Alignment of multiple configurations using hierarchical models

By Yann RUFFIEUX

FSB IMA STAT Station 8, EPFL - Swiss Federal Institute of Technology

1015 Ecublens, Switzerland

yann.ruffieux@epfl.ch

and Peter J. GREEN

School of Mathematics, University of Bristol, Bristol BS8 ITW, UK

P.J.Green@bristol.ac.uk

SUMMARY

We describe a method for aligning multiple unlabelled configurations simultaneously. Specifically, we extend the two-configuration matching approach of Green and Mardia (2006) to the multiple configuration setting. Our approach is based on the introduction of a set of hidden locations underlying the observed configuration points. A Poisson process prior is assigned to these locations, resulting in a simplified formulation of the model. We make use of a structure containing the relevant information on the matches, numerous types of which must be taken into account. Markov chain Monte Carlo-based inference can be made simultaneously on the matching and the relative transformations between the configurations. We focus on the particular case of rigid-body transformations and Gaussian observation errors. We apply our method to a problem taken from chemoinformatics: the alignment of steroid molecules.

Some key words: chemoinformatics, Markov chain Monte Carlo, matching, rigid-body transformation, shape analysis, steroids

1 Introduction

In many scientific disciplines one is confronted with the problem of comparing objects. Typically, the scientist locates a number of characteristic points, called landmarks, which correspond on the objects of a given population. For example, a landmark might be a recognisable location on a given organism, such as the corner of an eye, the tip of a finger, or the meeting of two sutures on a skull. Numerous techniques have been studied over the years for the geometrical comparison of objects when the landmarks are labelled, i.e. when the point correspondences between the objects under study have been established. Now if the landmark configurations are unlabelled, so that the correspondences between the points of each configuration are unknown, then our problem also becomes one of *matching*: identifying and labelling corresponding landmarks.

A number of methods have been developed for the alignment of unlabelled point configurations, in various contexts. In image analysis, for instance, Cross and Hancock (1998) use graph theory techniques for the matching of point sets representing two-dimensional images, while Chui and Rangarajan (2000) consider the use of non-rigid body transformations. The alignment problem has also attracted a lot of interest from the chemoinformatics community (Lemmen and Lengauer 2005). In drug design, for example, a subject of prime interest is the local interaction between a small molecule (the ligand) and a given protein receptor. If the geometrical structure of the receptor is known, then established methods such as docking can be applied so as to specify the protein-ligand interaction. However, in most cases this structure is unknown, meaning the drug designer must rely on a study of the similarity (or diversity) in available ligands. The alignment of the molecules is a first important step towards such a study.

We focus specifically on the work of Green and Mardia (2006), which describes a Bayesian methodology for aligning two point configurations. We wish to extend this methodology so as to deal with an arbitrary number of configurations. Independent pairwise comparison of the configurations could be an option, but would be very costly in terms of computation should the number of configurations be large. Furthermore, unless all of the configurations are treated simultaneously in a single model, there is loss of information, for example about the ‘noise-free’ locations of the matched points. In Section 3.4 we see some evidence of the empirical impact of this. The elegance of the pairwise model makes this extension natural and relatively straightforward. The problem of matching multiple configurations has also been addressed by Dryden et al. (2007), see the Discussion section of our paper for further details.

This paper is organised as follows. In Section 2 we treat the simultaneous alignment of multiple point configurations. We describe a hierarchical Bayesian model for this task, and propose a Markov chain Monte Carlo algorithm for making inference on the model in the case of rigid-body transformations between the configurations. In Section 3 we consider an application of our approach to the matching of three steroid molecules. Finally in Section 4 we make an assessment of our methods and suggest directions for future work.

2 Hierarchical modelling of multi-configuration alignment

In this section we consider a hierarchical model for matching multiple configurations simultaneously. We closely follow the two-configuration method of Green and Mardia (2006), though we must now allow for the possibility of many types of matches.

2.1 The alignment problem

Suppose we are given C configurations $x^{(1)}, x^{(2)}, \ldots, x^{(C)}$, whose points are recorded in d-dimensional real space: for $c = 1, 2, \ldots, C$, write $x^{(c)} = \left\{x_j^{(c)}, j = 1, 2, \ldots, n_c\right\}$, where $x_j^{(c)} \in \mathbb{R}^d$ and n_c is the number of points in configuration $x^{(c)}$. The labelling is assumed to be arbitrary, thus providing no initial information on the correspondences between points. We wish to align the C configurations simultaneously by establishing these correspondences and filtering out the relative transformations between the configurations.

We introduce a set of hidden locations $\mu = \{\mu_i\} \subset \mathbb{R}^d$. These can be seen as the 'true' locations of the configuration points, so that the latter are noisy observations of the former. Specifically, define the labelling arrays $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(C)}$, which link the index of an observation to that of its corresponding hidden point. In particular $\xi_j^{(c)}$ is the index of the μ–point underlying the observation $x_j^{(c)}$. Assume that a hidden location is observed at most once in each configuration, and that it may remain unobserved. Thus the elements within each $\xi^{(c)}$ are distinct, and a μ–point may generate anywhere between zero and C configuration points.

Now suppose each configuration goes through some transformation before being observed. For $c = 1, 2, \ldots, C$, let $\mathcal{A}^{(c)}$ be the transformation bringing the points of the $x^{(c)}$ configuration *back* to the reference frame defined by the μ–points. Our C-configuration alignment model can be expressed as:

$$\mathcal{A}^{(c)} x_j^{(c)} = \mu_{\xi_j^{(c)}} + \varepsilon_j^{(c)}, \quad \text{for } j = 1, 2, \ldots, n_c, \quad c = 1, 2, \ldots, C. \tag{1}$$

The random error vector $\varepsilon_j^{(c)}$ is assumed to have density $f^{(c)}$ and to be independent of the μ–points and of all the other errors.

Our primary objective with this model is to *match* observations of the same μ–point. Formally we wish to find maximal sets of configurations $\left\{x^{(i_1)}, x^{(i_2)}, \ldots, x^{(i_k)}\right\}$ and indices $\{j_1, j_2, \ldots, j_k\}$ such that $\xi^{(i_1)}_{j_1} = \xi^{(i_2)}_{j_2} = \cdots = \xi^{(i_k)}_{j_k}$. There is an abuse of language here in that a match may involve points taken from more than two configurations.

Below we will assume that each $\mathcal{A}^{(c)}$ is an affine transformation, made up of a linear transformation matrix $A^{(c)}$ and a translation vector $\tau^{(c)}$ so that $\mathcal{A}^{(c)} x^{(c)}_j = A^{(c)} x^{(c)}_j + \tau^{(c)}$ for $j = 1, \ldots, n_c$ and $c = 1, \ldots, C$. We will require constraints on the $\mathcal{A}^{(c)}$ to ensure identifiability of the model; we discuss this in Section 2.5.

2.2 Hierarchical modelling: preliminaries

We define a generic set $I \subset \{1, 2, \ldots, C\}$ of configuration indices, with $I \neq \emptyset$. The set I corresponds to a 'type' of match. For example if $C = 3$, then $I = \{2, 3\}$ refers to a match involving a point from the $x^{(2)}$ configuration and a point from the $x^{(3)}$ configuration but none from the $x^{(1)}$ configuration. We call *I–match* a match involving *exactly* the configurations whose index is included in I. If $I = \{i_1, i_2, \ldots, i_K\}$, an I–match can be represented by an index array $(j_1, j_2, \ldots, j_K)$ such that $\xi^{(i_1)}_{j_1} = \xi^{(i_2)}_{j_2} = \cdots = \xi^{(i_K)}_{j_K}$ and such that $c \notin I$ implies $\xi^{(c)}_l \neq \xi^{(i_k)}_{j_k}$ for all $l = 1, 2, \ldots, n_c$ and all $k = 1, 2, \ldots, K$. We write $|I|$ as the number of configuration indices in I. If $|I| = 1$ our I-match is in fact an unmatched point; this will also be treated as a type of match.

The totality of the matches is stored in a structure which we call $\mathcal{M}$. How these matches are represented is irrelevant for the moment; one might wish to use binary matrices, as do Dryden et al. (2007) and Green and Mardia (2006), or write each match as an index array containing the labels of the matched points and those of the configurations involved. This structure $\mathcal{M}$ is the main parameter of interest in our hierarchical model; we stress the fact that it contains no direct information on the μ–points themselves or on the actual values in the labelling vectors $\xi^{(c)}$. Note that for $\mathcal{M}$ to be consistent with the model described in Section 2.1, we must ensure that a given point is never involved in more than one (maximal) match. The directed acyclic graph (DAG) of our hierarchical model, including this new parameter $\mathcal{M}$, is displayed in Figure 1.

FIGURE 1 ABOUT HERE

2.3 Poisson process assumption and prior distribution for the matches

We make the prior assumption that the μ–points follow a multivariate Poisson process with constant rate λ over a region $V \subset \mathbb{R}^d$ of volume v. Recall that each μ–point generates a number of observations or remains unobserved. For I as defined in the previous section, let q_I be the probability that a given hidden location generates an I–match. For instance, if $C = 3$ then $q_{\{1,3\}}$ is the probability that a particular μ–point is observed in the $x^{(1)}$ and $x^{(3)}$ configurations but not in the $x^{(2)}$ configuration. Thus the probability of a hidden location remaining unobserved is $1 - \sum_I q_I$. Assume also that the matches are generated independently from μ–point to μ–point, based on the same probabilities q_I. A very useful consequence of these assumptions is that our global Poisson process can be partitioned into 2^C *thinned* Poisson processes: for fixed I, the set of μ–points which have generated an I–match is itself a Poisson process with rate λq_I; furthermore this process will be independent of the other processes of the partition.

We define the parametrisation

$$q_I \quad = \quad \rho_I \cdot \prod_{c \in I} q_{\{c\}}, \tag{2}$$

where $\rho_I = 1$ if $|I| = 1$. This type of parameterisation has been treated in other contexts, such as that of regression with binary response (Ekholm et al. 1995) and genetic map functions (Speed 2005). The parameter ρ_I is sometimes called the dependence ratio or coincidence coefficient. Here it can be seen as a relative measure of how likely an I–match is to occur *a priori*.

Now we wish to assign a prior distribution to the match structure $\mathcal{M}$, based on the Poisson process assumptions described above. Let L_I be the number of I–matches contained in $\mathcal{M}$. Given $\{n_c, c = 1, 2, \ldots, C\}$, we must have

$$L_{\{c\}} \quad = \quad n_c - \sum_{\{I:|I| \geq 2, I \ni c\}} L_I, \quad \text{for } c = 1, 2, \ldots, C. \tag{3}$$

The prior distribution for $\mathcal{M}$ can be decomposed as

$$p(\mathcal{M}) \quad = \quad p(\mathcal{M} \mid \{L_I\}) \cdot p(\{L_I\}). \tag{4}$$

From the Poisson process assumption on the μ–points, the counts L_I are independent Poisson variables with means $\lambda v q_I$. Using (2) and (3), we find that the prior distribution for the match counts has the form

$$p(\{L_I\}) \propto \prod_I (\lambda v q_I)^{L_I} \Big/ \prod_I L_I! \propto \prod_I \left\{ \frac{\rho_I}{(\lambda v)^{|I|-1}} \right\}^{L_I} \Big/ \prod_I L_I! \ , \tag{5}$$

so the q_I parameters conveniently disappear.

Now make the prior assumption that, conditional on the match counts $\{L_I\}$, the distribution for $\mathcal{M}$ is uniform. In other words, consider as equally likely each match arrangement which is consistent with the counts. The number of such arrangements is

$$\prod_{c=1}^{C} n_c! \Big/ \prod_I L_I! \ ,$$

as can be seen using a recursion argument. Using (4) and (5), it follows that the prior distribution for $\mathcal{M}$ has the form

$$p(\mathcal{M}) \propto \prod_I \left\{ \frac{\rho_I}{(\lambda v)^{|I|-1}} \right\}^{L_I} . \tag{6}$$

This distribution depends only on the number of matches of each type contained in $\mathcal{M}$, and is parametrised by the ratios $\rho_I/(\lambda v)^{|I|-1}$.

2.4 Joint model

Now we seek to compute the joint likelihood of $\mathcal{M}$ and $\mathcal{A} = \left\{\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(C)}\right\}$ given the set of configurations $X = \left\{x^{(1)}, x^{(2)}, \ldots, x^{(C)}\right\}$.

Fix $I = \{i_1, i_2, \ldots, i_K\}$ and let $\left\{x_{j_1}^{(i_1)}, x_{j_2}^{(i_2)}, \ldots, x_{j_K}^{(i_K)}\right\}$ be the points of a given I–match in $\mathcal{M}$, where of course $K = |I|$. From (1), we find that

$$p\left(x_{j_1}^{(i_1)}, x_{j_2}^{(i_2)}, \ldots, x_{j_K}^{(i_K)} \mid \mathcal{A}, \mu, \xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(C)}\right) = \prod_{k=1}^{K} \left|A^{(i_k)}\right| f^{(i_k)}\left(\mathcal{A}^{(i_k)} x_{j_k}^{(i_k)} - \mu_{\xi_{j_1}^{(i_1)}}\right),$$

where $|A|$ denotes the absolute value of the determinant of the matrix A. Now consider the set of μ–points which have generated an I–match — we mentioned that this set follows a Poisson process. Given $\mathcal{M}$, and therefore given L_I, the points of this set are uniformly distributed over the region V. As a result the contribution of the matched points defined above to our likelihood is

$$p\left(x_{j_1}^{(i_1)}, x_{j_2}^{(i_2)}, \ldots, x_{j_K}^{(i_K)} \mid \mathcal{A}, \mathcal{M}\right) = v^{-1} \int_V \prod_{k=1}^{K} \left|A^{(i_k)}\right| f^{(i_k)}\left(\mathcal{A}^{(i_k)} x_{j_k}^{(i_k)} - \mu\right) \mathrm{d}\mu.$$

The above integration will be carried out over $\mathbb{R}^d$. We are thus ignoring the edge effects from the boundary of V: this is valid if V is taken large enough relative to the support of the error densities $f^{(c)}$.

Suppose $I = \left\{i_1, i_2, \ldots, i_{|I|}\right\}$ and let S_I be the set of I–matches contained in $\mathcal{M}$. The elements of S_I are written as index arrays of the form $\left(j_1, j_2, \ldots, j_{|I|}\right)$, with the convention that $\left\{x_{j_1}^{(i_1)}, x_{j_2}^{(i_2)}, \ldots, x_{j_{|I|}}^{(i_{|I|})}\right\}$ is the corresponding set of matched points. The contribution of the I–matches to the likelihood is

$$v^{-L_I} \prod_{(j_1,\ldots,j_{|I|}) \in S_I} \int_{\mathbb{R}^d} \prod_{k=1}^{|I|} \left|A^{(i_k)}\right| f^{(i_k)}\left(\mathcal{A}^{(i_k)} x_{j_k}^{(i_k)} - \mu\right) \mathrm{d}\mu.$$

Multiplying over all match types, the full likelihood of $\mathcal{A}$ and $\mathcal{M}$ can be seen to be

$$p(X \mid \mathcal{A}, \mathcal{M}) = \left(v^{-\sum_I L_I} \prod_{c=1}^{C} \left|A^{(c)}\right|^{n_c}\right) \times \prod_I \prod_{(j_1,\ldots,j_{|I|}) \in S_I} \int_{\mathbb{R}^d} \prod_{k=1}^{|I|} f^{(i_k)}\left(\mathcal{A}^{(i_k)} x_{j_k}^{(i_k)} - \mu\right) \mathrm{d}\mu. \tag{7}$$

We introduce prior distributions $p\left(A^{(c)}\right)$ and $p\left(\tau^{(c)}\right)$ for the transformation parameters, for $c = 1, 2, \ldots, C$. These priors are left undefined for the time being. The parameters λ, v, and ρ_I are treated as fixed. From (6) and (7), the joint posterior distribution has the form

$$\begin{aligned} p(\mathcal{A}, \mathcal{M} \mid X) \propto & \prod_{c=1}^{C} \left\{p(A^{(c)}) p(\tau^{(c)}) \left|A^{(c)}\right|^{n_c}\right\} \\ & \times \prod_I \prod_{(j_1,\ldots,j_{|I|}) \in S_I} \frac{\rho_I}{\lambda^{|I|-1}} \int_{\mathbb{R}^d} \prod_{k=1}^{|I|} f^{(i_k)}\left(\mathcal{A}^{(i_k)} x_{j_k}^{(i_k)} - \mu\right) \mathrm{d}\mu. \end{aligned} \tag{8}$$

Here and elsewhere, the '$\propto$' symbol indicates proportionality with respect to the variables to the left of the conditioning sign. Thus the μ–points and labelling arrays have been effectively integrated out; the relevant information contained in these parameters is captured by the structure $\mathcal{M}$. Note also that the volume v plays no role in our posterior distribution.

Now assume the error densities $f^{(c)}$ are centred Gaussian densities with covariance matrices all equal to $\sigma^2 I_d$. In this case the integrals in (8) can be written in closed form: for a given set of points $\left\{x_{j_1}^{(i_1)}, x_{j_2}^{(i_2)}, \ldots, x_{j_{|I|}}^{(i_{|I|})}\right\}$, define

$$\gamma_{\mathcal{A}}\left(x_{j_1}^{(i_1)}, x_{j_2}^{(i_2)}, \ldots, x_{j_{|I|}}^{(i_{|I|})}\right) = \sum_{k=1}^{|I|} \left\| \mathcal{A}^{(i_k)} x_{j_k}^{(i_k)} - c \right\|^2,$$

where

$$c = \frac{1}{|I|} \sum_{k=1}^{|I|} \mathcal{A}^{(i_k)} x_{j_k}^{(i_k)}$$

and $\|\cdot\|$ is the Euclidean norm. Thus $\gamma_{\mathcal{A}}\left(x_{j_1}^{(i_1)}, \ldots, x_{j_{|I|}}^{(i_{|I|})}\right)$ is a measure of the deviation in the transformed points $\left\{\mathcal{A}^{(i_1)} x_{j_1}^{(i_1)}, \ldots, \mathcal{A}^{(i_{|I|})} x_{j_{|I|}}^{(i_{|I|})}\right\}$. With this notation and the Gaussian assumption for the errors, one finds that

$$\begin{aligned} \int_{\mathbb{R}^d} \prod_{k=1}^{|I|} f^{(i_k)}\left(\mathcal{A}^{(i_k)} x_{j_k}^{(i_k)} - \mu\right) \mathrm{d}\mu &= |I|^{-d/2}(2\pi\sigma^2)^{-d(|I|-1)/2} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma^2}\gamma_{\mathcal{A}}\left(x_{j_1}^{(i_1)}, x_{j_2}^{(i_2)}, \ldots, x_{j_{|I|}}^{(i_{|I|})}\right)\right\}. \end{aligned} \tag{9}$$

This identity is valid if $|I| = 1$, since $\gamma_{\mathcal{A}}\left(x_{j_1}^{(i_1)}\right) = 0$. Now a prior distribution $p\left(\sigma^2\right)$ can be introduced and the variance parameter σ^2 incorporated in the model (8).

2.5 Inference with Markov chain Monte Carlo

We wish to make inference on the parameters of the model (8), given the data configurations X. The parameters of interest are the error variance σ^2, the translations $\tau^{(c)}$, the transformation matrices $A^{(c)}$, and of course the matches $\mathcal{M}$. The ratios $\rho_I/\lambda^{|I|-1}$ will be considered as fixed hyperparameters, estimated through some other method.

The unwieldy aspect of the joint distribution (8) makes it difficult to use classical estimation methods in this context. An attractive possibility here is to use Markov chain Monte Carlo (MCMC) simulation. We simulate a Markov chain by updating the parameters in sweeps, in such a way that the underlying transition kernel of the chain verifies detailed balance, with (8) as the stationary, or limiting, distribution. The sampled chain can be used as a basis for inference, provided it has reached equilibrium. For an accessible introduction to MCMC methods, see for example Green (2001), while Robert and Casella (2004) give a more detailed account.

To simplify our method, we will make the assumption that the transformation matrices $A^{(c)}$ are rotation matrices. Thus we are concentrating on rigid-body transformations, and the point configurations can be seen as elements of a size-and-shape space (Dryden and Mardia, 1998).

The C++ implementation of the algorithm (with R interface), including instructions and functions for post-processing, can be found on the URL *http://ima.epfl.ch/∼ruffieux/multalign/*.

Updating the continuous parameters

For the parameters σ^2, $\tau^{(c)}$, and $A^{(c)}$, for $c = 1, 2, \ldots, C$, conditionally conjugate priors can be found which result in full conditional distributions of the same form. This will make updating these parameters relatively straightforward. The conjugacy assumptions are not particularly restrictive here: in practice we do not expect to make use of strong prior information on the continuous parameters.

We assign an inverse gamma prior distribution to σ^2; in particular we set $\sigma^{-2} \sim \Gamma(a, b)$, where a and b are respectively the shape and rate parameters of the gamma distribution. From (8) and (9), the full conditional distribution of σ^{-2} is

$$\left(\sigma^{-2} \mid \mathcal{A}, \mathcal{M}, X\right) \sim \Gamma(\tilde{a}, \tilde{b}),$$

where

$$\tilde{a} = a + \frac{d}{2} \sum_{I} L_I(|I| - 1),$$

and

$$\tilde{b} = b + \frac{1}{2} \sum_{I} \sum_{(j_1,\ldots,j_{|I|}) \in S_I} \gamma_{\mathcal{A}} \left(x_{j_1}^{(i_1)}, \ldots, x_{j_{|I|}}^{(i_{|I|})} \right).$$

Thus the error variance can be updated using a Gibbs sampler step, i.e. by simulating from the full conditional inverse gamma distribution.

Set $\mathcal{A}^{(-c)} = \{\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(c-1)}, \mathcal{A}^{(c+1)}, \ldots, \mathcal{A}^{(C)}\}$. We choose to assign Gaussian priors to the translation parameters. For $c = 1, 2, \ldots, C$, suppose *a priori* that $\tau^{(c)} \sim \mathcal{N}_d(\mu^{(c)}, \eta_c^2 I_d)$. Using (8),

$$\left(\tau^{(c)} \mid \sigma^2, \mathcal{A}^{(-c)}, A^{(c)}, \mathcal{M}, X\right) \sim \mathcal{N}_d \left(\frac{\frac{1}{\eta_c^2}\mu^{(c)} + \frac{1}{\sigma^2} m_c}{\frac{1}{\eta_c^2} + \frac{1}{\sigma^2} w_c}, \frac{1}{\frac{1}{\eta_c^2} + \frac{1}{\sigma^2} w_c} I_d \right),$$

with

$$m_c = \sum_{I: I \ni c} \sum_{(j_1,\ldots,j_{|I|}) \in S_I} \frac{1}{|I|} \left\{ \left(\sum_{k: i_k \neq c} \mathcal{A}^{(i_k)} x_{j_k}^{(i_k)} \right) - (|I| - 1) A^{(c)} x_{j_{k(c)}}^{(c)} \right\}$$

and

$$w_c = \sum_{I: I \ni c} \left(\frac{|I| - 1}{|I|} \right) L_I,$$

and where the sub-index $k(c)$ is such that $x^{(i_{k(c)})} = x^{(c)}$. The translation parameters are thus also updated using a Gibbs move.

We can also find conjugate priors for the rotation matrices $A^{(c)}$, though this is less obvious. For $c = 1, 2, \ldots, C$, set $p(A^{(c)}) \propto \exp\left\{\mathrm{tr}(F_c^T A^{(c)})\right\}$ for some $d \times d$ matrix F_c. The fact that we are concentrating on rotation matrices means that $\left|A^{(c)}\right| = 1$ and $\left(A^{(c)}\right)^{-1} = \left(A^{(c)}\right)^T$. A somewhat involved calculation yields

$$p(A^{(c)} \mid \sigma^2, \mathcal{A}^{(-c)}, \tau^{(c)}, \mathcal{M}, X) \propto \exp\left[\mathrm{tr}\left\{(F_c + S_c)^T A^{(c)}\right\}\right],$$

where $\mathrm{tr}(\cdot)$ is the trace operator and S_c the $d \times d$ matrix

$$S_c = \frac{1}{\sigma^2} \sum_{I:I \ni c} \sum_{(j_1,\ldots,j_{|I|}) \in S_I} \frac{1}{|I|} \left\{ \left(\sum_{k:i_k \neq c} \mathcal{A}^{(i_k)} x_{j_k}^{(i_k)} \right) - (|I|-1)\tau^{(c)} \right\} \left(x_{j_{k(c)}}^{(c)} \right)^T.$$

The conditionally conjugate distribution $p(A) \propto \exp\left\{\mathrm{tr}(F^T A)\right\}$ is called the matrix Fisher distribution, and is well-known in directional statistics (Mardia and Jupp 2003, p.289). Rather than updating the rotation matrices themselves, we will work on the corresponding rotation angles. For example if $d = 2$, we define the angle $\theta^{(c)}$, while if $d = 3$ we have the three generalised Euler angles $\theta_{12}^{(c)}, \theta_{23}^{(c)}$, and $\theta_{13}^{(c)}$. Green and Mardia (2006, pp. 241–242) describe how the angles can be updated when assuming a conjugate matrix Fisher prior for the rotation matrices, in the case where d is 2 or 3.

For simplicity, we consider here only the case where the $A^{(c)}$ are uniformly distributed and mutually independent. This is achieved by assigning zero matrices to the F_c above. It is then true that the relative rotations $\left(A^{(c_1)}\right)^T \cdot A^{(c_2)}$ are uniform and mutually independent for $c_2 \neq c_1$ and fixed c_1. So without loss of generality, we can impose the identifying constraint that $\mathcal{A}^{(1)}$ be fixed as the identity transformation. This is the same as saying that the first data configuration lies in the same frame as the hidden point locations.

Updating the matches

The matches will be updated using a Metropolis–Hastings jump. We write

$$\mathcal{M} = \left\{ \left(t_1^1, t_2^1, \ldots, t_C^1\right), \left(t_1^2, t_2^2, \ldots, t_C^2\right), \ldots, \left(t_1^K, t_2^K, \ldots, t_C^K\right) \right\},$$

with $K = \sum_I L_I$. Each C-tuple $(t_1^k, t_2^k \ldots, t_C^k)$ represents a match, t_c^k being the index of the point from the $x^{(c)}$ configuration involved in the match. If a given configuration is not involved in the match, a '$-$' flag is inserted at the appropriate position. For instance, if $C = 3$ the 3-tuple $(2, 4, 1)$ refers to a match between

$x_2^{(1)}, x_4^{(2)}$ and $x_1^{(3)}$, while $(-,2,1)$ is a match between $x_2^{(2)}$ and $x_1^{(3)}$, with no $x^{(1)}$–point involved. We also include unmatched points in this list: $(1,-,-)$ indicates that $x_1^{(1)}$ is unmatched, for example.

Suppose that $\mathcal{M}$ is the current list of matches in the MCMC algorithm. The jump proposal proceeds as follows:

- with probability q we choose to *split* a C-tuple; in this case we draw an element uniformly at random in the list $\mathcal{M}$.
  - If the C-tuple drawn corresponds to an unmatched point, we do nothing;
  - otherwise we split it into two C-tuples; for instance $(2,3,1)$ can be split into $(2,-,-)$ and $(-,3,1)$. In general there will be many potential splits: here we could have chosen to split $(2,3,1)$ into $(-,3,-)$ and $(2,-,1)$. Suppose the match to be split is an I–match. Then there are $B_I = 2^{|I|-1} - 1$ ways to split this match. We select one of these splits uniformly at random.
- With probability $1-q$ we choose to *merge* two C-tuples; in this case we select two distinct elements uniformly at random from $\mathcal{M}$.
  - If the two C-tuples drawn contain a common configuration, i.e. $(j_1,k,-)$ and $(j_2,-,-)$, then we do nothing;
  - otherwise we merge the C-tuples, for example $(j,k,-)$ and $(-,-,l)$ become (j,k,l), while $(-,k,-)$ and $(-,-,l)$ become $(-,k,l)$.

The split and merge operations defined above form a complementary reversible pair. Clearly, all possible match arrangements can be explored using these two operations only.

The acceptance probability of a jump is readily worked out from (8). Suppose the proposal is to split an I–match into an I'-match and an I''-match, such that $I = I' \cup I''$ and $I' \cap I'' = \emptyset$. Reverting to the algebraic representation, suppose we are splitting $\left(x_{j_1}^{(i_1)}, \ldots, x_{j_{|I|}}^{(i_{|I|})}\right)$ into $\left(x_{j'_1}^{(i'_1)}, \ldots, x_{j'_{|I'|}}^{(i'_{|I'|})}\right)$ and $\left(x_{j''_1}^{(i''_1)}, \ldots, x_{j''_{|I''|}}^{(i''_{|I''|})}\right)$. The acceptance probability for this proposal is $\min\{1, p_S\}$, where

$$p_S = \left(\frac{\rho_{I'}\rho_{I''}\lambda}{\rho_I}\right) \times \left(\frac{2\pi\sigma^2|I|}{|I'||I''|}\right)^{d/2} \times \frac{2(1-q)B_I}{q(K+1)} \times \frac{\exp\left\{-\frac{1}{2\sigma^2}\gamma_{\mathcal{A}}\left(x_{j'_1}^{(i'_1)}, \ldots, x_{j'_{|I'|}}^{(i'_{|I'|})}\right)\right\}\exp\left\{-\frac{1}{2\sigma^2}\gamma_{\mathcal{A}}\left(x_{j''_1}^{(i''_1)}, \ldots, x_{j''_{|I''|}}^{(i''_{|I''|})}\right)\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\gamma_{\mathcal{A}}\left(x_{j_1}^{(i_1)}, \ldots, x_{j_{|I|}}^{(i_{|I|})}\right)\right\}}.$$

This acceptance probability is also valid when at least one of the new matches after the split is an unmatched point — recall that $\gamma_{\mathcal{A}}\left(x_{j_1}^{(i_1)}\right) = 0$ and that $\rho_I = 1$ if $|I| = 1$. Now suppose we attempt to merge $\left(x_{j'_1}^{(i'_1)}, \ldots, x_{j'_{|I'|}}^{(i'_{|I'|})}\right)$ and $\left(x_{j''_1}^{(i''_1)}, \ldots, x_{j''_{|I''|}}^{(i''_{|I''|})}\right)$ into $\left(x_{j_1}^{(i_1)}, \ldots, x_{j_{|I|}}^{(i_{|I|})}\right)$. The acceptance probability for this jump is $\min\{1, p_M\}$, where

$$
\begin{aligned}
p_M \quad = \quad & \left(\frac{\rho_I}{\rho_{I'}\rho_{I''}\lambda}\right) \times \left(\frac{|I'||I''|}{2\pi\sigma^2|I|}\right)^{d/2} \times \frac{qK}{2(1-q)B_I} \\
& \times \frac{\exp\left\{-\frac{1}{2\sigma^2}\gamma_{\mathcal{A}}\left(x_{j_1}^{(i_1)}, \ldots, x_{j_{|I|}}^{(i_{|I|})}\right)\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\gamma_{\mathcal{A}}\left(x_{j'_1}^{(i'_1)}, \ldots, x_{j'_{|I'|}}^{(i'_{|I'|})}\right)\right\}\exp\left\{-\frac{1}{2\sigma^2}\gamma_{\mathcal{A}}\left(x_{j''_1}^{(i''_1)}, \ldots, x_{j''_{|I''|}}^{(i''_{|I''|})}\right)\right\}}.
\end{aligned}
$$

The simplicity of the jump proposal has a drawback: the change to $\mathcal{M}$ is very small relative to its parameter space. To speed up the exploration of this space, we will typically make several match jump proposals within each sweep of the MCMC algorithm.

3 Application: Aligning steroid molecules

3.1 The Data

For this example we select $C = 3$ steroid molecules from the CoMFA database, which can be accessed at *http://www2.ccc.uni-erlangen.de/services/steroids*. This database is frequently used as a benchmark for testing computer-assisted drug design methods (see Coats, 1998). The three molecules are aldosterone, cortisone, and prednisolone, which we label $x^{(1)}, x^{(2)}$, and $x^{(3)}$ respectively. Each of these molecules contain $n_1 = n_2 = n_3 = 54$ arbitrarily labelled atoms in $d = 3$ dimensional space. We wish to align these molecules using the methodology described in this paper.

Here we have seven types of matches to deal with, counting the unmatched types: $\{1\}$, $\{2\}$, $\{3\}$, $\{1, 2\}$, $\{2, 3\}$, $\{1, 3\}$, and $\{1, 2, 3\}$. For simplicity we will drop the brackets and commas from the I sets when appropriate, so that for example $L_{\{1,2\}}$ and $\rho_{\{2,3\}}$ are written as L_{12} and ρ_{23}, and a 13–match is a match involving the first and third configurations but not the second.

3.2 Results

The MCMC algorithm was launched with a ‘clean sheet’: no initial matches were assigned and no information about the atoms was considered. We set $\rho_{12}/\lambda = \rho_{23}/\lambda = \rho_{13}/\lambda = 31.25$, and $\rho_{123}/\lambda^2 = 3660$. By convention the values of ρ_1, ρ_2, and ρ_3 are fixed at one. The error variance parameters were set to $a = 1$ and $b = 0.1$. The transformation priors were rendered largely non-informative by setting $\mu^{(2)} = \mu^{(3)} = 0$ and $\eta_2 = \eta_3 = 10$ and by assigning the zero matrix to F_2 and F_3. The sampler was run for 50 000 sweeps, the first 10 000 being discarded as burn-in; 50 match proposals were made per sweep.

Figures 2 and 3 give the time series traces of the transformation parameters and match counts respectively. The rotation matrices $A^{(2)}$ and $A^{(3)}$ are each represented by three angles. From inspection of these traces as well as that of the posterior likelihood (not plotted here), we conclude the chain has reached equilibrium. The error variance and translations were estimated using the sample posterior means based on a subsample of 2000 after burn-in; we find $\bar{\sigma}^2 = 0.0075$, $\bar{\tau}^{(2)} = (-1.224, -0.639, -0.786)^T$, and $\bar{\tau}^{(3)} = (-0.796, -0.444, -0.640)^T$. The rotation matrices were estimated by taking their respective sampled polar parts (see Green and Mardia 2006, p.248), giving us the estimates

$$\widehat{A}^{(2)} = \begin{pmatrix} 0.967 & 0.136 & -0.216 \\ -0.166 & 0.977 & -0.131 \\ 0.193 & 0.163 & 0.968 \end{pmatrix}, \quad \widehat{A}^{(3)} = \begin{pmatrix} 0.888 & 0.186 & -0.420 \\ -0.141 & 0.980 & 0.137 \\ 0.438 & -0.063 & 0.897 \end{pmatrix}.$$

The posterior sample means of the match counts were $(\bar{L}_{12}, \bar{L}_{23}, \bar{L}_{13}, \bar{L}_{123}) = (4.59, 4.46, 1.14, 42.61)$.

FIGURES 2 AND 3 ABOUT HERE

To estimate $\mathcal{M}$, we rank the matches by order of their sample posterior probability, and select the k most frequent, say. Providing these are mutually compatible, the resulting estimator $\widehat{\mathcal{M}}$ can be considered as a Bayesian estimator under a certain loss function (again Green and Mardia 2006, pp.243-244). By mutually compatible matches we mean that 1) they do not imply that two different points in one configuration are matched to a single point in another configuration, 2) they do not imply that two different points are matched to a same point but are not matched to each other. Incompatible matches are possible, though quite rare in practice, and cannot happen if the selected matches have posterior probability larger than 0.5.

Here we find that 47 matches have probability higher than 0.9 and 54 have probability higher than 0.5. Of these 54 matches, 44 are 123–matches, 4 are 12–matches, 5 are 23–matches, and one is a 31–match. The three molecules are aligned in Figure 4. Notice that in the top right corner of the latter figure, there seems to be a non-random observation error in the matched points. This might be a result of assigning too large a value to ρ_{123}/λ^2, as will be seen in Section 3.3. It could also be the consequence of systematic model error: the assumption of rigid-body transformations might be invalid for instance.

FIGURE 4 ABOUT THERE

To study the vulnerability of the sampler to local modes, the following experiment was conducted: 100 independent MCMC runs were launched, all with the hyperparameters fixed at the values given above. After 50 000 sweeps of a run, the posterior likelihood (8) was computed and compared to a threshold value established from earlier runs, such as the one described above. If the likelihood was lower than this threshold, then the sampler was adjudged to have become trapped in a local mode. Of the 100 runs, 91 passed this test and thus were deemed to have found their way to the main mode of the distribution. Of course these favorable results might be more a consequence of the nature of the data than of the robustness of our method. In a different context one might need to devise a more elaborate MCMC algorithm to deal with multimodality.

3.3 Prior settings

We now briefly study the effect of the hyperparameter values on the MCMC inference. In the case where no prior information on the transformation parameters is available, it is convenient to set a uniform prior on the rotation matrix and on the directions of the translations, as was done in Section 3.2. Furthermore, we typically select the variances η_2^2 and η_3^2 to be large enough so that the resulting translations encompass the configurations. We study in a little more detail the effect of the hyperparameters a, b, and $\rho_I/\lambda^{|I|-1}$.

In the above runs we set $a = 1$, thus assigning an exponential prior distribution to $1/\sigma^2$. The second hyperparameter b determines the rate of this distribution: larger b should result in larger variance in the observation errors, and thus more variability in the matching. Conversely, the smaller the value of b, the closer together a set of (transformed) points will have to be in order to be considered as a candidate for a match. Increasing $b = 0.1$ by a factor of ten will double the posterior mean of σ^2. Inference on the matches is only slightly affected: a few two-way matches are replaced by 123–matches. Also, this increase generates a local mode problem, as some of the runs become entangled in a minor mode for an indefinite time. This is to be expected, since by increasing b we are allowing the sampler to explore additional alignments. Reducing $b = 0.1$ to $b = 0.01$ has little effect on either the matching or the posterior mean of the variance. However, reducing it further seems to create a second major mode in the posterior distribution, causing the algorithm to switch continually between two alignments. This new alignment is very similar to the first, except that approximately 10 of the 123–matches are replaced by 23–matches. The likely explanation for this is that in reducing b, we have become less tolerant towards matching, and thus have split several ‘borderline’

123–matches.

Now we study the influence of the ratios $r_I = \rho_I / \lambda^{|I|-1}$ on the matching. Recall that each hyperparameter r_I appears in the prior distribution (6) for the matches; we expect that increasing r_I will result in more I-matches being accepted in the algorithm. These ratios may be estimated by taking advantage of prior 'guesses' one might have on the number of matches of each type. When such information is available, as is often the case in practice, the argument of Green and Mardia (2006, p. 250) can be extended to the multivariate distribution (5). Suppose we have established the guesses $\{\tilde{L}_I\}$ for the match counts: if we set

$$r_I = \tilde{L}_I \cdot v^{|I|-1} \Big/ \prod_{c \in I} \tilde{L}_{\{c\}},$$

then the resulting prior distribution for the counts will have a unique mode in $\{\tilde{L}_I\}$. The value for the volume v must be determined from the data, but this is not usually difficult to do. For example the ratios chosen in Section 3.2 are based on the guesses $\tilde{L}_{12} = \tilde{L}_{23} = \tilde{L}_{13} = 8$ and $\tilde{L}_{123} = 30$, with v fixed at 250. The actual values given to the r_I appear less important than their *relative* values. For instance increasing (or decreasing) all the ratios by a factor of ten brings about very little change in the resulting MCMC inference. Now suppose the values of the r_I parameters are determined using the guesses $\tilde{L}_{12} = \tilde{L}_{13} = 5$, $\tilde{L}_{23} = 25$ and $\tilde{L}_{123} = 20$. The resulting inference is pictured in Figure 5, where again we have selected the 54 most probable matches. Clearly the alignment is very similar to the one displayed in Figure 4, except the 'borderline' 123–matches mentioned earlier have been replaced by 23–matches. The above simulations illustrate an interesting feature of our methodology, namely that prior information on the match counts may influence subtle aspects of the alignment inference.

FIGURE 5 ABOUT HERE

3.4 Multiple vs. pairwise matching

We briefly consider what can be gained by using our multiple matching approach rather than aligning the configurations independently by pairs. For this purpose we add two further steroid molecules 11-deoxycorticosterone and 17a-hydroxyprogesterone ($x^{(4)}$ and $x^{(5)}$ respectively) to the three described above.

First we treat the pairwise alignment of molecule $x^{(1)}$ to molecules $x^{(2)}, x^{(3)}, x^{(4)}$, and $x^{(5)}$ respectively. For each of these alignments we based the match prior on the 'guess' $\tilde{L}_{12} = 30$. The four MCMC means for the number of unmatched atoms in the first molecule are between 6 and 10. If we align the five molecules simultaneously, based namely on the guesses $\tilde{L}_{12345} = 30$ and $\tilde{L}_{2345} = 3$, we find a posterior mean of $\bar{L}_1 = 21$ unmatched atoms in the first molecule. Thus a fair portion of the first molecule has disengaged from the other four (note that this was already somewhat apparent when aligning three molecules, see Section 3.3). It is also worth mentioning that, relative to the pairwise cases, the posterior estimate for σ^2 decreases roughly by a factor of eight. For reference the pairwise and multiple alignments are displayed in Figures 6 and 7 respectively. It is conceivable that the difference in the inferences is a result of the prior match specifications. However, one would have to set $\tilde{L}_{12}$ to be as low as 10 to obtain pairwise alignments similar to the multiple alignment. This suggests that, in this context at least, the pairwise approach has a proclivity for overmatching.

FIGURE 6 AND FIGURE 7 ABOUT HERE

The above example illustrates that the inclusion of two or three additional configurations may have a strong impact on the alignment inference. One might understand this as a 'borrowing of strength' of sorts: further configurations provide further information on the number and location of implied μ-points, information which can in turn be exploited in the alignment of the initial configurations. Clearly, there is no way to take advantage of this information if the molecules are aligned by pairs.

4 Discussion

In this paper we have seen that the two-configuration Bayesian matching approach of Green and Mardia (2006) generalises readily to the multi-configuration context. The methodology was applied to the matching of three steroid molecules, with promising results: with this 'easy' dataset, the sampler seemed to have little difficulty avoiding the anticipated local mode problem.

The problem of aligning multiple molecules has also been treated by Dryden et al. (2007); their approach is similar to ours, in that a hierarchical model is constructed and a hidden reference molecule defined. However the hidden points are not integrated out, and the transformations are maximised out using Procrustean registration techniques. Furthermore, only C 'types' of matches are considered in their model (compared to our $2^C - C - 1$): the alignment is made pairwise between each observed point configuration and the hidden molecule. A result of this is that, in terms of computation speed, the methodology of Dryden et al. (2007) would probably be much faster than the one proposed in this paper when C is large. So the choice of method might depend upon the number of configurations to be aligned and the extent to which one wished to retain full statistical efficiency and control the prior match specifications.

An important aspect of alignment which is not addressed in this paper is that of *marking*. In many contexts, additional information on the observations is available. For example, Dryden et al. (2007) include 'marks' on each atom of the molecules to be aligned; these marks may contain information influencing the matching, such as partial charge and van der Waals radius. In a similar vein, Green and Mardia (2006) include the possibility of colouring the observations, in order to model the possibility that points of the same colour are more likely to be matched *a priori*. Thus knowledge of amino acid types can be used advantageously for the matching of active sites in proteins. Incorporating such information on the points may make the inference more clear-cut, by reducing multi-modality in the posterior distribution.

It would be interesting to consider applications which assume non-rigid or even non-linear transformations between the configurations. Our model allows for such transformations, but the implementation would have to be suitably adapted. The same can be said regarding the use of non-Gaussian observation errors and of different prior distributions for the parameters.

5 Acknowledgments

The authors wish to thank Kanti Mardia for introducing them to this problem and for stimulating discussions, as well as Anthony Davison and John "Mac" McDonald for their helpful comments and suggestions.

6 References

Coats, E. A. (1998) The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods. *Perspectives in Drug Discovery and Design*, **12/13/14**, 199–213.

Chui, H. & Rangarajan, A. (2000) A new algorithm for non-rigid point matching *IEEE Conference on Computer Vision and Patter Recognition*, **2**, 44-51.

Cross, A. D. J. & Hancock, E. R. (1998). Graph matching with a dual-step EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(11), 1236-1253.

Dryden, I. L., Hirst, J. D. & Melville, J. L. (2007). Statistical analysis of unlabelled point sets: comparing molecules in chemoinformatics. *Biometrics*, **63**, 237–251

Dryden, I. L. & Mardia, K. V. (1998). *Statistical Shape Analysis*. Wiley: Chichester.

Eckholm, A., Smith, P. W. F. & McDonald, J. W. (1995). Marginal regression analysis of multivariate binary response. *Biometrika*, **82**(4), 847–854.

Green, P. J. (2001). A Primer on Markov chain Monte Carlo. pp. 1–62 of *Complex Stochastic Systems*, Ed. O. E. Barndorff-Nielsen, D. R. Cox, & C. Klüppelberg. London: Chapman and Hall.

Green, P. J. & Mardia, K. V. (2006). Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika*, **93**(2), 234-254.

Lemmen, C. & Lengauer, T. (2000). Computational methods for the structural alignment of molecules. *Journal of Computer-Aided Molecular Design*, **14**, 215-232.

Mardia, K. V. & Kent, J. T. (2003). *Directional Statistics*. Chichester: Wiley.

Robert, C. P. & Casella, G. (2004). *Monte Carlo Statistical Metods*, second edn. Springer-Verlag, New York.

Speed, T. P. (2005) Genetic map functions. In *Encyclopedia of Biostatistics, 2nd Edition*. Wiley & Sons.

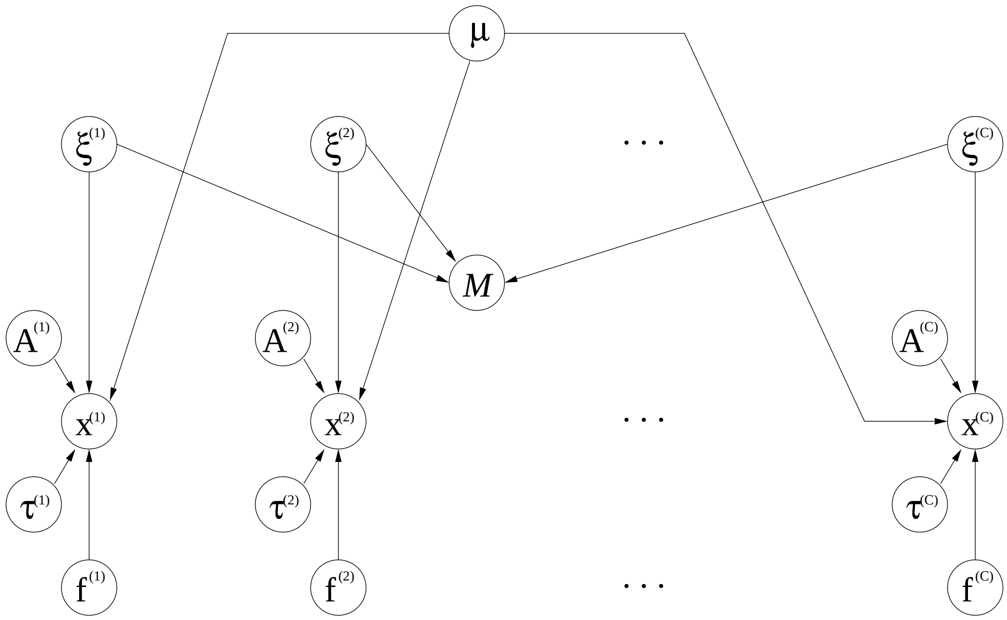

Figure 1: Directed acyclic graph of the hierarchical model.

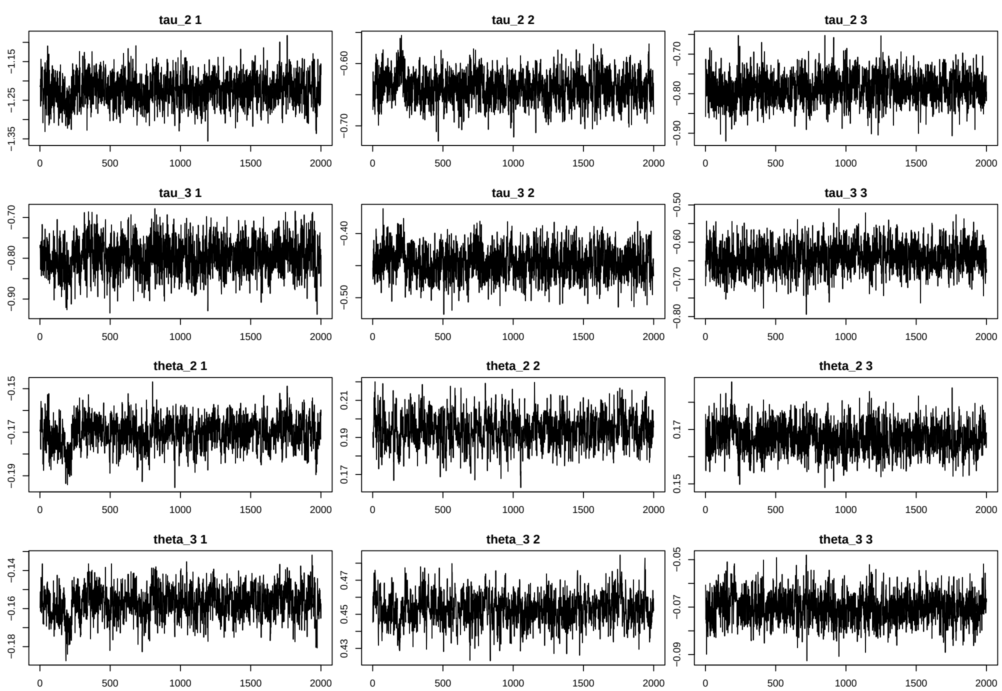

Figure 2: Time series traces of the transformation parameters, taken from a thinned sample of 2000 after burn-in.

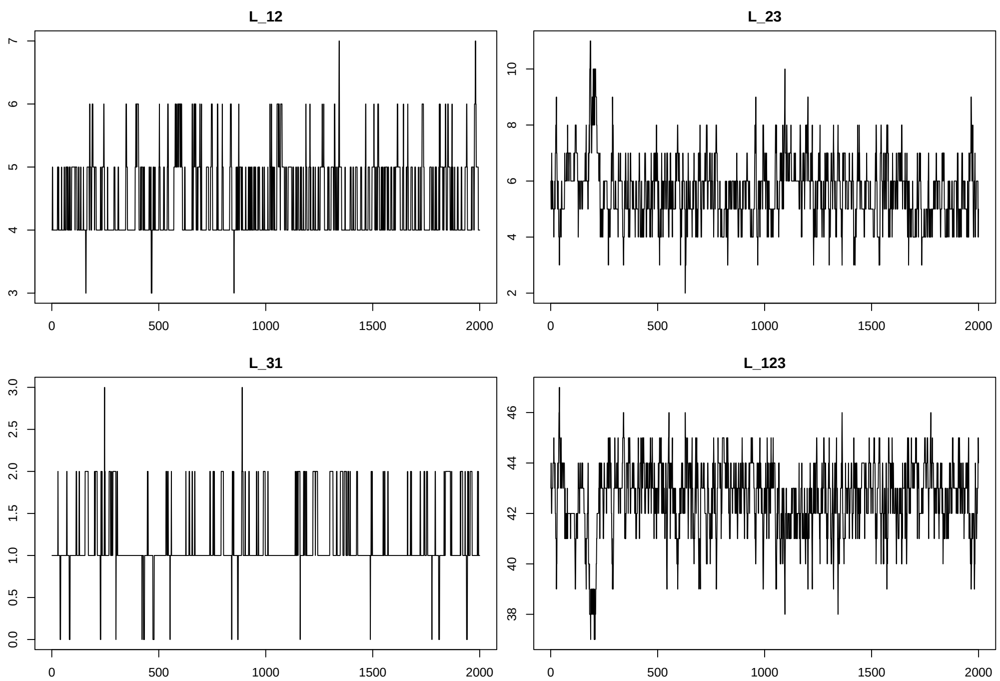

Figure 3: Time series traces of the match counts, taken from a thinned sample of 2000 after burn-in.

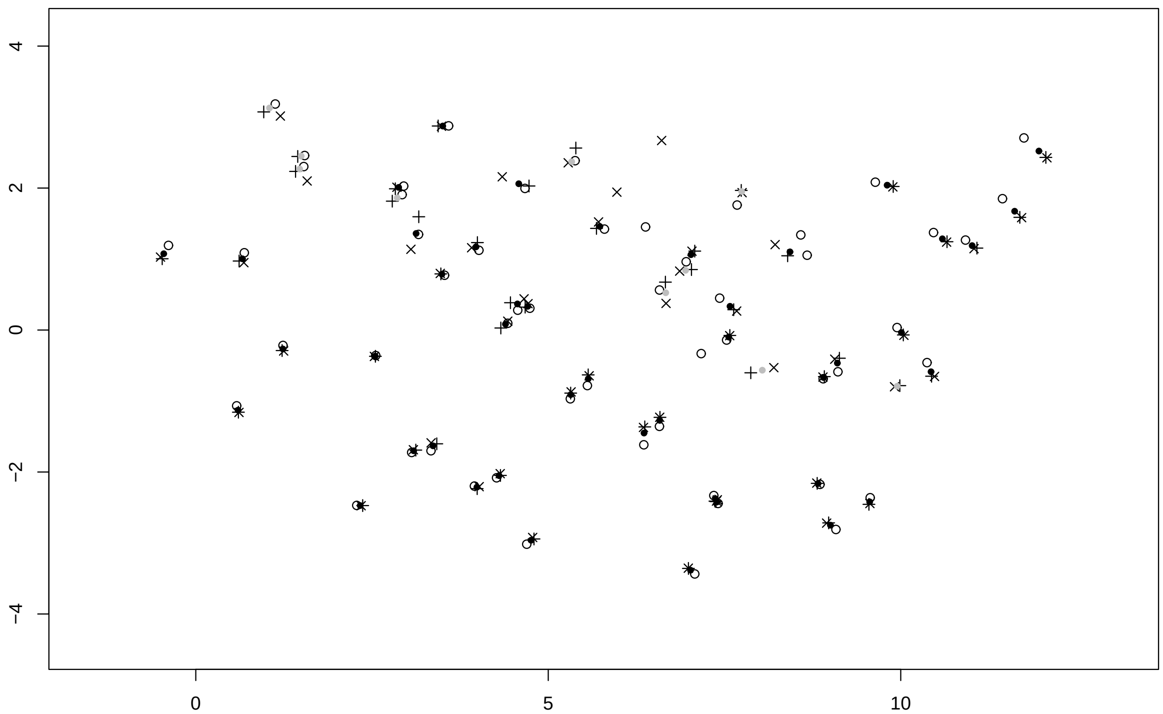

Figure 4: Aligned molecules from Section 3.2: the full transformations are estimated from a MCMC subsample of size 2000, and are filtered out from the data. The observations are then projected onto the principal components plane. The '∘' symbols represent the $x^{(1)}$ configuration (aldosterone), the '+' symbols the $x^{(2)}$ configuration (cortisone), and the '×' symbols the $x^{(3)}$ configuration (prednisolone). The solid dots correspond to the centres of the 123–matches (black) and jk–matches (grey).

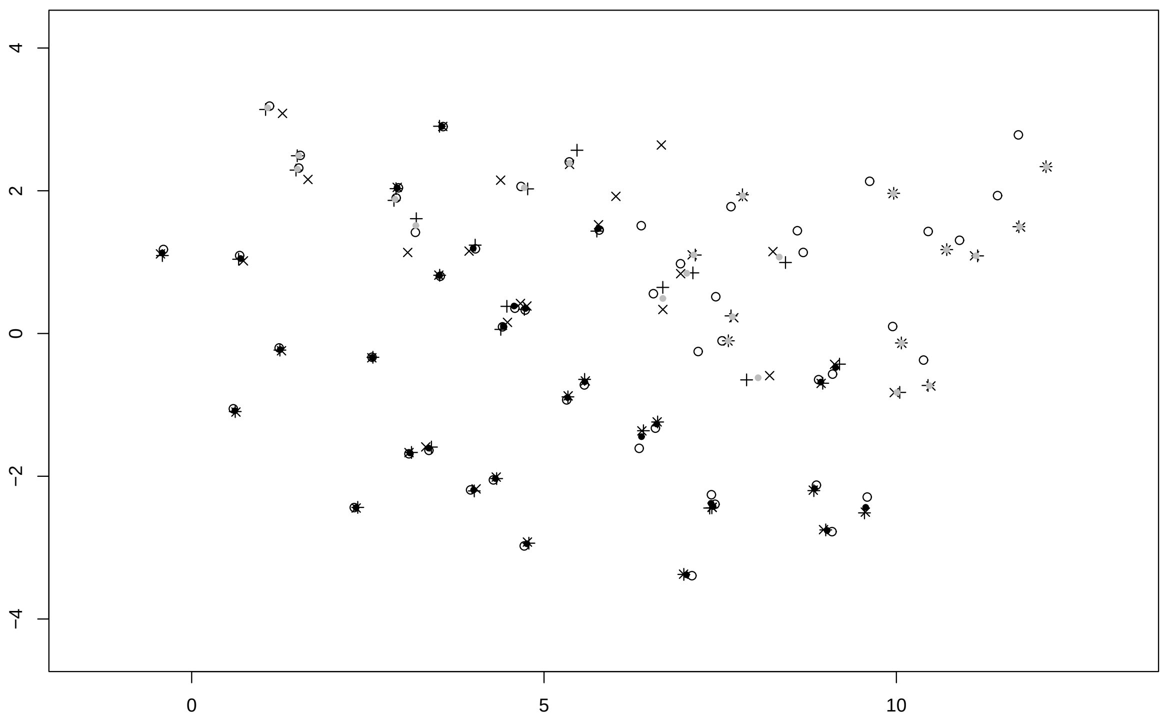

Figure 5: Aligned molecules from Section 3.3: the full transformations are estimated from a MCMC subsample of size 2000, and are filtered out from the data. The observations are then projected onto the principal components plane. The solid dots correspond to the centres of the 123–matches (black) and jk–matches (grey).

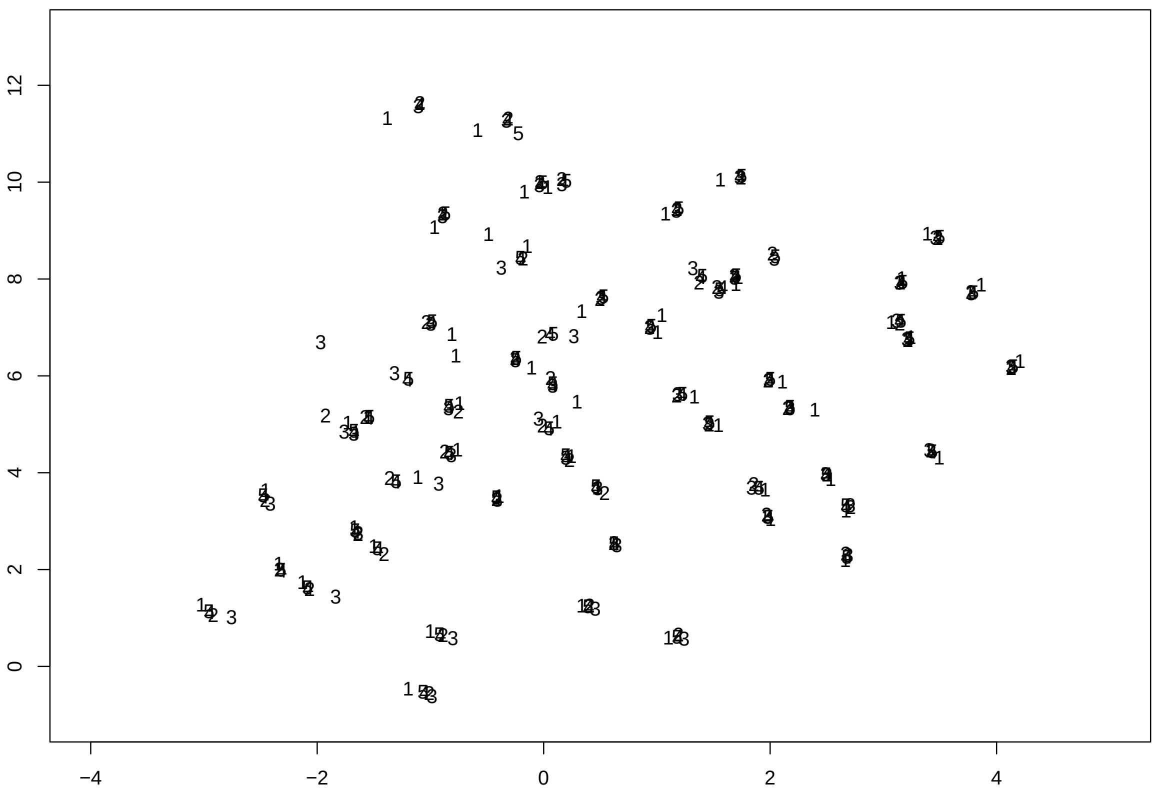

Figure 6: Pairwise alignments of molecules from Section 3.4: the full transformations are estimated from MCMC subsamples of size 2000, and are filtered out from the data. The points are projected onto the first two canonical axes, and are labelled according to the number of the configuration they belong to.

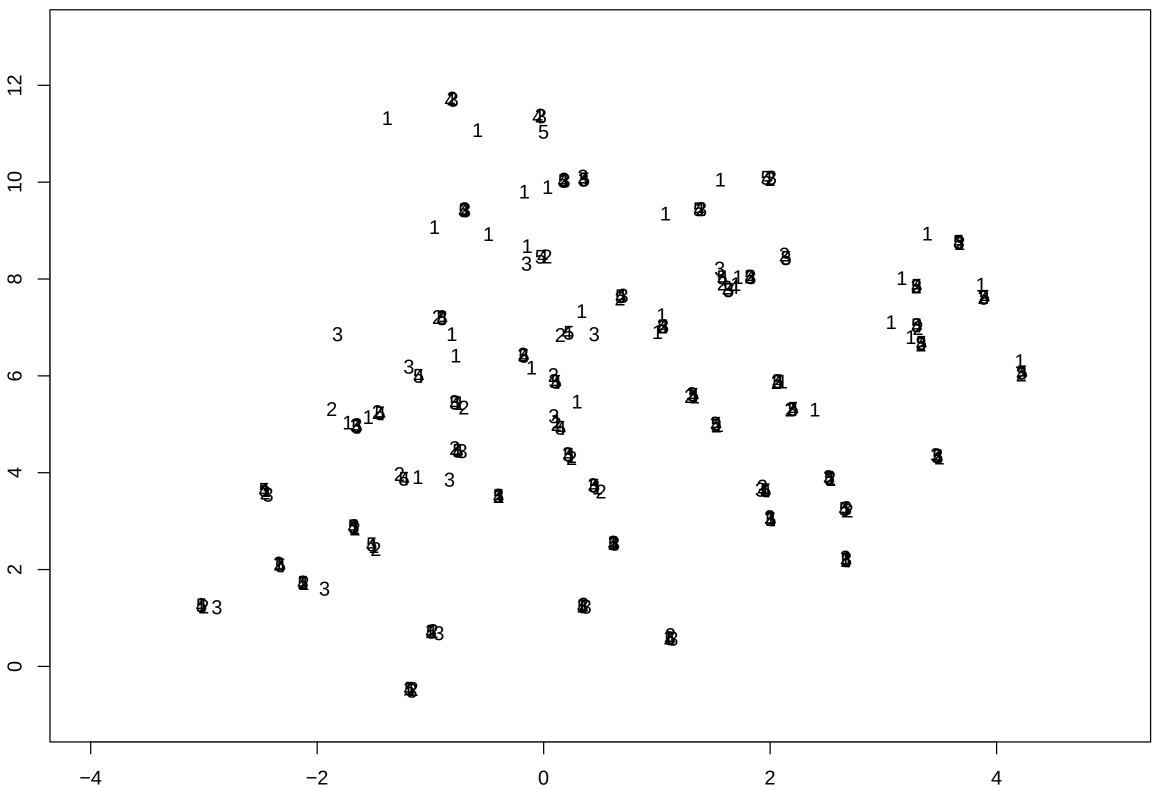

Figure 7: Multiple alignment of the five molecules from Section 3.4: the full transformations are estimated from a MCMC subsample of size 2000, and are filtered out from the data. The points are projected onto the first two canonical axes, and are labelled according to the number of the configuration they belong to.