasing and will share optimizing paths. Using this function with Equation (4.8), we begin the proof by focusing on μ we can easily see that $F_\mu = 0$ We now need

$$F_{\dot{\mu}} = \frac{\dot{\mu}}{\sigma^2}. \tag{4.55}$$

Using this



(a) $\sigma = 1$



(b) $\sigma = 50$

Figure 4.7: Visual comparison of normal distributions with different variances. Shown above in Figure 4.7a is two distributions: $N(-5,1)$ and $N(5,1)$. In Figure 4.7b are two distributions: $N(-5,50)$ and $N(5,50)$. Even though the difference of the means in each pair of distributions are the same, the distributions in Figure 4.7b appear more similar.

$$\begin{aligned}
\frac{d}{dx}F_{\dot{\mu}} &= \frac{\sigma^2\ddot{\mu} - 2\sigma\dot{\mu}\dot{\sigma}}{\sigma^4} \\
0 &= F_{\mu} - \frac{d}{dt}F_{\mu'} \\
&= 0 - \frac{\sigma^2\ddot{\mu} - 2\sigma\dot{\mu}\dot{\sigma}}{\sigma^4} \\
&= -\sigma^2\ddot{\mu} + 2\sigma\dot{\mu}\dot{\sigma} \\
\ddot{\mu} &= \frac{2\dot{\mu}\dot{\sigma}}{\sigma}.
\end{aligned} \tag{4.56}$$

Focusing on σ,

$$F_{\sigma} = -\frac{1}{\sigma^3}\left[\dot{\mu}^2 + 2\dot{\sigma}^2\right]$$

and

$$F_{\dot{\sigma}} = \frac{2\dot{\sigma}}{\sigma^2}. \tag{4.57}$$

Taking the derivative with respect to the parameter

$$\frac{d}{dx}F_{\dot{\sigma}} = \frac{2\sigma^2\ddot{\sigma} - 4\dot{\sigma}^2\sigma}{\sigma^4}.$$

With this

$$\begin{aligned} 0 &= -\frac{1}{\sigma^3}\left[\dot{\mu}^2 + 2\dot{\sigma}^2\right] - \frac{2\sigma^2\ddot{\sigma} - 4\dot{\sigma}^2\sigma}{\sigma^4} \\ &= -\dot{\mu}^2 - 2\dot{\sigma}^2 - 2\sigma\ddot{\sigma} + 4\dot{\sigma}^2 \\ &= -\dot{\mu}^2 - 2\dot{\sigma}^2 - 2\sigma\ddot{\sigma} + 4\dot{\sigma}^2 \\ \ddot{\sigma} &= \frac{2\dot{\sigma}^2 - \dot{\mu}^2}{2\sigma}. \end{aligned} \tag{4.58}$$

So, the Euler-Lagrange equation for the Gaussian are

$$\ddot{\mu} = \frac{2\dot{\mu}\dot{\sigma}}{\sigma} \tag{4.59}$$

and

$$\ddot{\sigma} = \frac{2\dot{\sigma}^2 - \dot{\mu}^2}{2\sigma}. \tag{4.60}$$

To solve this second order system of differential equations, we take advantage of a common change of variable technique that will convert this to a first order system by employing the following substitutions.

$$\begin{aligned} y_1 &= \mu \\ y_2 &= \dot{\mu} \\ y_3 &= \sigma \\ y_4 &= \dot{\sigma}. \end{aligned} \tag{4.61}$$

With these changes, the Euler-Lagrange equations become

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= \frac{2y_2y_4}{y_3} \\ y_3' &= y_4 \\ y_4' &= \frac{2y_4^2 - y_2'^2}{2y_3}, \end{aligned} \tag{4.62}$$

4.4.2 Example for Gaussian Variable Endpoint

If given fixed boundary conditions, all that is required to find the shortest path are Equations (4.59) and (4.60). However, the goal is to find the closest distribution on a given surface, the transversality conditions for the Gaussian need to be applied. For clarity, a specific example will be used.

What is the shortest path from the distributions $N(0, 0.5)$ to a distribution on the line $S(\mu, \sigma) = \sigma - \mu^2$=0?

Equation (4.47) gives us everything required to find the additional conditions to solve this. Considering the surface does not depend on the time parameter, the remaining requirements are $F_{\dot{\mu}} = S_{\mu}$ and $F_{\dot{\sigma}} = S_{\sigma}$, with the functional $F(\mu, \sigma)$ already defined in Equation (4.54).

First, focusing on μ, we already have $F_{\dot{\mu}} = \frac{\dot{\mu}}{\sigma^2}$. From our surface, we can see that

$$S_{\mu} = -2\mu \tag{4.63}$$

With these

$$\begin{aligned} F_{\dot{\mu}} &= S_{\mu} \\ \frac{\dot{\mu}}{\sigma^2} &= -2\mu \\ \sigma^2 &= -\frac{\dot{\mu}}{2\mu} \end{aligned} \tag{4.64}$$

Similarly, for σ

$$\begin{aligned} F_{\dot{\sigma}} &= S_{\sigma} \\ \frac{2\dot{\sigma}}{\sigma^2} &= 1 \\ \sigma^2 &= 2\dot{\sigma} \end{aligned} \tag{4.65}$$

Equating Equations (4.64) and (4.65), we can see that the transersality condition for the surface $S(\mu, \sigma) = \sigma - \mu^2$ is

$$2\sigma = -\frac{\dot{\mu}}{2\mu} \tag{4.66}$$

In summary, the shortest path from $N(0, 0.5)$ to the nearest distribution on the line $S(\mu, \sigma) = \sigma - \mu^2$ must satisfy the following:

$$\begin{aligned} 2\sigma &= -\frac{\dot{\mu}}{2\mu} \\ \sigma &= \mu^2 \qquad \text{(from the surface)} \\ \ddot{\mu} &= \frac{2\dot{\mu}\dot{\sigma}}{\sigma} \\ \ddot{\sigma} &= \frac{2\dot{\sigma}^2 - \dot{\mu}^2}{2\sigma} \end{aligned} \tag{4.67}$$

The results are contrary to usual Euclidean idioms suggesting that the shortest path between two points is a straight line. In Figures 4.8 and 4.9, it can be seen that the shortest path involves traversing through larger standard deviations, with more favorable conditions according to the Fisher metric tensor. The final distribution is $N(0.71, 0.50)$,

which is on the terminal prescribed terminal surface and the distance along this path is 1.306. Noteworthy is the fact that it appears that the almost horizontal line connecting the initial distribution and final distribution appears to be shorter that the above path. However, the metric tensor defined by the Fisher Information matrix heavily penalizes this path. Essentially, each unit parallel to the μ axis at $\sigma = 0.5$ has an arc length of much larger than its Euclidean distance.

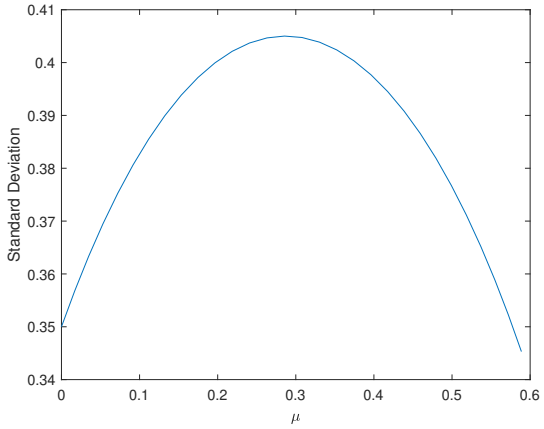


Figure 4.8: Shortest path from $N(0, 0.5)$ to $\sigma = \mu^2$.
Shortest path from $N(0, 0.5)$ to $\sigma = \mu^2$. The final distribution is $N(0.71, 0.50)$. The distance is 1.306.

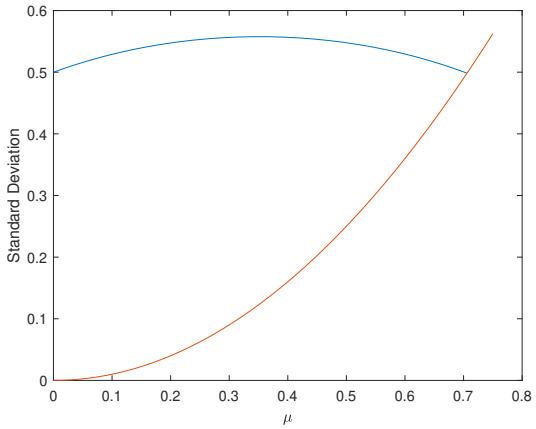


Figure 4.9: Shortest path from $N(0, 0.5)$ to $\sigma = \mu^2$ as terminal surface.
Shortest path from $N(0, 0.5)$ to $\sigma = \mu^2$. The blue path is identical to the path found in Figure 4.8 but the terminal surface, $S(\mu, \sigma) = \sigma - \mu^2 = 0$ is shown in red.

Figure 4.10 shows seven intermediate distribution along the geodesic. As expected by the path, the distributions along the path have a larger standard deviation. As already

established, larger standard deviations hide the uniqueness in univariate Gaussians, thus providing a logical bridge between the initial distribution and the terminal surface.

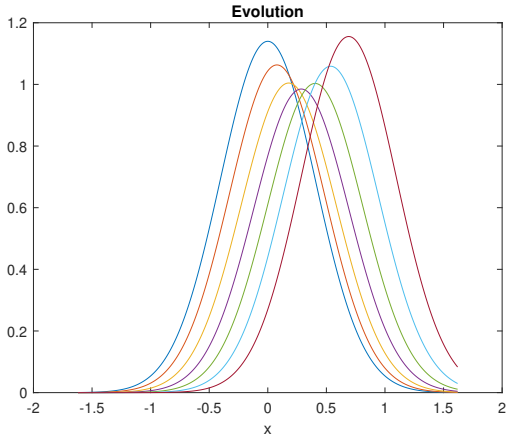


Figure 4.10: Evolution of Gaussian from N(0,0.5) to final distribution.
Showing the evolution of the seven Gaussian distributions as they transition from N(0,.5) (leftmost distribution) to the closest distributions on $\sigma = \mu^2$, N(.71,0.5) (rightmost distribution).

As an example to provide further insight, Figure 4.11 shows the shortest path from the distribution $N(5, .35)$ to the line $\mu = 7$. As expected, the path traverses through higher standard deviations. Furthermore, it achieves the highest standard deviation at its eventual terminal distribution. Logically, $\dot{\sigma} = 0$ at the final distribution since $\dot{\sigma} < 0$, at the terminal distribution the shortest path would start to be penalized by the tensor prior to achieving its destination. Conversely, if $\dot{\sigma} > 0$ at the terminal distribution. the shortest path would have not yet used the most beneficial standard deviations, according to the Fisher Information matrix.

The intermediate distributions shown in Figure 4.12 show the distributions change as they seek the final surface. As the path move towards the final surface $\mu = 7$, the spread of the distributions get larger, with the largest spread being the final distribution. This is expected, since there is no benefit to revisiting smaller standard deviation, according to the metric tensor.

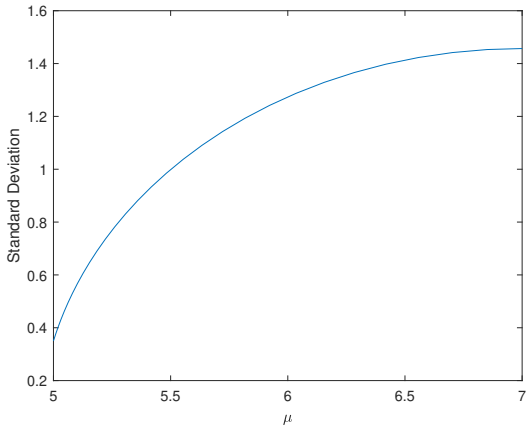


Figure 4.11: Shortest path to $S(\mu) = \mu - 7 = 0$.
Shortest path from $N(5, 0.35)$ to $\mu = 7$. The final distribution is $N(7, 1.46)$. The distance is 3.0.

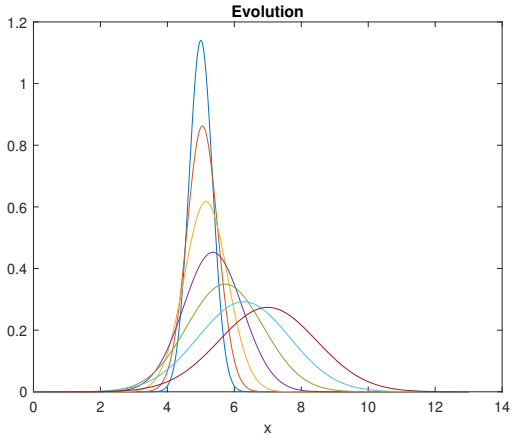


Figure 4.12: Evolution of distributions to $\mu = 7$.
Showing the evolution of seven Gaussian distributions as they transition from N(5,.35) (leftmost distribution) to the closest distributions on μ=7, N(7,1.46) (rightmost distribution).

4.5 Euler-Lagrange Equation for Multivariate Gaussian Distributions

The Euler-Lagrange equations are a system of second differential equations used to find extremals in calculus of variation problems. Euler's and Lagrange's original motivation was in solving the isochrone curve, but applications in many areas soon emerged. Here, we employ the Euler-Lagrange equations to find shortest path between two points on a Riemannian manifold.

To find any distance on a manifold, you need to define a distance concept using a metric tensor. Rao [80] was first to determine that the Fisher information matrix satisfies the conditions of a metric on a Riemannian manifold, and is widely used because of its invariance [30].

Recall that the Fisher information matrix is a measure of how much information about the parameter of interest from a multivariate distribution is revealed from collected data. Intuitively, it can be considered an indication of how "peaked" a distribution is around a parameter. If the distribution is sharply peaked, very few data points are required to locate it. As such, each data point carries a lot of information. For a multivariate probability distribution, the Fisher information matrix is given by we use the Equations (A.4) or (3.10). In our model selection criteria, the Fisher information was used to measure a ratio of of volumes in order to properly penalize models on a manifold. Now, the Fisher Information matrix is the metric tensor that will define distances on Riemannian manifolds. Given a distribution on a manifold, by use of this metric tensor, we can identify a closest second distribution residing on a surface within the manifold. Furthermore, finding the Fisher Information matrix for the multivariate Gaussian will allow exploration into the behaviour of geodesics on their manifold via the use of Equation 4.6.

Here, we consider the n-dimensional multivariate Gaussian with density given by:

$$f(x_n : \mu_n, \Sigma) = 2\pi^{-\frac{n}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp -\frac{(X-\mu)^T \Sigma^{-1} (X-\mu)}{2} \tag{4.68}$$

where X is a data vector, $\mu = \mu_1, \mu_2, ..., \mu_n$ is the n-dimensional mean vector of the distribution and Σ is the $n \times n$ covariance matrix.

Since the covariance matrix is symmetric, the number of unique parameters contained in it is the sum of the number of diagonal elements and the upper $\frac{(n+1)(n)}{2} = \frac{n(n+1)}{2}$. With the n-dimensional mean vector, the total number of scalar parameters in an n-dimensional multivariate Gaussian is $\frac{(n+3)n}{2}$, which will be the size of the Fisher Information matrix. For the purpose of this proof, we will collect all of these parameters as a single vector, θ such that

$$\theta = \{\underbrace{\mu_1, \mu_2, ...\mu_n}_{\theta_1, \theta_2, ..., \theta_n}, \underbrace{\sigma^2_{1,1}, \sigma^2_{1,2}, ..., \sigma^2_{n,n}}_{\theta_{n+1}, ...\theta_{\frac{(n+3)n}{2}}}\} \tag{4.69}$$

To clarify, this new parameter θ has the mean vector μ as its first n components and the resulting components are made up of the unique elements of the covariance matrix, starting with the first row, followed, by the second row but without the first entry, since $\Sigma_{1,2} = \Sigma_{2,1}$ and $\Sigma_{1,2}$ is already included in θ. We capture all of the parameters of the multivariate Gaussian distribution in this non-traditional vector form because it is more conceptually inline with the calculation of the Fisher Information matrix defined in Equation (A.4).

So using Equations (A.4) and 4.68, the Fisher Information for the general multivariate Gaussian distribution is

$$g_{ij}(\mu, \Sigma) = \frac{1}{2} tr \left[\left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \right) \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \right) \right] + \frac{\partial \mu}{\partial \theta_i}^T \Sigma^{-1} \frac{\partial \mu}{\partial \theta_j} \tag{4.70}$$

for which a very detailed proof can be found in the appendix of this work. In the case of the bivariate Gaussian distribution, this 5 x 5 matrix has only 15 unique elements, because of its symmetry. Once again, the detailed derivation of each of the elements is provided in the appendix. The results are

$$
\begin{aligned}
g_{11} &= \frac{\sigma_2^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\
g_{22} &= \frac{\sigma_1^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\
g_{33} &= \frac{1}{2}\left(\frac{\sigma_2^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}\right)^2 \\
g_{44} &= \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}\right)^2 \\
g_{55} &= \frac{\sigma_1^2\sigma_2^2 + \sigma_{12}^2}{\left(\sigma_1^2\sigma_2^2 - \sigma_{12}^2\right)^2} \\
g_{12} &= -\frac{\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} = g_{21} \\
g_{34} &= \frac{1}{2}\left(\frac{\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}\right)^2 = g_{43} \\
g_{35} &= -\frac{\sigma_{12}\sigma_2^2}{\left(\sigma_1^2\sigma_2^2 - \sigma_{12}^2\right)^2} = g_{53} \\
g_{45} &= -\frac{\sigma_{12}\sigma_1^2}{\left(\sigma_1^2\sigma_2^2 - \sigma_{12}^2\right)^2} = g_{54}
\end{aligned}
\tag{4.71}
$$

All elements dealing with the information between a component of μ and an element of Σ vanish, which is a property extended to every multivariate Gaussian distributions of higher dimensions than the bivariate case.

4.5.1 Euler-LaGrange equations for Bivariate Gaussian Distributions

In general, the Euler-Lagrange equations have proven very difficult to solve for the $n-$dimensional multivariate Gaussian distribution, in part because the derivatives required are elusive. However, restricting the multivariate Gaussian to one with a bivariate mean vector leaves us with problems that have both manageable solutions and interesting consequences.

With the Fisher information matrix defined in Equation 4.70 and more specifically for the bivariate Gaussian distribution discussed in the appendix, and the general form of the arc length functional from Equation 4.6, we can define the functional to be minimized in order to find the geodesic on a bivariate Gaussian distribution as

$$
\begin{aligned}
K(\theta) =& \frac{\sigma_2^2 \dot{\mu_1}^2}{k} + \frac{\sigma_1^2 \dot{\mu_2}^2}{k} + \frac{(\sigma_2^2)^2(\dot{\sigma}_1^2)^2}{2k^2} + \frac{(\sigma_1^2)^2(\dot{\sigma}_2^2)^2}{2k^2} + \frac{\sigma_1^2\sigma_2^2\dot{\sigma}_{12}^2}{k^2} + \frac{\sigma_{12}^2\dot{\sigma}_{12}^2}{k^2} \\
& - \frac{2\sigma_{12}\dot{\mu}_1\dot{\mu}_2}{k} + \frac{\sigma_{12}^2\dot{\sigma}_1^2\dot{\sigma}_2^2}{k^2} - \frac{2\sigma_{12}\sigma_2^2\dot{\sigma}_1^2\dot{\sigma}_{12}}{k^2} - \frac{2\sigma_{12}\sigma_1^2\dot{\sigma}_2^2\dot{\sigma}_{12}}{k^2}
\end{aligned} \tag{4.72}
$$

where $k = \sigma_1^2\sigma_2^2 - \sigma_{12}^2$.

After a considerable amount of algebra, Equation 4.72 and Equation 4.8 finally yield the system of second order differential equations with solutions that satisfy the Euler-LaGrange equation and reveal the shortest path between two distributions.

$$
\ddot{\mu}_1 = \frac{\dot{\mu}_1\dot{\sigma}_1^2\sigma_2^2 + \dot{\mu}_2\sigma_1^2\dot{\sigma}_{12} - \dot{\mu}_2\dot{\sigma}_1^2\sigma_{12} - \dot{\mu}_1\sigma_{12}\dot{\sigma}_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \tag{4.73}
$$

$$
\ddot{\mu}_2 = \frac{\dot{\mu}_2\sigma_1^2\dot{\sigma}_2^2 + \dot{\mu}_1\sigma_2^2\dot{\sigma}_{12} - \dot{\mu}_1\dot{\sigma}_2^2\sigma_{12} - \dot{\mu}_2\sigma_{12}\dot{\sigma}_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \tag{4.74}
$$

$$
\ddot{\sigma}_1^2 = \frac{\dot{\mu}_1^2\sigma_{12}^2 + \dot{\sigma}_1^2\sigma_2^2 + \sigma_1^2\dot{\sigma}_{12}^2 - \dot{\mu}_1^2\sigma_1^2\sigma_2^2 - 2\dot{\sigma}_1^2\sigma_{12}\dot{\sigma}_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \tag{4.75}
$$

$$
\ddot{\sigma}_2^2 = \frac{\dot{\mu_2}^2\sigma_{12}^2 + \dot{\sigma}_2^2\sigma_1^2 + \sigma_2^2\dot{\sigma}_{12}^2 - \dot{\mu}_2^2\sigma_1^2\sigma_2^2 - 2\dot{\sigma}_2^2\sigma_{12}\dot{\sigma}_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \tag{4.76}
$$

$$
\ddot{\sigma}_{12} = -\frac{\sigma_{12}\dot{\sigma}_{12}^2 - \dot{\mu}_2\dot{\mu}_2\sigma_{12}^2 - \sigma_1^2\dot{\sigma}_2^2\dot{\sigma}_{12}^2 - \dot{\sigma}_1^2\sigma_2^2\dot{\sigma}_{12}^2 + \dot{\sigma}_1^2\dot{\sigma}_2^2\sigma_{12}^2 + \dot{\mu}_2\dot{\mu}_2\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \tag{4.77}
$$

Along with satisfying this system of equations, the solutions presented here must satisfy transversality conditions at one or both the terminal and initial boundaries. Those conditions will be prescribed accordingly, considering the application of interest.

4.6 Results

4.6.1 Isotropic Terminal Distribution

Multivariate Gaussian distributions can be computationally expensive, especially when trying to use their Fisher Information matrix, considering that the number of parameters grows quadradically with dimension. However, isotropic Gaussian distributions, defined below in 4.78 grow only linearly with the mean vector which can be orders of magnitude more favorable. Accordingly, If given a general multivariate Gaussian distribution, it would be useful computationally to find the isotropic distribution that is most similar.

Let Σ_i be the covariance matrix of an multivariate Gaussian distribution with n mean components. This distribution is *isotropic* if

$$\Sigma_i = \sigma^2 I_n \tag{4.78}$$

where I_n is the $n-$dimensional identity matrix.

In addition to being computationally efficient, isotropic Gaussian distributions provide a convenient submanifold on which we can build a transversality condition. As such, we can start with any general bivariate Gaussian distribution and locate the closest member of isotropic constraint surface defined by the transversality condition.

Formally, let Θ capture all the parameters of a multivariate Gaussian distribution according to Equation 4.69. The functional to minimize is given be

$$\begin{aligned} \min \quad & K = \frac{1}{2}\int \dot{\Theta}^T g(\Theta)\dot{\Theta} dx \\ & \Theta_1 = [\mu_1, \Sigma_1] \qquad \Theta_2 = \phi(\mu_2, \Sigma_2) \end{aligned} \tag{4.79}$$

where μ_1 and Σ_1 are defined but μ_2 and Σ_2 must satisfy the condition in equation 4.78

For the bivariate Gaussian distribution, the terminal surface described in Equation 4.78 can be defined by the surface

$$\Phi(\sigma_1^2, \sigma_2^2) = \sigma_1^2 - \sigma_2^2 = 0 \tag{4.80}$$

with μ_1 free and $\sigma_{12} = 0$

Apply Equation (4.47) to this surface, we get the condition

$$(\sigma_2^2)^2 \dot{\sigma_1}^2 + (\sigma_1^2)^2 \dot{\sigma_2}^2 + \sigma_{12}^2 \dot{\sigma_2}^2 + \sigma_{12}^2 \dot{\sigma_1}^2 - 2\sigma_{12}\sigma_1^2 \dot{\sigma_{12}} - 2\sigma_{12}\sigma_2^2 \dot{\sigma_{12}} = 0 \tag{4.81}$$

So, in addition to the Euler-Lagrange equations in Equation 4.73 thru Equation 4.77, requiring the final distribution to be isotropic requires the geodesic to also satisfy Equa-

tion 4.81, along with with the terminal distribution satisfying the conditions of constraint surface in Equation 4.80, of course.

4.6.1.1 Constant Mean Vector

To introduce the application of this, we will start with an arbitrary non-isotropic distribution with no covariance between the variables. Additionally, we keep the mean vector of the initial and final distributions to be $\mu_0 = \mu_1 = [0, 0]$, thus isolating the motivation of the geodesic to only satisfying the isotropic constraint..

With this in hand, we can now find the closest isotropic distribution to any prescribed bivariate Gaussian distribution. Let's assume an initial distribution with

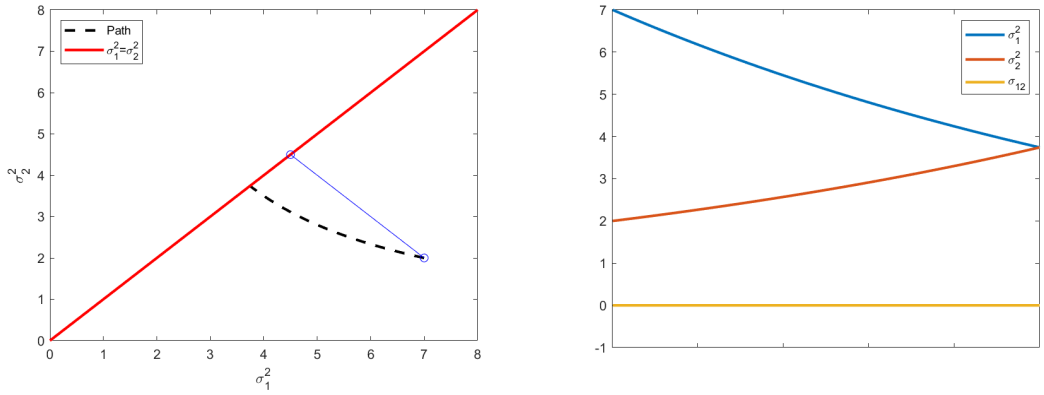
$$\mu_1 = [0, 0], \quad \Sigma_1 = \begin{bmatrix} 7 & 0 \\ 0 & 2 \end{bmatrix} \tag{4.82}$$

and the final distribution lie on the surface defined in Equation 4.80.

Applying the Euler-Lagrange equation with the transversality conditions to the problem above results in a final isotropic distribution is $\sigma^2 = 3, 74$. Figure 4.13a shows the path (dashed) from the initial distribution to the chosen distribution on the isotropic constraint. A Euclidean perpendicular distance would end with a distribution with an isotropic variance equivalent to the average of the original variances and the shortest path is shown with a dotted line. The actual path is an indication of the curvature of the manifold in this area. Here it is shown that it is the final distribution has variances closer to the smaller variance of the original distribution. In fact, with initial variances of σ_1^2, σ_2^2, it can be shown that the final variance, σ_f^2 is given by

$$\sigma_f^2 = \sqrt{\sigma_1^2 \sigma_2^2} \tag{4.83}$$

In Figure 4.13b, we can see the evolution of all the parameters of the distribution along the shortest path. Noteworthy is that, even though σ_{12} is not required to stay at 0, there is no benefit for it deviating from 0, as seen in Figure 4.13b. The Fisher Information Matrix is independent of the mean vector and, since the values of the mean vector are also not part of our isotropic constraint on the final distribution, the mean vector is not compelled to change from the original distribution not shown, justifying the exclusion of the mean vector's path in Figure 4.13



(a) path comparison

(b) parameter values

Figure 4.13: Isotropic example 1.

Shown above in 4.13a is the shortest path (dashed line) from a prescribed initial distribution with diagonal covariance matrix, $\sigma_1^2 = 7, \sigma_2^2 = 2$, to the closest isotropic distribution. The final distribution has $\sigma_1^2 = \sigma_2^2 = 3.74$. The solid line above is the transversality constraint $\sigma_1^2 = \sigma_2^2$, represents the isotropic submanifold. Also drawn is a line connecting the original distribution to the distribution on the constraint with the shortest Euclidean distance, showing the effects of the metric on the path. In 4.13b are the paths showing the value of each element of the Σ as the distributions move towards the transversality constraint.

4.6.1.2 Constant Mean Vector with Initial Covariance

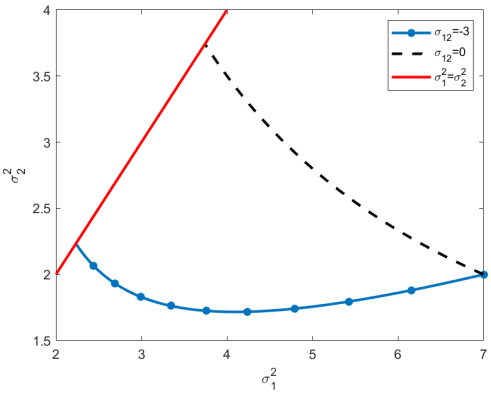
In the example above, the distributions are essentially moving along a $2-$dimensional manifold parameterized just by the individual variances of the variables. Neither the mean vector nor the off diagonal values of the covariance matrices are considered by the Euler-Lagrange equation and therefore remain static. However, starting with an off diagonal element of the covariance matrix changes the problem appreciably. For the reason of comparison, we will start with same diagonal elements of the covariance matrix but include a value for the off-diagonal element.

Here, we consider the problem outline in Equation 4.79 subject to the isotropic constraint in Equation 4.80. Furthermore, we define the initial distribution as

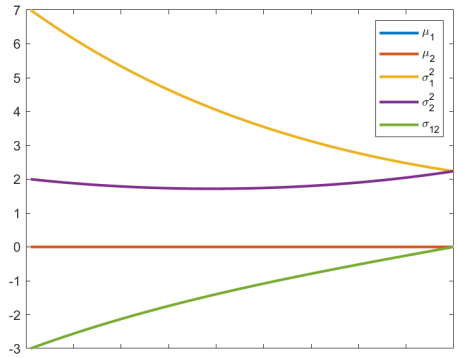
$$\mu_1 = [0, 0], \quad \Sigma_1 = \begin{bmatrix} 7 & -3 \\ -3 & 2 \end{bmatrix} \tag{4.84}$$

As seen in Figure 4.14, including $\sigma_{12} = -3$ alters the where the geodesic ends up on the transversality constraint. Now, the final covariance matrix has diagonal elements of $\sigma_1^2 = \sigma_2^2 = 2.24$. In Figure 4.14a, it is seen that how much influence including a value

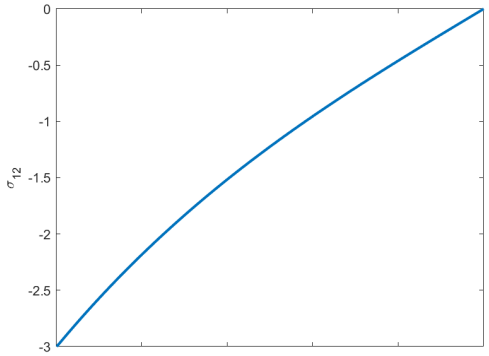
for σ_{12} on the path of the geodesic, as shown by the large amount of deviation from the previous path and the destination on the surface. Figure 4.14b shows all values of the parameters along the geodesic. As mentioned before the mean vector remains constant at all intermediate values of the distribution. Unlike before, the value for σ_{12} has to evolve in order to satisfy the terminal constraint surface, as shown in Figure 4.14b and emphasized in Figure 4.14c



(a) path comparison



(b) parameter values



(c) σ_{12} path

Figure 4.14: Isotropic example 2.

Shown above in 4.14a is the shortest path (blue line) from a prescribed initial distribution an off diagonal covariance element of $\sigma_{12} = -3$. Also shown for comparison is the path from Figure 4.14a (dashed) to see that the difference in variances of the terminal distribution.The final distribution has $\sigma_1^2 = \sigma_2^2 = 2.24$. The red line above is the transversality constraint $\sigma_1^2 = \sigma_2^2$, and represents the isotropic submanifold. Also drawn in Figure 4.14b is the path of all parameters from the initial distribution to the final distribution. Figure 4.14c highlights the values of σ_{12} for each distribution in the geodesic. along the manifold.

4.6.1.3 Starting on the Constraint

While searching for the isotropic boundary condition, interesting geodesics occur if we start with an initial isotropic distribution, but require the mean vector to change. That is, if the initial distribution already resides on the terminal constraint surface, it would seem counter-intuitive if the geodesic is compelled to leave this constraint. However, as seen in Figure 4.15, this is exactly what happens.

Here, once again, we consider the problem outline in Equation 4.79 subject to the isotropic constraint in Equation 4.80. Furthermore, we define the initial distribution as

$$\mu_1 = [-3, 3], \quad \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{4.85}$$

which is already isotropic. Here, we search for the closest final distribution that is isotropic but with a mean vector $\mu_2 = [3, -4]$. So, just considering the location of the mean vector, the distribution starts in Quadrant II and seeks a distribution in Quadrant IV. To the eye, the initial and final distributions are perfectly geometrically symmetric, though we do not know how wide the final distribution will look.

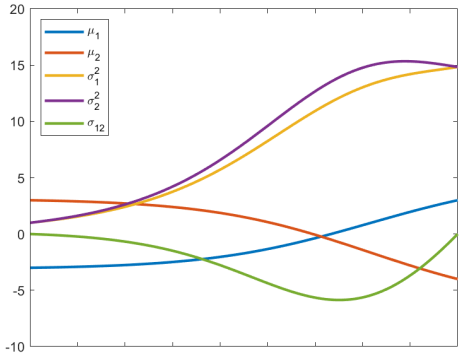
It would be reasonable to think that, on its way to satisfying its final value for μ, the shortest path would be one that maintains its current shape and just slide along the surface. After all, altering a perfectly good covariance matrix seems like more effort that just to change the mean vector.

However, as shown in Figure 4.15, this is not the case. In fact, intermediate distributions obtain covariance matrices with $\sigma_{12} < 0$. as emphasized in Figure 4.15c. Instead of just sliding, the distributions stretch in the direction of the desired mean, which explains negative values of the covariance between the variables, considering the relative locations of the initial mean vector (Quadrant II) and of the final mean vector (Quadrant IV), It is as if the distributions along the geodesic "reach" or "stretch" for its destination, as shown by the middle ellipse in Figure 4.16.

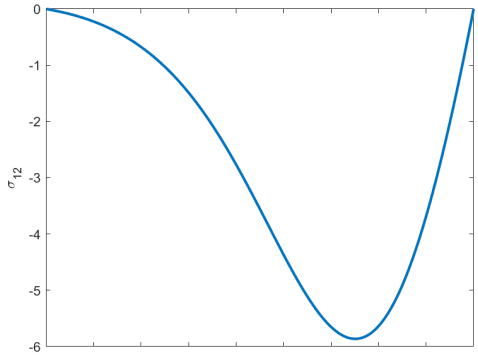
To emphasize how these intermediate distributions reach toward the final distribution, the problem above is repeated, but with a mean vector that starts in Quadrant III with a mean vector of $\mu_1 = [-3, -3]$ and seeks out a final mean vector in Quadrant I with a mean vector of $\mu_2 = [3, 4]$. The initial distribution is still isotropic and the requirement to end isotropic remains. However, as seen in Figure 4.17, the values of σ_{12} acquire positive values along the geodesic. The path of the variance of the variables remains unchanged.

4.6.2 Variable Initial and Final Conditions

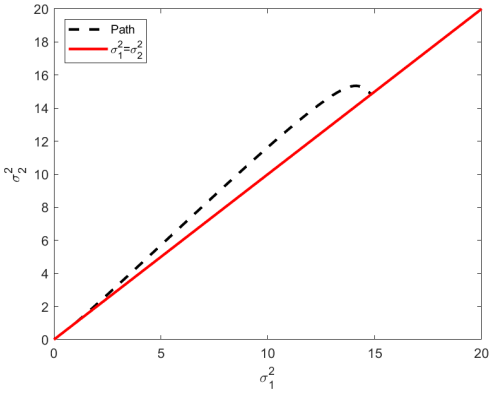
It is possible place transversality conditions on both the initial and final boundaries and these conditions can be entirely independent. Essentially, we are searching for a geodesic between two almost unknown distributions, with just a minimal amount of knowledge



(a) parameter values



(b) σ_{12}



(c) σ_i^2 paths

Figure 4.15: Starting on transversality constraint.
Shown above are representations of the geodesic from an initial isotropic distribution as it seeks a final isotropic distribution with a different mean vector. In Figure 4.15a, the values of all five parameters are shown at each iteration. Figure 4.15b highlights the values of σ_{12} showing that it leaves the constraint surface and acquires negative values. The individual variances of the variables also temporarily abandon their required isotropicity as seen in Figure 4.15c. In Figure 4.15c, the dotted line shows the path of the σ_1^2 and σ_2^2 and the solid line shows the isotropic constraint surface.

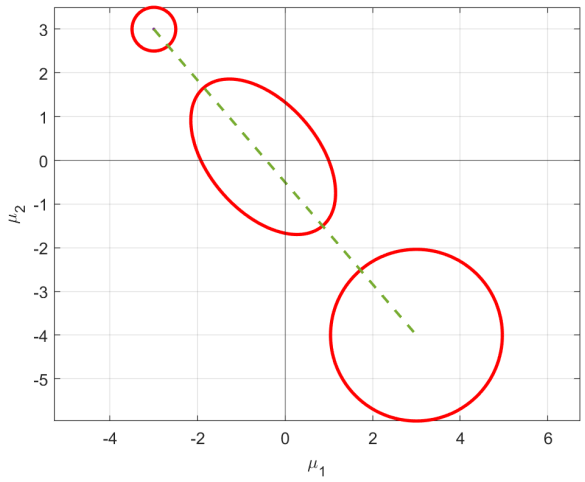


Figure 4.16: Error ellipses showing movement of distribution along constraint. Shown above are ellipses showing contour lines of three different density functions along the geodesic. The initial distribution shown in the top left and final distribution in the bottom right are isotropic. The intermediate distribution acquires negative covariance values, as if to reach towards the final distribution.

about their identities.

As an example, we will require that the final distribution be isotropic as before, but furthermore require that the initial distribution have a mean vector with equal components. In practice, this would further increase the efficiency of algorithms since it could greatly reduce the number of dimensions of the distributions used. The typical problem is slightly more involved so it is redefined as

$$\begin{aligned} \min \quad & K = \frac{1}{2}\int \dot{\Theta}^T g(\Theta)\dot{\Theta} dx \\ & \Theta_1 = \phi_0(\mu_1, \Sigma_1) \qquad \Theta_2 = \phi(\mu_2, \Sigma_2) \end{aligned} \tag{4.86}$$

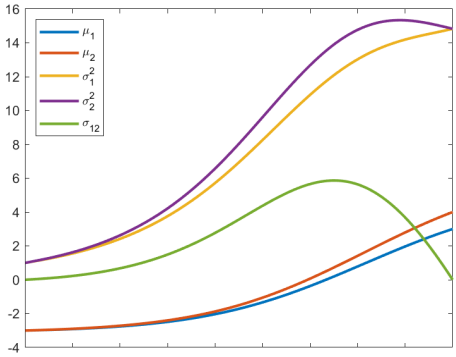
where ϕ_1 and ϕ_2 represent the initial and final transversality surface, such that

$$\phi_1(\mu_{11}, \mu_{12}) = \mu_{11} - \mu_{12} = 0 \quad \text{and} \quad \phi_2(\sigma_1^2, \sigma_2^2) = \sigma_1^2 - \sigma_2^2 = 0 \tag{4.87}$$

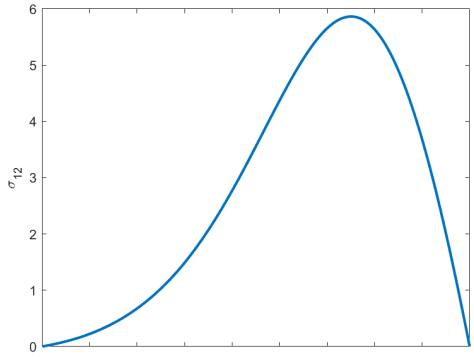
In a necessary arbitrary choice, the initial distribution with unknown mean vector must be prescribed with a covariance matrix. Here, we use

$$\Sigma_1 = \begin{bmatrix} 10 & 0 \\ 0 & 2 \end{bmatrix} \tag{4.88}$$

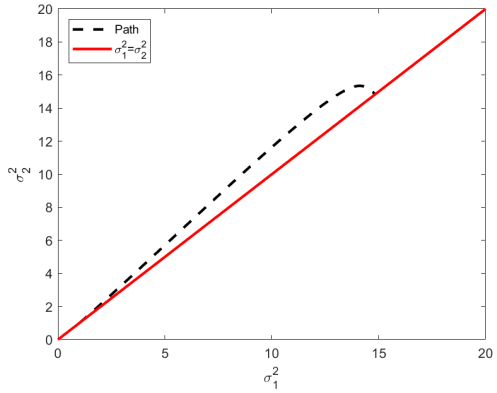
Similarly, the unknown isotropic terminal distribution must be prescribed with a mean vector. Here, we use $\mu_2 = [-3, 13]$



(a) parameter values



(b) σ_{12}



(c) σ_i^2 paths

Figure 4.17: Starting on constraint, example 2.
Shown above are representations of the geodesic from an initial isotropic distribution with a mean vector in Quadrant III as it seeks a final isotropic distribution with a different mean vector Quadrant I. In Figure 4.17a, the values of all five parameters are shown at each iteration. Figure 4.17b highlights the values of σ_{12} showing that it leaves the constraint surface and acquires positive values in contrast to the previous example. The individual variances of the variables also temporarily abandon their required isotropicity as seen in Figure 4.15c. In Figure 4.17c, the dotted line shows the path of the σ_1^2 and σ_2^2 and the solid line shows the isotropic constraint surface. This path is exactly equivalent to the path in Figure 4.15c

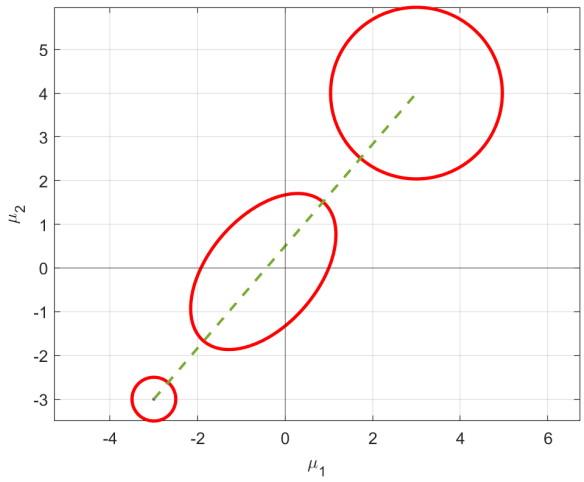
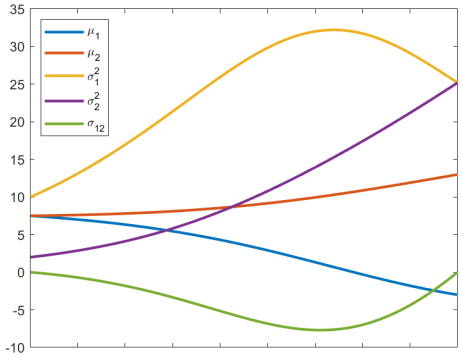


Figure 4.18: Error ellipses showing movement of distribution from Quadrant III to Quadrant I.
Similar to Figure 4.16, level curves of three densities along the geodesic are shown. This time, the intermediate distribution acquires $\sigma_{12} > 0$ along the geodesic as the distributions move from quadrant I to quadrant III along the dotted path.

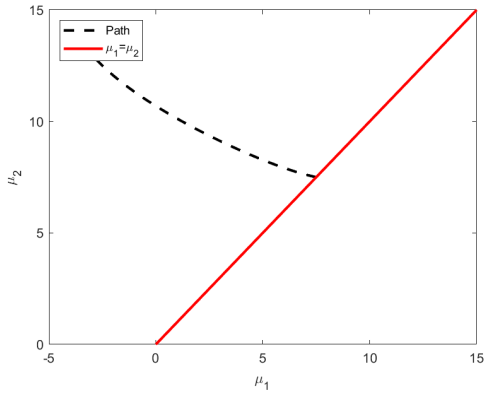
Using Equation (4.47), it can be shown that the requiring the initial distribution to reside on ϕ_1 further requires that the geodesic satisfy

$$(\sigma_2^2 - \sigma_{12})\dot{\mu}_1 + (\sigma_1^2 - \sigma_{12})\dot{\mu}_2 = 0 \tag{4.89}$$

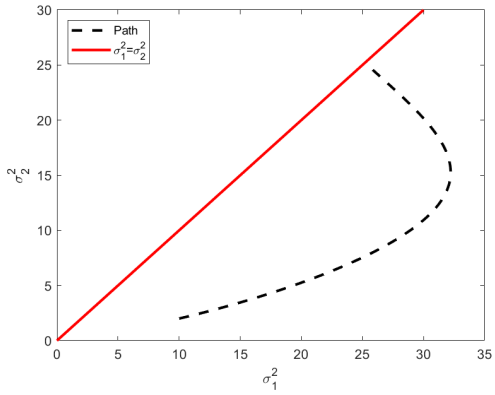
The unknown initial mean vector satisfying the ϕ_1 is $\mu_1 = [7.4, 7.4]$ and the final isotropic distribution has $\sigma_1^2 = \sigma_2^2 = 26.4$. The behavior of the geodesic is shown inn Figure 4.19.



(a) parameter values



(b) path of μ



(c) σ_i^2 paths

Figure 4.19: Initial and final variable endpoint.
Shown above are representations of the geodesic with both a variable initial and terminal boundary condition. Figure 4.19a shows the behaviour of all parameters along the geodesic. Figure 4.19b shows how the geodesic (dashed) seeks out the *initial* distribution on the constraint (solid) and Figure 4.19c shows how the geodesic (dashed) seeks out the *final* distribution on the isotropic constraint.

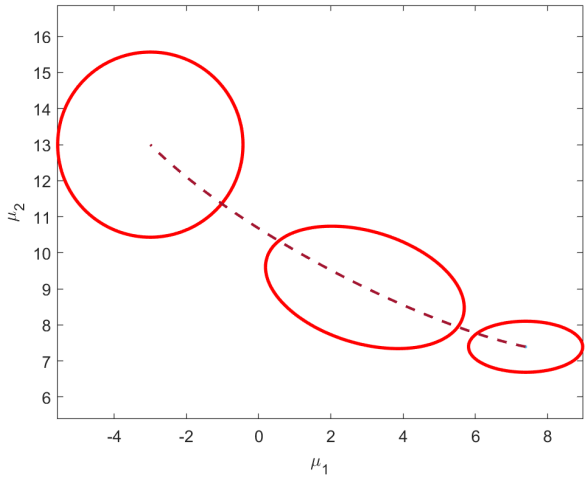


Figure 4.20: Error ellipses form initial and final variable endpoints.
Showing how the geodesic path of the mean vector (dashed) appears curved, but would be straight on the manifold. Additionally, a contour ellipse of the initial, intermediate and final distribution are included. The initial distribution represented by the bottom right ellipse, has components of the mean vector that are equal. The final distribution, at the top left of the path, isotropic.

4.7 Normal Approximation to the Poisson Distribution

Two of the most widely used discrete distributions in statistics are the binomial distribution and the Poisson distributions, with numerous applications of each and countless pages dedicated to them in statistics text books. Historically, especially when sample sizes are large, the calculations for the probabilities for these distributions are costly, considering that you cannot use integral calculus to calculate probabilities over an interval of the discrete random variable. To simplify these calculations, often a continuous distribution was used to approximate the probabilities, with the normal distribution being a popular and accurate choice.

With current statistical software packages, the calculation difficulty for discrete distributions is no longer an issue, making these approximations less important. However, these approximations are useful when teaching about the difference in behavior between discrete and continuous distributions, uses of the Central Limit Theorem in statistical inference and still to simplify large sample calculations. Additionally, in [31], the author shows that continuous approximation of discrete distributions are useful in various moment matching techniques.

Probably because its usefulness for statistical inference on population proportions, the normal approximation to the binomial distribution is well studied and explained. In first year statistics text books, students learn how to take a sample from a population that

is expected to follow a binomial distribution and use the normal approximation when appropriate to estimate confidence intervals and to perform t-tests.

Less common is an equal treatment for the data sampled from expected Poisson distributions. Nonetheless, the Poisson distribution has proven extremely valuable to model in a variety of fields. The simplicity of of its mass function and the lack of complexity that accompanies a single parameter distribution makes it attractive. From modeling DNA mutations [39], modeling mortality from COVID-19 [18] or modeling the clustering of bomb attacks on a city [34], the Poisson distribution is appropriate for numerous research applications.

The probability mass function for a Poisson random variable is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{4.90}$$

and is plotted in Figure 4.21, for $\lambda = 5$, $\lambda = 10$ and $\lambda = 30$. The distribution appears more normal as the value of λ increase. For clarity, the distribution has been plotted as if it is continuous, simply by connecting the value of the mass functions at each discrete value. Typically, the normal approximation is used when $\lambda > 20$, a rather arbitrary value with no mathematical justification. However, the errors in approximation shrink with larger values of the parameter. However, obviously absent from current research is an appropriate method of choosing the normal distribution

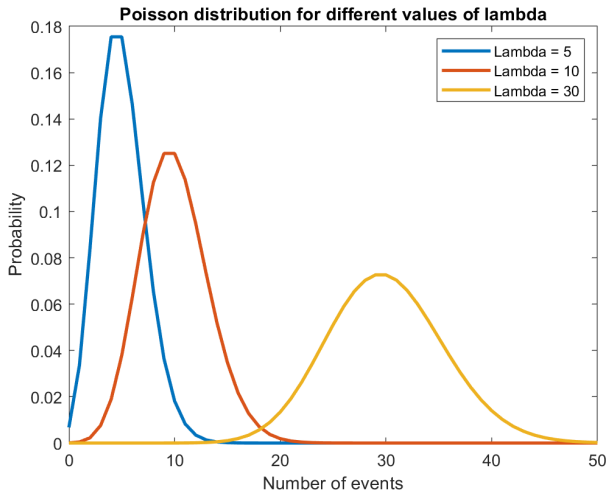


Figure 4.21: Three Poisson distributions.
Three Poisson distributions shown, with increasing values of λ. As the value of the parameter increases, the normal approximation becomes a better fit.

A classic problem involving the normal approximation for the Poisson distribution would give someone the value λ and simply use that value for the mean and variance

of a normal distribution to calculate probabilities that the random variable falls within certain values. However, the biggest setback of questions like this s is that, if someone were to attempt apply the normal approximation to a Poisson distribution, the original value of λ is most likely unknown. In fact when applied, all we would have to utilize is data sampled from a population. The challenge then would be how to choose a normal distribution using the data to best fit the original unknown distribution.

An obvious choice would be to use the MLE of data, absent of knowing any more information than just the realized sample. However, if we know that the data comes from a Poisson distribution, we can leverage this into finding a more suitable normal distribution that better approximates the original Poisson distribution from which the data was sampled. In the context of this research, the univariate Gaussian defined by the MLE resides on the 2-dimensional manifold of all univariate Gaussian distributions. Also on this manifold is a submanifold containing all univariate Gaussian distribution with equal mean and variance.

Choosing the MLE of the data may be over-penalizing an estimate. We are not suggesting that the normal distribution obtained from the MLE is the best possible approximation. However, the purpose of using it is to show how improvements can be made by using this as an initial distribution on the univariate Gaussian and the Laplace constraint surface as our transversality condition. Furthermore, as shown in Figure 4.23, the approximation obtained by using transversality conditions has very little error at the peek of the distribution. Considering that this point uniquely defines a particular univariate Gaussian distribution, it is unlikely that any other method of choosing a normal distribution outperforms the one chosen using the transversality contraint.

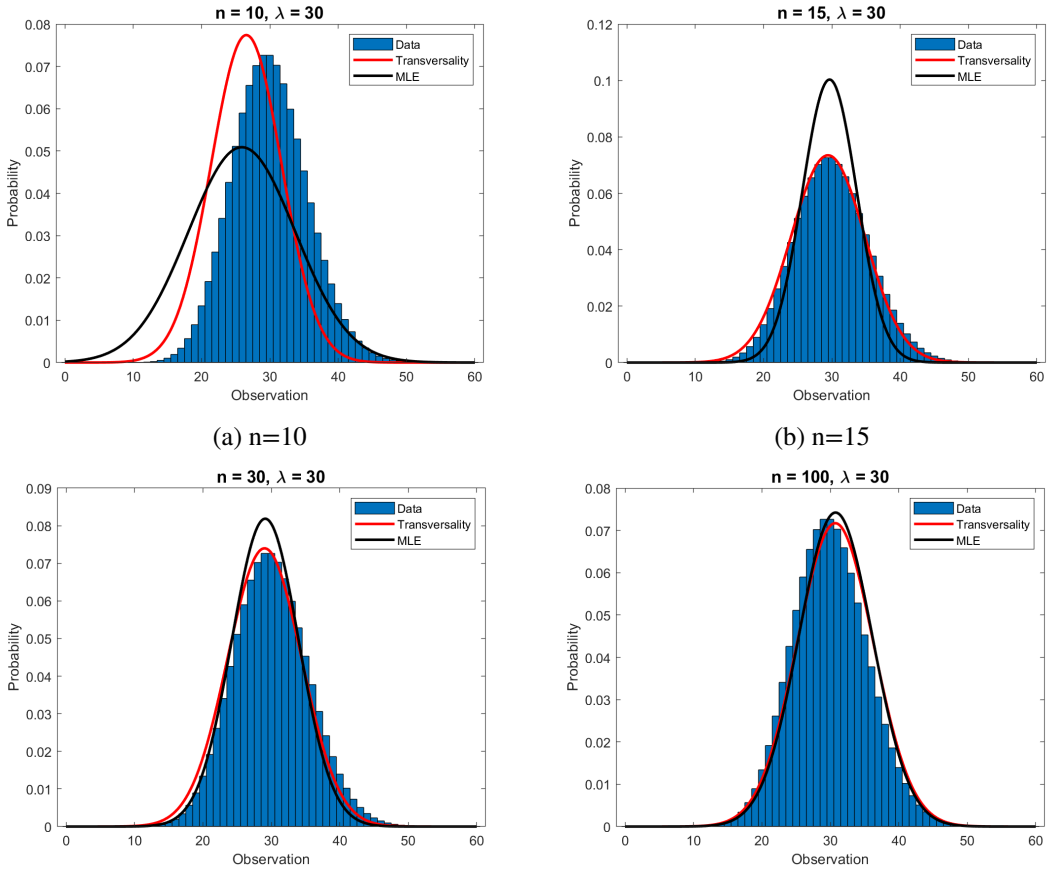
Formally, to choose the best univariate Gaussian to fit a sample from a Poisson distribution we would first find the normal distribution corresponding to the MLE of the data. Then, using $\mu = \sigma^2$ as a boundary surface, use the Euler-Lagrange equation with transversality constraints to identify a better approximation that satisfies the characteristics of the Poisson distribution.

The transversality conditions associated with using this surface as the boundary are

$$\begin{aligned} \mu - \sigma^2 &= 0 \\ \sigma\dot{\mu} + \dot{\sigma} &= 0 \end{aligned} \tag{4.91}$$

To show the relative accuracy of a normal approximation using the MLE and the distribution chosen from the transversality condition, various samples of different sizes, $n = 10, 15, 30, 100$, were drawn from a Poisson distribution with $\lambda = 30$. After sampling, both the

In Figure 4.22, we see the histogram of the original Poisson distribution in blue. In each case, both the MLE approximation and the transversality approximation are plotted. Even with small sample sizes, the distribution chosen from the transversality constraint outperforms the MLE distribution. As expected, the MLE distribution performs much better when the sample size is larger.



(a) n=10 (b) n=15

(c) N=30 (d) N=100

Figure 4.22: Normal approximation to the Poisson distribution.
Shown above are samples of different sizes drawn from a Poisson distribution with $\lambda = 30$. Shown in blue is the probability mass function of the original distribution. Shown in black is the normal distribution using the MLE of the drawn sample. Shown in red is the normal distribution resulting from finding the closest normal distribution on the manifold using the transversality condition that mean and the variance have the same value.

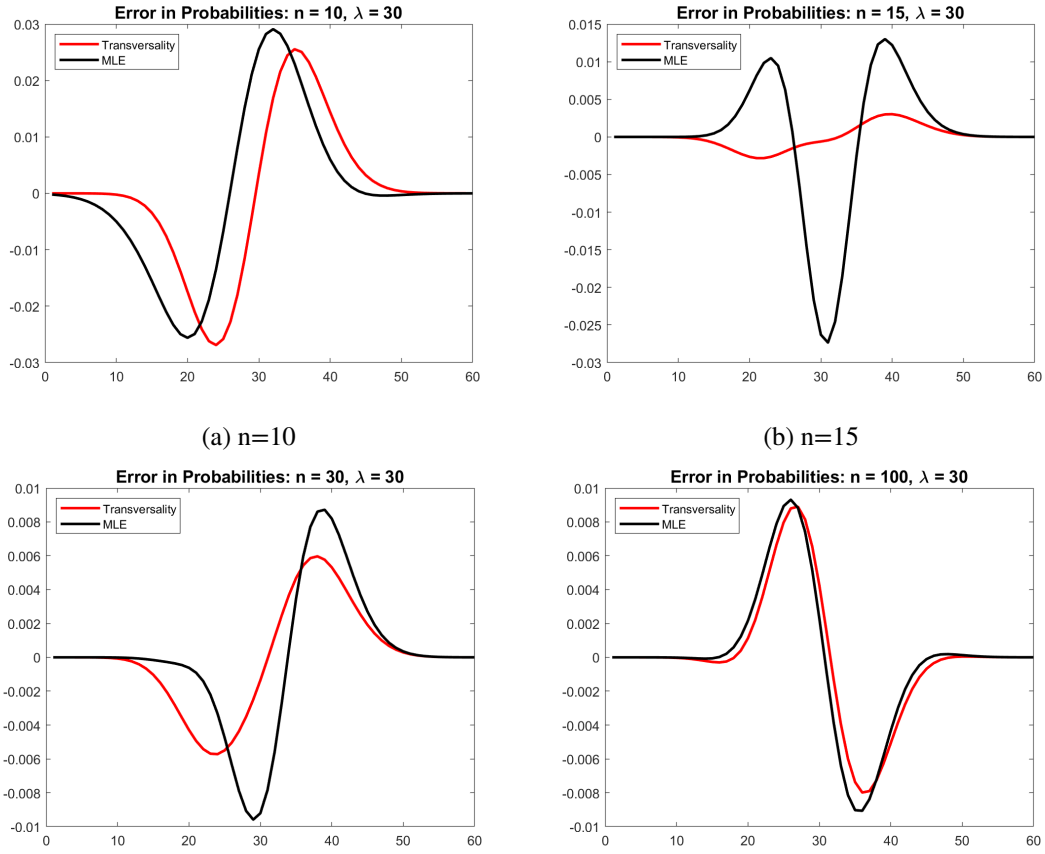
To show the efficacy of the approximations, we calculated the difference between the population probability mass function, the Poisson distribution with $\lambda = 5$ and both the univariate Gaussian constructed from using the MLE and the distribution chosen using the transversality condition. The results are shown in Figure 4.23. Except for large samples, the error from the MLE generated distribution fluctuates widely, with higher maximums and lower minimums. As expected by the law of large numbers, the MLE distribution is a better fit when sample sizes become larger. Except for instances where the sample size is large, the largest error for the MLE distribution occurs close to the mean of the distribution , $\lambda = 30$, Often, this corresponds to minimum error for the distribution from the transversality constraint. This is an indication the peak of the distribution is accurate. Errors in the tails are a result of the shape of the distribution, not the choice of parameter.

Visually, we can see that the distribution chosen by the transversality condition fits the original distribution better than the one chosen by the MLE. Researchers have quantified these errors in many ways [71, 83]. Here, we simply use the area of the absolute value of the difference of each normal probability density function from the probability mass function of the population. The results are summarized in Table 4.1. The errors, which are averaged over 1000 trials, show that the distribution chosen from the transversality condition outperforms the MLE across all sample sizes, with smaller samples showing the largest advantage.

$\lambda = 30$		
n	MLE difference	Transversality Difference
10	0.3325	0.2279
15	0.1875	0.1599
30	0.1670	0.1127
100	0.0701	0.0953

Table 4.1: Errors in normal approximation for $\lambda = 30$.
For each different sample size taken, the error of each distribution, measured by the area between the approximation and the mass function, was found. In all cases the Transversality outperforms the MLE, with both approximations improving with larger sample sizes.

Typically, the normal approximation is used only when $\lambda > 20$. The skew of Poisson distributions with $\lambda < 20$ is too severe for any normal approximation to fit well. However, if only given a sample, it could be uncertain as to what the actual value of the population parameter is, making the criteria for appropriateness elusive. To show that the distribution chosen from the transversality condition still outperforms the MLE, even when uncertain as to whether a normal approximation is advised, Table 4.2 shows the average errors over 1000 trials for both approximations with $\lambda = 5$. As expected,



(a) n=10

(b) n=15

(c) n=30

(d) n=100

Figure 4.23: Errors in the normal approximation to the Poisson distributions. Shown above are samples of different sizes drawn from a Poisson distribution with $\lambda = 30$. Shown in blue is the probability mass function of the original distribution. Shown in black is the normal distribution using the MLE of the drawn sample. Shown in red is the normal distribution resulting from finding the closest normal distribution on the manifold using the transversality condition that mean and the variance have the same value.

both fits improve as the sample size grows. However, even large sample sizes cannot overcome the skewness of the population's mass function. Regardless, the distribution chosen by the transversality condition still outperforms the MLE in all cases.

$\lambda = 5$		
n	MLE difference	Transversality Difference
10	0.2793	0.2137
15	0.2438	0.1965
30	0.1788	0.1489
100	0.1229	0.1084

Table 4.2: Errors in normal approximation for $\lambda = 5$.
Results showing the errors for normal approximations with $\lambda = 5$, as measured by the integral between the approximation and the pmf.

Chapter 5

Conclusions and Further Study

5.1 Conclusions

Model selection criteria seek to parsimoniously balance complexity and goodness of fit. Though many formulations exist, like the well-known AIC, BIC and ordinary MDL, most of them fail to appropriately consider the geometry of the parameter manifold when penalizing models. This always results in underpenalizing the complexity for AIC (for example). Here, we have revisited the MDL criterion from a geometric perspective and derived a new measure for spherical parameter spaces.

Spherical MDL incorporates appropriate asymptotic and geometric arguments to ensure the resulting criterion is intrinsic to the manifold. It was shown through experimental trials that, if regular MDL is used, the complexity penalty is small, resulting in choosing optimal models that are somewhat more complex than spherical MDL. The complexity penalty of the proposed spherical MDL measure employs corrections that take into consideration the shape of the manifold and mitigates the tendency to select unnecessarily complicated models. Using the histogram density estimator as proof of concept, spherical MDL proved that correctly applying the Laplace approximation to constrained parameters yields models that rival current criteria that are either too lenient or incorrectly calculate the volumes on the manifold.

Comparing two distributions is at the core of many statistical and differential geometry applications. If a symmetric difference is desirable, the most logical comparison would be finding a distance the distributions on the residing manifold, with the Fisher information matrix being a natural tensor on the manifold. Employing techniques from calculus of variations is an efficient way of finding this distance.

Finding the distance between a known a parameter and the closest parameter on a surface requires transversality conditions. Here, the implementation of transversality conditions applied to Gaussian distributions and spherical geometries is shown to yield results that are logical and consistent with current research.

Pursuit of the behavior of geodesics on Riemannian manifolds has been proven to be useful endeavor, both in application and expanding the body of research. With almost all of the current research focusing on geodesics between two known distribution on a manifold, obviously is any exploration of geodesics between surfaces on the manifold.

By placing transversality conditions on initial and final distributions, we can see how the geodesic interacts with the manifold as it seeks an unknown distribution. Though this present effort focused on just a small variety of constraints on Gaussian manifolds, this approach to geodesics can be applied to any manifold with any user prescribed constraints.

The efficacy of using transversality conditions was validated by searching for the optimal univariate normal distribution to model a random variable generated from a Poisson distribution. Even though limitations exist because of the nature of the approximation, using the distribution chosen by a transversality constraint outperformed the distribution suggested bt the MLE of the data.

5.2 Further Study

The research presented in this document builds on the current body of knowledge as well as lays a foundation for future research to build upon. Here, we outline some of these possible extensions.

5.2.1 Future in Model Selection

Spherical MDL is one of the few model selection criteria that approaches the field from a geometric perspective. Using vocabulary and ideas of information geometry to define complexity penalties for statistical models is rather novel. Correctly employing asymptotic approximations for curved manifolds has never been done.

We have shown that the standard Laplace approximation introduces inaccuracies when used for parameters residing on curved manifolds. Accordingly, the ideas used in the development of spherical MDL are applicable to other curved parameter spaces. Particularly if the manifold has a predictable geometry, the quadratic integral may have a known closed form solution. For example, for hyper-cylindrical surfaces, there is some indication that the closed form of the integral involves the normalizing constant of the normal von Mises distribution.

Furthermore, we have shown proof of concept by applying spherical MDL to the histogram density estimator, both because of how widely researched histograms are and the ease at which the parameters can be placed on a hypersphere. However, spherical MDL can be applied to other distributions with parameters appropriately reside on hyperspheres. For example, the transition probabilities of a Markov process can be placed on a unit hypersphere [12]. Spherical MDL can be applied to decide the optimal number of transitions of the state space to move from the initial state to the final state.

5.2.2 Geodesics

Employing transversality conditions is a technique that is missing from machine learning algorithms. Considering that machine learning focuses on building models using training data, using transversality conditions to aid in building some models seems a logical application.

The goal of all inference procedures is too generalize statistics to a larger population. However, in cases where that goal is too ambitious, if the sample isn't representative of a diverse population, the results of the generalization can be suspect. In extreme cases, researchers may question whether the results of one population can be generalized to an entirely different population. For example, how accurately can models of a spread of disease in the United States be used to predict the spread of a disease in South Africa. Such problems are known as domain adaptation problems [76, 13, 90]. The population of interest is known as the target domain and the population from which the original inference procedure was performed is known as the source domain.

In [8], the authors propose an information geometric approach to domain adaptation. Here it is shown that making careful use of the structure of the manifold, selected source samples can be trained using support vectors to obtain labels on target samples. The selection of source samples was based on the Hellinger distance to possible target distributions. Using this distance, it was shown that this approach can outperform current domain adaptation algorithms.

The Fisher Rao distance should be equally qualified in this approach to domain adaptation. Furthermore, one or both of the source distribution or target distribution belong to a family of distributions, using transversality conditions could build upon this research improving the performance of the algorithm. It is a sensible next step in researching the application of transversality conditions on manifolds.

However, transversality conditions may be able to expand machine learning algorithms in many statistical research areas. The geometry implied by the Fisher information metric has contributed to dimensional reduction [27], Monte Carlo sampling [93], statistical inference [58], among other explorations of statistical manifolds. Implementing transversality conditions into these fields is restricted by the creativity of the researcher.

Incremental learning [94] offers another problem to which transversality conditions could offer some further insight. Neural networks have the reputation of suffering from catastrophic forgetting, the phenomenon of putting higher priority on newly acquired data. The evolution of the distributions displayed in Chapter 4 offer solutions to this forgetting. Perhaps new data suggests a family of distributions on the same manifold as the current model. The work in Chapter 4 can both choose which model in the family is most like our current model and offer incremental steps towards that chosen distribution.

Regardless of the future applications of transversality conditions on geodesics, the largest obstacle to overcome is to obtain closed form solutions for the Euler-Lagrange equation and the transversality conditions. The mathematics involved in finding these

closed form solutions is extensive and, in many cases, an unsolved problem. But, as lofty of a goal that closed formed solution may be, it is an important goal and a best case scenario. Until then, employing transversality conditions on manifolds may required approximation methods, which currently has the focus of researchers in the field of information geometry.

Chapter 6

Publication

The bulk of the work detailed in Chapter 3 was published on August 3, 2018 in the special issue *Entropy: From Physics to Information Sciences and Geometry* and was presented at a poster in the 2018 *From Physics to Information Sciences and Geometry* conference in Barcelona, Spain.

The bulk of the work detailed in Chapter 4 was published on November 21, 2022 in the special issue of *Entropy: Information and Divergence Measures*.

Bibliography

[1] M. Abramowitz **and** I. A. Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover Publications, Mineola, NY, 1965.

[2] H. Akaike. "A new look at the statistical model identification". *IEEE Transactions on Automatic Control* 19.6 (1974), **pages** 716–723.

[3] Hirotogu Akaike. "Information theory and an extension of the maximum likelihood principle". *Selected Papers of Hirotugu Akaike*. New York, NY: Springer, 1998, **pages** 199–213. ISBN: 978-1-4612-1694-0.

[4] S-I. Amari. *Differential Geometric Methods in Statistics*. Springer, 1995.

[5] S-I. Amari **and** H. Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2001.

[6] Shun-Ichi Amari. "A foundation of information geometry". *Electronics and Communications in Japan (Part I: Communications)* 66.6 (1983), **pages** 1–10.

[7] Khadiga A Arwini **and** Christopher TJ Dodson. *Information Geometry*. Springer, 2008.

[8] Mahsa Baktashmotlagh, Mehrtash Harandi, Brian Lovell **and** Mathieu Salzmann. "Domain adaptation on the statistical manifold". June 2014.

[9] V. Balasubramanian. "Statistical inference, Occam's Razor, and statistical mechanics on the space of probability distributions". *Neural Computation* 9.2 (February 1997), **pages** 349–368.

[10] Ole Barndorff-Nielsen, David Cox **and** Nancy Reid. "The role of differential geometry in statistical theory". *International Statistics Review* 54.1 (April 1986), **pages** 83–96.

[11] A. Barron, J. Rissanen **and** B. Yu. "The minimum description length principle in coding and modeling." *IEEE Transaction on Information Theory* (1998).

[12] Jounghoon Beh, David K. Han, Ramani Durasiwami **and** Hanseok Ko. "Hidden Markov model on a unit hypersphere space for gesture trajectory recognition". *Pattern Recognition Letters* 36 (2014), **pages** 144–153.

[13] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira **and** Jennifer Vaughan. “A theory of learning from different domains”. *Machine Learning* 79 (2010), **pages** 151–175.

[14] Andrew C. Berry. “The Accuracy of the Gaussian approximation to the sum of Independent variates”. *Transactions of the American Mathematical Society* 49.1 (1941), **pages** 122–136.

[15] Dimitri P. Bertsekas. *Nonlinear programming*. 2nd. (pages 291–293). Belmont, MA: Athena Scientific, 1999.

[16] A. Bhattacharyya. “On a measure of divergence between two multinomial populations”. *Sankhyā: The Indian Journal of Statistics (1933-1960)* 7.4 (1946), **pages** 401–406.

[17] C. Bingham. “Distributions on the sphere and on the projective plane.” phdthesis. Yale University, 1964.

[18] Temesgen Birhanu Benti. *Modeling Mortality from COVID-19 Using Poisson Based Regressions: The Case of Sweden*. 2022.

[19] W. M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry*. San Diego: Academic Press, 2002.

[20] E. Bormashenko. “Contact angles of sessile droplets deposited on rough and flat surfaces in the presence of external fields.” *Mathematical Modelling of Natural Phenomena* 7.4 (2012), **pages** 1–5.

[21] Jacob Burbea **and** Dwyritinoren Statemen A. “Informative geometry of probability spaces”. *Exposition. Math* (1986).

[22] C.S.Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Springer, 2005.

[23] C.S.Wallace **and** D.M.Boulton. “An information measure for classification”. *Computer Journal* 11.2 (August 1968), **pages** 185–194.

[24] Carlo Cafaro **and** Stefano Mancini. “Quantifying the complexity of geodesic paths on curved statistical manifolds through information geometric entropies and Jacobi fields.” (2010).

[25] M. Calvo **and** J. M. Oller. “An explicit solution of information geodesic equations for the multivariate normal model”. *Statistics and Risk Modeling* 9.1-2 (1991), **pages** 119–138.

[26] X Cao **and** J Spall. “Comparison of expected and observed Fisher information in variance calculations for parameter estimates”. *Johns Hopkins APL technical digest* 28.3 (2010), **pages** 294–295.

[27] Kevin M. Carter, Raviv Raich, William G. Finn **and** Alfred O. Herro III. "Fisher information nonparametric embedding". *IEEE Transactions on Pattern Analysis and Machine Intelligence* (**march** 2009).

[28] Joseph E Cavanaugh. "Unifying the derivations for the Akaike and corrected Akaike information criteria". *Statistics & Probability Letters* 33.2 (1997), **pages** 201–208.

[29] Joseph E Cavanaugh **and** Robert H Shumway. "A bootstrap variant of AIC for state-space model selection". *Statistica Sinica* (1997), **pages** 473–496.

[30] N.N. Čencov. "Algebraic foundation of mathematical statistics". *Series Statistics* 9.2 (1978), **pages** 267–276.

[31] Ching-Hui Chang, Jyh-Jiuan Lin, Nabendu Pal **and** Miao-Chen Chiang. "A note on improved approximation of the binomial distribution by the skew-normal distribution". *The American Statistician* 62.2 (2008), **pages** 167–170.

[32] Alpha C. Chiang. *Elements of dynamic optimization / Alpha C. Chiang*. English. McGraw-Hill New York, 1992, xiii, 327 p. : ISBN: 0070109117.

[33] Florio M. Ciaglia, Fabio Di Cosmo, Domenico Felice, Stefano Mancini, Giuseppe Marmo **and** Juan M. Pérez-Pardo. "Aspects of geodesical motion with Fisher-Rao metric: classical and quantum". *Open Systems and Information Dynamics* 25.01 (2018), **page** 1850005.

[34] R. D. Clarke. "An application of the Poisson distribution". *Journal of the Institute of Actuaries* 72.3 (1946), **pages** 481–481.

[35] Sueli I.R. Costa, Sandra A. Santos **and** João E. Strapasson. "Fisher information distance: A geometrical reading". *Discrete Applied Mathematics* 197 (2015). Distance Geometry and Applications, **pages** 59–69.

[36] T.M. Cover **and** J.A. Thomas. *Elements of Information Theory*. 2nd. New York, NY: Wiley Interscience, 2006.

[37] Frank Critchley **and** Paul Marriott. "Information Geometry and Its Applications: An Overview". *Computational Information Geometry: For Image and Signal Processing*. Cham: Springer International Publishing, 2017, **pages** 1–31. ISBN: 978-3-319-47058-0.

[38] L. Davies, U. Gather, D. Nordman **and** H. Weinert. "A comparison of automatic histogram construction". *ESAIM: Probability and Statistics* (2009).

[39] Ira W Deveson, Binsheng Gong, Kevin Lai, Jennifer S LoCoco, Todd A Richmond, Jeoffrey Schageman, Zhihong Zhang, Natalia Novoradovskaya, James C Willey **and** Wendell Jones. "Evaluating the analytical validity of circulating tumor DNA sequencing assays for precision oncology". *Nature biotechnology* 39.9 (2021), **pages** 1115–1128.

[40] James G. Dowty. *Chentsov's theorem for exponential families*. 2017.

[41] Bradley Efron **and** David V. Hinkley. "Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information". *Biometrika* 65.3 (1978), **pages** 457–482.

[42] R. A. Fisher. "On the mathematical foundations of theoretical statistics". *Philosophical Transactions of the Royal Society of London, A* 222 (1922). **byeditor**R. A. Fisher, **pages** 309–368.

[43] Carl Friedrich Gauss. *Disquisitiones generales circa superficies curvas*. **volume** 1. Typis Dieterichianis, 1828.

[44] I.M. Gelfand **and** S.V. Fomin. *Calculus of Variations by I.M. Gelfand and S.V. Fomin*. Selected Russian publications in the mathematical sciences. Prentice-Hall, 1964.

[45] Enrico Giusti **and** Graham H Williams. *Minimal surfaces and functions of bounded variation*. **volume** 80. Springer, 1984.

[46] S. O. Gladkov. "A contribution to the computation of the Young Modulus". *Journal of Engineering Physics and Thermophysics* (2003).

[47] P. Grünwald. "A Tutorial introduction to the minimum description length principle". *Advances in Minimum Description Length: Theory and Applications*. MIT Press, 2005.

[48] Hubert Halkin. "Necessary conditions for optimal control problems with infinite horizons". *Econometrica* 42.2 (1974), **pages** 267–272.

[49] P. Hall **and** EJ. Hannan. "On stochastic complexity and nonparametric density estimation". *Biometrika* 75.4 (December 1988), **pages** 705–714.

[50] D. W. Heck, M. Moshagen **and** E. Erdfelder. "Model selection by minimum description length: Lower-bound sample sizes for the Fisher information approximation". *Journal of Mathematical Psychology* 60 (2014), **pages** 29–34.

[51] Trevor Herntier, Koffi Eddy Ihou, Anthony Smith, Anand Rangarajan **and** Adrian Peter. "Spherical minimum description length". *Entropy* 20.8 (2018).

[52] J. S. Hodges **and** D. J. Sargent. "Counting degree of freedom in hierarchical and other richly parameterized Models". *Biometrika* (1988).

[53] Harold Hotelling. "Spaces of statistical parameters". *Bulletin of American Mathematics Society* 36 (1930), **page** 191.

[54] Valerie Isham **and** MK Murray. *Differential Geometry and Statistics*. Routledge, 2017.

[55] H. Jeffreys. *Theory of Probability*. 3rd. New York: Oxford University Press, 1961.

[56] Michael I Jordan **and** Tom M Mitchell. "Machine learning: Trends, perspectives, and prospects". *Science* 349.6245 (2015), **pages** 255–260.

[57] Takashi Kamihigashi. “Transversality Conditions and Dynamic Economic Behaviour”. *The New Palgrave Dictionary of Economics*. London: Palgrave Macmillan UK, 2018, **pages** 13858–13862. ISBN: 978-1-349-95189-5.

[58] Robert E. Kass. “The geometry of asymptotic inference”. *Statistical Science* 4.3 (1989), **pages** 188–234.

[59] John T. Kent. “The Fisher-Bingham distribution on the sphere”. *Journal of the Royal Statistical Society. Series B (Methodological)* 44.1 (1982), **pages** 71–80.

[60] Petri Kontkanen. “Computational Efficient Methods for MDL-Optimal Desnisty Estimation and Data Clustering”. phdthesis. Univeristy of Helsinki, 2009.

[61] S. Kullback **and** R. A. Leibler. “On information and sufficiency”. *The Annals of Mathematical Statistics* 22.1 (1951), **pages** 79–86.

[62] A Kume. “Saddlepoint approximations for the Bingham and Fisher-Bingham normalising constants”. *Biometrika* (2005).

[63] P.S. Laplace. “Memoir on the probability of the causes of events”. *Statistical Science* 1.3 (1986). Translated from Mémoire sur la probabilité des causes par les événemens, Mémoires de l’Académie royale des sciences de Paris (Savants étrangers), t. VI. pp. 621-656; 1774. Oeuvres 8, pp. 27-65, **pages** 364–378.

[64] G. Lebanon. “Metric learning for text documents”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.4 (2006), **pages** 497–508.

[65] Yann LeCun, Yoshua Bengio **and** Geoffrey Hinton. “Deep learning”. *Nature* 521.7553 (2015), **pages** 436–444.

[66] C. Manté. “The Rao’s distance between negative binomial distributions for Exploratory Analyses and Goodness-of-Fit Testing”. *61st World Statistics Congress - ISI2017*. Marrakech, Morocco, 2017.

[67] S. J. Marron **and** M. P. Wand. “Exact mean integrated squared error”. *The Annals of Statistics* 20.2 (1992), **pages** 712–736.

[68] D.J.C. McKay. “A practical Bayesian framework for backpropation networks”. *Neural Computation* 4.3 (1992b), **pages** 448–472.

[69] Ronald E. Miller. *Dynamic Optimization and Economic Applications*. McGraw-Hill Inc., 1979. ISBN: 0-07-042810.

[70] Ronald E. Miller. *Dynamic Optimization and Economic Applications*. English. McGraw-Hill International Book Co New York : London, 1979, x, 332 p. : ISBN: 0070421803.

[71] Wouter Molenaar. “Simple Approximations to the Poisson, Binomial, and Hypergeometric Distributions”. *Biometrics* (1973), **pages** 403–407.

[72] D. J. Navarro. “A Note on the Applied Use of MDL Approximations”. *Neural Computation* 16 (2004), **pages** 1763–1768.

[73] Frank Nielsen. “The many faces of information feometry”. *Notices of the American Mathematical Society* 69 (January 2022), **pages** 36–45.

[74] W. Pan. “Bootstrapping likelihood for model selection with small samples”. *Jounral of Computational and Graphical Statistics* (1999), **pages** 687–698.

[75] B. Parthasarathy **and** J. B. Kadane. “Laplace approximations to posterior moments and marginal distributions on circles, spheres, and cylinders”. *The Canadian Journal of Statistics* 19.1 (March 1991), **pages** 67–77.

[76] Vishal M Patel, Raghuraman Gopalan, Ruonan Li **and** Rama Chellappa. “Visual domain adaptation: A survey of recent advances”. *IEEE Signal Processing Magazine* 32.3 (2015), **pages** 53–69.

[77] A. Peter **and** A. Rangarajan. “An information geometry approach to shape density minimum description length model selection.” *Computer Visions Workshop*. 2011.

[78] A. Peter **and** A. Rangarajan. “Information Ggometry for landmark shape analysis: unifying shape representation and deformation”. *IEEE Transactions on PAMI* 31.2 (February 2009), **pages** 337–350.

[79] A. Peter **and** A. Rangarajan. “Maximum likelihood Wavelet density estimation with applications to image and shape matching”. *IEEE Transactions on Image Processing* 17.4 (2008), **pages** 458–468.

[80] C. R. Rao. “Information and accuracy attainable in estimation of statistical parameters”. *Bulletin of the Calcutta Mathematical Society* 37 (1945), **pages** 81–91.

[81] Alfréd Rényi. “On measures of entropy and information”. *fourth Berkeley symposium on mathematical statistics and probability*. **volume** 1. 547-561. Berkeley, California, USA. 1961.

[82] F Reverter **and** J.M Oller. “Computing the Rao distance for gamma distributions”. *Journal of Computational and Applied Mathematics* 157.1 (2003), **pages** 155–167.

[83] Wesley Jacob Rich. “Examining the accuracy of the normal approximation to the Poisson random variable” (2009).

[84] J. Rissanen. “A Universal prior for integers and estimation by minimum description length”. *The Annals of Statistics* 11.2 (1983), **pages** 416–431.

[85] J. Rissanen. “Fisher information and stochastic complexity”. *IEEE Transactions on Information Theory* 42 (1996), **pages** 40–47.

[86] J. Rissanen. “MDL denoising”. *IEEE Transaction on Information Theory* (2000).

[87] J. Rissanen. “Modeling by shortest data description”. *Automatica* 14 (1978), **pages** 465–471.

[88] J. Rissanen. “Stochastic complexity”. *Journal of the Royal Statistical Society* 49.3 (1987), **pages** 223–239.

[89] Christian Robert **and** George Casella. *Monte Carlo Statistical Methods*. 2nd. Springer-Verlag New York, 2004.

[90] Kate Saenko, Brian Kulis, Mario Fritz **and** Trevor Darrell. "Adapting visual category models to new domains". *Computer Vision – ECCV 2010*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, **pages** 213–226. ISBN: 978-3-642-15561-1.

[91] G. Schwarz. "Estimating the dimension of a model". *The Annals of Statistics* 6.2 (1978), **pages** 461–464.

[92] C. E. Shannon. "A mathematical theory of communication". *The Bell System Technical Journal* 27.3 (1948), **pages** 379–423.

[93] Aaron Sim, Sarah Filippi **and** Michael P. H. Stumpf. *Information Geometry and Sequential Monte Carlo*. 2012.

[94] Christian Simon, Piotr Koniusz **and** Mehrtash Harandi. "On learning the geodesic path for incremental learning". *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 2021, **pages** 1591–1600.

[95] Lene Theil Skovgaard. "A Riemannian geometry of the multivariate normal model". *Scandinavian Journal of Statistics* 11.4 (1984), **pages** 211–223.

[96] James A Spall. "Model Selection and Statistical Information". *Introduction to Stochastic Search and Optimization*. John Wiley Sons, Ltd, 2003, **pages** 329–366. ISBN: 9780471722137.

[97] James C. Spall. "Monte Carlo Computation of the Fisher Information Matrix in Nonstandard Settings". *Journal of Computational and Graphical Statistics* 14.4 (2005), **pages** 889–909.

[98] A. Srivastava, I. Jermyn **and** S. Joshi. "Riemannian analysis of probability density functions with applications in vision". *IEEE Computer Vision and Pattern Recognition (CVPR)*. 2007, **pages** 1–8.

[99] Stephen M. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, 1986.

[100] C. Taylor. "Akaike's information criterion and the histogram". *Biometrika* (1987).

[101] C. S. Wallace **and** D. L. Dowe. "Refinements of MDL and MML coding". *Computer Jounral* (1999b)).

[102] M. P. Wand **and** M. C. Jones. "Comparison of smoothing parameterizations in bivariate kernel density estimation". *Journal of the American Statistical Association* 88.422 (1993), **pages** 520–528.

[103] Shintaro Yoshizawa **and** Kunio Tanabe. "Dual differential geometry associated with the Kullback-Leibler information on the Gaussian distributions and its 2-parameter deformations". *SUT Journal of Mathematics* 35.1 (1999), **pages** 113–137.

Appendix A

Proof

A.1 Fisher Information for Gaussian

Here, we consider the structure of the Fisher information matrix for n-dimensional multivariate Gaussian with density given by

$$f(x_n : \mu_n, \Sigma) = 2\pi^{-\frac{n}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp -\frac{(X-\mu)^T \Sigma^{-1}(X-\mu)}{2} \tag{A.1}$$

where X is a data vector, $\mu = [\mu_1, \mu_2, ..., \mu_n]$ is the n-dimensional mean vector of the distribution and Σ is the $n \times n$ covariance matrix.

These define an n-dimensional multivariate Gaussian. This distribution has $\frac{(n+3)n}{2}$ unique parameters, which we will capture as a single vector θ such that

$$\theta = \{\underbrace{\mu_1, \mu_2, ...\mu_n}_{\theta_1, \theta_2, ..., \theta_n}, \underbrace{\sigma^2_{1,1}, \sigma^2_{1,2}, ..., \sigma^2_{n,n}}_{\theta_{n+1}, ...\theta_{\frac{(n+3)n}{2}}}\}. \tag{A.2}$$

The log-likelihood associated with Equation (A.1) is

$$\log f = L(\theta) = \frac{n}{2} \ln 2\pi - \frac{1}{2} \ln \det \Sigma - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu). \tag{A.3}$$

As stated, we will find an equation that will yield each element of the Fisher information matrix using the following definition

$$g_{ij}(\theta) = E\left[\frac{\partial}{\partial \theta_i} \log f(x;\theta) \frac{\partial}{\partial \theta_j} \log f(x;\theta)\right]. \tag{A.4}$$

Equation (A.4) requires the partial derivative of each parameter in the log-likelihood defined in Equation (A.3).

$$\frac{\partial L}{\partial \theta_i} = -\frac{1}{2} tr \underbrace{\left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i}\right]}_{A} + \frac{1}{2}(x-\mu)^T \Sigma^{-1} \underbrace{\frac{\partial \Sigma}{\partial \theta_i}}_{B} \Sigma^{-1}(x-\mu) + \underbrace{\frac{\partial \mu}{\partial \theta_i}^T}_{C} \Sigma^{-1}(x-\mu) \tag{A.5}$$

Similarly, we can take the partial derivative with respect to a different parameter, θ_j and obtain the same result, indexed with j instead of i. The below equation labels $A.B, C$ will provide clarity in the proof.

To find each g_{ij} in the Fisher information matrix, we need to find the expectation of the product of every combination two partial derivatives which will result in nine terms. However, upon taking the expectation, some of these terms will vanish to 0, because the expectation of data vector x approaches the mean vector μ. Specifically, let us denote A_i, B_i, C_i to be the terms of $\frac{\partial L}{\partial \theta_i}$ and A_j, B_j, C_j to be the terms of $\frac{\partial L}{\partial \theta_j}$. Upon taking the expectation of the product, $A_i C_j = C_i A_j = B_i C_j = B_j C_i = 0$. Ignoring these, we will look individually at each of the remaining terms. Starting with $A_i A_j$

$$\begin{aligned} A_i A_j &= \left(-\frac{1}{2} tr\left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i}\right]\right)\left(-\frac{1}{2} tr\left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j}\right]\right) \\ &= \frac{1}{4} tr\left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i}\right] tr\left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j}\right] \end{aligned} \tag{A.6}$$

Next, calculating $B_i B_j$,

$$\begin{aligned} B_i B_j &= \left[\frac{1}{2}(x-\mu)^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1}(x-\mu)\right]\left[\frac{1}{2}(x-\mu)^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \Sigma^{-1}(x-\mu)\right] \\ &= \frac{1}{4}\left[(x-\mu)^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1}(x-\mu)(x-\mu)^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \Sigma^{-1}(x-\mu)\right] \end{aligned} \tag{A.7}$$

We are required to take the expectation of this, which is a fourth moment of the multivariable normal distribution. The result of this is

$$B_i B_j = \frac{1}{4}\left[tr\left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i}\right) tr\left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j}\right)\right] + 2tr\left[\left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i}\right)\left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j}\right)\right]. \tag{A.8}$$

Turning attention to $C_i C_j$,

$$\begin{aligned} C_i C_j &= \frac{\partial \mu}{\partial \theta_i}^T \Sigma^{-1}(x-\mu)(x-\mu)^T \Sigma^{-1} \frac{\partial \mu}{\partial \theta_j} \\ &= \frac{\partial \mu}{\partial \theta_i}^T \Sigma^{-1} \Sigma \Sigma^{-1} \frac{\partial \mu}{\partial \theta_j} \\ &= \frac{\partial \mu}{\partial \theta_i}^T \Sigma^{-1} \frac{\partial \mu}{\partial \theta_j}. \end{aligned} \tag{A.9}$$

The final set of non-vanishing terms, $A_i B_j + B_i A_j$ are considered simultaneously.

$$\begin{aligned} A_i B_j + B_i A_j &= \left(-\frac{1}{2} tr \left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \right] \right) \left(\frac{1}{2} (x-\mu)^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \Sigma^{-1} (x-\mu) \right) \\ &\quad + \left(\frac{1}{2} (x-\mu)^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1} (x-\mu) \right) \left(-\frac{1}{2} tr \left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \right] \right) \\ &= -\frac{1}{4} \Bigg\{ \left(tr \left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \right] \right) \left((x-\mu)^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \Sigma^{-1} (x-\mu) \right) \\ &\quad + \left((x-\mu)^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1} (x-\mu) \right) \left(tr \left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \right] \right) \Bigg\} \end{aligned} \tag{A.10}$$

Now, we use the identity

$$b^T A b = tr(b b^T A) \tag{A.11}$$

on all terms of Equation (A.10) that do not yet involve the trace of a matrix. Doing so,

Equation (A.10) becomes

$$
\begin{aligned}
A_i B_j + B_i A_j = &-\frac{1}{4}\Bigg\{ \left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\right]\right)\left(tr\left[(x-\mu)(x-\mu)^T\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\Sigma^{-1}\right]\right) \\
&+ \left(tr\left[(x-\mu)(x-\mu)^T\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1}\right]\right)\left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\right]\right)\Bigg\} \\
= &-\frac{1}{4}\Bigg\{ \left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\right]\right)\left(tr\left[\Sigma\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\Sigma^{-1}\right]\right) \\
&+ \left(tr\left[\Sigma\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1}\right]\right)\left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\right]\right)\Bigg\} \\
= &-\frac{1}{4}\Bigg\{ \left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\right]\right)\left(tr\left[\frac{\partial\Sigma}{\partial\theta_j}\Sigma^{-1}\right]\right) \\
&+ \left(tr\left[\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1}\right]\right)\left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\right]\right)\Bigg\}.
\end{aligned}
\tag{A.12}
$$

Once again, we took the expectation as required to find the Fisher information. Finally, we use the commutative property of trace to clean up the expression in Equation (A.12)

$$
\begin{aligned}
A_i B_j + B_i A_j = &-\frac{1}{4}\Bigg\{ \left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\right]\right)\left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\right]\right) \\
&+ \left(tr\left[\Sigma^{-1}[\frac{\partial\Sigma}{\partial\theta_i}\right]\right)\left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\right]\right)\Bigg\} \\
= &-\frac{1}{2}\left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\right]\right)\left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\right]\right).
\end{aligned}
\tag{A.13}
$$

Combining Equations (A.6), (A.7), (A.9) and (A.13) into Equation (A.5), we obtain

$$\begin{aligned} g_{i,j}(\theta) =& \frac{1}{4} tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\right] tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\right] \\ &+ \frac{1}{4}\left[tr\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\right) tr\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\right)\right] + 2tr\left[\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\right)\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\right)\right] \\ &+ \frac{\partial\mu}{\partial\theta_i}^T \Sigma^{-1}\frac{\partial\mu}{\partial\theta_j} \\ &+ -\frac{1}{2}\left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\right]\right)\left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\right]\right) \\ =& \frac{1}{2} tr\left[\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\right)\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\right)\right] + \frac{\partial\mu}{\partial\theta_i}^T \Sigma^{-1}\frac{\partial\mu}{\partial\theta_j}. \end{aligned} \tag{A.14}$$

A.2 Fisher Information of a 2-Dimensional Gaussian

The usefulness of Equation (A.14) lies in the ability of calculating the individual terms in the equation. Here, the inverse of a covariance matrix is extremely illusive for a high-dimensional Gaussian distribution, even after leveraging its symmetric properties.

Considering just a 2×2 covariance matrix, Equation (A.14) is tractable, since its inverse is known exactly and is reasonably manageable. Furthermore, we will collect all the parameters of a general bivariate Gaussian into a single vector, to facilitate the calculation of each element of the Fisher information matrix.

$$\begin{aligned} \theta &= \left[\theta_1, \theta_2, \theta_3, \theta_4, \theta_5\right] \\ &= \left[\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12}\right] \end{aligned} \tag{A.15}$$

Starting with the diagonal elements, g_{11} and g_{22} share similar structures. Focusing just on g_{11}, and consider a 2-dimensional Gaussian with mean vector $\mu^T = [\mu_1, \mu_2]$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

which will be indexed according to Equation (A.15). We now have, using the standard definition of the inverse

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}\begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix}.$$

For succinctness, we will let $k = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}$.

Conveniently, the means are not involved in the covariance matrix, so the first term of Equation (A.14) vanishes. To find g_{11} we need

$$\begin{aligned} g_{11} &= \frac{1}{k}\left[\frac{\partial \mu}{\partial \mu_1}^T \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix} \frac{\partial \mu}{\partial \mu_1}\right] \\ &= \frac{1}{k}\left[\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}\right] \\ &= \frac{\sigma_2^2}{k} \\ &= \frac{\sigma_2^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}. \end{aligned} \tag{A.16}$$

Finding g_{22} easily follows the above, resulting in

$$g_{22} = \frac{\sigma_1^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}. \tag{A.17}$$

the remaining diagonal elements involve the just the first term of Equation (A.14). For the variance of the first variable, we will need to find

$$\begin{aligned} g_{33} &= \frac{1}{2}\left(tr\left[\Sigma^{-1}\frac{\partial \Sigma}{\partial \sigma_1^2}\right]^2\right) \\ &= \frac{1}{2}\left(tr\left[\frac{1}{k}\begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\right]^2\right) \\ &= \frac{1}{2k^2}\left(tr\left[\begin{pmatrix} \sigma_2^2 & 0 \\ -\sigma_{12} & 0 \end{pmatrix}\right]^2\right) \\ &= \frac{1}{2k^2}\left(tr\left[\begin{pmatrix} (\sigma_2^2)^2 & 0 \\ (\sigma_{12}\sigma_2^2)^2 & 0 \end{pmatrix}\right]\right) \\ &= \frac{1}{2k^2}(\sigma_2^2)^2 \\ &= \frac{1}{2}\left(\frac{\sigma_2^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}\right)^2. \end{aligned} \tag{A.18}$$

Once again, the element of the Fisher information matrix for the variance of the second

variable mirrors the above exactly.

$$g_{44} = \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}\right)^2. \tag{A.19}$$

The covariance component will complete the diagonal elements of the Fisher information matrix.

$$\begin{aligned}
g_{55} &= \frac{1}{2}\left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\sigma_{12}}\right]^2\right) \\
&= \frac{1}{2}\left(tr\left[\frac{1}{k}\begin{pmatrix}\sigma_2^2 & -\sigma_{12}\\ -\sigma_{12} & \sigma_1^2\end{pmatrix}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\right]^2\right) \\
&= \frac{1}{2k^2}\left(tr\left[\begin{pmatrix}-\sigma_{12} & \sigma_2^2\\ \sigma_1^2 & -\sigma_{12}\end{pmatrix}\right]^2\right) \\
&= \frac{1}{2k^2}\left(tr\left[\begin{pmatrix}\sigma_1^2\sigma_2^2 + \sigma_{12}^2 & 2\sigma_2^2\sigma_{12}\\ 2\sigma_1^2\sigma_{12} & \sigma_1^2\sigma_2^2 + \sigma_{12}^2\end{pmatrix}\right]\right) \\
&= \frac{1}{2k^2}2(\sigma_1^2\sigma_2^2 + \sigma_{12}^2) \\
&= \frac{\sigma_1^2\sigma_2^2 + \sigma_{12}^2}{\left(\sigma_1^2\sigma_2^2 - \sigma_{12}^2\right)^2}.
\end{aligned} \tag{A.20}$$

The off-diagonal elements are only slightly more involved. However, because the terms in Equation (A.14) involve the partial derivatives, and because the mean vector and the covariance matrix have no overlapping terms, many of the off-diagonal elements vanish, specifically the ones that involve both a mean component and a variance component, i.e., $g_{ij} = 0$ for $i \in (1, 2)$ and $j \in (3, 4, 5)$. For the other off-diagonal components, we will employ all the conveniences of symmetry to complete the Fisher information matrix.

Turning our attention to the g_{12}, the element concerning the two means:

$$\begin{aligned}
g_{12} &= \frac{\partial\mu}{\partial\theta_1}^T\Sigma^{-1}\frac{\partial\mu}{\partial\theta_2} \\
&= \frac{1}{k}\begin{pmatrix}1 & 0\end{pmatrix}\begin{pmatrix}\sigma_2^2 & -\sigma_{12}\\ -\sigma_{12} & \sigma_1^2\end{pmatrix}\begin{pmatrix}0\\ 1\end{pmatrix} \\
&= -\frac{1}{k}\sigma_{12} \\
&= -\frac{\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} = g_{21}.
\end{aligned} \tag{A.21}$$

Next, we consider the elements of the Fisher information matrix involving both variances, g_{34}

$$\begin{aligned}
g_{34} &= \frac{1}{2}\left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\sigma_1^2}\Sigma^{-1}\frac{\partial\Sigma}{\partial\sigma_2^2}\right]\right) \\
&= \frac{1}{2k^2}\left(tr\left[\begin{pmatrix}\sigma_2^2 & -\sigma_{12}\\ -\sigma_{12} & \sigma_1^2\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}\sigma_2^2 & -\sigma_{12}\\ -\sigma_{12} & \sigma_1^2\end{pmatrix}\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}\right]\right) \\
&= \frac{1}{2k^2}\left(tr\left[\begin{pmatrix}\sigma_2^2 & 0\\ -\sigma_{12} & 0\end{pmatrix}\begin{pmatrix}0 & -\sigma_{12}\\ 0 & \sigma_1^2\end{pmatrix}\right]\right) \\
&= \frac{1}{2k^2}\left(tr\left[\begin{pmatrix}0 & -\sigma_2^2\sigma_{12}\\ 0 & \sigma_{12}^2\end{pmatrix}\right]\right) \\
&= \frac{1}{2}\left(\frac{\sigma_{12}}{\sigma_1^2\sigma_2^2-\sigma_{12}^2}\right)^2 = g_{43}.
\end{aligned} \tag{A.22}$$

The variance/covariance elements of the Fisher information matrix will all have similar structures. We calculate one of them below

$$\begin{aligned}
g_{35} &= \frac{1}{2}\left(tr\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\sigma_1^2}\Sigma^{-1}\frac{\partial\Sigma}{\partial\sigma_{12}}\right]\right) \\
&= \frac{1}{2k^2}\left(tr\left[\begin{pmatrix}\sigma_2^2 & -\sigma_{12}\\ -\sigma_{12} & \sigma_1^2\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}\sigma_2^2 & -\sigma_{12}\\ -\sigma_{12} & \sigma_1^2\end{pmatrix}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\right]\right) \\
&= \frac{1}{2k^2}\left(tr\left[\begin{pmatrix}\sigma_2^2 & 0\\ -\sigma_{12} & 0\end{pmatrix}\begin{pmatrix}-\sigma_{12} & \sigma_2^2\\ \sigma 1^2 & -\sigma_{12}\end{pmatrix}\right]\right) \\
&= \frac{1}{2k^2}\left(tr\left[\begin{pmatrix}-\sigma_{12}\sigma_2^2 & \sigma_{12}^2\\ \sigma_{12}^2 & -\sigma_{12}\sigma_2^2\end{pmatrix}\right]\right) \\
&= -\frac{\sigma_{12}\sigma_2^2}{\left(\sigma_1^2\sigma_2^2-\sigma_{12}^2\right)^2} = g_{53}.
\end{aligned} \tag{A.23}$$

Similarly, the element involving the second variance with the covariance is

$$g_{45} = g_{54} = -\frac{\sigma_{12}\sigma_1^2}{\left(\sigma_1^2\sigma_2^2-\sigma_{12}^2\right)^2}. \tag{A.24}$$

A.3 Euler-Lagrange on 2-Sphere

Before applying the results of Section 4.3 to Gaussian distributions, we demonstrate their application on the 2-sphere. The research of geodesics on the 2-sphere is extensive. However, most of it focuses on the path between 2 prescribed points, or the geodesic given an initial point and direction. Applying transversality conditions on a well studied surface reveals characteristics about the path as well as allows the exhibition of the techniques from Section 4.3 on a simple surface prior to using them on the manifold of Gaussians.

The examination of distributions on hyperspheres was the focus and motivation of spherical MDL. Choosing which distribution is most likely the data generating distribution is extremely important with regards to model selection. Slightly peripheral to this question would be the problem of finding which distribution on a submanifold of the hypersphere is closest to the chosen distribution. Calculus of variations is uniquely capable of answering these new questions.

What follows is the a detailed examination of geodesics on the 2-sphere, followed by examples with variable boundary conditions. Points on the 2-sphere are parameterized as (θ, ϕ), where θ is the azimuthal angle measured from the $x-axis$, ϕ is the polar angle measured from the $z-axis$. The arc length for a curve along the surface of a sphere is given as

$$L = \int \sqrt{\dot{\phi}^2 + \dot{\theta}^2 \sin^2 \phi} dx. \tag{A.25}$$

Once again, we simplify this to

$$K = \frac{1}{2}\left[\dot{\phi}^2 + \dot{\theta}^2 \sin^2 \phi\right] \tag{A.26}$$

with the factor of $\frac{1}{2}$ added to streamline future expressions without changing the extremal. Applying the Euler-Lagrange equations to each parameter, we end up with the following equations:

$$\frac{\partial K}{\partial \theta} - \frac{d}{dx}\frac{\partial K}{\partial \dot{\theta}} = 0, \text{and} \tag{A.27}$$

$$\frac{\partial K}{\partial \phi} - \frac{d}{dx}\frac{\partial K}{\partial \dot{\phi}} = 0. \tag{A.28}$$

The solution to this system of partial differential equations will define the shortest path between two points on a sphere. In solving this system, we look at the individual terms of Equations (A.27) and (A.28) separately. First, turning our attention to (A.27) consisting of two terms. The first of which is elementary, considering that K has no dependence on θ. Accordingly,

$$\frac{\partial K}{\partial \theta} = 0. \tag{A.29}$$

The second term of (A.27) becomes.

$$\begin{aligned} \frac{\partial K}{\partial \dot{\theta}} &= \dot{\theta} \sin^2 \phi \\ \frac{d}{dx}\frac{\partial K}{\partial \dot{\theta}} &= \ddot{\theta} \sin^2 \phi + \dot{\theta}\dot{\phi} \sin \phi \cos \phi \end{aligned} \tag{A.30}$$

where $\ddot{\theta} = \frac{d^2\theta}{dx^2}$.
With Equations (A.29) and (A.30), we can simplify (A.27),

$$0 - (\ddot{\theta} \sin^2 \phi + \dot{\theta}\dot{\phi} \sin \phi \cos \phi) = 0 \tag{A.31}$$

Dividing both sides by $\sin^2 \phi$ we end up with the final form of the Euler-Lagrange equation for ϕ

$$\ddot{\theta} = -2\dot{\phi}\dot{\theta} \cot \phi. \tag{A.32}$$

Now, turning our attention to Equation, we again solve for both terms individually, givin us (A.28)

$$\frac{\partial K}{\partial \phi} = \dot{\theta^2} \sin \phi \cos \phi, \tag{A.33}$$

$$\begin{aligned} \frac{\partial K}{\partial \dot{\phi}} &= \dot{\phi} \\ \frac{d}{dx}\frac{\partial K}{\partial \dot{\phi}} &= \ddot{\phi}. \end{aligned} \tag{A.34}$$

With Equations (A.33) and (A.34), we can simply (A.28),

$$\begin{aligned} \dot{\theta^2} \sin \phi \cos \phi - \ddot{\phi} &= 0 \\ 2\dot{\phi}\dot{\theta} \cot \phi + \ddot{\theta} &= 0 \end{aligned} \tag{A.35}$$

Now, our solution for the shortest path between two points must satisfy Equations (A.32) and (A.35). It will be useful to have this system of second order differential equations to be a system first order differential equation. To achieve this, we make the following substitutions.

$$
\begin{aligned}
y_1 &= \theta \\
y_2 &= y_1' = \dot{\theta} \\
y_2' &= \ddot{\theta} \\
y_3 &= \phi \\
y_4 &= y_3' = \dot{\phi} \\
y_4' &= \ddot{\phi}
\end{aligned}
\tag{A.36}
$$

With this, we can redefine Equations (A.32) and (A.35) as system of four first order differential equations.

$$
\begin{aligned}
y_1' &= y_2 \\
y_2' &= \sin(y_1)\cos(y_1)(y_3)^2 \\
y_3' &= y_4 \\
y_4' &= -2\cot(y_1)(y_2)(y_4)
\end{aligned}
\tag{A.37}
$$

Together with sometimes fixed boundary conditions, we can use the above to find the shortest path between two points on a sphere.

A.3.1 Transversality Conditions on the 2-Sphere

Exploring the transversality conditions on a 2-sphere will require our minimum path to satisfy both the Euler-Lagrange equations in (A.32) and (A.35), and the generic transversality conditions in (A.43).

A generic solution for an arbitrary surface offers little insight into how this shortest path behaves on the 2 sphere. Instead, the initial motivation will be provided by the following question:

What is the shortest path between the initial point $(\theta, \phi) = (\frac{3\pi}{2}, \frac{\pi}{6})$, to the line on the sphere defined by $\theta = \phi^2$?

We can choose to define this line as a level curve given by

$$
S(\theta, \phi) = \theta - \phi^2 = 0. \tag{A.38}
$$

According to Equation (4.46), the system of partial differential equations that must be satisfied are:

$$\begin{aligned} K - K_{\dot{Y}} \cdot \dot{Y} &= S_s \\ K_{\dot{\theta}} &= S_\theta \\ K_{\dot{\phi}} &= S_\phi. \end{aligned} \tag{A.39}$$

where K is defined in Equation (A.26) and the parameters are captured in the vector $Y = \begin{bmatrix} \phi & \theta \end{bmatrix}$.

We will look at each equation individually, starting with $K - K_{\dot{Y}} \cdot \dot{Y} = S_x$

K is defined in Equation (A.26) as

$$K = \frac{1}{2}\left[\dot{\phi}^2 + \dot{\theta}^2 \sin^2 \phi\right]$$

.

The next term is the product of two vectors.

$$\begin{aligned} \begin{bmatrix} K_{\dot{\theta}} & K_{\dot{\phi}} \end{bmatrix} \cdot \begin{bmatrix} \dot{\theta} & \dot{\phi} \end{bmatrix} &= \begin{bmatrix} \dot{\theta} \sin^2 \phi & \dot{\phi} \end{bmatrix} \cdot \begin{bmatrix} \dot{\theta} & \dot{\phi} \end{bmatrix} \\ &= \dot{\theta}^2 \sin^2 \phi + \dot{\phi}^2. \end{aligned}$$

Since the surface is independent of the path parameter, we have $S_X = 0$. This results in the first transversality condition

$$\begin{aligned} K - K_{\dot{Y}} \cdot \dot{Y} &= S_x \\ \frac{1}{2}\left[\dot{\phi}^2 + \dot{\theta}^2 \sin^2 \phi\right] - (\dot{\theta}^2 \sin^2 \phi + \dot{\phi}^2) &= 0 \\ 0 = \frac{1}{2}(\dot{\theta}^2 \sin^2 \phi + \dot{\phi}^2). \end{aligned} \tag{A.40}$$

The next transversality condition will be focus on the relationship between the dependence of the the extremal and the terminal surface on the parameter ϕ. Specifically

$$K_{\dot{\phi}} = S_\phi.$$

We have already defined $K_{\dot{\phi}} = \dot{\phi}$. In Equation (A.38), we define our terminal surface for this specific question. With that, we see

$$S_\phi = -2\phi$$

.

Giving us

$$\dot{\phi} = -2\phi. \tag{A.41}$$

Finally, we focus at the relationship between the dependence of the extremal and the terminal surface on the parameter θ. Specifically,

$$K_{\dot{\theta}} = S_{\theta}$$

.

We have already defined $K_{\dot{\theta}} = \dot{\theta}\sin^2\phi$. In Equation (A.38), we define our terminal surface for this specific question. with that, we see

$$S_{\theta} = 1$$

giving us

$$\dot{\theta} = \frac{1}{\sin^2\phi}. \tag{A.42}$$

Substituting Equations (A.41) and (A.42) into Equation (A.40), we get the transversality requirement that accounts for the geometrical relationship of our extremal with the terminal surface to be

$$\frac{1}{2}\left(\frac{1}{\sin^4\phi}\sin^2\phi + 4\dot{\phi^2}\right) = 0$$

or more simply

$$\sec^2\phi + 4\dot{\phi}^2 = 0. \tag{A.43}$$

Together, Equations (A.43) and (A.38) account for the geometry and location of our final boundary condition, respectively.

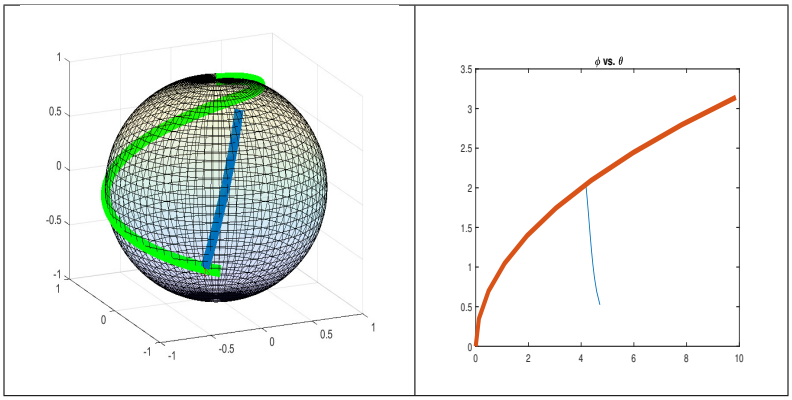


Figure A.1: Example of transversality on 2-sphere.
Two representation of the shortest path on a sphere between the initial point $(\frac{3\pi}{2}, \frac{\pi}{6})$ and the curve given by $\theta = \phi^2$.

Working first on spherical manifolds both offers a bridge between our research from Chapter 3 and future sections and provides proof of concept of the applications of transversality conditions on Riemannian manifolds. Regarding model selection, spherical MDL chooses the best model given sampled data. Here, an obvious extension would be to choose the best model given data and a user defined constraint. However, more importantly is applying the Fisher Information to transversality conditions in search for geodesics on Riemannian manifolds of different distributions.