A Model Selection Approach to Hierarchical Shape Clustering with an Application to Cell Shapes

Mina Mirshahi[1☯], Vahid Partovi-Nia[2‡], Masoud Asgharian[3],

1,2 Department of Mathematics and Industrial Engineering, Polytechnique Montreal, Canada.
3 Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada.

☯Manulife, Toronto, ON, Canada. ‡Noah's Ark Lab, Huawei Technologies, QC, Canada.
* mina.mirshahi@gerad.ca *vahid.partovinia@polymtl.ca
*masoud.asgharian-dastenei@mcgill.ca

Abstract

Shape is an important phenotype of living species that contain different environmental and genetic information. Clustering living cells using their shape information can provide a preliminary guide to their functionality and evolution. Hierarchical clustering and dendrograms, as a visualization tool for hierarchical clustering, are commonly used by practitioners for classification and clustering. The existing hierarchical shape clustering methods are distance based. Such methods often lack a proper statistical foundation to allow for making inference on important parameters such as the number of clusters, often of prime interest to practitioners. We take a model selection perspective to clustering and propose a shape clustering method through linear models defined on Spherical Harmonics expansions of shapes. We introduce a BIC-type criterion, called CLUSBIC, and study consistency of the criterion. Special attention is paid to the notions of over- and under-specified models, important in studying model selection criteria and naturally defined in model selection literature. These notions do not automatically extend to shape clustering when a model selection perspective is adopted for clustering. To this end we take a novel approach using hypothesis testing. We apply our proposed criterion to cluster a set of real 3D images from HeLa cell line.

Introduction

Shape modelling plays an important role in medical imaging and computer vision [1, 2]. There is, in particular, a vast literature on cell shape analysis where the shape can often provide important information about the functionality of the cell ([3, 4, 5, 6, 7]). Most of such analysis employ techniques that often lack a statistical error component. Variability among shapes has been studied using different approaches: active shape models [8], computational anatomy [9], planar shape analysis [10, 11], etc.

Hierarchical clustering is one of the most commonly used methods by practitioners to find unknown patterns in data [12]. There are several reasons why a tree based method is preferable: i) the closest objects merge earlier, which reflects the evolution as the hierarchies are built up ii) provides a visual guide through a binary tree, called dendrogram iii) For a chosen number of clusters, it is only required to cut the dendrogram to achieve the corresponding clustering iv) allows to choose the number of clusters visually so that corresponding clustering is meaningful to an expert.

The existing hierarchical shape clustering methods are distance based and often lack a proper statistical foundation for making statistical inference. Borrowing ideas from model selection and testing statistical hypothesis, we take a different perspective to hierarchical clustering. A statistical shape model built on a set of image data, using a common coordinate system, can be regarded as a random continuous curve. A 3D shape can then be represented as a linear function of some basis functions. Possible noises and fluctuations can be easily included in the linear model. One can therefore use the likelihood function to devise a shape clustering algorithm using a convenient metric. The main advantage of such approach is to acquire the probability distribution of the fitted function and to distinguish between different shapes using their probability distributions.

Having modelled shapes using linear models, clustering can be performed on the estimated coefficients of the basis functions; similar shapes are expected to have similar coefficients. The problem of assigning two shapes to the same cluster will then become a hypothesis testing problem. Having this formulation, choosing the optimal number of clusters can be treated as a model selection problem. To this end, we use BIC and devise a criterion, called CLUSBIC, whose consistency is established.

The rest of this manuscript is organized as follows. In Section we discuss statistical modeling of shapes in three-dimensions using spherical harmonics. In Section 1 we propose a new Bayesian information criterion, called CLUSBIC, for clustering shapes taking into account the likelihood function of their fitted models, and establish consistency of the proposed criterion. In Section 2 we apply our proposed method to 3D images from HeLa cell line captured by a laser-scanning 240 microscope. Section 3 is the conclusion. Proofs of the main results are given in the appendices, while the proofs of other auxiliary results and the lemmas are documented in the supplementary materials.

sectionShape Modelling A three-dimensional (3D) shape descriptor is highly beneficial in many fields such as biometrics, biomedical imaging, and computer vision. Double Fourier series and spherical harmonics are widely used for representing 3D objects. Here, we discuss spherical harmonics for 3D shape modelling. Spherical harmonics are a natural and convenient choice of basis functions for representing any twice differentiable spherical function.

Let x, y and z denote the Cartesian object space coordinates, and θ, ϕ, and r the spherical parameter space coordinates, where θ is taken as the polar (colatitudinal) coordinate with $\theta \in [0, \pi]$, and ϕ as azimuthal (longitudinal) coordinate with $\phi \in [0, 2\pi]$. Spherical harmonics is somehow equivalent to a 3D extension of the Fourier series, which models r as a function of θ and ϕ. The real basis for spherical harmonics $Y_l^m(\theta, \phi)$ of degree l and order m is defined as:

$$Y_l^m(\theta,\phi) = \begin{cases} \sqrt{2}\sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_l^m(\cos\theta)\cos(m\phi) & \text{for } m \geq 0, \\ \sqrt{2}\sqrt{\frac{2l+1}{4\pi}\frac{(l-|m|)!}{(l+|m|)!}}P_l^{|m|}(\cos\theta)\sin(|m|\phi) & \text{for } m < 0, \end{cases} \tag{1}$$

where l and m are integers with $|m| \leq l$, and the associated Legendre polynomial P_l^m [13].

Given a spherical function $r(\theta, \phi)$ and a specified maximum degree $L_{\max}$, one can write $r(\theta, \phi)$ as a linear expansion of Y_l^m's with possibly some measurement errors and find the coefficient of fit through the method of least squares. That is,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where

$$\mathbf{y}^{\mathrm{T}} = \begin{pmatrix} r_1 & r_2 & \dots & r_N \end{pmatrix}, \boldsymbol{\beta}^{\mathrm{T}} = \begin{pmatrix} \beta_1 & \beta_2 & \dots & \beta_K \end{pmatrix},$$

$$\mathbf{X} = \begin{pmatrix} Y_l^{-l}(\theta_1,\phi_1) & Y_l^{-(l-1)}(\theta_1,\phi_1) & \dots & Y_l^0(\theta_1,\phi_1) & \dots & Y_l^l(\theta_1,\phi_1) \\ Y_l^{-l}(\theta_2,\phi_2) & Y_l^{-(l-1)}(\theta_2,\phi_2) & \dots & Y_l^0(\theta_2,\phi_2) & \dots & Y_l^l(\theta_2,\phi_2) \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ Y_l^{-l}(\theta_N,\phi_N) & Y_l^{-(l-1)}(\theta_N,\phi_N) & \dots & Y_l^0(\theta_N,\phi_N) & \dots & Y_l^l(\theta_N,\phi_N) \end{pmatrix},$$

where N is the number of observations, $|m| \leq l \leq L_{\max}$, and $K = (L_{\max}+1)^2$ is the number of parameters. The quality of fit improves as $L_{\max}$ increases, i.e., the larger the number of expansion terms.

The model (2) is only suitable for surface modeling of stellar shapes. A different parametric form for surface modeling, regardless of the type of shape, is suggested by [14, 15]. This parametric form gives us three explicit functions defining the surface of shape as $x_s(\theta,\phi)$, $y_s(\theta,\phi)$, and $z_s(\theta,\phi)$, where each of the coordinates is modeled as a function of spherical harmonic bases. Accordingly, the following three linear models are generated,

$$\begin{aligned}\mathbf{x}_s &= \mathbf{X}\boldsymbol{\beta}_x + \boldsymbol{\varepsilon}_x,\\ \mathbf{y}_s &= \mathbf{X}\boldsymbol{\beta}_y + \boldsymbol{\varepsilon}_y,\\ \mathbf{z}_s &= \mathbf{X}\boldsymbol{\beta}_z + \boldsymbol{\varepsilon}_z.\end{aligned}$$

The set of expansion coefficients $(\boldsymbol{\beta}_x, \boldsymbol{\beta}_y, \boldsymbol{\beta}_z)$ defines the shape completely. Assuming $\mathbf{x}_s$, $\mathbf{y}_s$ and $\mathbf{z}_s$ to be independent of each other, the above three models are equivalent to the following model

$$\begin{bmatrix}\mathbf{x}_s\\ \mathbf{y}_s\\ \mathbf{z}_s\end{bmatrix} = \mathbf{X}\begin{bmatrix}\boldsymbol{\beta}_x\\ \boldsymbol{\beta}_y\\ \boldsymbol{\beta}_z\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_x\\ \boldsymbol{\varepsilon}_y\\ \boldsymbol{\varepsilon}_z\end{bmatrix}. \tag{3}$$

1 Shape Clustering

Having characterised shapes by functional forms, we can cluster shapes according to their estimated coefficients. Clustering shapes can be achieved simply by computing the Euclidean distance over their parameters of the models in equation (2). However, this heuristic method may not lead to proper clusters, so we aim to develop a methodology using fitted models. To this end, we present a likelihood-based approach for clustering shapes in this section. From the Bayesian point of view, the clustering procedure is enhanced by contemplating the distribution of parameters in equation (2). We assume some prior distributions over parameters of the fit $\boldsymbol{\beta}$, and calculate the marginal distribution of the model by integrating the likelihood with respect to the assumed prior. In this approach, the hierarchy of clusters is built up based on the marginal likelihoods [16, 17, 18, 19]. Agglomerative hierarchical clustering is used to establish the dendrogram in which curves are merged as long as the merge improves the marginal likelihoods, as discussed in [20].

This section is organized as follows. First, a brief sketch of likelihood calculation with classical assumptions is provided in Subsection 1.1. In Subsection 1.2, we introduce a new Bayesian information criterion for clustering curves, called CLUSBIC. In Subsection 1.3 the consistency of the CLUSBIC is proved.

1.1 Linear Models

We consider the following linear model in polar coordinates

$$\mathbf{y}_{N\times 1} = \mathbf{X}_{N\times K}\boldsymbol{\beta}_{K\times 1} + \boldsymbol{\varepsilon}_{N\times 1}, \tag{4}$$

where N and K indicate the number of observations and the attributes respectively. We assume $\boldsymbol{\varepsilon}$ is distributed according to the Gaussian distribution with mean $\mathbf{0}_{N\times 1}$ and covariance matrix $\sigma^2\mathbf{I}_N$, i.e. $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$, where $\mathbf{I}_N$ is the identity matrix of size N.

In case of spherical harmonic expansions, equation (3), we assume that the error terms $\boldsymbol{\varepsilon}_x$, $\boldsymbol{\varepsilon}_y$, and $\boldsymbol{\varepsilon}_z$ are independent of each other and each follows the Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$. Subsequently, the vector

$$(\boldsymbol{\varepsilon}_x, \boldsymbol{\varepsilon}_y, \boldsymbol{\varepsilon}_z) \sim \mathcal{N}_{3N}(\mathbf{0}, \mathbf{I}_3 \otimes \sigma^2\mathbf{I}_N), \tag{5}$$

where the symbol $\otimes$ is the Kronecker product.

Given D distinct shapes, the model associated with this set of shapes is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{6}$$

where

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_D \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} \mathbf{X}_1^1 & \mathbf{0}^2_{N_1 \times K} & \dots & \mathbf{0}^D_{N_1 \times K} \\ \mathbf{0}^1_{N_2 \times K} & \mathbf{X}_2^2 & \dots & \mathbf{0}^D_{N_2 \times K} \\ \vdots & \vdots & & \vdots \\ \mathbf{0}^1_{N_D \times K} & \mathbf{0}^2_{n_D \times K} & \dots & \mathbf{X}_D^D \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_D \end{bmatrix}, \ \mathbf{e} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_D \end{bmatrix},$$

We explain the methodology using the notation for equation (4). Similar methodology applies using equation (3), taking into account the property in equation (5).

Suppose $\mathbf{d} = (d_1, d_2, \dots, d_D)$ is a grouping vector, e.g. $\mathbf{d} = (1, 1, \dots, 1)$ assigns all D shapes to one group and $\mathbf{d} = (1, 2, \dots, D)$ assigns each shape to a singleton. The likelihood function for the model (6) given the grouping vector $\mathbf{d}$ is,

$$p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X}, \mathbf{d}) = \prod_{i=1}^{\mathcal{C}(\mathbf{d})} p(\mathbf{y}_{(i)} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X}_{(i)}),$$

where $\mathcal{C}(\mathbf{d})$ denotes the number of unique elements in $\mathbf{d}$ (the number of clusters) $N_{(i)}$, while $\mathbf{y}_{(i)}$ and $\mathbf{X}_{(i)}$ represent, respectively, the number of observations, vector of response, and matrix of covariates after combining the clusters. The vector of unknown parameters is denoted by $\boldsymbol{\beta}$.

For the sake of simplicity, we propose conjugate priors [21]. To begin with, σ^2 is assumed to be known. The standard conjugate prior imposed on $\boldsymbol{\beta}$, conditional on σ^2, is $\boldsymbol{\beta} \mid \sigma^2 \sim \mathcal{N}(\boldsymbol{\beta}_0, \sigma^2 \boldsymbol{V}_0)$, where $\boldsymbol{\beta}_0$, and $\boldsymbol{V}_0$ are the prior mean and prior covariance matrix for $\boldsymbol{\beta}$ respectively. For a detailed discussion of Bayesian methods see [22]. The marginal likelihood of $\mathbf{y}$ can be computed as follows:

$$p(\mathbf{y}) = \int p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta} \mid \sigma^2) d\boldsymbol{\beta}. \tag{7}$$

In case of conjugate priors, the model appears as the multivariate Gaussian distribution with mean $\mathbf{X}\boldsymbol{\beta}_0$ and covariance matrix of $\mathbf{I}_N + \mathbf{X}\boldsymbol{V}_0\mathbf{X}^{\mathrm{T}}$, where $N = \sum_{i=1}^{D} N_i$ denotes the number of observations. The curves are assigned to a group with maximum $p(\mathbf{d} \mid \mathbf{y})$. By the Bayes theorem,

$$\begin{aligned} p(\mathbf{d}|\mathbf{y}) &= \frac{p(\mathbf{y} \mid \mathbf{d}) p(\mathbf{d})}{p(\mathbf{y})} \\ &\propto \prod_{i=1}^{\mathcal{C}(\mathbf{d})} p(\mathbf{y}_{(i)}) p(\mathbf{d}). \end{aligned} \tag{8}$$

The dendrogram exploits the marginal likelihood to build the hierarchy. It is expected that the marginal likelihood reaches its maximum over a reasonable grouping. Thus, the logical cut-off on the dendrogram is when the marginal likelihood is maximized over the dendrogram.

In order to calculate the marginal likelihood $p(\mathbf{y} \mid \sigma^2)$ for each curve, one needs to compute the inverse of the covariance matrix $(\mathbf{I}_N + \mathbf{X}\boldsymbol{V}_0\mathbf{X}^{\mathrm{T}})$ which is of size N, the number of observations. The computation of the inverse has the computational complexity of $\mathcal{O}(N^{2.37})$ [23] for the best-case scenario. To circumvent the computational complexity involved in computing the inverse of the covariance matrix, we modify the Gaussian model by adding σ^2 to the hierarchy of parameters. Assuming an inverse-gamma distribution for σ^2,

transform the Gaussian model to Student's t-model which does not require any matrix inversion of size N, see [24]. If σ^2 is nearly degenerate, i.e, $\mathrm{E}(\sigma^2) = \mu_0$, and $\mathrm{Var}(\sigma^2) \approx 0$, this model is, asymptotically, the same as the Gaussian model. The computational complexity of this model is bounded by $\mathcal{O}(NK^2)$ which is a significant improvement over the Gaussian model. The proposed computational trick leads to a marginal likelihood which comes from a distribution with heavier tails than the Gaussian distribution. This trick is particularly helpful in modeling data containing outliers.

As [25] discussed, agglomerative clustering using Student's t-distribution suffers from some instabilities under the common settings of the hyper-parameters. Here, the hyper-parameters of the inverse-gamma distribution are determined such that it produces a nearly degenerate distribution.

Bayesian clustering according to the $p(\mathbf{d} \mid \mathbf{y})$, equation (8), requires modeling $p(\mathbf{y} \mid \mathbf{d})$ and $p(\mathbf{d})$. Here, we follow the structure suggested in [20]. The random vector $\mathbf{d}$ denotes the possible groupings for a set of shapes. Suppose $\mathcal{C}(\mathbf{d})$ is the total number of clusters at each step and $n_1, n_2, \ldots, n_{\mathcal{C}(\mathbf{d})}$ are the total number of shapes in each of the clusters. Suppose that $\mathcal{C}(\mathbf{d})$ is uniformly distributed over the set $\{1, 2, \ldots, D\}$, where D is the total number of shapes and the n_j, $j = 1, 2, \ldots, D$, is distributed according to multinomial-Dirichlet with parameter $\boldsymbol{\pi}$,

$$p(\mathbf{d} \mid \boldsymbol{\pi}) = \frac{1}{D} \frac{\Gamma(\sum_{i=1}^{\mathcal{C}(\mathbf{d})} \pi_i) \prod_{i=1}^{\mathcal{C}(\mathbf{d})} \Gamma(n_i + \pi_i)}{\prod_{i=1}^{\mathcal{C}(\mathbf{d})} \Gamma(\pi_i) \Gamma(D + \sum_{i=1}^{\mathcal{C}(\mathbf{d})} \pi_i)}.$$

A uniform setting on the parameter vector $\boldsymbol{\pi}$ leads to

$$p(\mathbf{d}) = \frac{(\mathcal{C}(\mathbf{d}) - 1)! n_1! n_2! \ldots n_{\mathcal{C}(\mathbf{d})}!}{D(D + \mathcal{C}(\mathbf{d}) - 1)!}.$$

1.2 Clustering Bayesian Information Criterion (CLUSBIC)

We introduce a criterion for clustering, based on the marginal likelihoods, called *Clustering Bayesian Information Criterion* (CLUSBIC). CLUSBIC is similar to BIC in nature, but it is designed for the purpose of hierarchical clustering. When all data fall into one cluster, CLUSBIC coincides with BIC.

[26] proposed a simple informative distribution on the coefficient $\boldsymbol{\beta}$ in Gaussian regression models. As a prior on the parameter verctor, given the covariates, he considered $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, g\sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})$, a conjugate Gaussian prior distribution. This covariance matrix is a scaled version of the covariance matrix of the maximum likelihood estimator of $\boldsymbol{\beta}$. In practice, $\boldsymbol{\beta}_0$ can be taken as zero and g is an overdispersion parameter to be estimated or manually tuned. The parameter g can be set according to various common model selection methods such as AIC, BIC and RIC [27]. [26, 28] suggested the use of the prior distribution on g as a fully Bayesian method (see also [29]). Various other methods have been recommended in the literature for finding an optimal value for g, such as the empirical Bayes methods [30, 31]. Choosing the number of clusters according to the marginal probability distribution is analogous to using the BIC criterion in a model selection problem if $g = N$ where N is the number of the observations.

Denote by $\mathcal{M}$ the set of all possible models. The set $\mathcal{M}$ contains all the possible ways that one can assign D distinguishable shapes into $D, D-1, D-2, \ldots, 1$ indistinguishable clusters. The cardinality of $\mathcal{M}$ is $\sum_{i=1}^{D} \left\{ {D \atop i} \right\}$ where $\left\{ {n \atop k} \right\}$ is the Stirling number of the second kind. The sum is equal to the Bell number $\mathcal{B}_D$.

The testing formulation of model selection is to test the following hypothesis

$$H_0 : \boldsymbol{\beta}_{j_1} = \boldsymbol{\beta}_{j_2} = \ldots = \boldsymbol{\beta}_{j_k} = \mathbf{0}, \quad \text{vs.} \quad H_1 : \boldsymbol{\beta}_{j_1} \neq \mathbf{0}, \ldots, \boldsymbol{\beta}_{j_k} \neq \mathbf{0}, \tag{9}$$

for $1 \leq j_1 < j_2 < \ldots < j_k \leq D$. Clustering, in contrast, is concerned with finding different ways of combining the covariates. Clustering may then be formulated as testing the hypothesis

$$H_0 : \mathbf{T}\boldsymbol{\beta} = \mathbf{0}, \quad \text{vs.} \quad H_1 : \mathbf{T}\boldsymbol{\beta} \neq \mathbf{0}, \tag{10}$$

where $\mathbf{T}_{qK\times DK}$ is the matrix of constraints such that $\mathbf{T}\boldsymbol{\beta}=\mathbf{0}$ is a set of q linear constraints on $\boldsymbol{\beta}_j$ for $1\leq j\leq D$. The following theorem provides an approximation of the marginal likelihood which we call CLUSBIC.

Theorem 1. *Suppose $p(\mathbf{y}\mid\boldsymbol{\beta},\mathbf{X},\mathbf{d},\sigma^2)$ is a C^3 function of $\boldsymbol{\beta}$. Then as $N_i\to\infty$, $\forall i$, we have*

$$-2\log\left[\frac{p(\mathbf{y}\mid\boldsymbol{\beta},\mathbf{X},\mathbf{d},\sigma^2)}{p(\mathbf{y}\mid\tilde{\boldsymbol{\beta}},\mathbf{X},\mathbf{d},\sigma^2)}\right]=K\mathcal{C}(\mathbf{d})\log(\sum_{i=1}^{\mathcal{C}(\mathbf{d})}N_i)+\mathcal{O}_p(||\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}||^3),$$

where $\tilde{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$ under the null hypothesis, $H_0:\mathbf{T}\boldsymbol{\beta}=\mathbf{0}$.

Proof. The proof is given in Section A of the Appendix. □

Theorem 1 shows that for large sample sizes our proposed clustering method using the marginal likelihood has the same premise as Ward's linkage [32]. Ward's linkage, $\Delta(A,B)$, is a popular method in hierarchical clustering which is based on minimizing variance after the merge. Ward's method merges two clusters A and B that minimizes the sum of squares after the merge.

Using CLUSBIC, two groups, say group i and group j, are merged together if it leads to a decrease in the CLUSBIC. The CLUSBIC is calculated for different combination of shapes. A pair of groupings that minimizes CLUSBIC is a merging candidate. For example, groups i and j are merged together in the second level of the hierarchy only if the merge decreases the total CLUSBIC compared to the level before. The following calculation shows that this idea of merging the groups is very much aligned with likelihood maximization.

$$\begin{aligned}\Delta_{21}(\text{CLUSBIC})&=\text{CLUSBIC}^{(2)}-\text{CLUSBIC}^{(1)}\\&=\left[-2\{\ell_1+\ell_2+\ldots+\ell_{(ij)}+\ldots+\ell_{\mathcal{C}(\mathbf{d})}\}+K\sum_{q=1}^{\mathcal{C}(\mathbf{d})-1}N_q\log(\sum_{q=1}^{\mathcal{C}(\mathbf{d})-1}N_q+1)\right]\\&-\left[-2\{\ell_1+\ell_2+\ldots+\ell_i+\ldots+\ell_j+\ldots+\ell_D\}\quad-K\sum_{i=1}^{\mathcal{C}(\mathbf{d})}N_i\log(\sum_{i=1}^{\mathcal{C}(\mathbf{d})}N_i+1)\right],\end{aligned}$$

where $\ell_i=\log\{p(\mathbf{y}_i\mid\boldsymbol{\beta}_i,\mathbf{X},\sigma^2)\}$ and $\ell_{(ij)}$ is the log-likelihood after merging group i with group j. As the total number of observations is fixed at each level of hierarchy, i,e., $\sum_{i=1}^{\mathcal{C}(\mathbf{d})}N_i=\sum_{q=1}^{\mathcal{C}(\mathbf{d})-1}N_q$,

$$\Delta_{21}(\text{CLUSBIC})=-2\{\ell_{(ij)}-(\ell_i+\ell_j)\}-k\log(\sum_{i=1}^{\mathcal{C}(\mathbf{d})}N_i+1).$$

For Gaussian models, Ward's linkage is closely related to CLUSBIC. Minimizing the CLUSBIC between each two consecutive levels of the dendrogram leads to minimizing the metric, $c_{ij}=\ell_i+\ell_j-\ell_{(ij)};\quad i,j=1,2,\ldots D$. Now, considering the model $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$,

$$\log\left\{p(\mathbf{y}_i\mid\hat{\boldsymbol{\beta}}_i,\mathbf{X},\sigma^2)\right\}=-\frac{N_i}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}||\mathbf{y}_i-\hat{\mathrm{E}}(\mathbf{y}_i)||_2^2,$$

where $\hat{\mathrm{E}}(\mathbf{y}_i)=\mathbf{X}\hat{\boldsymbol{\beta}}_i$. Substituting the above equation in the c_{ij} metric, we have

$$\begin{aligned}c_{ij}&=\frac{1}{2\sigma^2}\{||\mathbf{y}_{(ij)}-\hat{\mathrm{E}}(\mathbf{y}_{(ij)})||_2^2-||\mathbf{y}_i-\hat{\mathrm{E}}(\mathbf{y}_i)||_2^2-||\mathbf{y}_j-\hat{\mathrm{E}}(\mathbf{y}_j)||_2^2\},\\&\approx\Delta(i,j)\end{aligned}$$

When $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is a diagonal matrix, c_{ij} equals $\Delta(i,j)$.

1.3 Consistency of CLUSBIC

In this section, we show that the CLUSBIC criterion is consistent in choosing the true clustering as $N_i \to \infty$ for $i = 1, 2, \ldots, \mathcal{C}(\mathbf{d})$. Assuming that there exists a model $m_0 \in \mathcal{M}$ that represents the true clustering, the CLUSBIC, developed in Theorem 1, is said to be a consistent criterion if

$$\lim_{N\to\infty} p_N(m_0) = \lim_{N\to\infty} p_N(\hat{m} = m_0) = 1, \tag{11}$$

where $\hat{m}$ is the model selected by CLUSBIC.

In a regression setting, the space of models $\mathcal{M}$ can be partitioned into two sub-spaces of under-specified and over-specified models in a rather straightforward fashion. The space of under-specified models $\mathcal{M}_1$ contains all models that mistakenly exclude the attributes of the true models. On the other hand, the space of over-specified models $\mathcal{M}_2$ contains all models that include more attributes besides the true model's attributes. In other words, the sub-spaces $\mathcal{M}_1$ and $\mathcal{M}_2$ can be effortlessly established considering the presence or the absence of the attributes of the true model. Therefore, more formally, we have

$$\mathcal{M}_1 = \{m \in \mathcal{M} \mid m \not\supseteq m_0\}, \quad \text{and} \quad \mathcal{M}_2 = \{m \in \mathcal{M} \mid m \supseteq m_0\}.$$

Consequently, for each model that belongs to $\mathcal{M}_2$, the dimension of the model, K, is always greater than the true model.

The definition of under-specified and over-specified models needs to be adjusted for the clustering context. By definition, the over-specified models must include the attributes of the true model. The question that arises here is how the notion of attributes can be interpreted for the clustering problem. Suppose that each column of $\mathbf{X}$ represents an attribute; more clusters mean more attributes. There is some confusion as to whether the over-specified models should include the clusters of the true model (i.e., smaller number of clusters compared to the true model) or disjoint the clusters of the true model (i.e., having more number of clusters).

Consider the general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ in equation 6. Let $N = \sum_{i=1}^{\mathcal{C}(\mathbf{d})} N_i$ and $K_{\mathbf{d}} = K\mathcal{C}(\mathbf{d})$ be the total number of observations and the number of parameters in the model respectively. One can approach the problem of defining the model space of $\mathcal{M}_2$ from the hypothesis testing point of view (10). In this approach, the over-specified model is the one that the null hypothesis for that model holds true considering the assumptions on the true model. More formally, we define $\mathcal{M}_1$ and $\mathcal{M}_2$ as follows for the clustering problem,

$$\mathcal{M}_1 = \{m \in \mathcal{M} \mid \mathbf{T}_m\boldsymbol{\beta} \neq \mathbf{0}\}, \quad \text{and} \quad \mathcal{M}_2 = \{m \in \mathcal{M} \mid \mathbf{T}_m\boldsymbol{\beta} = \mathbf{0}\},$$

where $\mathbf{T}_m$ is the matrix of constraints for the model $m \in \mathcal{M}$. To establish the consistency of CLUSBIC, we need to prove the following statements,

a) $\lim_{N\to\infty} p_N(m) = 0$ for $m \in \mathcal{M}_1$.

b) $\lim_{N\to\infty} p_N(m) = 0$ for $m \in \mathcal{M}_2 - \{m_0\}$.

We in fact prove a stronger version of the first statement (a). We'll show that

a*) $\lim_{N\to\infty} N^h p_N(m) = 0$ for $m \in \mathcal{M}_1$, and for any positive h.

The consistency of the BIC in model selection has been developed in the literature considering different assumptions, see [33, 34, 35]. Estimating the number of parameters under the null hypothesis in (10) leads to constrained least squares. Consequently,

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}} - (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{T}[\mathbf{T}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{T}^{\mathrm{T}}]^{-1}\mathbf{T}\hat{\boldsymbol{\beta}},\\
\hat{\mathbf{y}} &= \mathbf{X}\tilde{\boldsymbol{\beta}}\\
&= \{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{T}^{\mathrm{T}}[\mathbf{T}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{T}^{\mathrm{T}}]^{-1}\mathbf{T}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y}\\
&= \{\mathbf{Q}_1 - \mathbf{Q}_2\}\mathbf{y}\\
&= \mathbf{Q}\mathbf{y}.
\end{aligned}$$

The matrix $\mathbf{Q}$ is the projection matrix which is symmetric and idempotent.

In the following theorem, we show that CLUSBIC is a consistent clustering measure for any arbitrary distribution on model (4). To prove the consistency of CLUSBIC, we rely on a quadratic approximation to the logarithm of the likelihood suggested by [36]. It can be shown that results can be exact for Gaussian models, see Section B.

Theorem 2. *The* CLUSBIC *is a consistent clustering criterion.*

Proof. The proof is given in Section B of the Appendix. □

2 Application

In this section, we apply our proposed method to the biological cell data obtained from the Murphy lab [1] ([37]). The database includes 3D images from HeLa cell line captured by a laser-scanning microscope. For this study, we consider images which are labeled as monoclonal antibody against an outer membrane protein of mitochondria. There are fifty data folders each representing the data from a distinct cell. Each folder contains four sub-folders and the data corresponding to the cell and crop image folders are used for this study. The cell folder has various images of a specific cell, taken at different depths called the confocal plane.

To better illustrate our proposed method of clustering, we start by an example of shape clustering in a two-dimensional (2D) space. For this purpose, a single stack, common over all cells, is selected. The chosen stack, thereafter, is segmented using the designed crop image such that there is only one cell per image. To obtain a true representation of a shape from each image, location, scale, and rotational effects should be filtered out. For this purpose, we aligned cells such that their centroids are located in the centre of images. In addition, the cells are rotated in the direction of their main principal component axes to assure a rotation-free analysis. Consequently, the coordinate of pixels on the boundary of the cell is extracted for modeling. We use MATLAB standard methods including segmentation, boundary detection and Savitzky–Golay smoothing filter functions from MATLAB toolboxes [38, 39] for detecting the boundary of the cell. We call the associated line the *oracular boundary* since it, supposedly, represents the true boundary for each cell. Besides, we propose another method for boundary detection which enables us to take into account the associated uncertainty.

For each observation on the oracular boundary, the uncertainty (variance) is calculated based on the data points on the lower and the upper boundaries in polar coordinates. Lower and upper boundaries are treated as a 95% confidence interval and

$$\hat{\sigma}_i \approx \frac{\mathrm{UCL}_i - \mathrm{LCL}_i}{4}, \quad \text{for} \quad i = 1, 2, \ldots, N,$$

gives an estimation of the standard deviation for point i, where UCL_i and LCL_i represent the lower boundary and upper boundary values at point i respectively. Then the median of the $\hat{\sigma}_i$ is treated as the common standard deviation to be used for all the computations throughout our clustering algorithm. A Gaussian sample centered around each observation on the oracular boundary with the so-called common variance is generated.

Note that in the case of 2D shape modeling, one can employ basis functions such as Fourier, splines or wavelets and then follow the same procedure for shape clustering as in Section 1. For the sake of readability, in Figure 1, the dendrogram is reported for only ten random cells. Each cell is assigned a number corresponding to its order in the database.

In order to obtain the 3D Cartesian coordinates associated to each voxel on the surface, we do as follows. First, the boundary data for each image stack is extracted separately. Second, the 2D coordinates of each stack are combined all together to create the 3D coordinates of the cell shapes depicted in Figure 2. We then take the image spacing information into acccount.

[1] http://murphylab.web.cmu.edu/data/#3DHeLa

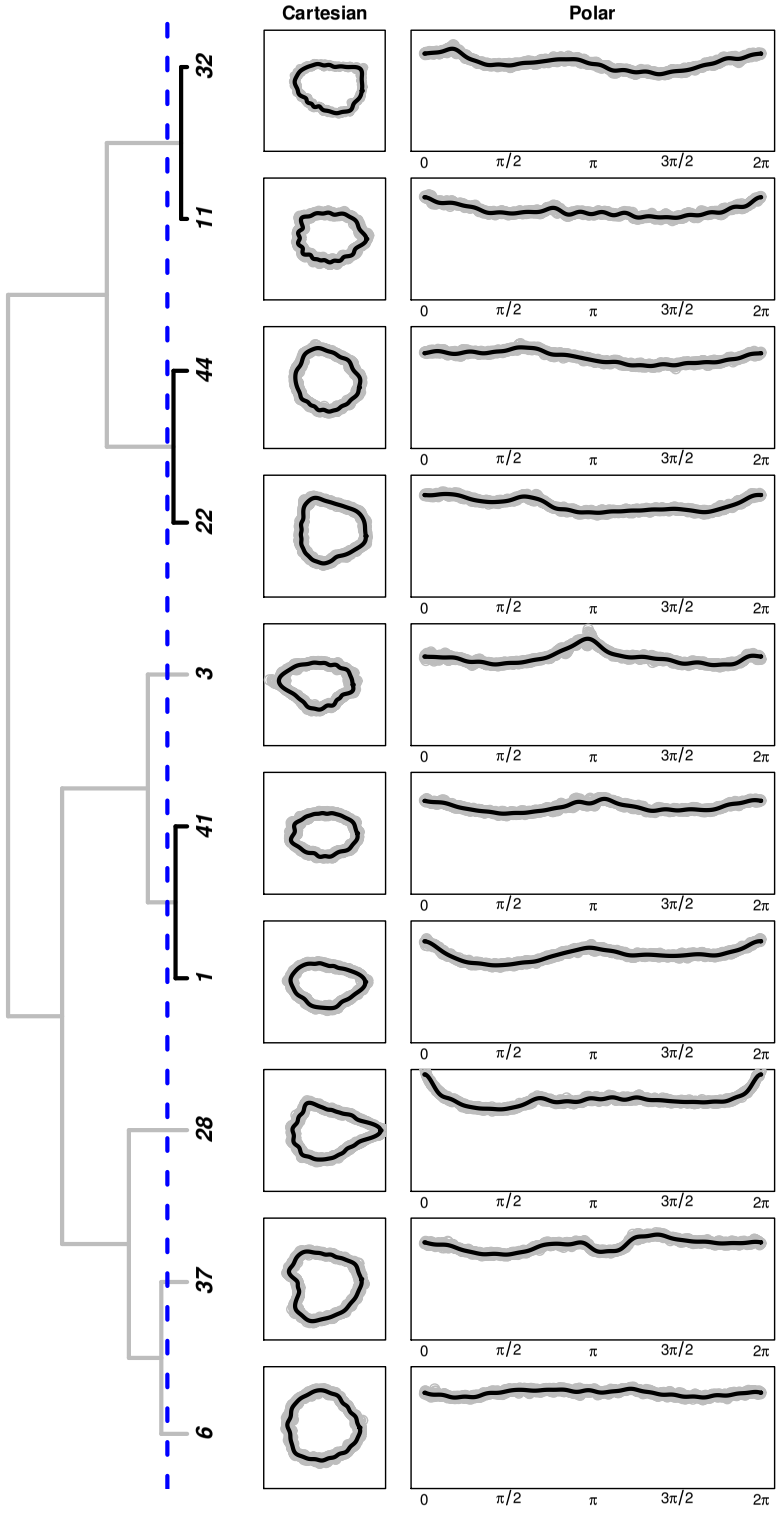

Fig 1. Top panel, dendrogram of the marginal likelihood associated with each cell using Fourier basis functions with $K = 33$ expansion terms. In this example, $D = 10$, $N_i = 1000$ for $i = 1, 2, \ldots, 10$, $d = (1, 3, 6, 11, 22, 28, 32, 37, 41, 44)$ and $\mathcal{C}(\mathbf{d}) = 10$. Black lines, in dendrogram, represent the improvement in the marginal likelihood and gray lines depict the deterioration in the marginal likelihood. The dashed blue line indicates the maximum a posteriori cutting point for the dendrogram, $d = (1, 3, 6, 11, 22, 28, 11, 37, 1, 22)$ and $\mathcal{C}(\mathbf{d}) = 7$. Middle panel, the fitted curves to each of ten random selected cells used in dendrogram in 2D space. The gray points represent the boundary data used in modeling. Bottom panel, the same curves as the middle panel are depicted a in 1D space.

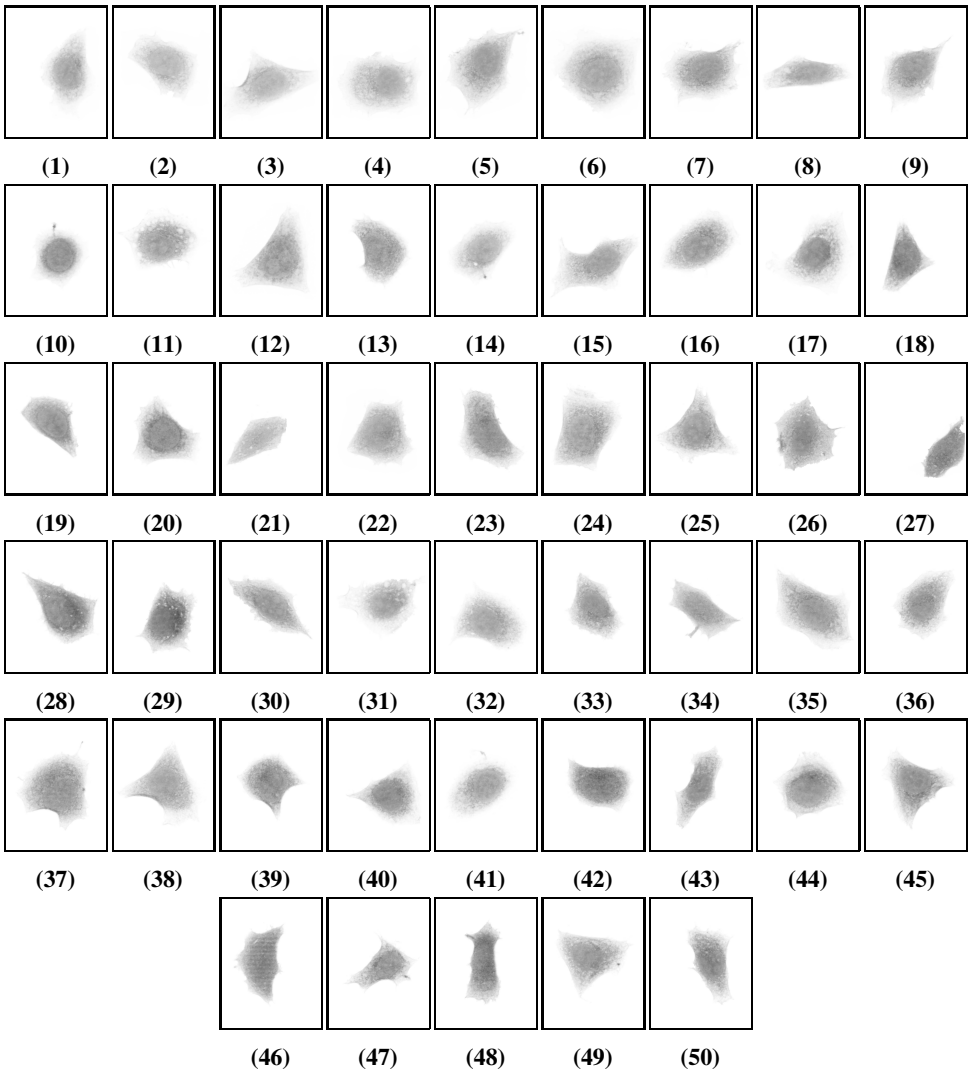

Fig 2. The raw images of fifty cells(D=50) used for clustering throughout this work. The number assigned to each cell matches with its order in the dataset.

For this dataset, the voxel spacing is $(0.049\mu m, 0.049\mu m, 0.203\mu m)$ [40]. It should be noted that the final extracted data must be within a sphere of unit radius to be suitable for modeling using spherical harmonics.

As image data involve noise, the least squares equation (3) is ill posed. A regularization approach can then be taken to estimate the model parameters. See [41] for details and further discussion. The result of clustering for $D = 10$ random cells is reported in Figure 3.

We repeat the same procedure considering all $D = 50$ cells. The clustering result is reported in Figure 4.

As we discussed in Section 1.2, the number of all possible groupings is $\sum_{k=1}^{50} \left\{ {50 \atop k} \right\} \approx 10^{47}$. In practice, it is not feasible to explore all possible groupings when D is relatively large. We ran the random Gibbs sampling for 8000 cycles as an example. The convergence behavior of the sampling throughout the 8000 cycles is reported in Figure 5.

The grouping generated from random Gibbs sampling after the 8000 cycles is as follows,

$$\text{Cluster } 1 = \{4, 12, 33, 38\}, \text{Cluster } 2 = \{1, 13, 26, 48\}, \text{Cluster } 3 = \{5, 8, 11, 21, 22, 32\},$$

Fig 3. Top panel, dendrogram of the marginal likelihood associated with each cell using spherical harmonics with $L_{\max} = 12$. In this example, $D = 10$, $N_i = 2000$ for $i = 1, 2, \ldots, 10$, $d = (1, 3, 6, 11, 22, 28, 32, 37, 41, 44)$ and $\mathcal{C}(\mathbf{d}) = 10$. Black lines represent the improvement in the marginal likelihood and gray lines depict the deterioration in the marginal likelihood. The dashed blue line indicates the maximum a posteriori cutting point for the dendrogram, $d = (1, 3, 6, 1, 22, 22, 3, 6, 41, 41)$ and $\mathcal{C}(\mathbf{d}) = 5$. Bottom panel, the 3D data and the corresponding fit in the Cartesian coordinates.

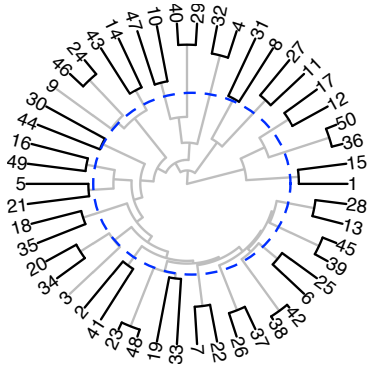

Fig 4. The dendrogram of the marginal likelihood associated with $D = 50$ cells using spherical harmonics with $L_{\max} = 12$. Black lines represent the improvement in the marginal likelihood and gray lines depict the deterioration in the marginal likelihood. The dashed blue circle indicates the maximum a posteriori cutting point for the dendrogram.

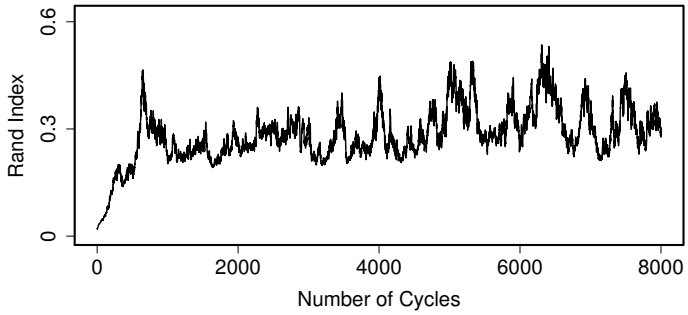

Fig 5. The result of random Gibbs sampling for the same cells as in Figure 4. The Rand index between the grouping suggested at each cycle of random Gibbs sampling with the grouping produced by the dendrogram in Figure 4.

$$\text{Cluster } 4 = \{2, 3, 14, 18, 20, 24, 27, 29, 30, 34, 35, 40, 43, 44, 45, 46, 47\},$$

$$\text{Cluster } 5 = \{6, 7, 9, 10, 15, 17, 19, 23, 25, 28, 31, 36, 37, 39, 41, 42, 49, 50\}.$$

3 Conclusion

Having regarded the surface of a shape as a continuous function, rather than discrete landmarks, we have proposed a simple method for surface modelling of shapes such as biological cells. We have also proposed a new information criterion, called CLUSBIC, for model-based clustering and have shown that the proposed criterion is consistent.

In this work, we considered the Gaussian conjugate priors to favor computational simplicity. We proved the consistency of CLUSBIC in clustering. Note that in our settings, the increase in number of observations N does not necessarily imply the increase in the number of clusters $\mathcal{C}(\mathbf{d})$ contrary to the classical clustering problem. Therefore, the consistency of CLUSBIC remains valid.

Investigation of the physical structure of cells, as simple closed shapes, can be highly useful in biology, specifically for the diagnosis of cancer. The result, in this preliminary work,

shows that our proposed methodology is quite applicable and can produce promising results.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada through Discovery Grants to M. Asgharian (NSERC RGPIN 217398-13).

A

Proof of Theorem 1. Consider the marginal posterior for a set of D shapes

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{d}, \sigma^2) = \int_{\boldsymbol{\beta}} p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{X}, \mathbf{d}, \sigma^2) p(\boldsymbol{\beta} \mid \mathbf{d}, \sigma^2) d\boldsymbol{\beta}.$$

For simplicity, we use the notation $p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)$ instead of $p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{X}, \mathbf{d}, \sigma^2)$. In order to obtain an approximation to this integral, we take a second order Taylor expansion of the log-likelihood at $\tilde{\boldsymbol{\beta}}$, the solution to the following constrained optimization problem,

$$\max \log\{p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)\}, \quad \text{subject to} \quad \mathbf{T}\boldsymbol{\beta} = \mathbf{0}. \tag{12}$$

The solution $\tilde{\boldsymbol{\beta}}$ can be found using the method of Lagrange multipliers. The Lagrangian function for this problem is

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \log\{p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)\} + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{T} \boldsymbol{\beta}, \tag{13}$$

where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers. Expanding $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda})$ about $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\lambda}}$.

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathcal{L}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\lambda}}) &+ \begin{bmatrix}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^{\mathrm{T}} & (\boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}})^{\mathrm{T}}\end{bmatrix} \begin{bmatrix} \frac{\partial \log\{p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\}}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} + \mathbf{T}^{\mathrm{T}}\boldsymbol{\lambda} \\ \mathbf{T}\tilde{\boldsymbol{\beta}} \end{bmatrix} \\ &+ \frac{1}{2}\begin{bmatrix}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^{\mathrm{T}} & (\boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}})^{\mathrm{T}}\end{bmatrix} \begin{bmatrix} \frac{\partial^2 \log\{p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}} & \mathbf{T}^{\mathrm{T}} \\ \mathbf{T} & \mathbf{0} \end{bmatrix} \begin{bmatrix}\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}} \\ \boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}}\end{bmatrix} \qquad (14) \\ &+ \mathcal{O}_p(||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||^3). \qquad (15)\end{aligned}$$

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \log\{p(\mathbf{y} \mid \tilde{\boldsymbol{\beta}}, \sigma^2)\} &+ (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^{\mathrm{T}} \frac{\partial^2 \log\{p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)\}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}\Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \\ &+ 2(\boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}})^{\mathrm{T}} \mathbf{T}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + \mathcal{O}_p(||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||^3). \qquad (16)\end{aligned}$$

Under the assumption that $\mathbf{T}\boldsymbol{\beta} = \mathbf{0}$,

$$\begin{aligned}\log\{p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)\} = \log\{p(\mathbf{y} \mid \tilde{\boldsymbol{\beta}}, \sigma^2)\} &+ (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^{\mathrm{T}} \frac{\partial^2 \log\{p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)\}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}\Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \\ &+ \mathcal{O}_p(||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||^3). \qquad (17)\end{aligned}$$

Defining the average observed Fisher information matrix as

$$\bar{\mathbf{J}}(\tilde{\boldsymbol{\beta}}, \mathbf{y}) = -\frac{1}{\sum_{i=1}^{\mathcal{C}(\mathbf{d})} N_i} \frac{\partial^2 \log\{p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)\}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}\Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}},$$

we have

$$\int_{\boldsymbol{\beta}} p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta} \mid \sigma^2) d\boldsymbol{\beta} = p(\mathbf{y} \mid \tilde{\boldsymbol{\beta}}, \sigma^2) \int_{\boldsymbol{\beta}} \exp\{-\frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^{\mathrm{T}} \sum_{i=1}^{\mathcal{C}(\mathbf{d})} N_i \bar{\mathbf{J}}(\tilde{\boldsymbol{\beta}}, \mathbf{y})(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\} \times p(\boldsymbol{\beta} \mid \sigma^2) d\boldsymbol{\beta} + \mathcal{O}_p(||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||^3).$$

Considering $\boldsymbol{\beta} \sim \mathcal{N}(\tilde{\boldsymbol{\beta}}, \bar{\mathbf{J}}^{-1}(\tilde{\boldsymbol{\beta}}, \mathbf{y}))$ where $\bar{\mathbf{J}}^{-1}(\tilde{\boldsymbol{\beta}}, \mathbf{y}) = \sigma^2 (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \sum_{i=1}^{\mathcal{C}(\mathbf{d})} N_i$,

$$\begin{aligned}
\int_{\boldsymbol{\beta}} p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta} \mid \sigma^2) d\boldsymbol{\beta} &= p(\mathbf{y} \mid \tilde{\boldsymbol{\beta}}, \sigma^2) \int_{\boldsymbol{\beta}} (2\pi)^{-\frac{K\mathcal{C}(\mathbf{d})}{2}} |\bar{\mathbf{J}}(\tilde{\boldsymbol{\beta}}, \mathbf{y})|^{\frac{1}{2}} \\
&\quad \times \exp\{-\frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathrm{T}} (\sum_{i=1}^{\mathcal{C}(\mathbf{d})} N_i + 1) \bar{\mathbf{J}}(\tilde{\boldsymbol{\beta}}, \mathbf{y})(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\} d\boldsymbol{\beta} + \mathcal{O}_p(||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||^3) \\
&= p(\mathbf{y} \mid \tilde{\boldsymbol{\beta}}, \sigma^2) |\bar{\mathbf{J}}(\tilde{\boldsymbol{\beta}}, \mathbf{y})|^{\frac{1}{2}} |(\sum_{i=1}^{\mathcal{C}(\mathbf{d})} N_i + 1) \bar{\mathbf{J}}(\tilde{\boldsymbol{\beta}}, \mathbf{y})|^{-\frac{1}{2}} + \mathcal{O}_p(||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||^3) \\
&= p(\mathbf{y} \mid \tilde{\boldsymbol{\beta}}, \sigma^2) (\sum_{i=1}^{\mathcal{C}(\mathbf{d})} N_i + 1)^{-\frac{K\mathcal{C}(\mathbf{d})}{2}} |\bar{\mathbf{J}}(\tilde{\boldsymbol{\beta}}, \mathbf{y})|^{\frac{1}{2}} |\bar{\mathbf{J}}(\tilde{\boldsymbol{\beta}}, \mathbf{y})|^{-\frac{1}{2}} + \mathcal{O}_p(||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||^3).
\end{aligned}$$

The desired result then follows upon noticing that $\tilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ [42] and $\log(\sum_{i=1}^{\mathcal{C}(\mathbf{d})} N_i) \approx \log(\sum_{i=1}^{\mathcal{C}(\mathbf{d})} N_i + 1)$ for large value of $\sum_{i=1}^{\mathcal{C}(\mathbf{d})} N_i$. □

B

Proof of Theorem 2. The proof comprises two steps. First, the consistency of CLUSBIC is established for Gaussian models. We then extend the result to smooth non-Gaussian models where by "smooth " we generally mean the likelihood is a C^3 function of the unknown parameter. The second step essentially follows from step one upon applying a quadratic approximation to the logarithm of the likelihood.

Step 1. Gaussian Model:

Suppose the error terms, (4), are distributed according to the Gaussian distribution, one can easily show that the CLUSBIC has the following form, similar to the BIC, see [43],

$$\begin{aligned}
\text{CLUSBIC}(m) &= N \log \hat{\sigma^2}(m) + K_{\mathbf{d}}(m) \log N \\
&= N \log \frac{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \{\mathbf{I} - \mathbf{Q}(m)\} \mathbf{X} \boldsymbol{\beta} + \mathbf{e}^{\mathrm{T}} \{\mathbf{I} - \mathbf{Q}(m)\} \mathbf{e} + 2\mathbf{e}^{\mathrm{T}} \{\mathbf{I} - \mathbf{Q}(m)\} \mathbf{X} \boldsymbol{\beta}}{N} \\
&\quad + K_{\mathbf{d}}(m) \log N. \qquad (18)
\end{aligned}$$

The estimate of the variance matrix for model m is,

$$\hat{\sigma^2}(m) = \frac{\mathbf{y}^{\mathrm{T}} \{\mathbf{I} - \mathbf{Q}(m)\} \mathbf{y}}{N}.$$

Suppose $K_{\mathbf{d}}$ and D are fixed, we follow the same setting as in [34] by modifying constraints of the form $\mathbf{T}\boldsymbol{\beta} = \mathbf{0}$. The following two assumptions are required for the proof,

1. $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is positive definite.
2. $\mathbf{H} = \lim_{N \to \infty} \frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{N}$ is positive definite.

The validity of these two assumptions relies on the validity of the following two assumptions for all models, i.e. $\forall i \in \{1, 2, \ldots, D\}$

1. $\mathbf{H}_i^{\mathrm{T}}\mathbf{H}_i$ is positive definite.

2. $\mathbf{H}_i = \lim_{N_i\to\infty} \frac{\mathbf{H}_i^{\mathrm{T}}\mathbf{H}_i}{N_i}$ is positive definite.

Lemma 1. *For $m \in \mathcal{M}_1$, and for any positive h, $\lim_{N\to\infty} N^h p_N(m) = 0$, using* CLUSBIC.

Proof. The proof is given in Supplementary Material. □

Lemma 2. *Here, we follow a similar approach as in Lemma 1. For $m \in \mathcal{M}_2 - \{m_0\}$, $\lim_{N\to\infty} p_N(m) = 0$, using* CLUSBIC.

Proof. The proof is given in Supplementary Material. □

Having established the above lemmas, the following theorem can be established for Gaussian models.

Theorem 3. *The* CLUSBIC *is a consistent clustering measure for Gaussian models.*

Proof. The proof is given in Supplementary Material. □

Step 2. non-Gaussian Model:
First note that for a general smooth likelihood we have

$$\mathrm{CLUSBIC}(m) = -2\log\{p_m(\mathbf{y} \mid \tilde{\boldsymbol{\beta}})\} + K_{\mathbf{d}}(m)\log(N).$$

We need to show that $\lim_{N\to\infty} p_N(m) = 0$ for $m \in \mathcal{M}_1$ and $m \in \mathcal{M}_2 - \{m_0\}$ or equivalently,

$$\lim_{N\to\infty} p[\mathrm{CLUSBIC}(m_0) < \mathrm{CLUSBIC}(m)] = 1,$$

for $m \in \mathcal{M}_1$ and $m \in \mathcal{M}_2 - \{m_0\}$. We now note that

$$\begin{aligned}\mathrm{CLUSBIC}(m_0) - \mathrm{CLUSBIC}(m) &= -2\log\{p_{m_0}(\mathbf{y} \mid \tilde{\boldsymbol{\beta}})\} + K_{\mathbf{d}}(m_0)\log(N) + 2\log\{p_m(\mathbf{y} \mid \tilde{\boldsymbol{\beta}})\} \\ &\quad - K_{\mathbf{d}}(m)\log(N) \\ &= -2[\log\{p_{m_0}(\mathbf{y} \mid \tilde{\boldsymbol{\beta}})\} - \log\{p_m(\mathbf{y} \mid \tilde{\boldsymbol{\beta}})\}] + [K_{\mathbf{d}}(m_0) - K_{\mathbf{d}}(m)]\log(N). \qquad (19)\end{aligned}$$

For any $m \in \mathcal{M}$ and the true value of parameter $\boldsymbol{\beta}_0$, one can write the following decomposition

$$\begin{aligned}\log\{p_m(\mathbf{y} \mid \tilde{\boldsymbol{\beta}})\} &= \underbrace{\log\{p_m(\mathbf{y} \mid \tilde{\boldsymbol{\beta}})\} - \log\{p_m(\mathbf{y} \mid \boldsymbol{\beta}_0)\}}_{①} \\ &\quad + \underbrace{\log\{p_m(\mathbf{y} \mid \boldsymbol{\beta}_0)\} - N\mathrm{E}\left(\log\{p_m(y \mid \boldsymbol{\beta}_0)\}\right)}_{②} \\ &\quad + \underbrace{N\mathrm{E}\left(\log\{p_m(y \mid \boldsymbol{\beta}_0)\}\right)}_{③}. \qquad (20)\end{aligned}$$

Applying the second order Taylor expansion to $\log\{p_m(\mathbf{y} \mid \boldsymbol{\beta}_0)\}$ at the point $\tilde{\boldsymbol{\beta}}$, equation (17),

$$\begin{aligned}\log\{p_m(\mathbf{y} \mid \tilde{\boldsymbol{\beta}})\} - \log\{p_m(\mathbf{y} \mid \boldsymbol{\beta}_0)\} &= -\frac{1}{2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}} \frac{\partial^2 \log\{p_m(\mathbf{y} \mid \boldsymbol{\beta})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ &= -\frac{1}{2}\sqrt{N}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}} \frac{1}{N} \frac{\partial^2 \log\{p_m(\mathbf{y} \mid \boldsymbol{\beta})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}} \\ &\quad \times \sqrt{N}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).\end{aligned}$$

Under the usual regularity conditions, see [44, page 209]), we have

$$\frac{1}{N}\frac{\partial^2 \log\{p_m(\mathbf{y} \mid \boldsymbol{\beta})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \xrightarrow{p} \mathrm{E}\{\frac{\partial^2 \log\{p_m(\mathbf{y} \mid \boldsymbol{\beta})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\} = -\mathrm{I}(\boldsymbol{\beta}_0), \tag{21}$$

$$\frac{1}{\sqrt{N}}\frac{\partial \log\{p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)\}}{\partial\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathrm{I}(\boldsymbol{\beta}_0)) \tag{22}$$

where $\mathrm{I}(\boldsymbol{\beta}_0)$ is the Fisher information matrix. On the other hand, by the definition of constrained optimization problem equations (12), and (13),

$$\begin{bmatrix} \frac{\partial \log\{p(\mathbf{y}|\boldsymbol{\beta})\}}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} + \mathbf{T}^{\mathrm{T}}\boldsymbol{\lambda} \\ \mathbf{T}\tilde{\boldsymbol{\beta}} \end{bmatrix} = \mathbf{0}. \tag{23}$$

Expanding the score function around the point $\boldsymbol{\beta}_0$, and $\boldsymbol{\lambda}_0 = \mathbf{0}$, we have

$$\begin{aligned}
\mathbf{0} &= \begin{bmatrix} \frac{\partial \log\{p(\mathbf{y}|\boldsymbol{\beta})\}}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} + \mathbf{T}^{\mathrm{T}}\boldsymbol{\lambda}_0 \\ \mathbf{T}\boldsymbol{\beta}_0 \end{bmatrix} + \begin{bmatrix} \frac{\partial^2 \log\{p(\mathbf{y}|\boldsymbol{\beta})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\Big|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}} & \mathbf{T}^{\mathrm{T}} \\ \mathbf{T} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0 \end{bmatrix} \\
\sqrt{N}\begin{bmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \tilde{\boldsymbol{\lambda}} \end{bmatrix} &= -\sqrt{N}\begin{bmatrix} \frac{\partial^2 \log\{p(\mathbf{y}|\boldsymbol{\beta})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\Big|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}} & \mathbf{T}^{\mathrm{T}} \\ \mathbf{T} & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial \log\{p(\mathbf{y}|\boldsymbol{\beta})\}}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \\ \mathbf{0} \end{bmatrix} \\
&= \begin{bmatrix} \left(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{A}^{-1}\right)\frac{1}{\sqrt{N}}\frac{\partial \log\{p(\mathbf{y}|\boldsymbol{\beta})\}}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \\ (\mathbf{T}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{A}^{-1}\frac{1}{\sqrt{N}}\frac{\partial \log\{p(\mathbf{y}|\boldsymbol{\beta})\}}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \end{bmatrix},
\end{aligned}$$

where $\mathbf{A} = \frac{\partial^2 \log\{p(\mathbf{y}|\boldsymbol{\beta})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\Big|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}$. Taking into account equations (21) and (22), one can show that

$$\begin{aligned}
\sqrt{N}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &\xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathrm{I}^{-1}(\boldsymbol{\beta}_0) - \mathrm{I}^{-1}(\boldsymbol{\beta}_0)\mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathrm{I}^{-1}(\boldsymbol{\beta}_0)\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathrm{I}^{-1}(\boldsymbol{\beta}_0)), \\
&\xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}),
\end{aligned} \tag{24}$$

where $\mathrm{I}^{-1}(\boldsymbol{\beta}_0)$ is the inverse of the Fisher information matrix. By equations (24) and (21), and the fact that $\mathrm{I}(\boldsymbol{\beta}_0)\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}}$ is an idempotent matrix, one can conclude that

$$-\frac{1}{2}\sqrt{N}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}}\frac{1}{N}\frac{\partial^2 \log\{p_m(\mathbf{y} \mid \boldsymbol{\beta})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}\sqrt{N}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \frac{1}{2}\chi_{\dim(\boldsymbol{\beta}_0)}, \tag{25}$$

where χ has a chi-squared distribution with $\dim(\mathbf{b}_0)$ degrees of freedom. As the convergence in distribution implies boundedness in probability, component (1) in (20) is of order $\mathcal{O}_p(1)$.

As for component (2) in (20), by the central limit theorem,

$$\frac{1}{\sqrt{N}}\left[\log\{p_m(\mathbf{y} \mid \boldsymbol{\beta}_0)\} - N\mathrm{E}(\log\{p_m(y \mid \boldsymbol{\beta}_0)\})\right] \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}^*), \tag{26}$$

where $\boldsymbol{\Sigma}^* = \mathrm{Var}\left\{\frac{1}{\sqrt{N}}\left[\log\{p_m(\mathbf{y} \mid \boldsymbol{\beta}_0)\} - N\mathrm{E}(\log\{p_m(y \mid \boldsymbol{\beta}_0)\})\right]\right\}$. Accordingly, component (2) in (20) is of order $\mathcal{O}_p(\sqrt{N})$.

Now coming back to the equation (19), for both $m \in \mathcal{M}_1$ and $m \in \mathcal{M}_2 - \{m_0\}$,

$$\begin{aligned}\text{CLUSBIC}(m_0) - \text{CLUSBIC}(m) = -2[\mathcal{O}_p(1) + \mathcal{O}_p(\sqrt{N}) + N[\text{E}\left(\log\{p_{m_0}(y \mid \boldsymbol{\beta}_0)\}\right) \\ - \text{E}\left(\log\{p_m(y \mid \boldsymbol{\beta}_0)\}\right)] + [K_{\mathbf{d}}(m_0) - K_{\mathbf{d}}(m)]\log(N).\end{aligned} \tag{27}$$

Using Jensen's inequality,

$$\begin{aligned}\text{E}\left(\log\{p_m(y \mid \boldsymbol{\beta}_0)\}\right) - \text{E}\left(\log\{p_{m_0}(y \mid \boldsymbol{\beta}_0)\}\right) &= \text{E}\left(\log\{\frac{p_m(y \mid \boldsymbol{\beta}_0)}{p_{m_0}(y \mid \boldsymbol{\beta}_0)}\}\right) \\ &\leq \log\left\{\text{E}\left(\frac{p_m(y \mid \boldsymbol{\beta}_0)}{p_{m_0}(y \mid \boldsymbol{\beta}_0)}\right)\right\} \\ &\leq \log\left\{\int p_m(y \mid \boldsymbol{\beta}_0)dy\right\} \\ &\leq 0\end{aligned}$$

and hence

$$\text{CLUSBIC}(m_0) - \text{CLUSBIC}(m) = -\mathcal{O}(N). \tag{28}$$

Therefore, as N tends to infinity

$$\Pr\{\text{CLUSBIC}(m_0) < \text{CLUSBIC}(m)\} = 1,$$

for models in both $\mathcal{M}_1$ and $\mathcal{M}_2 - \{m_0\}$. This completes the proof. □

C Supplementary Material

Here, we provide some supplementary technical materials useful in the proof of the Lemma 1, Lemma 2, and Theorem 3 in the Section B.

(i) The column space of a matrix $\mathbf{A}$ is denoted by $\text{C}(\mathbf{A})$, and defined as the space spanned by the columns of $\mathbf{A}$.

(ii) The rank of $\mathbf{A}$ is defined to be the dimension of $\text{C}(\mathbf{A})$, $\dim\{\text{C}(\mathbf{A})\}$, i,e, the number of linearly independent columns of $\mathbf{A}$.

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^{\text{T}}) = \text{rank}(\mathbf{A}^{\text{T}}\mathbf{A}) = \text{rank}(\mathbf{A}\mathbf{A}^{\text{T}}).$$

(iii) Orthogonal complement of the sub-space is defined as

$$\mathbf{V}^{\perp} = \{\mathbf{v} \in \mathbf{R}^n \mid \mathbf{v} \perp \mathbf{V}\}.$$

(iv) If $\mathbf{V} \subset \mathbf{W}$, then $\mathbf{V}^{\perp} \cap \mathbf{W} = \{\mathbf{v} \in \mathbf{W} \mid \mathbf{v} \perp \mathbf{V}\}$ is called the orthogonal complement of $\mathbf{V}$ with respect to $\mathbf{W}$.

$$\text{rank}(\mathbf{W}) = \text{rank}(\mathbf{V}) + \text{rank}(\mathbf{V}^{\perp} \cap \mathbf{W}).$$

(v) $\mathbf{Q}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^{\text{T}}\mathbf{X})^{-1}\mathbf{X}^{\text{T}}$ is called projection matrix onto $\text{C}(\mathbf{X})$. The matrix $\mathbf{Q}_{\mathbf{X}}$ is symmetric ($\mathbf{Q}_{\mathbf{X}} = \mathbf{Q}_{\mathbf{X}}^{\text{T}}$) and idempotent ($\mathbf{Q}_{\mathbf{X}}\mathbf{Q}_{\mathbf{X}} = \mathbf{Q}_{\mathbf{X}}$).

(vi) If $\mathbf{Q}_{\mathbf{X}}^1$ and $\mathbf{Q}_{\mathbf{X}}^2$ are projection matrices with $\text{C}(\mathbf{Q}_{\mathbf{X}}^1) \subset \text{C}(\mathbf{Q}_{\mathbf{X}}^2)$, then $\mathbf{Q}_{\mathbf{X}}^2 - \mathbf{Q}_{\mathbf{X}}^1$ is also a projection matrix onto the orthogonal complement of $\text{C}(\mathbf{Q}_{\mathbf{X}}^1)$ with respect to $\text{C}(\mathbf{Q}_{\mathbf{X}}^2)$.

(vii) Let $\mathbf{A}$ be $k \times k$ matrix of constants and $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. If $\mathbf{A}$ is idempotent with rank p, then

$$\frac{\mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{y}}{\sigma^2} \sim \chi^2_{p,\lambda},$$

where $\lambda = \frac{\boldsymbol{\mu}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}}{\sigma^2}$.

of Lemma 1 . In order to prove the lemma, we work with the difference between the CLUSBIC of the true model and any arbitrary model in $\mathcal{M}_1$. We decompose this difference into several random variables. Taking into account the properties of multivariate Gaussian distribution and the quadratic forms from the same family, we proceed with the proof. Let $m \in \mathcal{M}_1$,

$$\begin{aligned} p_N(m) &= \Pr\{\mathrm{CLUSBIC}(m) < \mathrm{CLUSBIC}(m^*); m^* \in \mathcal{M}\} \\ &\leq \Pr\{\mathrm{CLUSBIC}(m) < \mathrm{CLUSBIC}(m_0)\} \\ &= \Pr\{X + Y_N + N^{\frac{1}{2}}c_N - \sigma^2(\lambda_N N)^{-\frac{1}{2}}b_N \leq Z_N\}, \end{aligned} \tag{29}$$

where,

$$\begin{aligned} X &= 2(\lambda_N N)^{-\frac{1}{2}}\mathbf{e}^{\mathrm{T}}\mathbf{Q}^*\mathbf{X}\boldsymbol{\beta}, \\ Y_N &= (\lambda_N N)^{-\frac{1}{2}}\mathbf{e}^{\mathrm{T}}\mathbf{Q}^*\mathbf{e}, \\ Z_N &= b_N(\lambda_N N)^{-\frac{1}{2}}N^{-1}[\mathbf{e}^{\mathrm{T}}\{\mathbf{I} - \mathbf{Q}(m_0)\}\mathbf{e} - \sigma^2 N], \\ c_N &= \lambda_N^{-\frac{1}{2}}N^{-1}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Q}^*\mathbf{X}\boldsymbol{\beta}, \\ \lambda_N &= 4\sigma^2 N^{-1}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Q}^{*^2}\mathbf{X}\boldsymbol{\beta}, \\ b_N &= N\left\{\exp\left(\frac{\log(N)\mathrm{p}}{N}\right) - 1\right\}, \end{aligned}$$

and $\mathrm{p} = K_{\mathbf{d}}(m_0) - K_{\mathbf{d}}(m)$, $\mathbf{Q}^* = \mathbf{Q}(m_0) - \mathbf{Q}(m)$. Since $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, the following properties can be easily verified.

1. $$\mathbf{e}^{\mathrm{T}}\mathbf{Q}^*\mathbf{X}\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2\boldsymbol{\beta}^T\mathbf{X}^{\mathrm{T}}\mathbf{Q}^{*^2}\mathbf{X}\boldsymbol{\beta}) \Longrightarrow X \sim \mathcal{N}(0,1).$$
2. Y_N is a quadratic form from the Gaussian distribution.
3. Since $\{\mathbf{I} - \mathbf{Q}(m_0)\}$ is a symmetric, idempotent matrix,

$$\frac{\mathbf{e}^{\mathrm{T}}\{\mathbf{I} - \mathbf{Q}(m_0)\}\mathbf{e}}{\sigma^2} \sim \chi^2_{N-[D-q(m_0)]K},$$

where χ^2_b is the chi-squared distribution with b degrees of freedom.

The validity of equation (29) can be easily verified through the definition of CLUSBIC in equation (18). By the Fréchet inequality,

$$\max\{0, p(A_1) + \ldots + p(A_n) - (n-1)\} \leq p(A_1 \cap \ldots \cap A_n) \leq \min\{p(A_1), \ldots, p(A_n)\},$$

where A_i's are some events, the equation (29) is bounded by,

$$\Pr(X \leq -N^{\frac{1}{2}}c_N + \sigma^2(\lambda_N N)^{-\frac{1}{2}}b_N + 2N^{\frac{1}{4}}) + p(-Y_N > N^{\frac{1}{4}}) + p(Z_N > N^{\frac{1}{4}}). \tag{30}$$

Using the assumptions 1 and 2,

$$\begin{aligned}\lim_{N\to\infty}\lambda_N &= 4\sigma^2(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\beta})^{\mathrm{T}}\{\mathbf{Q}_H^*\}^2\mathbf{H}^{\frac{1}{2}}\boldsymbol{\beta},\\ &= \lambda,\\ \lim_{N\to\infty}c_N &= \lambda^{-\frac{1}{2}}(\mathbf{H}^{\frac{1}{2}}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{Q}_H^*\mathbf{H}^{\frac{1}{2}}\boldsymbol{\beta}\\ &= c,\\ b_N &= \mathcal{O}(\log N),\end{aligned}$$

where $\mathbf{Q}_H^* = \mathbf{Q}_H(m_0) - \mathbf{Q}_H(m)$, and

$$\begin{aligned}\mathbf{Q}_H(m) = \mathbf{H}^{\frac{1}{2}}(m_0)\{\mathbf{H}(m_0)\}^{-1}\mathbf{H}^{\frac{1}{2}}(m_0) - \mathbf{H}^{\frac{1}{2}}(m_0)\{\mathbf{H}(m_0)\}^{-1}\mathbf{T}^{\mathrm{T}}(\mathbf{T}\{\mathbf{H}(m_0)\}^{-1}\mathbf{T}^{\mathrm{T}})^{-1}\mathbf{T}\\ \{\mathbf{H}(m_0)\}^{-1}\mathbf{H}^{\frac{1}{2}}(m_0), \quad \text{for } m\in\mathcal{M}.\end{aligned}$$

Let $d_N = c_N - 2N^{-\frac{1}{4}} - \sigma^2(\lambda_N)^{-\frac{1}{2}}N^{-1}b_N$. One can easily show that $d_N = \mathcal{O}(1)$ as $\lim_{N\to\infty}c_N = c$ and $\sigma^2(\lambda_N)^{-\frac{1}{2}}N^{-1}b_N = o(1)$. Using the characteristics of the standard Gaussian distribution function,

$$\begin{aligned}\Pr(X \le -N^{\frac{1}{2}}c_N + \sigma^2(\lambda_N N)^{-\frac{1}{2}}b_N + 2N^{\frac{1}{4}}) &= p(X \le -N^{\frac{1}{2}}d_N)\\ &\le N^{-\frac{1}{2}}d_N^{-1}\phi(N^{\frac{1}{2}}d_N)\\ &= o\left(\exp\left\{-\frac{Nc_N^2}{2}\right\}\right),\end{aligned} \tag{31}$$

where $\phi(.)$ is the density function of the standard Gaussian distribution.

Given that Y_N is a quadratic form, using the definition of moment generating functions for quadratic forms [45], we have

$$\begin{aligned}\Pr(-Y_N > N^{\frac{1}{4}}) &\le \exp\{-N^{\frac{1}{4}}\}\mathrm{E}(\exp\{-Y_N\}),\\ &= \exp\{-N^{\frac{1}{4}}\}|\mathbf{I} + 2\sigma^2(N\lambda_N)^{-\frac{1}{2}}\mathbf{Q}^*|^{-\frac{1}{2}}\\ &= o(\exp\{-N^{\frac{1}{4}}\}).\end{aligned} \tag{32}$$

By the property 3,

$$\begin{aligned}\Pr(Z_N > N^{\frac{1}{4}}) &\le \exp\{-N^{\frac{1}{4}}\}\mathrm{E}(\exp\{Z_N\}),\\ &= \exp\{-N^{\frac{1}{4}} - \sigma^2 b_N(\lambda_N N)^{-\frac{1}{2}}\}[1 - 2\sigma^2 b_N\lambda_N^{-\frac{1}{2}}N^{-\frac{3}{2}}]^{-\frac{N-K_{\mathbf{d}}(m_0)}{2}}\\ &= o(\exp\{-N^{\frac{1}{4}}\}).\end{aligned} \tag{33}$$

The equations (31), (32) and (33) complete the proof. □

of Lemma 2. For $m \in \mathcal{M}_2 - \{m_0\}$,

$$\begin{aligned}p_N(m) &\le \Pr\{\mathrm{CLUSBIC}(m) < \mathrm{CLUSBIC}(m_0)\}\\ &= \Pr(\chi \ge N^{-1}b_N\chi_N)\\ &\le \Pr\left(\chi \ge b_N\left[1 - \frac{1}{\sqrt{\log N}}\right]\right) + \Pr\left(\chi_N \le N\left[1 - \frac{1}{\sqrt{\log N}}\right]\right),\end{aligned} \tag{34}$$

where

$$\begin{aligned}\chi &= \frac{N\{\hat{\sigma^2}(m_0) - \hat{\sigma^2}(m)\}}{\sigma^2} = \frac{\mathbf{e}^{\mathrm{T}}\{\mathbf{Q}(m) - \mathbf{Q}(m_0)\}\mathbf{e}}{\sigma^2}\\ \chi_N &= \frac{N\hat{\sigma^2}(m)}{\sigma^2} = \frac{\mathbf{e}^{\mathrm{T}}\{\mathbf{I} - \mathbf{Q}(m)\}\mathbf{e}}{\sigma^2}.\end{aligned}$$

By definition, we have

$$\begin{aligned}\mathbf{Q}(m) - \mathbf{Q}(m_0) &= \mathbf{Q}_1 - \mathbf{Q}_2(m) - [\mathbf{Q}_1 - \mathbf{Q}_2(m_0)] \\ &= \mathbf{Q}_2(m_0) - \mathbf{Q}_2(m) \\ &= \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{T}_{m_0}^{\mathrm{T}}[\mathbf{T}_{m_0}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{T}_{m_0}^{\mathrm{T}}]^{-1}\mathbf{T}_{m_0}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} \\ &\quad - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{T}_{m}^{\mathrm{T}}[\mathbf{T}_{m}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{T}_{m}^{\mathrm{T}}]^{-1}\mathbf{T}_{m}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}.\end{aligned}$$

Under H_0, for an arbitrary model we have, $\mathbf{T}\boldsymbol{\beta} = \mathbf{0}$,

$$\begin{aligned}\implies \mathbf{T}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\boldsymbol{\beta} &= \mathbf{0} \\ \mathbf{T}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\mu} &= \mathbf{0} \\ \mathbf{T}^{*^{\mathrm{T}}}\boldsymbol{\mu} &= \mathbf{0}.\end{aligned}$$

Under H_0, $\boldsymbol{\mu} \in \mathrm{C}(\mathbf{X}) = \mathbf{V}$ and $\boldsymbol{\mu} \perp \mathrm{C}(\mathbf{T}^*)$, or $\boldsymbol{\mu} \in \mathrm{C}(\mathbf{T}^*)^{\perp} \cap \mathrm{C}(\mathbf{X}) = \mathbf{V}_0$ which is the orthogonal complement of $\mathrm{C}(\mathbf{T}^*)$ with respect to $\mathrm{C}(\mathbf{X})$.

$$\mathrm{rank}(\mathbf{T}^*) = \mathrm{rank}(\mathbf{T}^{*^{\mathrm{T}}}) \geq \mathrm{rank}(\mathbf{T}^{*^{\mathrm{T}}}\mathbf{X}) = \mathrm{rank}(\mathbf{T}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}) = \mathrm{rank}(\mathbf{T}) = qK,$$

$$\mathrm{rank}(\mathbf{T}^*) = \mathrm{rank}(\mathbf{T}^{*^{\mathrm{T}}}\mathbf{T}^*) = \mathrm{rank}(\mathbf{T}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{T}^{\mathrm{T}}) \leq qK.$$

Therefore, $\mathrm{rank}(\mathbf{T}^*) = qK$. By definition of projection matrix (v),

$$\mathbf{Q}_{\mathbf{V}_0} = \mathbf{Q}_{\mathrm{C}(\mathbf{X})} - \mathbf{Q}_{\mathrm{C}(\mathbf{T}^*)}.$$

$$\dim(\mathbf{V}_0) = \dim\{\mathrm{C}(\mathbf{X})\} - \dim\{\mathrm{C}(\mathbf{T}^*)\} = \mathrm{rank}(\mathbf{X}) - \mathrm{rank}(\mathbf{T}^*) = DK - qK. \quad (35)$$

By the property (vi), $\mathbf{Q}(m) - \mathbf{Q}(m_0)$ is a projection matrix. Thus, using equation (35), χ has chi-squared distribution with p degrees of freedom,

$$\begin{aligned}\mathrm{p} = \mathrm{rank}(\mathbf{Q}(m) - \mathbf{Q}(m_0)) &= DK - q_m K - [DK - q_{m_0}K] \\ &= [q_{m_0} - q_m]K > 0,\end{aligned}$$

since $q_{m_0} > q_m$. Similarly, χ_N has chi-squared distribution with $N - [DK - q_m K]$ degrees of freedom.

Back to equation (34), since $\lim_{N\to\infty} b_N[1 - \frac{1}{\sqrt{\log N}}] = \infty$,

$$\Pr\left(\chi \geq b_N\left[1 - \frac{1}{\sqrt{\log N}}\right]\right) = o(1).$$

For the second term in equation (34), one can show that

$$\Pr\left(\chi_N \leq N\left\{1 - \frac{1}{\sqrt{\log N}}\right\}\right) \leq \exp\left\{-\frac{1}{4}\frac{N}{\log N}\right\},$$

using an inequality on chi-squared distribution, see [46]. This completes the proof. □

of Theorem B 3. The equation (11) follows from Lemma 1 and Lemma 2.

The risk, or expected loss, for the model is

$$\begin{aligned}\mathrm{R}_N &= \mathrm{E}||\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2 \\ &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\{\mathbf{I} - \mathbf{Q}(m)\}\mathbf{X}\boldsymbol{\beta}.p_N(m) + \mathrm{E}[\mathbf{e}^{\mathrm{T}}\mathbf{Q}(m)\mathbf{e}.\mathrm{I}_{\hat{m}=m}] \\ &= A_1 + A_2,\end{aligned}$$

where $\mathrm{I}_{\hat{m}=m}$ is the indicator function. By the Cauchy-Schwartz's inequality,

$$\begin{aligned} A_2 &\leq \sqrt{\mathrm{E}\{\mathbf{e}^{\mathrm{T}}\mathbf{Q}(m)\mathbf{e}\}^2}\sqrt{p_N(m)} \\ &= \sigma^2\sqrt{2(D-q_m)K+(D-q_m)^2K^2}\sqrt{p_N(m)}. \end{aligned}$$

For $m \in \mathcal{M}_1$

$$\lim_{N\to\infty}\frac{1}{N}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\{\mathbf{I}-\mathbf{Q}(m)\}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}^{\mathrm{T}}\{\mathbf{H}-\mathbf{H}^{\frac{1}{2}^{\mathrm{T}}}\mathbf{Q}_H(m)\mathbf{H}^{\frac{1}{2}}\}\boldsymbol{\beta} > 0.$$

Equivalently,

$$\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\{\mathbf{I}-\mathbf{Q}(m)\}\mathbf{X}\boldsymbol{\beta} = \mathcal{O}(N).$$

Therefore, A_1 and A_2 tend to 0 as $N \to \infty$ by condition (a).

For $m \in \mathcal{M}_2 - \{m_0\}$, $A_2 \to 0$ as $N \to \infty$ by condition (b). In addition,

$$\begin{aligned} A_1 &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\{\mathbf{I}-\mathbf{Q}(m)\}\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\{\mathbf{I}-\mathbf{Q}_1+\mathbf{Q}_2(m)\}\mathbf{X}\boldsymbol{\beta}, \\ &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\{\mathbf{Q}_2(m)\}\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{T}_m^{\mathrm{T}}[\mathbf{T}_m(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{T}_m^{\mathrm{T}}]^{-1}\mathbf{T}_m\boldsymbol{\beta} \\ &= 0, \end{aligned}$$

since $\mathbf{T}_m = \mathbf{0}$ for models in $\mathcal{M}_2$.

Consequently, $\lim_{N\to\infty}\mathrm{R}_N = 0$ for both $m \in \mathcal{M}_1$ and $m \in \mathcal{M}_2 - \{m_0\}$. Therefore, if conditions (a) and (b) are satisfied,

$$\lim_{N\to\infty}\mathrm{R}_N = \lim_{N\to\infty}\mathrm{R}_N(m_0) = \sigma^2[D-q_{m_0}]K.$$

□

This completes the proof of Theorem.

References

1. Krim H, Yezzi A, editors. Statistics and Analysis of Shapes. Birkhaüser; 2006.
2. Dryden IL, Mardia KV. Statistical Shape Analysis. Wiley; 1998.
3. Lehmussola A, Selinummi J, Ruusuvuori P, Niemisto A, Yli-Harja O. Simulating fluorescent microscope images of cell populations. Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference. 2005; p. 1–4.
4. Zhao T, Murphy RF. Automated learning of generative models for subcellular location: building blocks for system biology. Cytometry. 2007;Part A(71A):978–990.
5. Khairy K, Howard J. Minimum-energy vesicle and cell shapes calculated using spherical harmonics parameterization. Soft Matter. 2010;7:2138–2143.
6. Lee AM, Berney-Lang MA, Liao S, Kanso E, Kuhn P. A low-dimensional deformation model for cancer cells in flow. Physics Fluids. 2012;24(081903).
7. Johnson G, Buck T, Sullivan D, Rhode G, RF M. Joint modelling of cell and nuclear shape variation. Molecular Biology of Cell. 2015;26(22):40476–4056.
8. Cootes TF, Taylor CJ, Cooper DH, Graham J. Active shape models: their training and application. Computer Vision and Image Understanding. 1995;61:38–59.

9. Grenander U, Miller MI. Computational anatomy: an emerging discipline. Quarterly of Applied Mathematics. 1995;LVI:617–694.

10. Srivastava A, Joshi S, Kaziska D, Wilson D. In: Paragios N, Chen Y, Faugeras O, editors. Planar Shape Analysis and Its Applications in Image-Based Inferences. Boston, MA: Springer US; 2006. p. 189–203.

11. Klassen E, Srivastava A, Mio W, Joshi S. Analysis of planar shape using geodesic paths on shape spaces. IEEE Pattern Analysis and Machiner Intelligence. 2004;26:372–383.

12. Altman N, Krzywinski M. Points of significance: clustering; 2017.

13. Abramowitz M, Stegun IA, et al. Handbook of mathematical functions. Applied mathematics series. 1966;55(62):39.

14. Brechbühler C, Gerig G, Kübler O. Parametrization of closed surfaces for 3-D shape description. Computer vision and image understanding. 1995;61(2):154–170.

15. Duncan BS, Olson AJ. Approximation and characterization of molecular surfaces. Biopolymers. 1993;33(2):219–229.

16. Hartigan JA. Partition models. Communications in Statistics-Theory and Methods. 1990;19:2745–2756.

17. Booth JG, Casella G, Hobert JP. Clustering using objective functions and stochastic search. Journal of the Royal Statistical Society, Series B. 2008;70:119–139.

18. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association. 2002;97:611–631.

19. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. Bioinformatics. 2001;17:977–987.

20. Heard NA, Holmes CC, Stephens DA. A quantitative study of gene regulation involved in the immune response of *Anopheline* mosquitoes: An application of Bayesian hierarchical clustering of curves. Journal of the American Statistical Association. 2006;101(473):18–29.

21. Bernardo JM, Smith AFM. Bayesian Theory. New York: Wiley; 1994.

22. O'Hagan A, Forster JJ. Kendall's advanced theory of statistics, volume 2B: Bayesian inference. vol. 2. Arnold; 2004.

23. Davie AM, Stothers AJ. Improved bound for complexity of matrix multiplication. Proceedings of the Royal Society of Edinburgh. 2013;143A:351–370.

24. Murphy KP. Conjugate Bayesian analysis of the Gaussian distribution. def. 2007;1($2\sigma 2$):16.

25. Smith JQ, Anderson PE, Liverani S. Separation measures and the geometry of Bayes factor selection for classification. Journal of the Royal Statistical Society, Series B. 2008;70:957–980.

26. Zellner A. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti North-Holland Elsevier. 1986; p. 233–243.

27. George EI, Foster DP. Calibration and empirical Bayes variable selection. Biometrika. 2000;87:731–747.

28. Zellner A. Application of Bayesian analysis with g-prior distributions. The Statistician. 1983;32(1):23–34.

29. Liang F, Paulo R, Molina G, Clyde MA, Berger JO. Mixtures of g-priors for Bayesian variable selection. Journal of the American Statsitical Association. 2008;103(481):410–423.

30. Clyde M, George EI. Flexible empirical Bayes estimation for wavelets. Journal of Royal Statistical Society Series B. 2000;62:681–698.

31. Hansen MH, Yu B. Model selection and principle of minimum description length. Journal of American Statistical Association. 2001;96:746–774.

32. Ward JH. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association. 1963;58:236–244.

33. Shao J. An asymptotic theory for linear model selection. Statistica Sinica. 1997; p. 221–264.

34. Nishi R. Aysmptotic properties of citeria for selection of variables in multiple regression. The Annals of Statistics. 1984;12:758–765.

35. Rao CR, Wu Y. A strongly consistent procedure for model selection in a regression problem. Biometrika. 1989;76:369–374.

36. Hong H, Preston B. Bayesian averaging, prediction and nonnested model selection. Journal of Econometrics. 2012;167:358–369.

37. Velliste M, Murphy RF. Automated Determination of Protein Subcellular Locations from 3D Fluorescence Microscope Images. Proceedings of the 2002 IEEE International Symposium on Biomedical Imaging (ISBI 2002). 2002;33:867–870.

38. MATLAB. Image Processing Toolbox R2013b. The MathWorks Inc , Natick, Massachusetts, United States. 2013;.

39. MATLAB. Signal Processing Toolbox R2013b. The MathWorks Inc , Natick, Massachusetts, United States. 2013;.

40. Peng T, Murphy RF. Image-derived, three-dimensional generative models of cellular organization. Cytometry Part A. 2011;79(5):383–391.

41. Wahba G. In: Spline models for observational data. The Society for Industrial and Applied Mathematics; 1990.

42. Newey W, McFadden D. Large sample estimation and hypothesis testing. Handbook of Econometrics. 1994;4:2113–2245.

43. Priestley MB. Spectral analysis and time series. No. v. 1-2 in Probability and mathematical statistics. Academic Press; 1982.

44. Sen PK, Singer JM. Large Sample Methods in Statistics: An Introduction with Applications. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis; 1994.

45. Mathai AM, Provost SB. Quadratic Forms in Random Variables, Theory and Applications. Marcel Dekker, New York; 1992.

46. Shibata R. An optimal selection of regression variables. Biometrika. 1981;68:45–54.