Full Bayesian analysis of the generalized non-isotropic Procrustes problem with scaling

Douglas L. Theobald[1*] and Kanti V. Mardia [2]

[1] Department of Biochemistry, Brandeis University
[2] Department of Statistics, University of Leeds

1 Introduction

An object's shape is defined as the geometrical information about an object that remains when translational, rotational, and scaling effects are removed. Hence, shape is invariant under the corresponding similarity transformations. To compare and contrast the shapes of different objects, it is first necessary to superposition them in some optimal fashion.

An object can often be reduced to a set of k landmarks (Cartesian points in d dimensions) and represented as a $k \times d$ matrix (a *configuration* of landmarks for a geometrical object). For example, in the classical molecular biology approach, the structures of biological macromolecules are described concisely by the three-dimensional coordinates of their atoms. Structures are conventionally referred to a common reference frame using the statistical optimization method of ordinary least-squares (OLS) (Flower(1999); Dryden and Mardia(1998)). The least squares criterion stipulates that the optimal transformations are those that minimize the sum of squared distances among corresponding landmarks in the objects. OLS assumes that all landmarks have the same variance and are uncorrelated, yet both conditions are frequently violated in real data.

In previous work, we relaxed the assumptions of homoscedasticity and non-correlation by treating the superposition problem within a likelihood framework where multiple structures are distributed normally (Theobald and Wuttke(2006a); Theobald and Wuttke(2006b); Theobald and Wuttke(2008)); some theoretical work on such models appeared in Goodall and Mardia (1993). This ML method effectively accounts for uneven variances and correlations in the landmarks by weighting by the inverse of the covariance matrix. Notably, this work did not address issues of scaling, as the size of a macromolecule is fixed by the physics of chemical bonding.

Simultaneous point estimation of the sample covariance matrix and the translations is generally impossible due to indentifiability issues, which has been a significant impediment to a viable non-isotropic Procrustes analysis (Dryden and Mardia(1998); Lele and Richtsmeier(1990); Lele(1993); Lele and Richtsmeier(2001); Glasbey et al.,(1995); Goodall(1991b)). We allow joint identifiability by regularizing the covariance matrix using a hierarchical, empirical Bayes approach in which the eigenvalues of the covariance matrix are treated as variates from an inverse gamma distribution. This hierarchical method is analogous to putting a conjugate inverse Wishart prior on the covariance matrix. An expectation-maximization implementation of this method performs well in practice.

Here we describe a Bayesian extension of this matrix normal model for the generalized non-isotropic Procrustes problem with scaling. Building on previous work, this analysis applies to multiple configurations (as opposed to pairwise Procrustes), uses an arbitrary covariance matrix (as well as a diagonal and isotropic covariance matrix as special cases), and allows for proper conjugate priors for all parameters. Generalizing to full shape analysis with scaling is non-trivial, as the conditional posterior distribution of the scale factors turns out to be a non-standard form (which we christen the halfnormal-gamma). We have developed rejection and Metropolis-Hastings algorithms for simulating from this distribution.

2 Methods

2.1 A matrix normal probability model for the macromolecular superposition problem

Consider n structures ($\mathbf{X}_i$, $i = 1 \ldots n$), each with k labelled landmarks, where each structure is described by a $k{\times}d$. We assume that each structure $\mathbf{X}_i$ is normally distributed and is observed in an arbitrary coordinate system (Dryden and Mardia(1998); Goodall(1991a); Goodall and Mardia(1993)). Heterogeneous variances and correlations among the landmarks are described by a $k{\times}k$ covariance matrix $\mathbf{\Sigma}$ (isotropic in the d-dimensional space). Hence each $\mathbf{X}_i$ can be considered to be an arbitrarily scaled, rotated, and translated zero-mean normal matrix displacement $\mathbf{E}_i \sim N_{K,D}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I})$ of the mean structure M,

$$\mathbf{X}_i = \frac{1}{\beta_i} \left(\mathbf{M} + \mathbf{E}_i\right) \mathbf{R}_i' - \mathbf{1}_K \boldsymbol{t}_i' \tag{1}$$

where β_i is a global scale factor, $\mathbf{R}_i$ is a $d \times d$ orthogonal rotation matrix, $\boldsymbol{t}_i$ is a $d \times 1$ column vector for the translational offset, and $\mathbf{1}_K$ denotes the $k \times 1$ column vector of ones.

2.2 A Procrustes matrix normal likelihood function

The full joint likelihood function for the model given in (1) is obtained from a matrix normal distribution (Dawid(1981)). Define

$$\mathbf{Y}_i = (\beta_i \mathbf{X}_i + \mathbf{1}_K \boldsymbol{t}_i')\mathbf{R}_i$$

then the PDF for the likelihood function is:

$$\mathrm{p}\left(\mathbf{X}|\mathbf{R}, \boldsymbol{t}, \beta, \mathbf{M}, \mathbf{\Sigma}\right) = C \exp\left(-\frac{1}{2}\sum_i^n \mathrm{tr}\left\{[\mathbf{Y}_i - \mathbf{M}]'\mathbf{\Sigma}^{-1}[\mathbf{Y}_i - \mathbf{M}]\right\}\right), \tag{2}$$

with normalization constant:

$$C = (2\pi)^{-\frac{kdn}{2}} \left(\prod_i^n \beta_i^{kd}\right) |\mathbf{\Sigma}|^{-\frac{dn}{2}} . \tag{3}$$

2.3 A Bayesian extension

The likelihood analysis described above does not provide ready estimates of the uncertainty in the estimated parameters. In an earlier presentation at this conference (Theobald, 2009), a Bayesian extension was described allowing for the incorporation of other prior data. For the Bayesian analysis we assume that $\mathbf{\Sigma}, \mathrm{M}, \mathbf{R}, \boldsymbol{t}, \beta$ are all independent, so that

$$\mathrm{p}\left(\mathbf{\Sigma}, \mathbf{M}, \mathbf{R}, \boldsymbol{t}, \beta|\mathbf{X}\right) \propto \mathrm{p}\left(\mathbf{X}|\mathbf{\Sigma}, \mathbf{M}, \mathbf{R}, \boldsymbol{t}, \beta\right) \mathrm{p}\left(\mathbf{\Sigma}\right) \mathrm{p}\left(\mathbf{M}\right) \mathrm{p}\left(\mathbf{R}\right) \mathrm{p}\left(\boldsymbol{t}\right) \mathrm{p}\left(\beta\right). \tag{4}$$

We will also assume a hierarchical prior for $\mathbf{\Sigma}$:

$$\mathrm{p}\left(\mathbf{\Sigma}\right) \propto \mathrm{p}\left(\mathbf{\Sigma}|\delta, n\right) \mathrm{p}\left(\delta\right). \tag{5}$$

Conditional distributions and MAP estimates for the parameters of this Bayesian superposition model have been solved using proper conjugate priors. We have also coded a hybrid Gibbs-MCMC sampling algorithm for the full Bayesian solution. When using improper reference priors it is critical to establish the propriety of the posterior. We have shown that the posterior is proper in the special isotropic, non-correlated case (corresponding to the classic OLS assumptions) when using uniform priors on M and the translations $\boldsymbol{t}$ and placing a standard improper reference prior on the variance. However, in the nonisotropic treatment with an arbitrary diagonal or non-diagonal covariance matrix, the standard reference priors on the variance hyperparameters lead to an improper posterior, and so proper priors are necessary. Conditional distributions for the unknown parameters (other than scale factors) have been presented previously (e.g., Theobald (2009)).

2.3.1 Conditional probability of the scale factors β

The conditional probability density function of β_i is given by

$$\mathrm{p}\left(\beta_i|\mathbf{X},\boldsymbol{\Sigma},\mathbf{M},\boldsymbol{t}_i,\mathbf{R}\right)=C{\beta_i}^{m-1}\exp\left\{-\frac{\phi_i}{2}{\beta_i}^2-\beta_i\gamma_i\right\}, \tag{6}$$

$$C=\frac{2{\phi_i}^{-\frac{m}{2}}e^{-\frac{\gamma_i^2}{8\phi_i}}}{\Gamma(m)\,\mathrm{D}_{-m}\left(\frac{\gamma_i}{\sqrt{2\phi_i}}\right)}, \tag{7}$$

$$\phi_i={\beta_i}^2\,\mathrm{tr}\left(\check{\mathbf{X}}_i'\boldsymbol{\Sigma}^{-1}\check{\mathbf{X}}_i\right), \tag{8}$$

$$\gamma_i=\beta_i\,\mathrm{tr}\left(\mathbf{M}'\boldsymbol{\Sigma}^{-1}\check{\mathbf{X}}_i\mathbf{R}_i\right), \tag{9}$$

$$m=kd+1, \tag{10}$$

where $\mathrm{D}_p(z)$ is a “parabolic cylinder function”, a type of confluent hypergeometric function, defined in Gradshteyn and Ryzhik p 1028, section 9.24-9.25 (also described as the Whitaker function in Chapter 19 of Abramowitz and Stegun). This distribution (halfnormal-gamma) has been studied in Mardia *et al.,* (2011), where the isotropic single global scaling case (pairwise superposition) has been treated as well.

References

A. P. Dawid. Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274, 1981.

I. L. Dryden and K. V. Mardia. *Statistical shape analysis*. Wiley series in probability and statistics. Probability and statistics. John Wiley & Sons, Chichester ; New York, 1998.

D. R. Flower. Rotational superposition: A review of methods. *J Mol Graph Model*, 17(3-4): 238–244, 1999.

C. Glasbey, G. Horgan, G. Gibson, and D. Hitchcock. Fish shape analysis using landmarks. *Biometrical J*, 37:481–495, 1995.

C. Goodall. Procrustes methods in the statistical analysis of shape. *J Roy Stat Soc B Met*, 53 (2):285–321, 1991a.

C. Goodall. Procrustes methods in the statistical analysis of shape: Rejoinder to discussion. *J Roy Stat Soc B Met*, 53(2):334–339, 1991b.

C. R. Goodall and K. V. Mardia. Multivariate aspects of shape theory. *Annals of Statistics*, 21 (2):848–866, June 1993.

S. Lele. Euclidean distance matrix analysis (EDMA) - estimation of mean form and mean form difference. *Math Geol*, 25(5):573–602, 1993.

S. Lele and J. T. Richtsmeier. Statistical models in morphometrics: Are they realistic? *Systematic Zoology*, 39:60–69, 1990.

S. Lele and J. T. Richtsmeier. *An invariant approach to statistical analysis of shapes*. Interdisciplinary statistics. Chapman and Hall/CRC, Boca Raton, Fla., 2001.

K. V. Mardia, C. J. Fallaize, S. Barber, R. M. Jackson, and D. L. Theobald. Bayesian alignment of similarity shapes and halfnormal-gamma distributions. *Annals of Applied Statistics*, submitted, 2011.

D. L. Theobald. A nonisotropic bayesian approach to superpositioning multiple macromolecules. In A. Gusnanto, K. V. Mardia, and C. J. Fallaize, editors, *Statistical Tools for Challenges in Bioinformatics*, Proceedings of the 28th Leeds Annual Statistical Research (LASR) Workshop, pages 55–59, UK, 2009. Department of Statistics, University of Leeds.

D. L. Theobald and D. S. Wuttke. Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem. *Proc Natl Acad Sci U S A*, 103(49):18521–18527, 2006a. ISSN 0027-8424 (Print).

D. L. Theobald and D. S. Wuttke. THESEUS: Maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics*, 22(17):2171–2172, 2006b. ISSN 1460-2059 (Electronic).

D. L. Theobald and D. S. Wuttke. Accurate structural correlations from maximum likelihood superpositions. *PLoS Comput Biol*, 4(2):e43, 2008. ISSN 1553-7358 (Electronic). doi:10.1371/journal.pcbi.0040043.