A nonisotropic Bayesian approach for superpositioning multiple macromolecules

Douglas L. Theobald

Department of Biochemistry, Brandeis University

1 Introduction

Superpositioning is a key technique in structural biology that enables the comparison and analysis of various conformational differences and commonalities among macromolecules with similar structures. As such, superpositioning is used routinely in the fields of NMR, X-ray crystallography, protein folding, molecular dynamics, rational drug design, and structural evolution (Bourne and Shindyalov, 2003; Flower, 1999). The interpretation of a superposition relies upon the validity of the estimated orientations, and hence accurate superpositioning tools are an essential component of modern structural analysis.

In the classical molecular biology approach, structures are referred to a common reference frame using the conventional statistical optimization method of ordinary least-squares (OLS) (Flower, 1999). The OLS criterion stipulates that the optimal rotations and translations are those that minimize the squared distances among corresponding atoms in the structures being analyzed. As a theoretical justification for OLS, the Gauss-Markov theorem requires both homoscedastic and uncorrelated data. However, both requirements are generally violated in the case of macromolecules. The variances of backbone atoms in NMR superpositions, for instance, commonly span greater than three orders of magnitude. Furthermore, adjacent atoms typically covary strongly due to covalent chemical bonds and other physical interactions. In practice, researchers usually perform a OLS superposition, identify regions that do not "superposition well", and calculate a new superposition in which the more variable regions have been subjectively excluded from the analysis.

In previous work, we relaxed the assumptions of homoscedasticity and noncorrelation by treating the superposition problem within a likelihood framework where the macromolecular structures are distributed normally (Theobald and Wuttke, 2006a,b, 2008). In our non-isotropic likelihood treatment, superpositioning requires estimating five classes of parameters: (1) a mean structure, (2) a global covariance matrix describing the variance and correlations for each atom in the structures, (3) hierarchical parameters describing the covariance matrix, and, for each structure in the analysis, (4) a proper orthogonal rotation matrix and (5) a translation vector. Our ML method accounts for uneven variances and correlations in the structures by weighting by the inverse of the covariance matrix.

Estimation of the covariance matrix has been a significant impediment to a viable non-isotropic likelihood-based Procrustes analysis (Dryden and Mardia, 1998; Lele and Richtsmeier, 1990; Lele, 1993; Lele and Richtsmeier, 2001; Glasbey et al., 1995; Goodall, 1991b, 1995). Simultaneous estimation of the sample covariance matrix and the translations is generally impossible. We permit joint identifiability by regularizing the covariance matrix using a hierarchical, empirical Bayes treatment in which the eigenvalues of the covariance matrix are themselves distributed according to an inverse gamma pdf.

In general all the estimates of the unknown parameters are interdependent and cannot be solved for analytically. Furthermore, the smallest eigenvalues of the sample covariance matrix are zero due to colinearity imparted by the centering operation necessary to estimate the unknown translations. We treat these smallest eigenvalues as "missing data" using an expectation-maximization algorithm. For simultaneous estimation, we use iterative conditional maximization of the joint likelihood augmented by the expectation-maximization algorithm. This method works very well in practice, with excellent convergence properties for the many hundreds of real cases analyzed to date. An example of a conventional LS superposition compared with our ML estimate is shown in Figure 1.

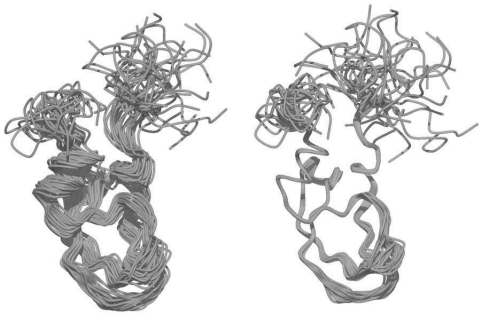

Figure 1: A conventional LS superposition *vs* the ML superposition (*center* and *right*) of 30 NMR models of the 71 amino acid Kunitz domain 2 of Tissue Factor Pathway Inhibitor (PDB ID: 1adz). All C_αs were included in the calculations.

2 A matrix normal probability model for the macromolecular superposition problem

Consider N structures ($\mathbf{X}_i$, $i = 1 \ldots N$), each with K corresponding atoms (landmarks), where each structure is defined as a $K \times D$ matrix holding K rows of D-dimensional coordinates. We assume a probabalistic model for the Procrustes problem in which each form $\mathbf{X}_i$ is distributed according to a Gaussian probability density and is observed in a different unknown coordinate system (Dryden and Mardia, 1998; Goodall, 1991a, 1995). We allow heterogeneous variances and correlations among the atoms in the structures, as described by a $K \times K$ covariance matrix $\mathbf{\Sigma}$ for the atoms. Under this Gaussian model, each $\mathbf{X}_i$ can be considered as an arbitrarily scaled, rotated, and translated zero-mean Gaussian matrix displacement $\mathbf{E}_i \sim N_{K,D}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{\Xi})$ of the mean structure $\mathbf{M}$,

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

where $\boldsymbol{t}_i$ is a $D \times 1$ column vector for the translational offset, and $\mathbf{1}_K$ denotes the $K \times 1$ column vector of ones.

3 A Procrustes matrix normal likelihood equation

The full joint likelihood equation for the model given in (1) is obtained from a multivariate matrix normal distribution Arnold, 1981;Dutilleul, 1999). Define

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

then

$$\mathrm{p}\left(\mathbf{X}|\mathbf{R}, \boldsymbol{t}, \mathbf{M}, \mathbf{\Sigma}\right) = (2\pi)^{-\frac{3KN}{2}} |\mathbf{\Sigma}|^{-\frac{3N}{2}} \exp\left(-\frac{1}{2}\sum_i^N \mathrm{tr}\left\{[\mathbf{Y}_i - \mathbf{M}]'\mathbf{\Sigma}^{-1}[\mathbf{Y}_i - \mathbf{M}]\right\}\right) \tag{2}$$

4 A Bayesian extension

The likelihood analysis described above does not provide ready estimates of the uncertainty in the estimated parameters. Hence, an extension of the method to a full Bayesian analysis would be useful, and would also allow for the incorporation of other prior data (e.g., B-factors, an empirical measure of uncertainty, attached to each atom in a crystal structure).

For our Bayesian analysis we assume that $\mathbf{\Sigma}, \mathbf{M}, \mathbf{R}, \boldsymbol{t}$ are all independent, so that

$$\mathrm{p}\left(\mathbf{\Sigma}, \mathbf{M}, \mathbf{R}, \boldsymbol{t} | \mathbf{X}\right) \propto \mathrm{p}\left(\mathbf{X} | \mathbf{\Sigma}, \mathbf{M}, \mathbf{R}, \boldsymbol{t}\right) \mathrm{p}\left(\mathbf{\Sigma}\right) \mathrm{p}\left(\mathbf{M}\right) \mathrm{p}\left(\mathbf{R}\right) \mathrm{p}\left(\boldsymbol{t}\right) \tag{3}$$

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

$$\mathrm{p}\left(\mathbf{\Sigma}\right) \propto \mathrm{p}\left(\mathbf{\Sigma} | \phi, n\right) \mathrm{p}\left(\phi\right) \tag{4}$$

We have solved the MAP estimates for our Bayesian superposition model, and we are currently in the process of coding a Gibbs 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 isotropic case (corresponding to the classic OLS assumptions) when using uniform priors on $\mathbf{M}$ and the translations $\boldsymbol{t}$ and placing a standard improper reference prior on the variance. However, in the nonisotropic treatment, the standard reference priors on the variance hyperparameters lead to an improper posterior, and so here we use vague proper priors. In the following we present the conditional distributions of the unknown parameters.

4.1 Mean

The conditional distribution of the mean $\mathbf{M}$, given a flat prior ($\mathrm{p}\left(\mathbf{M}\right) \propto C$), is:

$$\mathrm{p}\left(\mathbf{M} | \mathbf{X}, \mathbf{\Sigma}, \mathbf{R}, \boldsymbol{t}\right) \propto \exp\left(-\frac{1}{2} \sum_{i}^{N} \mathrm{tr}\left\{[\mathbf{Y}_i - \mathbf{M}]' \mathbf{\Sigma}^{-1} [\mathbf{Y}_i - \mathbf{M}]\right\}\right) \tag{5}$$

which is a matrix normal distribution.

4.2 Translations

The conditional distribution of the translations $\boldsymbol{t}$, given a flat prior ($\mathrm{p}\left(\boldsymbol{t}\right) \propto C$), is:

$$\mathrm{p}\left(\boldsymbol{t}_i | \mathbf{X}_i, \mathbf{M}, \mathbf{\Sigma}, \mathbf{R}_i\right) \propto \exp\left(-\mathrm{tr}\left\{\mathbf{X}_i' \mathbf{\Sigma}^{-1} \mathbf{1}_K \boldsymbol{t}_i'\right\} - \frac{1}{2} \mathrm{tr}\left\{\boldsymbol{t}_i \mathbf{1}_K' \mathbf{\Sigma}^{-1} \mathbf{1}_K \boldsymbol{t}_i'\right\}\right) \tag{6}$$

assuming $\mathbf{R}_i \mathbf{M}' \mathbf{\Sigma}^{-1} \mathbf{1}_K = \mathbf{0}$. This is a multivariate normal distribution.

4.3 Rotations

The conditional distribution of the rotations $\mathbf{R}_i$, given a proper uniform prior ($\mathrm{p}\left(\mathbf{R}_i\right) \propto C$):

$$\mathrm{p}\left(\mathbf{R}_i | \mathbf{X}_i, \mathbf{\Sigma}, \mathbf{M}, \boldsymbol{t}_i\right) \propto \exp\left(-\frac{1}{2} \mathrm{tr}\left\{\mathbf{M}' \mathbf{\Sigma}^{-1} \check{\mathbf{X}}_i \mathbf{R}_i\right\}\right) \tag{7}$$

which is a matrix von Mises-Fisher distribution.

4.4 Isotropic covariance matrix

The conditional distribution of an isotropic covariance matrix ($\mathbf{\Sigma}_{\text{iso}} = \phi\mathbf{I}$), given a reference prior ($\mathrm{p}(\phi) \propto \frac{1}{\phi}$), is:

$$\mathrm{p}(\phi|\mathbf{X}, \mathbf{\Sigma}, \mathbf{M}, \mathbf{R}, \boldsymbol{t}) \propto \phi^{-\left(\frac{3NK}{2}+1\right)} \exp\left(-\frac{1}{2\phi}\sum_{i}^{N} \mathrm{tr}\left\{[\mathbf{Y}_i - \mathbf{M}]'[\mathbf{Y}_i - \mathbf{M}]\right\}\right) \tag{8}$$

which is an inverse gamma distribution. This solution corresponds to the Bayesian version of the traditional OLS superposition solution.

4.5 A Diagonal Inverse Wishart prior for a diagonal covariance matrix (multivariate scaled inverse chi square)

In the following we assume the covariance matrix is diagonal ($\mathbf{\Sigma}$ diagonal):

$$\mathrm{p}(\mathbf{\Sigma}|\mathbf{\Psi}, n, K) = \frac{|\mathbf{\Psi}|^{\frac{n}{2}}}{2^{\frac{nK}{2}}\,|\mathbf{\Sigma}|^{\left(\frac{n}{2}+1\right)}\,\Gamma\left(\frac{n}{2}\right)^{K}} \exp\left\{-\frac{1}{2}\,\mathrm{tr}\left(\mathbf{\Psi}\mathbf{\Sigma}^{-1}\right)\right\} \tag{9}$$

4.6 Conditional probability for the covariance matrix $\mathbf{\Sigma}$

If we further assume that the hyperparameter $\mathbf{\Psi}$ is isotropic (i.e., $\mathbf{\Psi} = \phi\mathbf{I}$), then we have a simple expression for the conditional distribution of the covariance matrix:

$$\mathrm{p}(\mathbf{\Sigma}|\mathbf{X}, \mathbf{M}, \mathbf{R}, \boldsymbol{t}, \phi) \propto |\mathbf{\Sigma}|^{-\frac{3N+n+2}{2}}\, e^{-\frac{1}{2}\,\mathrm{tr}\left(\mathbf{\Sigma}^{-1}[\mathbf{S}+\phi\mathbf{I}]\right)} \tag{10}$$

where

$$S = \frac{1}{2}\sum_{i}^{N} \mathrm{tr}\left\{[\mathbf{Y}_i - \mathbf{M}]'[\mathbf{Y}_i - \mathbf{M}]\right\} \tag{11}$$

This is simply another multivariate scaled inverse chi-square distribution.

4.7 Conditional probability of the variance hyperparameter ϕ

The "conventional" reference prior for the variance hyperparameter ($\mathrm{p}(\phi) \propto \frac{1}{\phi^{\alpha}}$) leads to an improper posterior, so we adopt a vague proper conjugate prior on ϕ (a scaled chi-square with parameters α, m):

$$\mathrm{p}(\phi|\alpha, m) \propto \phi^{\frac{m-2}{2}} \exp\left\{-\frac{\phi}{2\alpha}\right\} \tag{12}$$

$$\mathrm{p}(\phi|\mathbf{X}, \mathbf{\Sigma}, \mathbf{M}, \mathbf{R}, \boldsymbol{t}, n) \propto \phi^{\frac{nK+m-2}{2}} \exp\left\{-\frac{\phi}{2}\left[\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\right) + \frac{1}{\alpha}\right]\right\} \tag{13}$$

which is a gamma distribution (or a scaled chi-square).

References

S. F. Arnold.*The Theory of Linear Models and Multivariate Analysis*. Wiley, New York, 1981.

P. E. Bourne and I. N. Shindyalov. Structure comparison and alignment. In P. E. Bourne and H. Weissig, editors, *Structural Bioinformatics*, volume 44 of *Methods of Biochemical Analysis*, pages 321–337. Wiley-Liss, Hoboken, N.J., 2003.

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

P. Dutilleul. The MLE algorithm for the matrix normal distribution. *J Stat Comput Sim*, 64:105–123, 1999.

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. Goodall. Procrustes methods in the statistical analysis of shape revisited. In K. V. Mardia and C. A. Gill, editors, *Proceedings in current issues in statistical shape analysis*, pages 18–33. Leeds University Press, Leeds, 1995.

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? *Syst. Zool*, 39(1):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.

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.