BayesSizeAndShape: a Julia package for Bayesian estimation of Size and Shape data regression models

BayesSizeAndShape: a Julia package for Bayesian estimation of Size and Shape data regression models

Gianluca Mastrantonio[1*] and Giovanna Jona Lasinio[2]

[1] *Department of Mathematical Sciences, Polytechnic of Turin, Corso Duca degli Abruzzi 24,10129 Turin, Italy; gianluca.mastrantonio@polito.it*
[2] *Department of Statistical Sciences Sapienza University; Piazzale Aldo Moro 5, 00185, Rome, Italy; giovanna.jonalasinio@uniroma1.it*
[*] *Corresponding author*

Abstract. *Size and shape data are of interest in automatic object identification, life-form studies, and any time the shape of a statistical unit is relevant. This paper briefly introduces the Bayesian estimation of regression models for size and shape data while presenting a Julia package implementing the MCMC estimation required.*

Keywords. *Size and shape data; Monte Carlo Markov Chain; Bayesian models, Julia package.*

---

1 Introduction

Shape is all the geometrical information that remains when location, scale and rotational effects are removed from an object. Modeling shape data has recently gained interest in many research fields. For example, in biology, there is great interest in modeling the shape features of organisms, and relating them to environmental conditions. In [8], a study about the sex differences in the crania of a macaque species is reported, and environmental features could be included as well. Magnetic Resonance data can be seen as size-and-shape data. Their analysis is helpful to assess specific features such as the Fetal alcohol spectrum disorder or Schizophrenia Magnetic Resonance images (see [11]). Image analysis and computer vision are other natural fields of application. Generally speaking, image recognition could be approached in different ways, and, among others, the shape analysis approach has often given successful results, e.g., digit recognition [2]. In genetics, it is common to use electrophoretic gel images [10], and shape analysis could be successfully adopted to analyze them efficiently. In chemistry, it is fundamental to assess the geometric structure of molecules using three-dimensional coordinates, e.g., it has been used to evaluate the shape features of steroid molecules [3]. And in bio-informatics, an important task is Protein matching, i.e., aligning molecules to find common geometrical structures in them. Shape analysis could be adopted to deal successfully with these tasks [9]. Theoretically, the involved data spaces are not Euclidean, and differential geometric tools should be adopted to develop a proper statistical shape theory. Some descriptive and basic inference tools have been proposed in the scientific literature, but to our current knowledge, only two papers introduced a regression model [6, 5]. In [6] the authors propose a marginal likelihood approach to implement a size-and-shape response regression with Gaussian landmarks. Likelihood-based approaches show some difficulties in dealing with complex covariance structures and are affected by numerical stability issues in numerical optimization algorithms.

In [5], a Bayesian approach is proposed in a very simplified setting. Here we move to an alternative easily-interpretable approach that starts from the Bayesian latent variable models framework. It potentially allows for the specification of complex covariance structures, with the advantage that numerical issues are solved through MCMC algorithms efficiently implemented. Hence, building on [6, 5], this work presents the Bayesian estimation of a regression model for size-and-shape response variables with Gaussian landmarks. Our proposal fits into the framework of Bayesian latent variable models and defines a highly flexible modeling approach.

2 Bayesian Size and Shape

Adopting the same notation as in [4], let $\tilde{\mathbf{X}}_i \in \mathbb{R}^{(k+1)\times p}$, with $i = 1,\dots,n$ be a $(k+1)\times p$ dimensional configuration matrix, with $k \geq p$, that represents the Euclidean coordinates of $k+1$ landmarks (points of interest an object) in dimension p for the $i-$th recorded object, where p is usually equal to 2 or 3. To perform any size-and-shape inference, we must remove information about the objects' location and orientation to model the data. Location information is usually removed post-multiplying by the Helmert submatrix $\mathbf{H}$ [see 7], obtaining the Helmertized configuration

$$\check{\mathbf{X}}_i = \mathbf{H}\tilde{\mathbf{X}}_i, \tag{1}$$

where $\mathbf{H}$ has dimension $k\times(k+1)$, and its j-th row is equal to $(-d_j,-d_j,\dots,-d_j,jd_j,0,\dots,0)$ where $d_j = 1/\sqrt{j(j+1)}$. The matrix $\check{\mathbf{X}}_i \in \mathbb{R}^{k\times p}$ is also called *pre-form* matrix. Following [6], if we decompose $\check{\mathbf{X}}_i$ using the singular value decomposition, i.e. according to

$$\check{\mathbf{X}}_i = \mathbf{U}_i\boldsymbol{\Delta}_i\check{\mathbf{R}}_i^\top \tag{2}$$

where $\check{\mathbf{R}}_i \in O(p)$ i.e. belongs to the space of the $p\times p$ orthogonal matrices, it could be readily proven that $\check{\mathbf{R}}_i$ contains all the information about orientation and $\mathbf{Y}_i = \mathbf{U}_i\boldsymbol{\Delta}_i$ is the *size-and-shape* version of the original configuration $\check{\mathbf{X}}_i$, it represents the object of the inference and the data we are modelling. An important point is that allowing $\check{\mathbf{R}}_i \in O(p)$ the reflection information is lost, and if we want to retain it, we have to assume that $\check{\mathbf{R}}_i \in \mathrm{SO}(p)$, i.e., $\check{\mathbf{R}}_i$ is a rotation matrix with $|\check{\mathbf{R}}_i| = 1$, and $\mathrm{SO}(p)$ is the p-dimensional Special Orthogonal group or rotation group.

Let us suppose that for each $\mathbf{Y}_i$ we have an associated vector of d covariates $\mathbf{z}_i = (z_{i1}, z_{i2},\dots,z_{id})^\top$ and we are interested in the relation between $\mathbf{z}_i$ and $\mathbf{Y}_i$. Unfortunately, the size-and-shape space, where $\mathbf{Y}_i$ lives, is a non-Euclidean manifold with a very complicated geometric structure that is not easy to handle, and even a distribution for $\mathbf{Y}_i$ cannot be easily specified. The idea proposed in [6], which here we extend to a Bayesian setting, is to model this relation using the latent variable $\mathbf{R}_i \in \mathrm{SO}(p)$ and to define a regressive-type relation between $\mathbf{X}_i = \mathbf{Y}_i\mathbf{R}_i$ and $\mathbf{z}_i$ in the following way:

$$\mathrm{vec}(\mathbf{X}_i) \sim \mathcal{N}_{kp}\left(\mathrm{vec}\left(\sum_{h=1}^{d} z_{ih}\mathbf{B}_h\right), \mathbf{I}_p \otimes \boldsymbol{\Sigma}\right), i = 1,\dots,n, \tag{3}$$

with $\mathbf{X}_i \perp \mathbf{X}_{i'}$ if $i \neq i'$, where $\mathrm{vec}(\cdot)$ indicates the vectorization of a matrix, $\boldsymbol{\Sigma}$ is a non singular $k\times k$ covariance matrix and $\mathbf{B}_h$, $h = 1,\dots,d$, is a $k\times p$ matrix of regressive coefficients. It should be noted that $\mathbf{R}_i$ is latent and hence a non-observable variable, which must not be confused with $\check{\mathbf{R}}_i$, which is the rotation matrix of the original data. In [4], the set of full conditional is found when $\boldsymbol{\Sigma}$ is a generic covariance matrix, and this is the model implemented in the Julia package.

3 Julia implementation

Julia is a high-level, general-purpose dynamic programming language. Its features are well-suited for numerical analysis and computational science. Work on Julia was started in 2009 by Jeff Bezanson, Stefan Karpinski, Viral B. Shah, and Alan Edelman, who set out to create a free language that was both high-level and fast. The first website with a blog post explaining the language's mission was out on February 2012, and in the past 10 years, the community has grown. The Julia package ecosystem has over 11.8 million lines of code (including documentation and tests) [1]. The JuliaCon academic conference for Julia users and developers has been held annually since 2014 with an increasing public. The popularity of the language is mostly due to its efficiency, it is a compiled language, whereas R and Python are interpreted ones. Using the appropriate interface, the user can handle data with R and process estimation using Julia (we use VScode and its extensions).

3.1 The *BayesSizeAndShape* package

The *BayesSizeAndShape* package (directly available in Julia) implements the model sketched in section 2. Currently, a generic $\boldsymbol{\Sigma}$ covariance matrix with inverse Wishart prior is implemented. In [4], a small simulation experiment using this package was carried out. At the same time, here we present the estimation of the regression model using the well-known data on rats skull [see 6] to allow comparison with the likelihood proposal. The dataset includes rat skull data from X-rays. Eight landmarks in two dimensions for eighteen individuals were observed at 7, 14, 21, 30, 40, 60, 90, and 150 days from birth see figure 1. To compare our proposal with [6], we estimate two regression models, both of which include time as an independent variable; in M1, the log-transformed time is linear, while in M2 it appears as a linear and quadratic term. We run the MCMC estimation with 100000 iterations, discarding half of them and keeping every 5 samples for inferential purposes. To assess the model's goodness of fit, we compute the Riemannian distance between observed and predicted landmarks (mean of predictive samples), both treated as curves on a Riemannian manifold. M2 (0.375) performs slightly better than M1 (0.430). In [6], a parametric bootstrap approach is built to test for the model's goodness of fit. In the Bayesian setting, we can use the posterior samples and the samples from the predictive distribution for all inferences. For example, we can compute a distance between curves using each realization from the predictive distribution and the data. Using the Riemannian distance, again, all realizations from M2 are closer to the data than those from M1.

References

[1] (2021). Julia newsletter, August.

[2] Anderson, C. R. (1997). *Object recognition using statistical shape analysis*. PhD thesis, University of Leeds.

[3] Czogiel, I., Dryden, I. L., and Brignell, C. J. (2011). Bayesian matching of unlabeled marked point sets using random fields, with an application to molecular alignment. *Annals of Applied Statistics*, 5:2603–2629.

[4] Di Noia, A., Mastrantonio, G., and Jona-Lasinio, G. (2022). Bayesian Size-and-Shape regression modelling. *arXiv:2303.06661*.

[5] Dryden, I. L., Kim, K.-R., and Le, H. (2019). Bayesian linear size-and-shape regression with applications to face data. *Sankhya A*, 81(1):83–103.

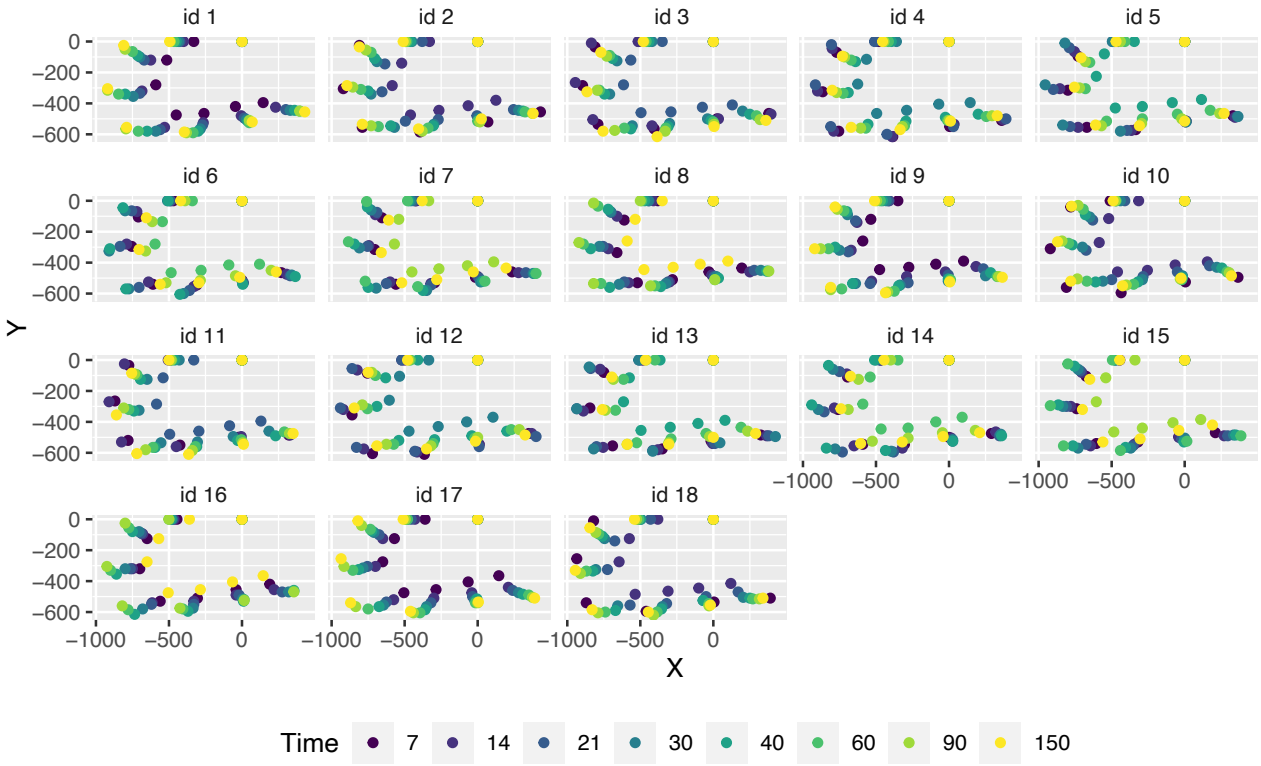
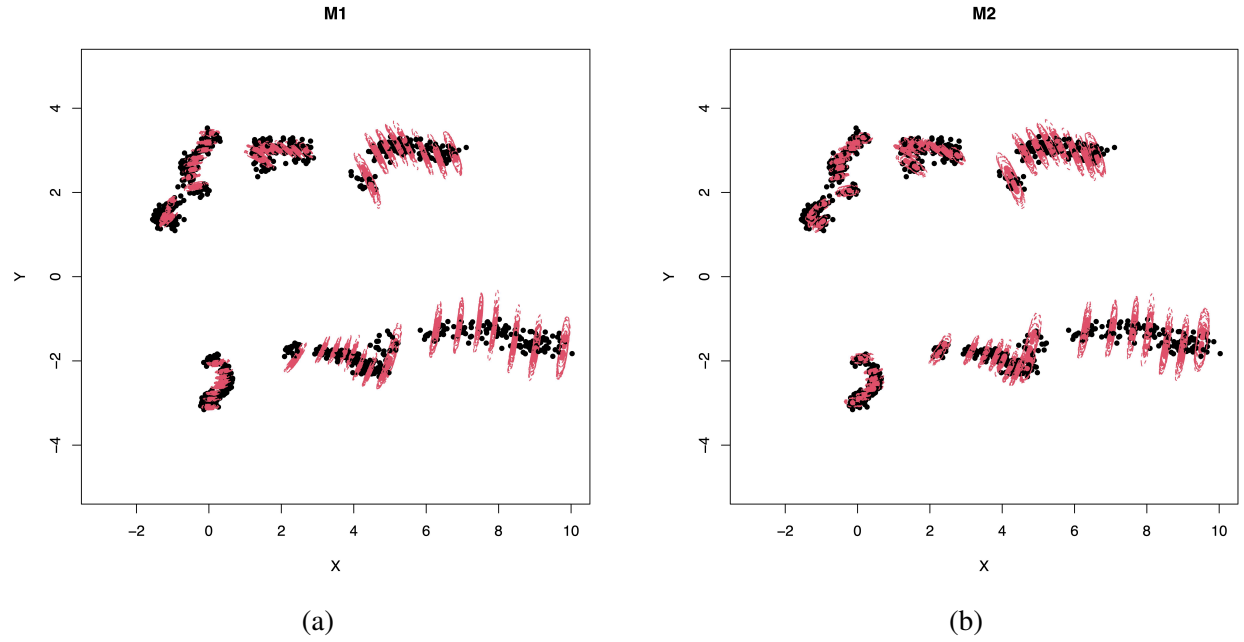


Figure 1: Rats Skull data: original landmarks



(a) (b)

Figure 2: Rats Skull data: Scaled helmertized observed (black) and predicted (red) landmarks with 95% highest posterior density regions for M1 (a) and M2 (b)

[6] Dryden, I. L., Kume, A., Paine, P. J., and Wood, A. T. A. (2021). Regression modeling for size-and-shape data based on a gaussian model for landmarks. *Journal of the American Statistical Association*, 116(534):1011–1022.

[7] Dryden, I. L. and Mardia, K. (2016). *Statistical Shape Analysis: With Applications in R*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, 2nd edition.

[8] Dryden, I. L. and Mardia, K. V. (1993). Multivariate shape analysis. *Sankhya Series A*, 55:460–480.

[9] Green, P. J. and Mardia, K. V. (2006). Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika*, 93:235–254.

[10] Horgan, G. W., Creasey, A., and Fenton, B. (1992). Superimposing two-dimensional gels to study genetic variation in malaria parasites. *Electrophoresis*, 13:871–875.

[11] Mardia, K. V., Bookstein, F. L., and Kent, J. T. (2013). Alcohol, babies and the death penalty: Saving lives by analysing the shape of the brain. *Significance*, 10(3):12–16.