An Information Geometry Approach to Shape Density Minimum Description Length Model Selection

Adrian M. Peter
Florida Institute of Technology
Melbourne, FL
apeter@fit.edu

Anand Rangarajan
University of Florida
Gainesville, FL
anand@cise.ufl.edu

Abstract

For advantages such as a richer representation power and inherent robustness to noise, probability density functions are becoming a staple for complex problems in shape analysis. We consider a principled and geometric approach to selecting the model order for a class of shape density models where the square-root of the distribution is expanded in an orthogonal series. The free parameters associated with these estimators can then be rigorously selected using the Minimum Description Length (MDL) criterion for model selection. Under these models, it is shown that the MDL has a closed-form representation, atypical for most applications of MDL in density estimation. We provide a straightforward application of our derivations by using this closed-from MDL criterion to select the optimal multiresolution level(s) for a class of square-root, wavelet density estimators. Experimental evaluation of our technique is conducted on one and two dimensional density estimation problems in shape analysis, with comparative analysis against other popular model selection criteria such as Bayesian and Akaike information criteria.

1. Introduction

Shape analysis is a key ingredient to many computer vision and medical imaging applications that seeks to study the intimate relationship between the form and function of natural, cultural, medical and biological structures. Ever increasingly, probability density functions are being utilized to model and analyze shapes [21, 8, 14, 5, 4], a growth driven by advantages such as richer representation power, inherent robustness to noise, and alleviation of the correspondence problem. Here we consider a specific class of probabilistic shape densities—those that arise from expanding the square-root of the density in an orthogonal series. Such models have been used in the past [15] without consideration as to how one chooses the order of the series expansion. For the first time, we present a principled and geometric approach to selecting the model order of all orthogonal series estimators using the Minimum Description Length (MDL) criterion. Specifically, we focus on wavelet density estimators and derive several insightful model selection properties motivated by the information geometry of such spaces.

Several model selection criteria have been proposed, but arguably the following are the most commonly used: Akaike information criterion (AIC) [1], Bayesian information criterion (BIC) [19] and Minimum Description Length (MDL) [17, 18]. A fourth—Bayesian model selection (BMS) [9]—has been proven to be asymptotically equivalent to MDL [3]. The basic premise behind the resulting functional form of these criteria is to assign a goodness-of-fit measure (via the likelihood of the observed data sample) and a complexity penalty that can depend on the number of parameters in the model as well as the sample size. The AIC criterion is given by

$$AIC = -2\ln p(E|\hat{\boldsymbol{\Theta}}) + 2k \tag{1}$$

and BIC

$$BIC = -2\ln p(E|\hat{\boldsymbol{\Theta}}) + k\ln(N), \tag{2}$$

where E is the evidence (current observed data samples), $\hat{\boldsymbol{\Theta}}$ the maximum likelihood estimate (MLE) of the parameters, N the number of samples, and k the cardinality of the model parameters. For example, $k = 2$ for a linear model where the parameters correspond to (m, b), i.e. the slope and intercept of the line. In the context of wavelet density estimation (WDE) addressed in this paper, k will represent the number of coefficients per the multiresolution decomposition structure. For each criterion, the best model is the minimizer of these measures. Both AIC and BIC reward paucity of parameters as a penalty is paid for large values of k. Since BIC's second term also incorporates the sample size, it tends to prefer smaller complexity models (versus AIC) for sample sizes greater than eight. After eight samples, the second term of BIC, $k\ln(N)$, always has a lower value than AIC's second term, $2k$.

The complexity of a model under AIC and BIC is only measured by the cardinality of the parameters. This is the basic departure point of these (and others) versus MDL: they fail to take into account the functional form (how the parameters interact in the model) of the models. The MDL criterion given by

$$MDL = -\ln p(E|\hat{\boldsymbol{\Theta}}) + \frac{k}{2}\ln\left(\frac{N}{2\pi}\right) + \ln\int\sqrt{\det g_{ij}(\boldsymbol{\Theta})}d\boldsymbol{\Theta} \tag{3}$$

has an extra term (the third term) which penalizes based on the volume occupied by the model's manifold in the space of probability distributions (more on this in §3). Here, g_{ij} is the expected Fisher information of the parametric distribution, a.k.a. Fisher-Rao metric tensor. Chronologically, eq. 3 is the more recent version of MDL [18]. The original MDL [17] was similar to AIC and BIC in that it only contained the first two terms in (3), thus lacking a penalty based on the functional form. Our experiments in §5 will assess the usefulness of incorporating the additional volume term.

In practically all useful models, the Riemannian volume term in (3) must be computed by truncating the parameter space and using numerical techniques such as Monte Carlo integration. It will be shown in §4 that when one uses an orthogonal series density estimator (OSDE), this term is known in closed-form. MDL was originally developed using coding theory arguments that are based on the notion of finding the shortest code to describe the observed data [6], the more regularity in the data the shorter the code. Shorter code lengths can be shown to be inversely proportional to the likelihood of observing the data, i.e. higher probabilities are associated with shorter code lengths and smaller probabilities with large code lengths. Hence the use of the terminology 'minimum description length' to find the best model. The criterion as given in (3) is an approximation to the code length for the maximum-likelihood code [18]. In §3, we illustrate how MDL can be re-derived using differential geometry. It will allow us to transition from describing the second and third terms of eq. (3) as penalties for the number of parameters and functional form, respectively. Instead, we will see that together they determine a volume ratio designed to measure the ellipsoidal volume around the maximum likelihood estimate relative to the total volume occupied by the model in the space of probability densities.

2. Square-root Density Estimation

2.1. General Orthogonal Series Expansions

The square-root model of density estimation estimates $\sqrt{p(x)}$ and then obtains a bona fide density as $\left(\sqrt{p(x)}\right)^2$. This has several advantages over estimating $p(x)$ directly such as guaranteeing non-negativity. An insightful geometric advantage is gained if one expands $\sqrt{p(x)}$ in an orthogonal series—the unit integrability requirement of all probability densities translates to a spherical constraint on the coefficients; i.e. for

$$\sqrt{p(x)} = \sum_{j=0}^{\infty}\alpha_j\phi_j(x) \tag{4}$$

where $\int\phi_i\phi_j dx = \delta_{ij}$, we have

$$\int\left(\sqrt{p(x)}\right)^2 dx = \sum_j \alpha_j^2 = 1. \tag{5}$$

In computational implementations, the infinite expansion is truncated to some finite value J. This immediately leads to the interpretation that the basis coefficients—which are unique to a particular density since we have assumed the orthogonal series serve as a true basis for the space of continuous distributions—give the coordinates for a position on the unit hypersphere. The ordering of the coefficients in the coordinate vector can be taken in any arrangement but it must be consistent across all densities. The dimensionality of the hypersphere is determined by the cardinality of the set containing all the coefficients. The hypersphere geometry of the densities can be more rigorously justified when we analyze the $\sqrt{p(x)}$ representation under the theoretical basis of information geometry [2, 21, 14]. In this context, the Fisher information matrix (FIM) serves as the metric tensor on the manifold of a parametric family of distributions. One of the algebraic forms of the FIM is given by

$$g_{u,v} = 4\int\frac{\partial\sqrt{p(x|\boldsymbol{\Theta})}}{\partial\theta^u}\frac{\partial\sqrt{p(x|\boldsymbol{\Theta})}}{\partial\theta^v}dx \tag{6}$$

where $\boldsymbol{\Theta} = \left\{\theta^1,\ldots,\theta^m\right\}$ denotes the parameters of the distribution and u and v indicate the row and column index, i.e. for a family with m parameters the FIM is $m\times m$. Under an orthonormal expansion of $\sqrt{p(x|\boldsymbol{\Theta})}$, eq. (6) reduces to the canonical metric tensor of a unit hypersphere embedded in an $m+1$ Hilbert space. Rather than use the metric tensor to intrinsically compute geodesics on the hypersphere (an undertaking which would require us to parametrize the manifold), we can accomplish the same computation by realizing that the constraint $\sum_{i=1}^{m+1}\left(\theta^i\right)^2 = 1$ also implies the unit hypersphere geometry. Hence, closed-form geodesics distances can be simply computed using the usual angle measure between two unit vectors. In the present context, the coefficients of the expansion serve as the parameters of the density, i.e. $\boldsymbol{\Theta} = \{\alpha_j\}_{j=1}^J$. In §3, this geometric identification of the space of orthogonal series estimators with the unit hypersphere will enable the closed-form computation of the full MDL criterion (eq. 3).

2.2. Wavelet Density Estimators

We now focus on a specific class of orthogonal series estimators, representing densities which are expanded in multiresolution wavelet bases. Here the wavelet expansion of

$\sqrt{p(x)}$ is given by

$$\sqrt{p(x)} = \sum_{j_0,k} \alpha_{j_0,k}\phi_{j_0,k}(x) + \sum_{j\geq j_0,k}^{\infty} \beta_{j,k}\psi_{j,k}(x), \quad (7)$$

where $\alpha_{j_0,k}$ and $\beta_{j,k}$ are coefficients for the father $\phi(x)$ and mother $\psi(x)$ basis functions; the j-index represents the current scale level and the k-index the integer translation value. (Note: $\phi(x)$ and $\psi(x)$ are also referred to as the scaling and wavelet functions respectively.) For numerical implementation, the infinite expansion in (7) is truncated to some n set of scale levels and we must also select a starting scale level j_0. The goal is to estimate the set of coefficients $\{\alpha_{j_0,k}, \beta_{j,k}\}$ and reconstruct the density using (7). An efficient maximum likelihood method to estimate them, with fast convergence, is discussed in [13]. As before, the unit integrability requirement of all probability densities translates to a constraint on the wavelet coefficients

$$\int \left(\sqrt{p(x)}\right)^2 dx = \sum_{j_0,k} \alpha_{j_0,k}^2 + \sum_{j\geq j_0,k}^{j_1} \beta_{j,k}^2 = 1. \quad (8)$$

Orthonormal wavelet bases with compact support consist of families such as Daubechies, Coiflets and Symlets.

For applications to shape analysis, one begins from a point-set representation of the shapes and then proceeds to estimate their corresponding density function. All further analysis is carried out using the coefficients of the expansion. It is worth noting that another related approach is to perform the analysis on the space of $\sqrt{p(x)}$ directly, as [21] first introduced in the context of shape analysis. This surprisingly results in a similar unit hypersphere geometry, but whereas this spherical model arises from the functional values of $\sqrt{p(x)}$, the orthogonal series spherical manifold is constructed on the coefficients of the expansion.

3. MDL from Differential Geometry

In this section we briefly recap the geometric development of MDL as first presented by Balasubramanian [3]. The author refers to the model selection criterion as the *razor*. It is asymptotically equivalent to MDL. The derivations begin from a Bayesian approach by considering the posterior of a parametric model class $\mathcal{M}$

$$p(\mathcal{M}|E) = \frac{p(\mathcal{M})\int p(\boldsymbol{\Theta})p(E|\boldsymbol{\Theta})d\Theta}{p(E)} \quad (9)$$

where $\boldsymbol{\Theta} \in \mathbb{R}^d$ are the parameters of the model class. Hence, $p(\boldsymbol{\Theta})$ is the prior distribution on the parameters and $p(E|\boldsymbol{\Theta})$ is the likelihood. When comparing two candidate model classes $\mathcal{M}_1$ and $\mathcal{M}_2$, we can drop $p(E)$ since it is common factor and we can also omit the prior on the models, $p(\mathcal{M}_i)$, by assuming they are equally likely. (To avoid aberrant cases, we assume throughout that the parameter spaces of candidate models are compact.) These assumptions reduce (9) to $p(\mathcal{M}|E) \propto \int p(\boldsymbol{\Theta})p(E|\boldsymbol{\Theta})d\Theta$. It was show in [3, 11] that the Jeffrey's prior [7]

$$p(\boldsymbol{\Theta}) = \frac{\sqrt{\det g_{ij}(\boldsymbol{\Theta})}}{\int \sqrt{\det g_{ij}(\boldsymbol{\Theta})}d\boldsymbol{\Theta}} \quad (10)$$

is the appropriate prior to choose when the desire is to: treat all parameters equally (uniform), be invariant to reparametrizations of the parameter space, and geometrically count only *distinguishable* volumes on the parameter domain. (The notion of distinguishability was rigorously derived in [3].) Finally we assume the observed data $E = \{x_i\}_{i=1}^N$ are i.i.d., hence $p(E|\boldsymbol{\Theta}) = \prod_{i=1}^N p(x_i|\boldsymbol{\Theta})$. With the aforementioned substitutions, the razor is given as

$$R(\mathcal{M}) = \frac{\int \sqrt{\det g_{ij}(\boldsymbol{\Theta})} \exp\left\{-N\left(\frac{-\ln p(E|\boldsymbol{\Theta})}{N}\right)\right\} d\boldsymbol{\Theta}}{\int \sqrt{\det g_{ij}(\boldsymbol{\Theta})}d\boldsymbol{\Theta}}. \quad (11)$$

In order to use the razor for practical evaluation of candidate models, the integral in the numerator of eq. (11) must be approximated around the maximum likelihood estimate of the parameters, $\hat{\boldsymbol{\Theta}}$. (The integral approximation technique is commonly referred to as the Laplace approximation.) To a second order approximation, this yields the final version of the razor

$$\rho(\mathcal{M}) = -\ln R(\mathcal{M}) = -\ln p(E|\hat{\boldsymbol{\Theta}}) + \frac{k}{2}\ln\left(\frac{N}{2\pi}\right) + \ln\int \sqrt{\det g_{ij}(\boldsymbol{\Theta})}d\boldsymbol{\Theta} + \frac{1}{2}\ln\left(\frac{\det \tilde{g}_{ij}(\boldsymbol{\Theta})}{\det g_{ij}(\boldsymbol{\Theta})}\right) \quad (12)$$

where $\tilde{g}_{ij}$ is the *empirical Fisher information* computed from our observed sample values. Notice that the first three terms of (12) correspond to the MDL criterion in (3). The last term considers the ratio of the expected Fisher to the empirical Fisher, which has the property that as $N \to \infty$, $\tilde{g}_{ij} \to g_{ij}$ (empirical Fisher approaches expected Fisher), so this term vanishes, giving us back the MDL eq. (3).

To better understand the connection of MDL to the Riemannian volumes associated with a model class, we can rewrite (12) as

$$\rho(\mathcal{M}) = -\ln p(E|\hat{\boldsymbol{\Theta}}) + \ln\left(\frac{\mathcal{V}_{\mathcal{M}}}{\mathcal{V}_{\hat{\boldsymbol{\Theta}}}}\right). \quad (13)$$

The numerator of the second term is the total Riemannian volume, $\mathcal{V}_{\mathcal{M}} = \int \sqrt{\det g_{ij}(\boldsymbol{\Theta})}d\boldsymbol{\Theta}$, of the probabilistic manifold (i.e. total volume of the model class). The denominator $\mathcal{V}_{\hat{\boldsymbol{\Theta}}} = \left(\frac{2\pi}{N}\right)^{\frac{k}{2}} G(\hat{\boldsymbol{\Theta}})$, where $G(\hat{\boldsymbol{\Theta}}) = \left(\frac{\det g_{ij}(\hat{\boldsymbol{\Theta}})}{\det \tilde{g}_{ij}(\hat{\boldsymbol{\Theta}})}\right)^{\frac{1}{2}}$, is a term that measures appreciable volume of distinguishable distributions around the maximum likelihood estimate

that comes close to the truth (close in the sense that the model is able to predict the evidence E with high probability). As observed above, this data dependent term has the property that $G(\hat{\mathbf{\Theta}}) \rightarrow 1$ as $N \rightarrow \infty$. Hence the ellipsoidal volume around the MLE can be approximated by $\tilde{\mathcal{V}}_{\hat{\mathbf{\Theta}}} \approx \left(\frac{2\pi}{N}\right)^{\frac{k}{2}}$. Given this approximation, we have

$$\rho(\mathcal{M}) = MDL = -\ln p(E|\hat{\mathbf{\Theta}}) + \ln\left(\frac{\mathcal{V}_{\mathcal{M}}}{\tilde{\mathcal{V}}_{\hat{\mathbf{\Theta}}}}\right). \quad (14)$$

Hence it can be seen that MDL penalizes models that have excessively small distinguishable volumes close to the truth (small $\mathcal{V}_{\hat{\mathbf{\Theta}}}$) or those that occupy a large volume in the space of distributions (large $\mathcal{V}_{\mathcal{M}}$). The volumes in the second term of eq. (14) are an intrinsic property of the model and together are often referred to as the *geometric complexity* of the model. MDL selects those models that have a low geometry complexity by picking those models with "*highest maximum likelihood per the relative ratio of the distinguishable distributions*" [11].

4. MDL and the Geometry of Square-Root Wavelet Densities

Up to now we have discussed the derivation and interpretations of the MDL criterion for an arbitrary parameter manifold of a probabilistic model class. We now turn our attention to the application of the MDL criterion to select the decomposition levels for our wavelet density estimation framework described in §2. It is worth reiterating that the fundamental idea of the closed-form MDL criterion holds true for all valid orthogonal series expansions, and not just the present focus on wavelets. Hence, *we would like to be able to use eq.* (3) *to decide how to pick the best* j_0 *and* j_1. The number of parameters, k in (3), for a particular choice of j_0 and j_1 is given by the cardinality of the coefficient set over all levels of the decomposition, i.e. $k = \#\{\mathbf{\Theta}\} = \#\{\alpha_{j_0,l}, \beta_{j,l}\}$. As discussed in §2, the coefficients are coordinates for the location of the density on the unit hypersphere embedded in a k-dimensional space. Thus each candidate model, given by choice of j_0 and j_1, is a unit hypersphere and computing the Riemannian volume $\mathcal{V}_{\mathcal{M}}$ in (14) amounts to calculating the *surface area* of a unit hypersphere. With this understanding, we now have a systematic procedure to select the best j_0 and j_1:

1. For each value of j_0 and j_1 estimate the wavelet density coefficients of the expansion [16, 13]. This will give you the likelihood term needed for (14) .
2. The cardinality of the coefficient set resulting for the selection of j_0 and j_1 will provide the value of k needed to compute volumes $\mathcal{V}_{\mathcal{M}}$ and $\mathcal{V}_{\hat{\mathbf{\Theta}}}$ (the remaining terms of the MDL).
3. The optimal $\{j_0^*, j_1^*\}$ is the one that minimizes (14).

Though systematic, the above process fails to take full advantage of the theoretical consequences associated with the use of wavelets. For example, there are significant computational savings by leveraging the nested subspace structure of wavelet bases. Another issue is that we must address an anomaly that arises when computing the volume of a unit hypersphere as the dimensions increase: $\mathcal{V}_{\mathcal{M}} \rightarrow 0$ as $k \rightarrow \infty$. The following subsections expand on these topics.

4.1. MDL is Invariant to Multiresolution Analysis

The first observation we make is that the *MDL criterion is invariant to multiresolution decompositions (consisting of scaling and wavelet functions) in comparison to their corresponding single level scaling counterparts*. This is a significant result that enables us to perform our model search over j_0 instead of j_0 and j_1. This result directly follows from the nested subspace property of wavelet bases and the dyadic relationship of the basis functions at different levels.

In order to establish the invariance of MDL to multiresolution analysis (MRA) versus an appropriate single level scaling-function expansion, we have to establish that the goodness-of-fit and geometric complexity terms are identical for both. First let us establish equivalence of the goodness-of-fit as measured by the log likelihood. Consider a wavelet density estimate using only scaling functions from an arbitrary level j. These form a basis for V_j. However, functions expanded using scaling functions from level j can be equivalently represented using both scaling and wavelet bases that span level $j-1$, V_{j-1} and W_{j-1} respectively. Then V_{j-1} can be recursively broken down again and again. The recursive decomposition relationship is given by

$$\begin{aligned} V_j &= V_{j-1} \bigoplus W_{j-1} \\ &= V_{j-2} \bigoplus W_{j-2} \bigoplus W_{j-1} \\ &= V_{j_0} \bigoplus \bigoplus_{l=j_0}^{j-1} W_l \end{aligned} . \quad (15)$$

Hence, densities estimated using only scaling functions have an equivalent representation in a multiresolution hierarchy. Since the estimated densities (either from only level j or MRA from j_0 to $j-1$) are equivalent, their corresponding log likelihoods would be the same. So two models, one with only scaling functions and one with an equivalent MRA representation, give the same goodness-of-fit measure for the MDL criterion.

To show that geometric complexities are identical, we have to establish that an expansion using only scaling functions has the same number of coefficients as its corresponding MRA. This is clearly true by the very nature of the dyadic relationships between levels in a MRA: *basis functions at a coarser level* $j-1$ *have twice the support of those at level* j*, hence half the number of coefficients*. The number of coefficients at a particular level is associated with the

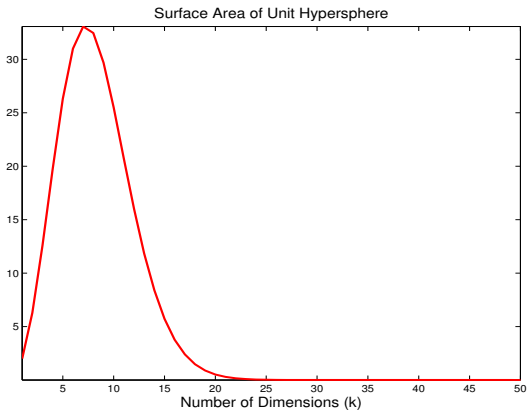

Figure 1: Surface area of unit hypersphere. Maximum surface area is at dimension seven.

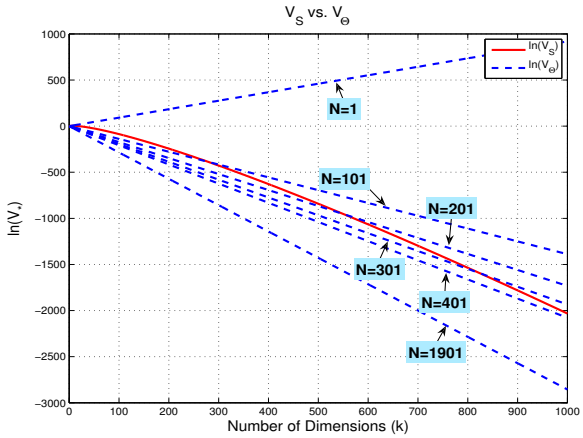

Figure 2: Riemannian volume comparisons, $\ln\left(\mathcal{V}_{\mathcal{S}}\right)$ (solid line) versus $\ln\left(\mathcal{V}_{\hat{\Theta}}\right)$ (dashed line). Misspecified models occur when $\mathcal{V}_{\hat{\Theta}} > \mathcal{V}_{\mathcal{S}}$. For sufficiently high number of samples we see that $\mathcal{V}_{\hat{\Theta}} < \mathcal{V}_{\mathcal{S}}$ as desired.

number of translations needed to span a defined spatial support. Theoretically, an infinite number of translations are used, but for any finite sample set the span of translations needed to cover the data will also be finite. The cardinality of the coefficient set from a level j with only scaling functions would equal the cardinality of coefficients from the coarser level $j-1$ that has both scaling and wavelet bases, i.e.

$$\begin{aligned} k &= \#\{V_j\} \\ &= \#\{V_{j-1}\} + \#\{W_{j-1}\} = \tfrac{k}{2} + \tfrac{k}{2} \\ &= \#\{V_{j-2}\} + \#\{W_{j-2}\} + \#\{W_{j-1}\} = \tfrac{k}{4} + \tfrac{k}{4} + \tfrac{k}{2} \\ &= \#\{V_{j_0}\} + \textstyle\sum_{l=j_0}^{j-1} \#\{W_l\} \end{aligned}, \tag{16}$$

where we have slightly abused the notation $\#\{\cdot\}$ to count the number of coefficients for a chosen basis level's function space. Since the value of k essentially determines the geometric complexity, it will be identical for single level decomposition at level j or a MRA from j_0 up to $j-1$. (The number of samples N is also a factor in the $\mathcal{V}_{\hat{\Theta}}$ term of geometric complexity, but it will be the same for all models so can be ignored in this analysis.)

With both the goodness-of-fit and geometric complexity shown to be the same for MRA versus single-level scaling function bases, *it is sufficient for density estimation to use only scaling functions* and to search for the best model by iterating over various starting levels j_0. So is MRA for wavelet density estimation not needed? It depends. If your goal is to simply obtain a reconstruction of the density, then it can be argued that scaling functions alone are enough. But if one's goal is sparsity among the coefficients (which is what MRA is designed for), then a different mechanism that measures this property must be incorporated into the model selection framework. Such a measure would include wavelet thresholding as part of the criterion for selecting the model. This is an avenue of future research.

4.2. Closed-Form Computation of $\mathcal{V}_{\mathcal{M}}$

In practice, the application of the MDL eq. (14) almost always requires numerical integration to compute $\mathcal{V}_{\mathcal{M}}$, the Riemannian volume of the statistical manifold. This involves derivation of the Fisher information metric (FIM), appropriate truncation of the parameter space to perform the integration and other numerical adjustments to ensure that the FIM does not become singular. For very high-dimensional parameter spaces, one has to employ Monte Carlo integration methods. Only for very simple models is $\mathcal{V}_{\mathcal{M}}$ in an analytic form; sometimes even the FIM is not in closed-form and may require an additional numerical integration step. One significant advantage of our wavelet density estimation framework is that all of our models have a unit hypersphere geometry (again this is true for all orthogonal series expansions). Hence, $\mathcal{V}_{\mathcal{M}}$ is known in closed-form. It is merely the surface area ($\mathcal{V}_{\mathcal{S}}$) of a unit hypersphere of dimension $k-1$ where $k = \#\{\boldsymbol{\Theta}\} = \#\{\alpha_{j_0,l}\}$. (Choosing the j_0 decomposition level determines the coefficient set, the cardinality of which is k.) One would intuitively expect the volume of a manifold to increase as the number of dimensions increase. However, the unit hypersphere exhibits an odd property in that it decreases in volume (and surface area) as the dimensions increase [20].

The surface area of a unit hypersphere $\mathcal{S}$ is given by

$$\mathcal{V}_{\mathcal{S}} = \begin{cases} \frac{k\pi^{\frac{k}{2}}}{\left(\frac{k}{2}\right)!} & , k \text{ even} \\ 2^k \pi^{\frac{k-1}{2}} \frac{\left(\frac{k-1}{2}\right)!}{(k-1)!} & , k \text{ odd} \end{cases}. \tag{17}$$

As shown in Figure 1, the maximum surface area is reached at dimension seven, and then the surface area rapidly decreases to zero. Recall that the geometric complexity assesses a cost based on the ratio of the manifold volume to the ellipsoidal volume around the MLE, i.e. the penalty is

$\ln\left(\frac{\mathcal{V}_S}{\mathcal{V}_{\hat{\boldsymbol{\Theta}}}}\right)$. If $\mathcal{V}_S$ shrinks to zero so fast that it is smaller than $\mathcal{V}_{\hat{\boldsymbol{\Theta}}}$, then our penalty term is not valid since it would become negative. Having $\mathcal{V}_{\hat{\boldsymbol{\Theta}}} > \mathcal{V}_S$ tells us that the model is misspecified [12]. Geometrically we can visualize this as the ellipsoidal volume around the MLE protruding out of the smaller model manifold. In practice, one has to be careful to consider the trade-off between the number of samples and the number of parameters. A valid region of well-specified models is easily achieved when we consider $\mathcal{V}_{\hat{\boldsymbol{\Theta}}} = \left(\frac{2\pi}{N}\right)^{\frac{k}{2}}$. Once we reach above seven samples, i.e. $N \geq 7$, the ellipsoidal volume starts to decline exponentially as the number of parameters k increases. Since we need the number of samples to be generally greater than the number of parameters to avoid an ill-posed density estimation problem, we can easily satisfy our requirement of needing $\mathcal{V}_{\hat{\boldsymbol{\Theta}}} < \mathcal{V}_S$. In Figure 2, we illustrate $\ln(\mathcal{V}_S)$ versus $\ln\left(\mathcal{V}_{\hat{\boldsymbol{\Theta}}}\right)$ over a range of sample cardinalities and number of parameters. Notice that for a sufficiently high number of samples relative to the number of parameters (i.e. dimensions of the unit hypersphere), there is a sharp decrease of $\mathcal{V}_{\hat{\boldsymbol{\Theta}}}$ as desired. It is worth noting that to guarantee uniqueness of the estimated density, the coefficients of the expansion should be restricted to the positive orthant of the unit hypersphere. This requires the volume term be divided by a 2^k factor. We can easily account for misspecified models under this restriction by simply increasing the number of samples.

5. Experiments

The experiments evaluated the capability of the model selection methods to adequately select the best probability density for a given set of sample data, while judiciously balancing the desire for accuracy and model complexity. We first validate our approach on a complex set of 1D densities as in [10] and utilize a variety of compactly supported wavelet families in the density estimation. From each 1D density, 2000 samples were drawn and used in the parameter estimation process. For shape analysis, we illustrate the utility of MDL criterion to select the optimal density function representation and matching of shape point sets.

The MDL criterion of eq. (3) (denoted MDL-3 in results) was applied to the selection of the best j_0 level for the wavelet density estimator. For comparative analysis, we computed several other information-theoretic model selection criteria: the original two-term MDL (MDL-2) which lacks the model class Riemannian volume, AIC and BIC. In addition, since the true densities are known, we calculated three standard discrepancy measures: mean-squared error (MSE), Hellinger divergence (HELL) and L_1 loss. The best starting level j_0^* was selected as the minimum of these measures for $j_0 \in [-1, 6]$. A larger value of j_0 indicates a more complex model since it corresponds to a finer resolution level in the wavelet decomposition. Extensive comparisons

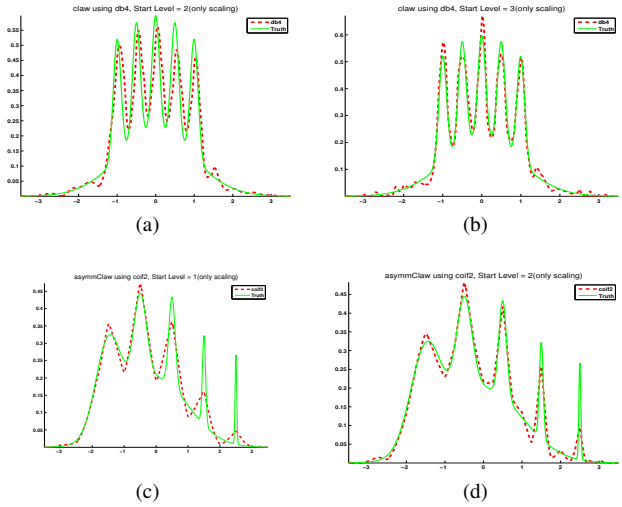

Figure 3: Model selection using MDL-3 versus MDL-2 . MDL-3 is able to select more complex models than MDL-2. a) $j_0^* = 2$ by MDL-2. b) $j_0^* = 3$ by MDL-3. c) $j_0^* = 1$ by MDL-2. d) $j_0^* = 2$ by MDL-3. Wavelet family DB4 used for (a) and (b), COIF2 used for (c) and (d).

were conducted using basis functions from the Daubechies, Symlet and Coiflet families for each of the criteria. MDL-3 and MDL-2 generally agreed on best levels across densities and families. There are a few cases in which MDL-3 (with the additional volume term) selected more complex models than MDL-2. In each of these cases, the selection of the higher complexity model was justified by the need to accurately capture the abrupt variations of the true data-generating densities. A high-level summarization of the general trends in the experiments is visually illustrated in Figures 3-5. In Figure 3, we see examples of two cases where the MDL-3 selected value of j_0 provides a better suited model. Thus the inclusion of the full geometric complexity $\ln\left(\frac{\mathcal{V}_S}{\mathcal{V}_{\hat{\boldsymbol{\Theta}}}}\right)$ can aid in the selection of more accurate models.

In general, our MDL criterion also agrees with AIC and BIC. As expected, AIC tends to pick slightly more complex models than MDL and BIC. This is because AIC does not incur a penalty dependent on the sample size. This slight over estimation can be a benefit when considering complex densities but it can also over compensate, see Figure 4. AIC selects a more complex model than necessary for the bimodal density [see (A) and (B)]. It starts to favor trends in the data, degrading its generalization capability. However, for a complex density like the asymmetric double claw, the AIC selection is a better model. BIC tends to somewhat underestimate the models, selecting less complex models than necessary to represent the densities [see Figure 5 (A) and (B)].

In real-world applications, the MSE, HELL, and L_1 are

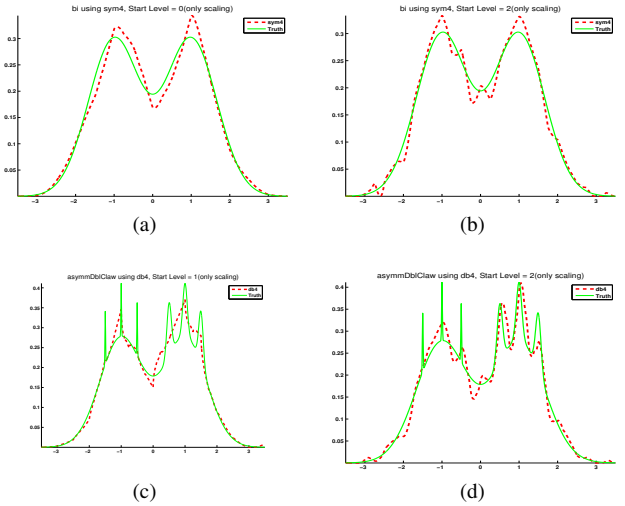

Figure 4: Model selection using MDL-3 versus AIC. AIC generally selects more complex models than MDL-3. This can be helpful for complex densities like in (c) and (d), but can also over estimate smooth ones like in (a) and (b). a) $j_0^* = 0$ by MDL-3. b) $j_0^* = 2$ by AIC. c) $j_0^* = 1$ by MDL-3. d) $j_0^* = 2$ by AIC. Wavelet family SYM4 used for (a) and (b), DB4 used for (c) and (d).

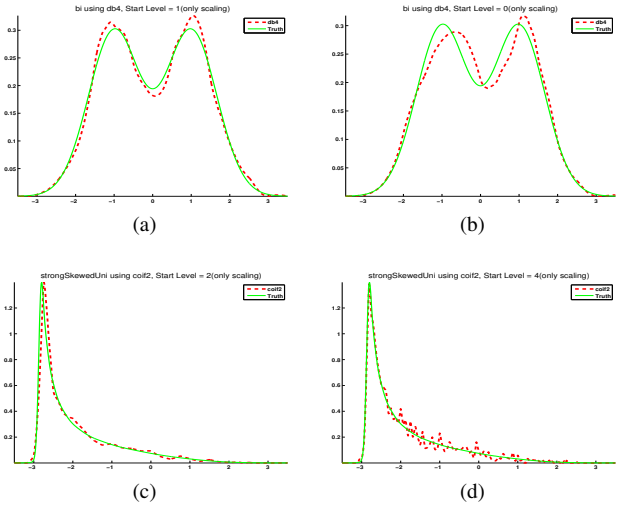

Figure 5: Model selection using MDL-3 versus BIC and MSE. BIC tends to favor less complex models than MDL-3, see (a) and (b). The MSE generally overfits since it is only a goodness-of-fit measure. a) $j_0^* = 1$ by MDL-3. b) $j_0^* = 0$ by BIC. c) $j_0^* = 2$ by MDL-3. d) $j_0^* = 4$ by MSE. Wavelet family DB4 used for (a) and (b), COIF2 used for (c) and (d).

not useful model selection criteria since the true underlying densities are not accessible or unknown. They also lack the trade-off between goodness-of-fit and complexity, only using the former as the performance measure. Hence, biased error measures like the MSE, tend to pick more complex models, Figure 5 (C) and (D). Since we know the true densities, the global agreement between these error measures and the information-theoretic model selection criteria showcases the power of these methods—without knowledge of true densities, they are able to select models that best describe the data while balancing the complexity of the model.

For applications to 2D shape analysis, Figure 6 illustrates the j_0 levels chosen by MDL-3, MDL-2, AIC and BIC for three MPEG-7 shapes (only a subset of the entire experiment). Here the density functions were estimated from point set representation of the shapes. The general approach of estimating wavelet densities from this shape data set was first demonstrated in [15]. They selected the optimal j_0 decomposition level by heuristic visual inspection and then showcased very favorable shape retrieval results utilizing the arc-length geodesic distance on the unit hypersphere of wavelet densities. Here we have rigorously compared several model selection criteria and interestingly enough, the optimal level $j_0^* = 1$ selected by both MDL-3 and AIC precisely corresponded exactly to the heuristically chosen level in [15].

6. Discussion

In the realm of model selection, there are a plethora of criteria (e.g. AIC, BIC, BMS, etc.) that are designed to pick the best model from a set of competing ones while taking into account a balance of goodness-of-fit and complexity. Of these MDL is the only one that takes into account the notion of model volumes and their occupation in the space of all distributions.

We have illustrated the use of MDL for square-root probability densities expanded in an orthogonal series. In this framework, we are able to compute the Riemannian volume term of the MDL criterion in closed-form. Closed-form computation of $\mathcal{V}_{\mathcal{M}}$ is a rarity for such a rich and flexible density estimation model. We applied this procedure to estimating the MRA structure for the wavelet expansion of the square-root density. For the first time, we proved the invariance property of MDL to MRA. This allowed the search for the best density model, for a given sample, to be conducted over just scaling functions at different levels. Validating the use of the MDL volume term also illustrated cases where it was useful in picking a more suited model than competing methods.

Square-root wavelet densities have already demonstrated their effectiveness in shape analysis [21, 14, 5, 4]. In this work, we have shown that information geometry allows us to perform model selection intrinsically on the space of the distributions, taking us one step closer to the goal of turnkey shape density estimation.

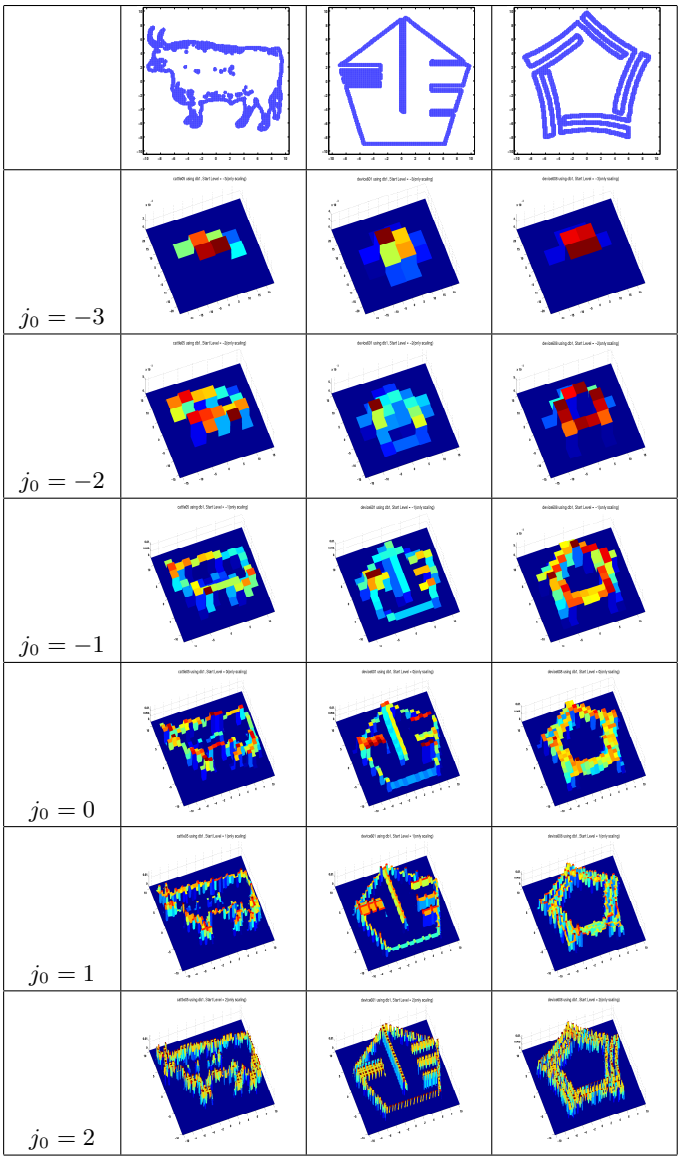

Figure 6: Model selection for 2D densities using MDL-3, MDL-2, AIC and BIC. Row 1 represents MPEG-7 shapes Cattle-05 (8,671 points), Device6-01 (8,947 points), and Device6-08 (11,301 points), respectively. Remaining rows show densities estimated from these point sets at different j_0 levels. For all three shapes, MDL-3 and AIC selected $j_0 = 1$, while MDL-2 and BIC selected $j_0 = 0$.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. NSF 1143963.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Intl. Symposium on Information Theory*, pages 267–281, 1973. 1

[2] S.-I. Amari and H. Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2001. 2.1

[3] V. Balasubramanian. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation*, 9(2):349–368, Feb. 1997. 1, 3, 3, 3

[4] J. Cheng, A. Ghosh, T. Jiang, and R. Deriche. A Riemannian framework for orientation distribution function computing. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 911–918, 2009. 1, 6

[5] A. Goh and R. Vidal. Unsupervised Riemannian clustering of probability density functions. In *European Conf. on Machine Learning and Knowledge Discovery in Databases (ECML KDD)*, pages 377–392, 2008. 1, 6

[6] P. Grünwald. A tutorial introduction to the minimum description length principle. In P. Grünwald, I. Myung, and M. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*. MIT Press, 2005. 1

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

[8] B. Jian and B. C. Vemuri. A robust algorithm for point set registration using mixture of Gaussians. In *IEEE Intl. Conf. on Computer Vision (ICCV)*, pages 1246–1251, 2005. 1

[9] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995. 1

[10] S. J. Marron and M. P. Wand. Exact mean integrated squared error. *The Annals of Statistics*, 20(2):712–736, 1992. 5

[11] I. J. Myung, V. Balasubramanian, and M. A. Pitt. Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, 97:11170–11175, 2000. 3, 3

[12] D. J. Navarro. A note on the applied use of MDL approximations. *Neural Computation*, 16:1763–1768, 2004. 4.2

[13] A. Peter and A. Rangarajan. Maximum likelihood wavelet density estimation with applications to image and shape matching. *IEEE Trans. on Image Processing*, 17(4):458–468, April 2008. 2.2, 1

[14] A. Peter and A. Rangarajan. Information geometry for landmark shape analysis: Unifying shape representation and deformation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(2):337–350, February 2009. 1, 2.1, 6

[15] A. Peter, A. Rangarajan, and J. Ho. Shape L'Âne Rouge: Sliding wavelets for indexing and retrieval. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2008. 1, 5

[16] A. Pinheiro and B. Vidakovic. Estimating the square root of a density via compactly supported wavelets. *Computational Statistics and Data Analysis*, 25(4):399–415, 1997. 1

[17] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978. 1, 1

[18] J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. on Information Theory*, 42:40–47, 1996. 1, 1

[19] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. 1

[20] D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley-Interscience, 2001. 4.2

[21] A. Srivastava, I. Jermyn, and S. Joshi. Riemannian analysis of probability density functions with applications in vision. In *IEEE CVPR*, pages 1–8, 2007. 1, 2.1, 2.2, 6