entropy

MDPI

Article

Spherical Minimum Description Length

Trevor Herntier [1,*], Koffi Eddy Ihou [2], Anthony Smith [1], Anand Rangarajan [3] and Adrian Peter [1]

1 Department of Computer Engineering and Sciences, Florida Institute of Technology, Melbourne, FL 32940, USA; anthonysmith@fit.edu (A.S.); apeter@fit.edu (A.P.)
2 Concordia Institute for Information Systems Engineering, Concordia University, Montreal, QC H3G 1M8, Canada; eddyihou@gmail.com
3 Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611, USA; anand@cise.ufl.edu
* Correspondence: therntier2007@my.fit.edu; Tel.: +1-321-536-0317

Received: 8 June 2018; Accepted: 30 July 2018; Published: 3 August 2018

Abstract: 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. Current model selection criteria either ignore the shape of the parameter space or incorrectly penalize the complexity of the model, largely because typical Laplace approximation techniques yield inaccurate results for curved spaces. 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.

Keywords: model selection; MDL; information geometry; von Mises–Fisher distribution; Fisher–Bingham distribution; Fisher information; Laplace approximation; Jeffreys prior

1. Introduction

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 goodness 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 work, we revisit the geometry associated with the complexity measure for the Minimum Description Length (MDL) criterion [1,2]. We note that almost all of this previous development is restricted to unconstrained parameter spaces. In this work, 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 work, we mainly

focus on unit vector or hypersphere geometry constraints. 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 criterion, 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 [3–5], Bayesian information criterion (BIC) [6] and Minimum Description Length (MDL). The AIC criterion is given by

$$AIC = -2\log f(X;\hat{\theta}) + 2K \tag{1}$$

and BIC

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

where $X = \{x_i\}_{i=1}^{N}$ 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) and (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), 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 [7]).

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 [1,8,9] with the introduction of the Minimum Description Length principle

$$MDL = -\log f(X;\hat{\theta}) + \frac{K}{2}\log\left(\frac{N}{2\pi}\right) + \log\int\sqrt{\det I(\theta)}d\theta, \tag{3}$$

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 [10] 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 [10].

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 [11] 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 [12] 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 the paper, 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.

The remainder of this article is organized as follows: Section 2 covers the works most relevant to the present development, including some historical reflection on various model selection criteria. With an ambitious goal of making the article more self-contained, we briefly recap the background on a Bayesian perspective to model selection in Section 3, paying specific attention to geometric motivations behind the use of the Fisher information matrix. Section 4 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 5. 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 6, 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. We conclude in Section 7 with a summary of the developments and future extensions.

2. Related Works

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

$$MDL = -\log f(X;\theta) + \frac{K}{2}\log\left(\frac{N}{2\pi}\right) \tag{4}$$

and later evolved to become the three part code seen in Equation (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 (3) penalizes a model for geometric complexity by incorporating the Riemannian volume [13] of the parameter manifold. MDL deviates from AIC [3,4] and BIC [6] 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 [14], 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

$$p(X) = \frac{f(X|\hat{\theta}_X)}{\int f(Y|\hat{\theta}_Y)dY}, \tag{5}$$

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 (5)

$$SC = -\log f(X|\hat{\theta}_X) + \log \int f(Y|\hat{\theta}_Y)dY \tag{6}$$

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* [10]. Several solutions have been proposed to fix the infinity problem [15,16], 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 [11]. A brief proof of this formula will be provided in Section 4. 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 5.

Prior to Rissanen's MDL, the Minimum Message Length (MML) was introduced in [17]. 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 [18]. 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.

Our model selection criterion is specifically suited for distributions that have parameters residing on a spherical manifold. In [19,20], it was shown that the parameters for histogram density estimation can appropriately be placed on the hypersphere. In [21,22], 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 [23], 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 [24], 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, the work in reference [25] offers efficient numerical ways to estimate the value of this normalizing constant.

As anecdotal empirical evidence of our theoretical development of spherical MDL, Section 6 evaluates its use for histogram optimal bin width selection. The authors in [26] 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 formulation of 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 [27] 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 [28], 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.

3. Bayesian Approach to Model Selection

3.1. Comparing Models

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

$$l(\theta; X) = \prod_{i=1}^{N} f(x_i;\theta). \tag{7}$$

As in [11], 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 6, 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

$$\Pr(A|X) = \frac{\Pr(A)}{\Pr(X)} \int_{\mathbb{S}^{K-1}} l(\theta; X)\pi(\theta)d\theta. \tag{8}$$

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

$$\mathcal{I}(X) = \int_{\mathbb{S}^{K-1}} l(\theta; X)\pi(\theta)d\theta \tag{9}$$

over all valid models.

3.2. An Inappropriate Prior

Before evaluating Equation (9), we need to define a prior probability in the parameter space. While a uniform prior seems to be a logical choice [29]—following Laplace's principle of insufficient reason [30]—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

$$p(\theta) = 1,\ \theta \in [0, 1]. \tag{10}$$

Now assume a second parametrization of the parameters, ψ along with a monotonic transformation $\theta \to \psi$, 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 (10) is expressed as

$$p(\psi) = p(\theta)\left|\frac{\partial r^{-1}(\psi)}{\partial \psi}\right| \neq 1 \tag{11}$$

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.

3.3. Geometry of Probabilistic Models

Applying geometrical constructs to statistical models is not a new idea. Rao [31] and Jeffreys [32] 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 [33,34]. 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 [35].

The manifold associated with a family of models is populated with many distributions. Let a sample set $X = \{x_i\}_{i=1}^N$ 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]:

$$D(f\|g) = \int f(x) \ln\left(\frac{f(x)}{g(x)}\right) dx. \tag{12}$$

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 [11] in Equation (12) results in

$$D(f\|g) \approx \frac{1}{2}\Delta\theta^T I(\theta)\Delta\theta, \tag{13}$$

where $\Delta\theta = \theta_f - \theta_g$ and $I(\theta)$ 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 [11]. 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.

3.4. 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

$$I_{i,j}(\theta) = \int f(x;\theta) \frac{\partial}{\partial\theta_i} \log f(x;\theta) \frac{\partial}{\partial\theta_j} \log f(x;\theta) dx, \tag{14}$$

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, provided certain regularity conditions are satisfied. Firstly, we can compute the Fisher information matrix from the expectation of the Hessian of the log likelihood. Specifically,

$$I_{i,j}(\theta) = -\mathcal{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log f(x;\theta)\right] = -\mathcal{E}\left[H\right], \tag{15}$$

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

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

$$I(\theta) = \mathrm{Var}(S_f(x;\theta)), \tag{16}$$

where

$$S_f(x;\theta) = \nabla \log f(x;\theta). \tag{17}$$

The Fisher information matrix can be used to define the Cramer–Rao lower bound of unbiased maximum likelihood estimators, determining the optimal sample size in a statistical experiment. 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 3.2. 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 paper.

3.5. 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

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

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

$$J_r(\theta)J_{r^{-1}}(\psi) = \mathtt{I}_K, \tag{18}$$

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

The non-informative prior, which is tantamount to Jeffreys prior [32] 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

$$g(x;\psi) = f(x;r^{-1}(\psi))\,|\det(J_{r^{-1}}(\psi))| \tag{19}$$

and a new score function

$$S_g(x;\psi) = (J_{r^{-1}}(\psi))^T S_f(x;r^{-1}(\psi)). \tag{20}$$

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

$$\begin{aligned}\tilde{I}(\psi) &= \operatorname{cov}(S_g(x;\psi)) \\ &= \operatorname{cov}\left(J_{r^{-1}}(\psi)^T S_f(x;r^{-1}(\psi))\right) \\ &= (J_{r^{-1}}(\psi))^T I(r^{-1}(\psi))J_{r^{-1}}(\psi).\end{aligned} \tag{21}$$

The Jeffreys prior for the ψ parametrization is

$$\begin{aligned}\tilde{\pi}(\psi) &\propto \sqrt{\det(\tilde{I}(\psi))} \\ &= \sqrt{\det\left[(J_{r^{-1}}(\psi))^T I(r^{-1}(\psi))J_{r^{-1}}(\psi)\right]} \\ &= \det(J_{r^{-1}}(\psi))\det\sqrt{I_\theta(r^{-1}(\psi))} \\ &= \det(J_{r^{-1}}(\psi))\pi(r^{-1}(\psi)).\end{aligned} \tag{22}$$

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

$$\tilde{\pi}(\psi)d\psi = \pi(\theta)d\theta. \tag{23}$$

Therefore, the Jeffreys prior remains unchanged under the reparametrization $\psi = r(\theta)$ and the value of Equation (9) 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 (9). Specifically, Jeffreys prior is

$$\pi(\theta) = \frac{\sqrt{\det(I(\theta))}}{\int\sqrt{\det(I(\theta))}d\theta}, \tag{24}$$

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 (9) 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 (13) 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,

$$V_{\mathcal{M}} = \int \sqrt{\det\left(I(\theta)\right)}d\theta \tag{25}$$

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 (24) represents these volumes on the parameter space. More specifically,

$$V_d = \sqrt{\det\left(I(\theta)\right)}d\theta \tag{26}$$

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 [33] for a more detailed discussion on the appropriateness of Jeffreys prior for spherical distributions.

4. Asymptotic MDL in $\mathbb{R}^K$

In [11], 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 (9). Again, given a set of data $X = \{x_i\}_{i=1}^N$, the likelihood of any given model with density $f(x;\theta)$ and a K-dimensional parameter vector is given in Equation (7) and its average negative log-likelihood is given by

$$L(\theta) = -\frac{1}{N}\log(l(\theta;X)). \tag{27}$$

Our goal is still to evaluate Equation (9):

$$\mathcal{I}(X) = \int \exp\left\{-NL(\theta)\right\}\pi(\theta)d\theta. \tag{28}$$

To evaluate the integral in Equation (28), we employ standard Laplace approximation techniques [37]. 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

$$\mathcal{I}(X) \approx \exp\left\{-NL(\hat{\theta})\right\}\pi(\hat{\theta})\int \exp\left\{-\frac{N}{2}\left(\theta-\hat{\theta}\right)^T H_I\left(\theta-\hat{\theta}\right)\right\}d\theta, \tag{29}$$

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,

$$RZR = \exp\left\{-NL(\hat{\theta})\right\} \pi(\hat{\theta}) \left(\frac{\left(\frac{2\pi}{N}\right)^K}{\det(H_I)} \right)^{\frac{1}{2}}, \tag{30}$$

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 evaluation of the prior from Equation (24) at the maximum likelihood estimate, the final form of MDL is found by taking the negative log of the razor and is given by

$$MDL = -\log(RZR) = -\log f(X;\hat{\theta}) + \frac{K}{2}\log\left(\frac{N}{2\pi}\right) + \log \int \sqrt{\det\left(I(\theta)\right)} d\theta. \tag{31}$$

The first term in Equation (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 (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 [24] suggested that this approximation needs modification in order to be used on curved spaces. Thus, if Equation (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.

5. Spherical MDL

5.1. 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 (9) 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

$$M(\theta, \lambda) = L(\theta) + \frac{\lambda}{N}(\theta^T\theta - 1) \tag{32}$$

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 (9) as

$$\mathcal{I}(X) = \int_{\mathbb{S}^{K-1}} \exp\left\{-NM(\theta, \hat{\lambda})\right\} \pi(\theta) d\theta \tag{33}$$

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

To evaluate Equation (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

$$M(\theta) = M(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T H (\theta - \hat{\theta}) + \mathcal{O}(\|\theta - \hat{\theta}\|_2^3), \tag{34}$$

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 (9). 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 (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 [38].

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

$$O(\theta_R) \equiv L(\theta_R, \theta_K(\theta_R)), \tag{35}$$

where the Kth parameter θ_K has been explicitly written out as a function of the remaining parameters $\theta_R \equiv \{\theta_1, \theta_2, \ldots, \theta_{K-1}\}$. 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

$$\frac{\partial O}{\partial \theta_k} = \frac{\partial L}{\partial \theta_k} + \frac{\partial \theta_K}{\partial \theta_k}\frac{\partial L}{\partial \theta_K}, \tag{36}$$

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

$$\frac{\partial^2 O}{\partial \theta_k \partial \theta_l} = \frac{\partial^2 L}{\partial \theta_k \partial \theta_l} + \frac{\partial \theta_K}{\partial \theta_l}\frac{\partial^2 L}{\partial \theta_k \partial \theta_K} + \frac{\partial \theta_K}{\partial \theta_k}\frac{\partial^2 L}{\partial \theta_l \partial \theta_K} + \frac{\partial \theta_K}{\partial \theta_k}\frac{\partial \theta_K}{\partial \theta_l}\frac{\partial^2 L}{\partial \theta_K^2} + \frac{\partial^2 \theta_K}{\partial \theta_k \partial \theta_l}\frac{\partial L}{\partial \theta_K}. \tag{37}$$

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

$$\begin{aligned}\sum_{kl} u_k u_l \frac{\partial^2 O}{\partial \theta_k \partial \theta_l} = \sum_{kl} u_k u_l \frac{\partial^2 L}{\partial \theta_k \partial \theta_l} + 2\left(\sum_{k=1}^{K-1} u_k \frac{\partial \theta_K}{\partial \theta_k}\right)\left(\sum_{l=1}^{K-1} u_l \frac{\partial^2 L}{\partial \theta_l \partial \theta_K}\right) \\ + \left(\sum_{k=1}^{K-1} u_k \frac{\partial \theta_K}{\partial \theta_k}\right)^2 \frac{\partial^2 L}{\partial \theta_K^2} + \frac{\partial L}{\partial \theta_K}\sum_{kl} u_k u_l \frac{\partial^2 \theta_K}{\partial \theta_k \partial \theta_l}.\end{aligned} \tag{38}$$

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

$$M(\theta, \hat{\lambda}) = L(\theta) + \frac{\hat{\lambda}}{N}\left(\sum_{k=1}^{K}\theta_k^2 - 1\right), \tag{39}$$

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

$$\frac{\partial M}{\partial \theta_k} = \frac{\partial L}{\partial \theta_k} + 2\theta_k\frac{\hat{\lambda}}{N} = 0, \tag{40}$$

with the optimal value of the Lagrange parameter

$$\hat{\lambda} = -\frac{N}{2}\sum_{k=1}^{K}\hat{\theta}_k\frac{\partial L}{\partial \theta_k}\bigg|_{\theta=\hat{\theta}} \tag{41}$$

obtained by multiplying Equation (40) by θ_k, summing over all $k \in \{1,\ \ldots,\ K\}$ and enforcing the constraint $\theta^T\theta = 1$. Furthermore, Equation (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

$$2\theta_k + 2\theta_K\frac{\partial \theta_K}{\partial \theta_k} = 0. \tag{42}$$

This relation holds for all θ on the hypersphere unlike Equation (40) which is valid only at the MLE. Taking second derivatives, we obtain

$$2\delta_{kl} + 2\frac{\partial \theta_K}{\partial \theta_k}\frac{\partial \theta_K}{\partial \theta_l} + 2\theta_K\frac{\partial^2 \theta_K}{\partial \theta_k \partial \theta_l} = 0. \tag{43}$$

We now have all the ingredients necessary to evaluate Equation (38) for the constrained problem. Substituting Equations (40), (42) and (43) into Equation (38), we get

$$\begin{aligned}\sum_{kl}u_k u_l\frac{\partial^2 O}{\partial \theta_k \partial \theta_l} = \sum_{kl}u_k u_l\frac{\partial^2 L}{\partial \theta_k \partial \theta_l} - \frac{2}{\theta_K}\left(\sum_{k=1}^{K-1}u_k\theta_k\right)\left(\sum_{l=1}^{K-1}u_l\frac{\partial^2 L}{\partial \theta_l \partial \theta_K}\right)\\ + \frac{1}{\theta_K^2}\left(\sum_{k=1}^{K-1}u_k\theta_k\right)^2\frac{\partial^2 L}{\partial \theta_K^2} + \frac{2\hat{\lambda}}{N}\sum_{kl}u_k u_l\left(\delta_{kl} + \frac{\theta_k\theta_l}{\theta_K^2}\right).\end{aligned} \tag{44}$$

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 (39). We get

$$\begin{aligned}\sum_{kl}u_k u_l\frac{\partial^2 O}{\partial \theta_k \partial \theta_l} &= \sum_{kl}u_k u_l\left(\frac{\partial^2 L}{\partial \theta_k \partial \theta_l} + \frac{2\hat{\lambda}}{N}\delta_{kl}\right) - \frac{2}{\theta_K}\left(\sum_{k=1}^{K-1}u_k\theta_k\right)\left(\sum_{l=1}^{K-1}u_l\frac{\partial^2 L}{\partial \theta_l \partial \theta_K}\right)\\ &+ \frac{1}{\theta_K^2}\left(\sum_{k=1}^{K-1}u_k\theta_k\right)^2\left(\frac{\partial^2 L}{\partial \theta_K^2} + \frac{2\hat{\lambda}}{N}\right)\\ &= \sum_{kl}u_k u_l\frac{\partial^2 M}{\partial \theta_k \partial \theta_l} - \frac{2}{\theta_K}\left(\sum_{k=1}^{K-1}u_k\theta_k\right)\left(\sum_{l=1}^{K-1}u_l\frac{\partial^2 M}{\partial \theta_l \partial \theta_K}\right) + \frac{1}{\theta_K^2}\left(\sum_{k=1}^{K-1}u_k\theta_k\right)^2\frac{\partial^2 M}{\partial \theta_K^2},\end{aligned} \tag{45}$$

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 $v \equiv \left[u_1,\ u_2,\ \ldots,\ u_{(K-1)}, -\frac{1}{\theta_K}\sum_{k=1}^{K-1}u_k\theta_k\right]^T$. 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 (39), we get

$$\sum_{kl} u_k u_l \frac{\partial^2 O}{\partial \theta_k \partial \theta_l}\bigg|_{\theta=\hat{\theta}} = \sum_{k=1}^{K}\sum_{l=1}^{K} v_k v_l \frac{\partial^2 M}{\partial \theta_k \partial \theta_l}\bigg|_{\theta=\hat{\theta}}, \tag{46}$$

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

$$\sum_{k=1}^{K} \theta_k d\theta_k = 0, \tag{47}$$

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 $[\Delta\theta_1,\ \Delta\theta_2,\ \ldots,\ \Delta\theta_K]^T$ is orthogonal to the gradient of the constraints, equal to $[2\theta_1,\ 2\theta_2,\ \ldots,\ 2\theta_K]^T$. 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 (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 (34), where the (k, l) element of the Hessian is

$$H_{kl} = \frac{\partial^2 M}{\partial \theta_k \partial \theta_l}\bigg|_{\theta=\hat{\theta}} = \frac{\partial^2 L}{\partial \theta_k \partial \theta_l}\bigg|_{\theta=\hat{\theta}} + \frac{2\hat{\lambda}}{N}\delta_{kl}, \tag{48}$$

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.,

$$\sum_{k=1}^{K} 2(\theta_k - \hat{\theta}_k)\hat{\theta}_k = 0, \tag{49}$$

then the quadratic form $(\theta - \hat{\theta})^T H (\theta - \hat{\theta})$ is equivalent to the independent one as shown above in Equation (46). 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 (49), 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 (33) can be found. Since Equation (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

$$\pi(\theta) = \pi(\hat{\theta}) + (\theta - \hat{\theta})^T \nabla \pi(\hat{\theta}) + \mathcal{O}(\|\theta - \hat{\theta}\|_2^2). \tag{50}$$

We now rewrite Equation (33) as a product of Equations (34) and (50) to get

$$\begin{aligned}\mathcal{I}(X) &= \int_{\mathbb{S}^{K-1}} \exp\left\{-NM(\theta,\hat{\lambda}\right\}\pi(\theta)d\theta \\ &\approx \int_{\mathbb{S}^{K-1}} \exp\left[-NM(\hat{\theta}) - \frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right]\left[\pi(\hat{\theta}) + (\theta-\hat{\theta})^T\nabla\pi(\hat{\theta}) + \dots\right]d\theta \\ &\approx \exp\left\{-N(M(\hat{\theta})\right\}\pi(\hat{\theta})\int_{\mathbb{S}^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\}d\theta.\end{aligned} \tag{51}$$

Here, we have used the fact that $\theta \to \hat{\theta}$ as $N \to \infty$ on the order of $N^{-\frac{1}{2}}$ [39] which makes

$$\int_{\mathbb{S}^{K-1}} (\theta-\hat{\theta})^T\nabla\pi(\hat{\theta})d\theta = 0. \tag{52}$$

The evaluation of the integral of the quadratic term in Equation (51), 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 Fisher information matrix, as is the case in the histogram, we can efficiently evaluate this integral. This assertion will be expanded upon in Section 6.

We continue solving Equation (51), and using the Jeffreys prior as the appropriate prior, we get

$$\mathcal{I}(X) = \exp\left\{-NM(\hat{\theta})\right\}\int_{\mathbb{S}^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\}d\theta\frac{\sqrt{\det(I(\hat{\theta}))}}{\int\sqrt{\det(I(\theta))}d\theta}. \tag{53}$$

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 (53). Hence,

$$\begin{aligned}MDL_{\mathbb{S}^{K-1}} &= NM(\hat{\theta}) - \log\sqrt{\det(I(\hat{\theta}))} + \log\int\sqrt{\det(I(\theta))}d\theta \\ &\quad - \log\int_{\mathbb{S}^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\}d\theta \\ &= N\left[L + \hat{\lambda}(\hat{\theta}^T\hat{\theta}-1)\right] - \log\sqrt{\det(I(\hat{\theta}))} + \log\int\sqrt{\det(I(\theta))}d\theta \\ &\quad - \log\int_{\mathbb{S}^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\}d\theta \\ &= -\log l(\hat{\theta}) - \log\sqrt{\det(I(\hat{\theta}))} + \log\int\sqrt{\det(I(\theta))}d\theta \\ &\quad - \log\int_{\mathbb{S}^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\}d\theta.\end{aligned} \tag{54}$$

The first term in Equation (54) is the log-likelihood and rewards a model for goodness of fit. The last three terms represent the parametric complexity penalty in spherical MDL:

$$C = -\log\sqrt{\det(I(\hat{\theta}))} + \log\int\sqrt{\det(I(\theta))}d\theta - \log\int_{\mathbb{S}^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\}d\theta. \tag{55}$$

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 (55) 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 (55) can be represented as a ratio of two terms

$$C = -\log\left[\frac{\sqrt{\det(I(\hat{\theta}))}\int_{\mathbb{S}^{K-1}}\exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\}d\theta}{\int\sqrt{\det(I(\theta))}d\theta}\right]. \tag{56}$$

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

$$C_{\mathbb{R}^K} = \frac{K}{2}\log\left(\frac{N}{2\pi}\right) + \log\int\sqrt{\det\left(I(\theta)\right)}d\theta. \tag{57}$$

5.2. 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 [40]. 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

$$V_{\mathcal{M}} = \begin{cases} K\pi^{K/2}, & K \text{ even}, \\ 2^K\pi^{\frac{K-1}{2}}\frac{(\frac{K-1}{2})!}{(K-1)!}, & K \text{ odd}. \end{cases} \tag{58}$$

Figure 1 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 [41]. Since spherical MDL represents an asymptotic approximation of NML, it is more suitable when applied to large sample data. We refer the reader to [42,43] for more details regarding the issue of misspecification for high-dimensional models.

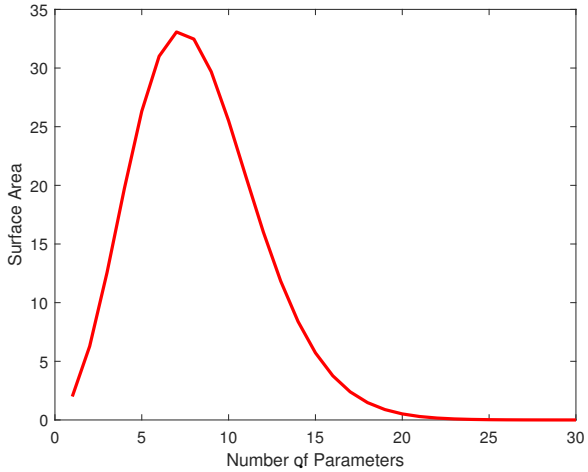

Figure 1. Riemannian Volume of Hypersphere. The surface area of a hyperspherical manifold plotted against the cardinality of the parameters. Interestingly, the surface area grows to a maximum at seven dimensions and then monotonically decreases. Accordingly, a seven-dimensional model family requires a relatively large ellipsoid around the MLE in order to avoid excessive penalties for complexity.

6. 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 5 to reach the final criterion given in Equation (54).

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 [28,44]. Here, we consider histograms with equal bin width, also known as regular histograms. When doing so, we can use Equation (54) 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.

6.1. 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.,

$$f(x) = \begin{cases} c_i, & \text{if } x \text{ is in interval } i, \\ 0, & \text{otherwise.} \end{cases} \tag{59}$$

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

$$l(c) = \prod_{i=1}^{K} c_i^{v_i}, \tag{60}$$

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

$$L(c) = -\frac{1}{N}\sum_{i=1}^{K} v_i \log c_i. \tag{61}$$

As in [20], 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)

$$L(u) = -\frac{1}{N}\sum_{i=1}^{K} 2v_i \log u_i. \tag{62}$$

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

$$\sum_{i=1}^{K} u_i^2 = h^{-1}, \tag{63}$$

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

$$\begin{aligned} M(u,\lambda) &= L(u) + \frac{\lambda}{N}(\sum_{i=1}^{K} u_i^2 - \frac{K}{R}) \\ &= -\frac{1}{N}\left[\sum_{i=1}^{K} 2v_i \log u_i - \lambda(\sum_{i=1}^{K} u_i^2 - \frac{K}{R})\right]. \end{aligned} \tag{64}$$

Minimizing $M(u,\lambda)$ with respect to 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 (48) for the histogram density log-likelihood function Equation (64). Starting with Equation (62),

$$L(u) = -\frac{1}{N}\sum_{i=1}^{K} 2v_i \log u_i, \tag{65}$$

the resulting gradient is

$$\frac{\partial L}{\partial u_k} = -\frac{1}{N}\frac{2v_k}{u_k} \tag{66}$$

and the Hessian

$$\frac{\partial^2 L}{\partial u_k^2} = \frac{1}{N}\frac{2v_k}{u_k^2} \tag{67}$$

with all other mixed partials equal to zero.

Next, we evaluate the Hessian of the average negative log likelihood at the MLE. Using Equation (48),

$$\begin{aligned} H_{kk} &= \frac{2v_k}{\frac{Kv_k}{R}} + 2\frac{R}{K} \\ &= 4\frac{R}{K}. \end{aligned} \tag{68}$$

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 (48). To evaluate the Fisher information, we take the expectation of Equation (68) to get

$$\begin{aligned} I_{kk}(\theta) &= \int H_{kk} f(x) dx \\ &= H_{kk} \int f(x) dx \\ &= 4\frac{R}{K}, \end{aligned} \tag{69}$$

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 (69) 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

$$V_H = 2^K V_{\mathcal{M}}, \tag{70}$$

where $V_{\mathcal{M}}$ is the hypersphere volume given in Equation (58). 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 (70) 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 (54). 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

$$\begin{aligned} MDL_{sphere} &= -\sum_{i=1}^{K} 2v_i \log u_i - \log \sqrt{\det(I(\hat{\theta}))} + \log V_H \\ &\quad - \log \int_{S^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\} d\theta \\ &= -\sum_{i=1}^{K} 2v_i \log u_i - \frac{K}{2}\log\left(4\frac{R}{K}\right) + \log V_H \\ &\quad - \log \int_{S^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\} d\theta. \end{aligned} \tag{71}$$

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

$$\begin{aligned} Q(\hat{\theta}) &= \int_{S^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\} d\theta \\ &= \int_{S^{K-1}} \exp\left\{-2Nh(\theta-\hat{\theta})^T(\theta-\hat{\theta})\right\} d\theta \\ &= \int_{S^{K-1}} \exp\left\{-2Nh(\theta^T\theta - 2\theta^T\hat{\theta} + \hat{\theta}^T\hat{\theta})\right\} d\theta. \end{aligned} \tag{72}$$

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

$$\begin{aligned} Q(\hat{\theta}) &= \int_{S^{K-1}} \exp\left\{-2Nh(h^{-1} - 2\theta^T\hat{\theta} + h^{-1})\right\} d\theta \\ &= \int_{S^{K-1}} \exp\left\{-4N + 4Nh\theta^T\hat{\theta})\right\} d\theta \\ &= \exp\{-4N\} \int_{S^{K-1}} \exp\left\{4Nh\theta^T\hat{\theta})\right\} d\theta. \end{aligned} \tag{73}$$

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:

$$\frac{x_i}{\sqrt{h}} = \theta_i, \frac{\hat{x}_i}{\sqrt{h}} = \hat{\theta}_i \text{ and } d\theta = \frac{dx}{\sqrt{h}}. \tag{74}$$

Once again, to more clearly show that complexity increases as the number of parameters increases, we make the substitution $h = \frac{R}{K}$. Equation (73) becomes (after a minor abuse of notation)

$$\begin{aligned} Q(\hat{x}) &= \exp(-4N) \int_{S^{K-1}} \exp\left\{4Nh\frac{x^T}{\sqrt{h}}\frac{\hat{x}^T}{\sqrt{h}}\right\} dx \frac{1}{\sqrt{h}^K} \\ &= \left(\frac{K}{R}\right)^{\frac{K}{2}} \exp(-4N) \int_{S^{K-1}} \exp\left\{4Nx^T\hat{x}\right\} dx. \end{aligned} \tag{75}$$

The integral in Equation (75) is now in the form of a von Mises distribution. In general, the von Mises distribution is

$$f_K(x, u, \kappa) = \frac{\exp(\kappa u^T x)}{C_K(\kappa)}, \tag{76}$$

where $\kappa \geq 0$, $\|u\| = 1$ and x is random unit vector. The distribution in Equation (76) must integrate to one, so

$$\int_{S^{K-1}} \exp(\kappa u^T x) dx = C_K(\kappa), \tag{77}$$

where

$$C_K(\kappa) = \left(\frac{2\pi}{\kappa}\right)^{\frac{K}{2}} \kappa\, I_{\frac{K}{2}-1}(\kappa) \tag{78}$$

and $I_\zeta(\kappa)$ is the modified Bessel function of order ζ [45]. The right side of Equation (78) will be used to determine the value of the integral in Equation (75).

By comparing the integral in Equation (75) to Equation (77), we can see that $\kappa = 4N$. With this, Equation (75) then becomes

$$Q(\hat{x}) = \left(\frac{K}{R}\right)^{\frac{K}{2}} \exp(-4N) \left(\frac{\pi}{2N}\right)^{\frac{K}{2}} 4N\, I_{\frac{K}{2}-1}(4N), \tag{79}$$

which is independent of $\hat{x}$. Substituting this into Equation (71), we obtain that spherical MDL will choose a histogram with the number of bins that minimizes

$$\begin{aligned} MDL_{sphere} &= -\sum_{i=1}^{K} 2v_i \log u_i - \frac{K}{2}\log\left(4\frac{R}{K}\right) + \log V_H \\ &\quad - \log \int_{S^{K-1}} \exp\left\{-\frac{N}{2}(\theta-\hat{\theta})^T H(\theta-\hat{\theta})\right\} d\theta \\ &= -\sum_{i=1}^{K} 2v_i \log u_i - \frac{K}{2}\log\left(4\frac{R}{K}\right) + \log V_H \\ &\quad - \log\left[\left(\frac{K}{R}\right)^{\frac{K}{2}} \exp\left(-4N\right)\left(\frac{\pi}{2N}\right)^{\frac{K}{2}} 4N\, I_{\frac{K}{2}-1}(4N)\right], \end{aligned} \tag{80}$$

where V_H is defined in Equation (70). 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 (80) simplifies to

$$\begin{aligned} MDL_{sphere} &= -\sum_{i=1}^{K} 2v_i \log u_i + \frac{K}{2}\log\left(\frac{K}{4R}\right) + \log V_H \\ &\quad + \frac{K}{2}\log\left(\frac{R}{K}\right) + \frac{K}{2}\log\left(\frac{2N}{\pi}\right) - \log\left(I_{\frac{K}{2}-1}(4N)\right) \\ &= -\sum_{i=1}^{K} 2v_i \log u_i + \log(V_H) + \frac{K}{2}\log\left(\frac{N}{2\pi}\right) - \log\left(I_{\frac{K}{2}-1}(4N)\right). \end{aligned} \tag{81}$$

Spherical MDL closely tracks ordinary MDL when it comes to asymptotics. The modified Bessel function in Equation (81) can be considerably simplified as $N \to \infty$:

$$I_{\frac{(K-1)}{2}}(4N) \approx \frac{\exp\{4N\}}{\sqrt{2\pi}}\left[\frac{1}{(4N)^{\frac{1}{2}}} + \frac{(4K-3-K^2)}{8(4N)^{\frac{3}{2}}} + \mathcal{O}(\frac{1}{(4N)^{\frac{5}{2}}})\right]. \tag{82}$$

Since the leading term in Equation (82) is independent of K, we obtain that spherical MDL and ordinary MDL converge to the same complexity (after ignoring terms independent of K) as $N \to \infty$.

6.2. 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 [46,47]. 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 2. 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 (4), Balasubramanian's asymptotic MDL from Equation (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 1.

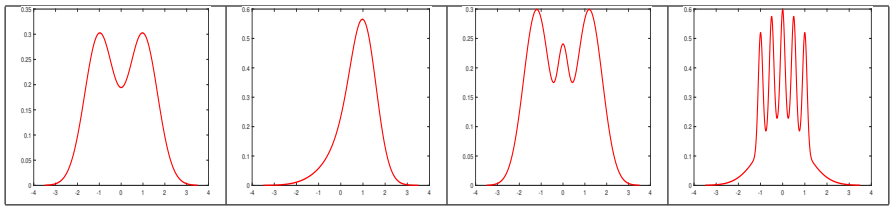

Figure 2. Four different densities selected for varying characteristics. Bimodal (left), skewed (center left), trimodal (center right) and claw (right).

Table 1. 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 (MDL2) and asymptotic MDL (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.

	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

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, 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.

7. 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. Though this present effort focused on spherical parameter manifolds, our geometric approach to model selection can be generalized to other curved domains.

Author Contributions: Conceptualization: T. Herntier, A. Rangarajan, A. Peter; Investigation: T. Herntier, K. Ihou, A. Rangarajan, A. Peter; Formal Analysis: T. Herntier, K. Ihou, A. Rangarajan, A. Peter; Methodology: T. Herntier, A. Smith, A. Rangarajan, A. Peter; Validation and Software: T. Herntier, A. Smith, A. Peter; Original

Draft Preparation: T. Herntier; Writing (Review and Editing): T. Herntier, K. Ihou, A. Smith, A. Rangarajan, A. Peter; Funding Acquisition: A. Smith, A. Rangarajan, A. Peter

Funding: The authors acknowledge support from National Science Foundation (NSF) Grant Nos. 1263011, 1560345, and IIS-1743050. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rissanen, J. Modeling by shortest data description. *Automatica* **1978**, *14*, 465–471. [CrossRef]
2. Rissanen, J. A universal prior for integers and estimation by minimum description length. *Ann. Stat.* **1983**, *11*, 416–431. [CrossRef]
3. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*; Parzen, E., Tanabe, K., Kitagawa, G., Eds.; Springer: New York, NY, USA, 1998; pp. 199–213.
4. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
5. Hodges, J.S.; Sargent, D.J. Counting degrees of freedom in hierarchical and other richly parameterised models. *Biometrika* **1988**, *88*, 367–379. [CrossRef]
6. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
7. Pan, W. Bootstrapping likelihood for model selection with small samples. *J. Comput. Gr. Stat.* **1999**, *8*, 687–698.
8. Rissanen, J. Stochastic complexity. *J. R. Stat. Soc.* **1987**, *49*, 223–239.
9. Rissanen, J. Fisher information and stochastic complexity. *IEEE Trans. Inf. Theory* **1996**, *42*, 40–47. [CrossRef]
10. Grünwald, P. A tutorial introduction to the minimum description length principle. In *Advances in Minimum Description Length: Theory and Applications*; Grünwald, P., Myung, I., Pitt, M., Eds.; The MIT Press: Cambridge, MA, USA, 2005; pp. 55–58.
11. Balasubramanian, V. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Comput.* **1997**, *9*, 349–368. [CrossRef]
12. Kent, J.T. The Fisher–Bingham distribution on the sphere. *J. R. Stat. Soc.* **1982**, *44*, 71–80.
13. Boothby, W.M. *An Introduction to Differentiable Manifolds and Riemannian Geometry*; Academic Press: San Diego, CA, USA, 2002.
14. Barron, A.; Rissanen, J.; Yu, B. The minimum description length principle in coding and modeling. *IEEE Trans. Inf. Theory* **1998**, *44*, 2743–2760. [CrossRef]
15. Rissanen, J. MDL denoising. *IEEE Trans. Inf. Theory* **2000**, *46*, 2537–2543. [CrossRef]
16. Wallace, C.S.; Dowe, D.L. Refinements of MDL and MML coding. *Comput. J.* **1999**, *42*, 330–337. [CrossRef]
17. Wallace, C.S.; Boulton, D.M. An information measure for classification. *Comput. J.* **1968**, *11*, 185–194. [CrossRef]
18. Wallace, C.S. *Statistical and Inductive Inference by Minimum Message Length*; Information Science and Statistics; Springer: New York, NY, USA, 2005.
19. Lebanon, G. Metric learning for text documents. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 497–508. [CrossRef] [PubMed]
20. Srivastava, A.; Jermyn, I.; Joshi, S. Riemannian analysis of probability density functions with applications in vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA , 17–22 June 2007; IEEE Press: Piscataway, NJ, USA, 2007; pp. 1–8.
21. Peter, A.; Rangarajan, A. Information geometry for landmark shape analysis: Unifying shape representation and deformation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 337–350. [CrossRef] [PubMed]
22. Peter, A.; Rangarajan, A. Maximum likelihood wavelet density estimation with applications to image and shape matching. *IEEE Trans. Image Process.* **2008**, *17*, 458–468. [CrossRef] [PubMed]
23. Bingham, C. Distributions on the Sphere and on the Projective Plane. Ph.D. Thesis, Yale University, New Haven, CT, USA, 1964.
24. Parthasarathy, B.; Kadane, J.B. Laplace approximations to posterior moments and marginal distributions on circles, spheres, and cylinders. *Can. J. Stat.* **1991**, *19*, 67–77.
25. Kume, A. Saddlepoint approximations for the Bingham and Fisher-Bingham normalising constants. *Biometrika* **2005**, *92*, 465–476. [CrossRef]

26. Hall, P.; Hannan, E. On stochastic complexity and nonparametric density estimation. *Biometrika* **1988**, *75*, 705–714. [CrossRef]
27. Kontkanen, P. Computationally Efficient Methods for MDL-Optimal Density Dstimation and Data Clustering. Ph.D. Thesis, University of Helsinki, Helsinki, Finland, 2009.
28. Davies, L.; Gather, U.; Nordman, D.; Weinert, H. A comparison of automatic histogram constructions. *ESAIM* **2009**, *13*, 181–196. [CrossRef]
29. McKay, D. A practical Bayesian framework for backpropation networks. *Neural Comput.* **1992**, *4*, 448–472. [CrossRef]
30. Stigler, S.M. *The History of Statistics: The Measurement of Uncertainty before 1900*; Harvard University Press: Cambridge, MA, USA, 1986.
31. Rao, C.R. Information and accuracy attainable in estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–91.
32. Jeffreys, H. *Theory of Probability*, 3rd ed.; Oxford University Press: New York, NY, USA, 1961.
33. Kass, R.E. The geometry of asymptotic inference. *Stat. Sci.* **1989**, *4*, 188–234. [CrossRef]
34. Barndorff-Nielsen, O.; Cox, D.; Reid, N. The role of differential geometry in statistical theory. *Int. Stat. Rev.* **1986**, *54*, 83–96. [CrossRef]
35. Amari, S.I.; Nagaoka, H. *Methods of Information Geometry*; American Mathematical Society: Providence, RI, USA, 2001.
36. Cover, T.; Thomas, J. *Elements of Information Theory*, 2nd ed.; Wiley Interscience: New York, NY, USA, 2006.
37. Laplace, P. Memoir on the probability of the causes of events. *Stat. Sci.* **1986**, *1*, 364–378. 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. [CrossRef]
38. Bertsekas, D.P. *Nonlinear Programming*, 2nd ed.; Athena Scientific: Belmont, MA, USA, 1999; pp. 291–293.
39. Berry, A.C. The accuracy of the Gaussian approximation to the sum of independent variates. *Trans. Am. Math. Soc.* **1941**, *49*, 122–136. [CrossRef]
40. Robert, C.; Casella, G. *Monte Carlo Statistical Methods*, 2nd ed.; Springer: New York, NY, USA, 2004.
41. Heck, D.W.; Moshagen, M.; Erdfelder, E. Model selection by minimum description length: Lower-bound sample sizes for the Fisher information approximation. *J. Math. Psychol.* **2014**, *60*, 29–34. [CrossRef]
42. Navarro, D.J. A note on the applied use of MDL approximations. *Neural Comput.* **2004**, *16*, 1763–1768. [CrossRef] [PubMed]
43. Peter, A.; Rangarajan, A. An information geometry approach to shape density minimum description length model selection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Barcelona, Spain, 6–13 November 2011; IEEE Press: Piscataway, NJ, USA, 2011; pp. 1432–1439.
44. Taylor, C. Akaike's information criterion and the histogram. *Biometrika* **1987**, *74*, 636–639. [CrossRef]
45. Abramowitz, M.; Stegun, I.A. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*; Dover Publications: Mineola, NY, USA, 1965.
46. Marron, S.J.; Wand, M.P. Exact mean integrated squared error. *Ann. Stat.* **1992**, *20*, 712–736. [CrossRef]
47. Wand, M.P.; Jones, M.C. Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Am. Stat. Assoc.* **1993**, *88*, 520–528. [CrossRef]