UNIVERSIDAD CARLOS III DE MADRID
working papers

Working Paper 07-09
Statistic and Econometric Series 03
March 2007

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34-91) 6249849

SPATIAL MATCHING OF M CONFIGURATIONS OF POINTS WITH A BIOINFORMATICS APPLICATION[*]

Juan Miguel Marín [1] and Carmen Nieto [2]

Abstract

In this paper, we present a model to deal with the problem of matching *M* objects or configurations of points. This is a generalization of the model proposed by Green and Mardia (2006). We consider, as a direct and simple application, the case of three configurations with labelled and with unlabelled points. In both cases, we consider data from a microarray experiment of gorilla, bonobo and human cultured fibroblasts published by Karaman et al. (2003). We find out the matchings and the best affine transformation between the projections of genes in a two dimensional space, obtained by a Multidimensional Scaling technique.

Keywords*:* MCMC computation; Microarray analysis; Multidimensional Scaling; Spatial Bayesian methods; Spatial Poisson Process; Von Mises distribution.
JEL Classification*:* Primary 62-02; secondary 62E10, 62F15.

[*]

[1] *Departamento de Estadística, Universidad Carlos III, 28903 Getafe (Madrid), Spain*
[2] *Departamento de Estadística e I. O. III, E. U. E., Universidad Complutense, 28040 Madrid, Spain.*

Spatial matching of M configurations of points with a Bioinformatics application

J.M. Marín.
Departamento de Estadística,
Universidad Carlos III, 28903 Getafe (Madrid), Spain

C. Nieto.
Departamento de Estadística e I. O. III
E. U. E., Universidad Complutense, 28040 Madrid, Spain

Abstract

In this paper, we present a model to deal with the problem of matching M objects or configurations of points. This is a generalization of the model proposed by Green and Mardia (2006). We consider, as a direct and simple application, the case of three configurations with labelled and with unlabelled points. In both cases, we consider data from a microarray experiment of gorilla, bonobo and human cultured fibroblasts published by Karaman et al. (2003). We find out the matchings and the best affine transformation between the projections of genes in a two dimensional space, obtained by a Multidimensional Scaling technique.

Keywords: MCMC computation; Microarray analysis; Multidimensional Scaling; Spatial Bayesian methods; Spatial Poisson Process; Von Mises distribution.
AMS 2000 subjet classifications. Primary 62-02; secondary 62E10, 62F15.

1 Introduction

In Shape Analysis a challenging problem is how to match two or more configurations with labelled points or landmarks (see Dryden and Mardia (1998)) after filtering out some kind of transformation. Nowadays, new problems are been considered where the points are not labelled, so it is necessary to find out which points of each configuration are matched with. A problem about matching two configurations, under Bayesian hierarchical modeling, is described in Green and Mardia (2006). In this paper, we generalize that model by considering not only two configurations, but M configurations under a full Bayesian approach. First, we focus on the case where the points are labelled and check the accuracy of the model with a simulated sample of three configurations; then we consider an application in Bioinformatics, where data are collected from a microarray experiment of gorilla (*Gorilla gorilla*), bonobo (*Pan paniscus*) and human (*Homo sapiens*) cultured fibroblasts, done by Karaman et al. (2003). We have used them in order to compare a set of genes in several samples of these species. Finally, we apply the model to some unlabelled genes of the three species.

In Section 2 we present the problem of matching M configurations, focusing on the case of $M = 3$ by making assumptions of spherical normality. In Section 3 we apply the model to labelled points of three configurations and we present an application in Bioinformatics. Last, in Section 4 we describe the model in the case of unlabelled points and we consider its application in Bioinformatics, by using an empirical Bayes step.

2 The problem of matching M configurations of points

Let consider $M \geq 2$ configurations of points located in $\mathbb{R}^d$ where $d \geqq 2$. Each of them has n_i points $(i = 1, \ldots M)$, such that

$$\mathbf{x}_i = \{x_{ij}, j = 1, \ldots, n_i\} \quad x_{ij} \in \mathbb{R}^d \qquad i = 1, \ldots, M.$$

The points are arbitrarily labelled for identification. We try to determine which point of each configurations of points matches with other points of the rest of configurations, finding the geometrical transformations which relates them. There may be matchings from order two up to order M. The sets of points are regarded as noisy observations from a set of unknown points $\{\mu_l\} \in \mathbb{R}^d$, where each μ_l can generate not more than one point of one configuration but more than one point of different configurations. We do not know which points are generated from each of these μ_l and the points from different configurations, generated from the same μ_l, are considered as matched. This model is a straightforward generalization of Green and Mardia's (2006) model.

We denote the geometrical transformation between configurations $\mathbf{x}_1$ and $\mathbf{x}_i$ by $\boldsymbol{\Psi}_i$ $i = 2, \ldots, M$, and we assume that configuration $\mathbf{x}_1$ is obtained from $\{\mu_l\}$ plus a random term. The relation among subindexes of $\{\mu_l\}$ and points $\{x_{ij}\}$ $(i = 1, \ldots, M,\ j = 1 \ldots, n_i)$ is denoted by a matrix $\left\{\xi_{ij}\right\}$, where ξ_{ij} is the subindex of μ_l that generates the point j of the configuration i, so that the point x_{ij} is generated by $\mu_{\xi_{ij}}$.

The full model is

$$\begin{array}{ll} x_{1j} = \mu_{\xi_{1j}} + \varepsilon_{1j} & j = 1, \ldots, n_1 \\ \boldsymbol{\Psi}_1(x_{2j}) = \mu_{\xi_{2j}} + \varepsilon_{2j} & j = 1, \ldots, n_2 \\ \ldots & \\ \boldsymbol{\Psi}_{r-1}(x_{rj}) = \mu_{\xi_{rj}} + \varepsilon_{rj} & j = 1, \ldots, n_r \\ \ldots & \\ \boldsymbol{\Psi}_{M-1}(x_{Mj}) = \mu_{\xi_{Mj}} + \varepsilon_{Mj} & j = 1, \ldots, n_M \end{array} \tag{1}$$

where $\{\varepsilon_{ij}\}$ has a density function f_i $(i = 1, \ldots, M,\ j = 1, \ldots, n_i)$. We assume that for a fixed i, $\xi_{i1} \neq \xi_{i2} \neq \ldots \neq \xi_{in_i}$, and the random variables $\{\varepsilon_{ij}\}$ are independent among them and independent of $\{\mu_l\}$.

2.1 Spatial Poisson Process

Let assume that $\{\mu_l\}$ are distributed as a homogeneous Poisson process with λ rate in $V \subset \mathbb{R}^d$, and there are N observations in V. Each μ_l can generate independently: zero points, one point of one configuration, two points of two different configurations (double matching) up to M points, one from each configuration (M order matching). Hence $\{\mu_l\}$ can be classified in $M+1$ groups depending on the type of matching they can produce. We assume that the rate of matching is independent of the type of configuration and can vary depending on the matching order. So, each μ_l belongs to one of the $M+1$ possible groups:

μ_l **generates**		**with probability**
One point of one configuration	→	p
Two points of two configurations	→	$\rho_2 p^2$
$\cdots$	$\cdots$	$\cdots$
M points, one of each configuration	→	$\rho_M p^M$
Zero points	→	$1-\sum_{j=1}^{M}\rho_j p^j$ with $\rho_1 = 1$

where ρ_j, $j = 2, \dots, M$ are the prior matching rates of j order.

In order to describe completely the model, we define the affinity and matching matrices.

Definition 1. *The affinity matrix of k order $\mathbf{M}^{(k)}$, with $M \times \overset{k\ times}{\dots} \times M$ dimensions, is defined as a matrix whose elements are $\left[M^{(k)}_{i_1 \dots i_k}\right]$, namely, the number of k order matchings among configurations $\mathbf{x}_{i_1}$, $\mathbf{x}_{i_2}$ $\dots$ $\mathbf{x}_{i_k}$.*

Thence, the number of k order matchings L_k is expressed as

$$L_k = \sum_{i_1=1}^{M}\sum_{i_2>i_1}^{M}\cdots\sum_{i_k>i_{k-1}}^{M} M^{(k)}_{i_1\dots i_k}.$$

Definition 2. *For all $i_1 < i_2 < \dots < i_{k-1} < i_k$ such that $M^{(k)}_{i_1\dots i_k} \neq 0$, the matching matrix of k order $\mathbf{S}^{(i_1,\dots,i_k)}$, with $n_{i_1} \times \dots \times n_{i_k}$ dimensions, is defined as a matrix whose elements are*

$$S^{(i_1,\dots,i_k)}_{j_1\cdots j_k} = \begin{cases} 1 \quad \textit{if} \quad \xi_{i_1 j_1} = \xi_{i_2 j_2} = \cdots = \xi_{i_k j_k} \\ \\ 0 \qquad\qquad \textit{otherwise} \end{cases}.$$

Now, we consider the prior joint distribution of the affinity matrices. Let assume that the distribution of the k order affinity matrix, given L_k k order matchings, is uniform,

$$P(\mathbf{M}^{(k)}|L_k) = \frac{1}{\binom{\binom{M}{k}+L_k-1}{L_k}},$$

because the number of k order affinity matrices, given L_k matchings of k order, is a combination with repetition of $\binom{M}{k}$ configurations chosen in subsets of L_k.

Hence, assuming independence, the joint distribution of all matrices, given the number of k order matchings, is

$$P(\mathbf{M}^{(2)}, \mathbf{M}^{(3)}, \dots, \mathbf{M}^{(M)}|L_2, L_3, \dots, L_M) = \prod_{k=1}^{M}\frac{1}{\binom{\binom{M}{k}+L_k-1}{L_k}}.$$

In the same way, we consider the prior joint distribution of the number of the matchings. In the next scheme we show the frequencies of each class of μ_l with L_2 double matchings, L_3 triple matchings, up to L_M of M order matchings when $N, n_1, \dots, n_M$ are known and $n = \sum_{i=1}^{M} n_i$.

Furthermore, under the previous assumptions, these $M+1$ counts will be independent Poisson distributed variables.

μ_l **generates**	**Frequency**	**Poisson rate**
- One point of one configuration	$n-2L_2-\cdots-ML_M$	$\lambda v p$
- Two points of two configurations	L_2	$\lambda v \rho_2 p^2$
$\cdots$	$\cdots$	$\cdots$
- M points, one of each configuration	L_M	$\lambda v \rho_M p^M$
- Zero points	$N-n-L_2-\cdots$ $-(M-1)L_M$	$\lambda v(1-$ $\sum_{j=1}^{M}\rho_j p^j\Big)$

Hence,

$$\begin{aligned} P(L_2,L_3,\ldots,L_M) &\propto \frac{e^{-\lambda v(\sum_{i=1}^{M}\rho_i p^i)}(\lambda v)^{n-\sum_{i=2}^{M}(i-1)L_i}p^n\rho_2^{L_2}\rho_3^{L_3}\ldots\rho_M^{L_M}}{(n-\sum_{i=2}^{M}iL_i)!L_2!L_3!\ldots L_M!} \\ &\propto \frac{(\frac{\rho_2}{\lambda v})^{L_2}\ldots(\frac{\rho_M}{\lambda v})^{L_M}}{(\lambda v)^{L_3+2L_4+\ldots+(M-2)L_M}(n-\sum_{i=2}^{M}iL_i)!L_2!L_3!\ldots L_M!}, \end{aligned}$$

where $\sum_{i=2}^{M}iL_i\leq n$. Hence,

$$\begin{aligned} &P(\mathbf{M}^{(2)},\mathbf{M}^{(3)},\ldots,\mathbf{M}^{(M)})=P(\mathbf{M}^{(2)},\mathbf{M}^{(3)},\ldots,\mathbf{M}^{(M)}|L_2,L_3,\ldots,L_M)P(L_2,L_3,\ldots,L_M) \\ \propto\ &\prod_{k=1}^{M}\frac{1}{\binom{\binom{M}{k}+L_k-1}{L_k}}\cdot\frac{(\frac{\rho_2}{\lambda v})^{L_2}\ldots(\frac{\rho_M}{\lambda v})^{L_M}}{(\lambda v)^{\sum_{k=3}^{M}(k-2)L_k}(n-\sum_{i=2}^{M}iL_i)!L_2!L_3!\ldots L_M!} \\ \propto\ &\frac{(\frac{\rho_2}{\lambda v})^{L_2}\ldots(\frac{\rho_M}{\lambda v})^{L_M}}{(\lambda v)^{\sum_{k=3}^{M}(k-2)L_k}(n-\sum_{i=2}^{M}iL_i)!\prod_{k=2}^{M}[\binom{M}{k}+L_k-1]!} \end{aligned}$$

For all $k=2,\ldots M$, by considering the affinity matrix of k order as known, we obtain by Combinatorics that the prior joint distribution of all the matching matrices of k order is,

$$P(\mathbf{S}^{(1,\ldots,k)},\ldots,\mathbf{S}^{(M-(k-1),\ldots,M)}|\mathbf{M}^{(k)})=\frac{1}{\prod\limits_{\{(i_1,\ldots,i_k)|\mathbf{M}^{(k)}_{i_1,\ldots,i_k}\neq 0\}}\binom{n_{i_1}}{\mathbf{M}^{(k)}_{i_1,\ldots,i_k}}\cdots\binom{n_{i_k}}{\mathbf{M}^{(k)}_{i_1,\ldots,i_k}}\left(\mathbf{M}^{(k)}_{i_1,\ldots,i_k}!\right)^{k-1}}.$$

2.2 Likelihood of data

We will henceforth assume affine transformations among configurations. Then, (1) is simplified as

$$\begin{aligned} x_{1j} &= \mu_{\xi_{1j}}+\varepsilon_{1j} \qquad j=1,\ldots,n_1 \\ A_1x_{2j}+\tau_1 &= \mu_{\xi_{2j}}+\varepsilon_{2j} \qquad j=1,\ldots,n_2 \\ &\cdots \\ A_{M-1}x_{Mj}+\tau_{M-1} &= \mu_{\xi_{Mj}}+\varepsilon_{Mj} \qquad j=1,\ldots,n_M \end{aligned} \tag{2}$$

where $\{\varepsilon_{ij}\}$ are independent with f_i density for all $i=1,\ldots,M$ and $j=1,\ldots,n_i$.

From (2), the density of x_{ij}, conditional on A_{i-1}, ξ_{ij}, τ_{i-1} and $\{\mu_i\}$ (for all $i = 1, \ldots, M$ and $j = 1, \ldots, n_i$) is

$$f(x_{ij}) = f_i(A_{i-1}x_{ij} + \tau_{i-1} - \mu_{\xi_{ij}}) \left| A_{i-1} \right| , \tag{3}$$

denoting as $A_0 = I$ and $\tau_0 = 0$.

The variables $\{\mu_l\}$ that generate non matched points are uniformly distributed over V. From (3),

$$f(x_{ij}) = \int_V f(x_{ij}/\mu) f(\mu) d\mu = \left| A_{i-1} \right| \frac{1}{v} \int_V f_i(A_{i-1}x_{ij} + \tau_{i-1} - \mu) d\mu.$$

If we denote E^0 as the set of unmatched points, their likelihood contribution is

$$\begin{aligned} \prod_{i=1}^{M} \prod_{\{j, x_{ij} \in E^0\}} f(x_{ij}) &= \prod_{i=1}^{M} \prod_{\{j, x_{ij} \in E^0\}} \frac{1}{v} \left| A_{i-1} \right| \int_V f_i(A_{i-1}x_{ij} + \tau_{i-1} - \mu) d\mu \\ &= \left(\frac{1}{v} \right)^{n - \sum\limits_{i=2}^{M} iL_i} \prod_{i=1}^{M} \prod_{\{j, x_{ij} \in E^0\}} \left| A_{i-1} \right| \int_V f_i(A_{i-1}x_{ij} + \tau_{i-1} - \mu) d\mu. \end{aligned}$$

In the case of locations $\{\mu_l\}$ that generate double matchings, for all (i_1, i_2) such that $M^{(2)}_{i_1,i_2} \neq 0$ and for all (j_1, j_2) such that $S^{(i_1,i_2)}_{j_1 j_2} = 1$,

$$\begin{aligned} f(x_{i_1 j_1}, x_{i_2 j_2}) &= \int_V f(x_{i_1 j_1} \mid \mu) f(x_{i_2 j_2} \mid \mu) f(\mu) d\mu = \\ & \frac{1}{v} \left| A_{i_1 - 1} \right| \left| A_{i_2 - 1} \right| \int_V f_{i_1}(A_{i_1 - 1} x_{i_1 j_1} + \tau_{i_1 - 1} - \mu) f_{i_2}(A_{i_2 - 1} x_{i_2 j_2} + \tau_{i_2 - 1} - \mu) d\mu. \end{aligned}$$

Then, their likelihood contribution is

$$\prod_{\{(i_1,i_2) | M^{(2)}_{i_1,i_2} \neq 0\}} \prod_{\{(j_1,j_2) | S^{(i_1,i_2)}_{j_1 j_2} = 1\}} f(x_{i_1 j_1}, x_{i_2 j_2}) = \left(\frac{1}{v} \right)^{L_2} \prod_{\substack{\{(i_1,i_2) | \\ M^{(2)}_{i_1,i_2} \neq 0\}}} \prod_{\substack{\{(j_1,j_2) | \\ S^{(i_1,i_2)}_{j_1 j_2} = 1\}}} \left| A_{i_1 - 1} \right| \left| A_{i_2 - 1} \right| \cdot$$

$$\cdot \int_V f_{i_1}(A_{i_1 - 1} x_{i_1 j_1} + \tau_{i_1 - 1} - \mu) f_{i_2}(A_{i_2 - 1} x_{i_2 j_2} + \tau_{i_2 - 1} - \mu) d\mu.$$

In general, the likelihood contribution of the points with matchings of k order, ($k = 2, \ldots, M$) is

$$\left(\frac{1}{v} \right)^{L_k} \prod_{\substack{\{(i_1,\ldots,i_k) | \\ M^{(k)}_{i_1,\ldots,i_k} \neq 0\}}} \prod_{\substack{\{(j_1,\ldots,j_k) | \\ S^{(i_1,i_2)}_{j_1 \cdots j_k} = 1\}}} \left| A_{i_1 - 1} \right| \ldots \left| A_{i_k - 1} \right| \cdot$$

$$\int_V f_{i_1}(A_{i_1 - 1} x_{i_1 j_1} + \tau_{i_1 - 1} - \mu) \ldots f_{i_k}(A_{i_k - 1} x_{i_k j_k} + \tau_{i_k - 1} - \mu) d\mu.$$

Previous expressions can be approximated if we consider $V \subset \mathbb{R}^d$ large enough and we extend it to all $\mathbb{R}^d$:

$$\left(\frac{1}{v}\right)^{L_k} \prod_{\substack{\{(i_1,\cdots i_k)| \\ M^{(k)}_{i_1\cdots i_k}\neq 0\}}} (|A_{i_1-1}|\cdots|A_{i_k-1}|)^{M^{(k)}_{i_1\cdots i_k}} \cdot$$

$$\prod_{\substack{\{(j_1\cdots j_k)| \\ S^{(i_1\cdots i_k)}_{j_1\cdots j_k}=1\}}} g_{i_1.i_2,\ldots,i_k}(A_{i_1-1}x_{i_1j_1}+\tau_{i_1-1}-A_{i_2-1}x_{i_2j_2}-\tau_{i_2-1},$$

$$\ldots, A_{i_1-1}x_{i_1j_1}+\tau_{i_1-1}-A_{i_k-1}x_{i_kj_k}-\tau_{i_k-1}) \tag{4}$$

where

$$g_{i_1.i_2,\ldots,i_k}(z_1,\cdots,z_{k-1}) = \int_{\mathbf{R}^d} f_{i_1}(w)f_{i_2}(w-z_2)f_{i_3}(w-z_3)\ldots f_{i_k}(w-z_k)dw$$

represents the joint distribution of $(\varepsilon_{i_1j_1}-\varepsilon_{i_2j_2}, \varepsilon_{i_1j_1}-\varepsilon_{i_3j_3}, \ldots, \varepsilon_{i_1j_1}-\varepsilon_{i_kj_k})$. Hence, the likelihood of all points is the product of expressions (4) for every $k = 2, \ldots, M$.

2.3 Matching of $M = 3$ configurations of points under normality

We consider, as a practical application, the particular case of $M = 3$ configurations, where $\{\varepsilon_{ij}\}$ are normally distributed. The joint prior distribution of the affinity and matching matrices is

$$P(\mathbf{M}^{(2)}, \mathbf{M}^{(3)}, \mathbf{S}^{(1,2)}, \mathbf{S}^{(1,3)}, \mathbf{S}^{(2,3)}, \mathbf{S}^{(1,2,3)}) \propto$$

$$\frac{(\frac{\rho_2}{\lambda v})^{L_2}(\frac{\rho_3}{\lambda v})^{L_3}(n_1-L_3)!(n_2-L_3)!(n_3-L_3)!}{(\lambda v)^{L_3}(n-2L_2-3L_3)!(2+L_2)! \displaystyle\prod_{\{(i_1,i_2)|\mathbf{M}^{(2)}_{i_1,i_2}\neq 0\}} \binom{n_{i_1}}{\mathbf{M}^{(2)}_{i_1,i_2}}\binom{n_{i_2}}{\mathbf{M}^{(3)}_{i_1,i_2}}\left(\mathbf{M}^{(2)}_{i_1,i_2}!\right)}$$

The contribution to the global likelihood of unmatched points, double matchings points and triple matchings points are

$$\left(\frac{1}{v}\right)^{n-2L_2-3L_3} |A_1|^{n_2-M^{(2)}_{23}-M^{(2)}_{21}-M^{(3)}_{123}}\, |A_2|^{n_3-M^{(2)}_{13}-M^{(2)}_{23}-M^{(3)}_{123}}$$

$$\left(\frac{1}{v}\right)^{L_2} \prod_{\substack{\{(i_1,i_2)| \\ M^{(2)}_{i_1,i_2}\neq 0\}}} (|A_{i_1-1}|\,|A_{i_2-1}|)^{M^{(2)}_{i_1,i_2}} \cdot$$

$$\prod_{\substack{\{(j_1,j_2)| \\ S^{(i_1,i_2)}_{j_1j_2}=1\}}} \left(\frac{1}{\sigma\sqrt{2}}\right)^d \varphi_d(\frac{A_{i_1-1}x_{i_1j_1}+\tau_{i_1-1}-A_{i_2-1}x_{i_2j_2}-\tau_{i_2-1}}{\sigma\sqrt{2}}).$$

$$\left(\frac{1}{v}|A_1|\,|A_2|\right)^{L_3} \prod_{\substack{\{(j_1j_2,j_3)| \\ S^{(1,2,3)}_{j_1,j_2,j_3}=1\}}} \left(\frac{1}{\sigma^2\sqrt{3}}\right)^d \cdot$$

$$\cdot\varphi_{2d}\left(\frac{x_{1j_1}-A_1x_{2j_2}-\tau_1}{\sigma\sqrt{2}}, \frac{\sqrt{6}}{6\sigma}(x_{1j_1}+A_1x_{2j_2}+\tau_1-2A_2x_{3j_3}-2\tau_2)\right)$$

where $\varphi_d(z)$ is the standard normal density in $\mathbb{R}^d$.

3 Matching $M = 3$ labelled configurations of points in $\mathbb{R}^2$

In this section we consider that the points of the three configurations are labelled. Each configuration has m points and we suppose that there are m triple matchings knowing how the points are matched. The problem reduces to determine the parameters A_1, A_2, τ_1, τ_2 and σ^2 of the affine transformation among points. We restrict attention to rotations, orthogonal matrices A_1, A_2, with $A_1^{-1} = A_1^T$, $A_2^{-1} = A_2^T$ and $|A_1| = |A_2| = 1$. For $j = 1, \ldots m$, we denote as x_{1j}, x_{2j} and x_{3j} the points involved in m triple matchings, namely, $\xi_{1j} = \xi_{2j} = \xi_{3j}$.

In this case, all $\mathbf{M}^{(2)}$ are zero and

$$
\begin{aligned}
M^{(3)}_{jkl} &= \begin{cases} m \quad \text{if} \quad j \neq k \neq l \in \{1,2,3\} \\ \\ 0 \qquad\qquad \text{otherwise} \end{cases} \\
S^{(1,2,3)}_{jkl} &= \begin{cases} 1 \quad \text{if} \quad j = k = l \\ \\ 0 \qquad \text{otherwise} \end{cases}
\end{aligned}
$$

Hence the global likelihood is

$$
\begin{aligned}
& P(A_1, A_2, \tau_1, \tau_2, \sigma^2, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \propto P(A_1)P(A_2)P(\tau_1)P(\tau_2)P(\sigma^2) \cdot \\
& \left(\frac{1}{\sigma^2\sqrt{3}}\right)^{md} \cdot \exp\left\{ -\frac{1}{2}\sum_{j=1}^{m} \left\| \frac{x_{1j} - A_1x_{2j} - \tau_1}{\sigma\sqrt{2}}, \frac{\sqrt{6}}{6\sigma}(x_{1j} + A_1x_{2j} + \tau_1 - 2A_2x_{3j} - 2\tau_2 \right\|^2 \right\} \\
= \; & P(A_1)P(A_2)P(\tau_1)P(\tau_2)P(\sigma^2)\left(\frac{1}{\sigma^2\sqrt{3}}\right)^{md} \exp\left\{ -\frac{1}{\sigma^2}\left(\sum_{j=1}^{m} \frac{1}{4}\left\| x_{1j} - A_1x_{2j} - \tau_1 \right\|^2 + \right.\right. \\
& \left.\left. + \frac{1}{3}\sum_{j=1}^{m} \left\| \frac{x_{1j} + A_1x_{2j} + \tau_1}{2} - A_2x_{3j} - \tau_2 \right\|^2 \right)\right\}.
\end{aligned}
$$

3.1 Posterior Distributions of the parameters

In this section we show the prior and posterior distributions of the parameters of the model.

Assume that the prior distribution is $\tau_i \sim N_d(\mu_i, \sigma_i^2\mathbf{I}_d)$ $(i = 1, 2)$ then, the posterior distributions of τ_1 and τ_2 are

$$
(\tau_1 \;\; | \;\; \cdots) \sim N_d\left(\frac{\frac{\mu_1}{\sigma_1^2} + \frac{1}{3\sigma^2}\sum\limits_{j=1}^{m}(x_{1j} - 2A_1x_{2j} + A_2x_{3j} + \tau_2)}{\frac{1}{\sigma_1^2} + \frac{2m}{3\sigma^2}}, \frac{1}{\frac{1}{\sigma_1^2} + \frac{2m}{3\sigma^2}}\mathbf{I}_d \right)
$$

$$
(\tau_2 \;\; | \;\; \cdots) \sim N_d\left(\frac{\frac{\mu_1}{\sigma_2^2} + \frac{1}{3\sigma^2}\sum\limits_{j=1}^{m}(x_{1j} + A_1x_{2j} + \tau_1 - 2A_2x_{3j})}{\frac{1}{\sigma_2^2} + \frac{2m}{3\sigma^2}}, \frac{1}{\frac{1}{\sigma_2^2} + \frac{2m}{3\sigma^2}}\mathbf{I}_d \right)
$$

Assume that the prior distribution is $\sigma^{-2} \sim gamma(\alpha, \beta)$, then the posterior distribution of σ^{-2} is $gamma(\alpha^*, \beta^*)$ where

$$\begin{aligned} \alpha^* &= \alpha + md \\ \beta^* &= \beta + \frac{1}{4}\left\{ \sum_{j=1}^{m} \|x_{1j} - A_1 x_{2j} - \tau_1\|^2 + \frac{1}{3}\sum_{j=1}^{m} \|x_{1j} + A_1 x_{2j} + \tau_1 - 2A_2 x_{3j} - 2\tau_2\|^2 \right\} \end{aligned}$$

We will henceforth focus on $\mathbb{R}^2$ and we assume that the prior distributions of A_1 and A_2 are von Mises distributions $A_i \sim M(\nu_i,k_i)$, with ν_i and $k_i > 0$ parameters (see, e.g. Mardia and Jupp (2000)), that is,

$$\begin{aligned} P(A_1) &\propto \exp\{tr\,(F_1' A_1)\} \\ P(A_2) &\propto \exp\{tr\,(F_2' A_2)\} \end{aligned}$$

where

$$F_1 = \frac{k_1}{2}\begin{pmatrix} \cos\nu_1 & -\sin\nu_1 \\ \sin\nu_1 & \cos\nu_1 \end{pmatrix} \text{ and } F_2 = \frac{k_2}{2}\begin{pmatrix} \cos\nu_2 & -\sin\nu_2 \\ \sin\nu_2 & \cos\nu_2 \end{pmatrix}.$$

Alternatively the distributions can be expressed in terms of angles θ_i

$$\begin{aligned} P(\theta_1) &\propto \exp\{k_1 \cos\nu_1 \cos\theta_1 + \sin\nu_1 \sin\theta_1\} \\ P(\theta_2) &\propto \exp\{k_2 \cos\nu_2 \cos\theta_2 + \sin\nu_2 \sin\theta_2\}. \end{aligned}$$

Then, if we denote

$$\begin{aligned} S &= \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \frac{1}{3\sigma^2}\sum_{j=1}^{m}(x_{1j} - 2\tau_1 + A_2 x_{3j} + \tau_2)x_{2j}' \\ T &= \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} = \frac{1}{3\sigma^2}\sum_{j=1}^{m}(x_{1j} + A_1 x_{2j} + \tau_1 - 2\tau_2)x_{3j}', \end{aligned}$$

the posterior distributions of A_1 and A_2 are von Mises distributions where

$$\begin{aligned} F_1^* &= F_1 + S \\ F_2^* &= F_2 + T \end{aligned}$$

or, alternatively, $A_i \sim M(\nu_i^*,k_i^*)$, $i = 1, 2$, where

$$\begin{aligned} k_1^* &= \left[(k_1\cos\nu_1 + S_{11} + S_{22})^2 + (k_1 \sin\nu_1 + S_{21} - S_{12})^2\right]^{1/2} \\ \nu_1^* &= a\cos\left[\frac{k_1\cos\nu_1 + S_{11} + S_{22}}{k_1^*}\right] \\ k_2^* &= \left[(k_2\cos\nu_2 + T_{11} + T_{22})^2 + (k_2 \sin\nu_2 + T_{21} - T_{12})^2\right]^{1/2} \\ \nu_2^* &= a\cos\left[\frac{k_2\cos\nu_2 + T_{11} + T_{22}}{k_2^*}\right]. \end{aligned}$$

Notice that, if $\sin\left[\frac{k_1\cos\nu_1 + S_{11} + S_{22}}{k_1^*}\right] < 0$ then $\nu_1^* = 2\pi - a\cos\left[\frac{k_1\cos\nu_1 + S_{11} + S_{22}}{k_1^*}\right]$,and if $\sin\left[\frac{k_2\cos\nu_2 + T_{11} + T_{22}}{k_2^*}\right] < 0$ then $\nu_2^* = 2\pi - a\cos\left[\frac{k_2\cos\nu_2 + T_{11} + T_{22}}{k_2^*}\right]$.

3.2 Simulated data

We consider, to check the procedure, a group of simulated data. We apply a Gibbs sampler (with 60000 observations, 20000 to burn-in) to render samples from the posterior distributions of the parameters $\sigma, \tau_1, \tau_2, A_1$ and A_2. We have chosen as estimates of A_1 and A_2, the rotation matrices of the mean posterior angles. Results are shown in table 1.

$\sigma = 1$	$\sigma = 8$
$\tau_1 = [0,0] \quad \tau_2 = [0,0]$ $A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, A_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ $\theta_1 = 0 \quad \theta_2 = 0$ **Estimates** $\sigma = 0.9557$ $\tau_1 = [0.0184, 0.0966]$ $\tau_2 = [-0.0976, -0.1008]$ $A_1 = \begin{pmatrix} 0.9999 & -0.0044 \\ 0.0044 & 0.9999 \end{pmatrix}$ $A_2 = \begin{pmatrix} 0.9998 & -0.0188 \\ 0.0188 & 0.9998 \end{pmatrix}$	$\tau_1 = [0,0] \quad \tau_2 = [0,0]$ $A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, A_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ $\theta_1 = 0 \quad \theta_2 = 0$ **Estimates** $\sigma = 7.9878$ $\tau_1 = [0.3919, 0.5455]$ $\tau_2 = [1.2254, 0.6294]$ $A_1 = \begin{pmatrix} 0.9818 & 0.1551 \\ -0.1551 & 0.9818 \end{pmatrix}$ $A_2 = \begin{pmatrix} 0.9834 & -0.1473 \\ 0.1473 & 0.9834 \end{pmatrix}$
$\tau_1 = [-10, -10] \quad \tau_2 = [10, 5]$ $A_1 = \begin{pmatrix} 0.7071 & -0.7071 \\ 0.7071 & 0.7071 \end{pmatrix}, A_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ $\theta_1 = \pi/4 \quad \theta_2 = 3\pi/2$ **Estimates** $\sigma = 1.0741$ $\tau_1 = [-10.0600, 9.7949]$ $\tau_2 = [9.8684, 5.0166]$ $A_1 = \begin{pmatrix} 0.9999 & -0.0044 \\ 0.00440 & 0.9999 \end{pmatrix}$ $A_2 = \begin{pmatrix} 0.0092 & 0.9999 \\ -0.9999 & 0.0092 \end{pmatrix}$	$\tau_1 = [-10, -10] \quad \tau_2 = [10, 5]$ $A_1 = \begin{pmatrix} 0.7071 & -0.7071 \\ 0.7071 & 0.7071 \end{pmatrix}, A_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ $\theta_1 = \pi/4 \quad \theta_2 = 3\pi/2$ **Estimates** $\sigma = 8.0049$ $\tau_1 = [-7.3327, -9.3902]$ $\tau_2 = [10.1808, 6.5261]$ $A_1 = \begin{pmatrix} 0.6384 & -0.7667 \\ 0.7667 & 0.6384 \end{pmatrix}$ $A_2 = \begin{pmatrix} 0.0318 & 0.9978 \\ -0.9978 & 0.0318 \end{pmatrix}$

Table 1.

3.3 An application in Bioinformatics

We consider data from a microarray experiment (*Affymetrix Genechip* 5.0) with gorilla (*Gorilla gorilla*), bonobo (*Pan paniscus*) and human (*Homo sapiens*) cultured fibroblasts done by Karaman et al. (2003). Data consist of expression scores for 12625 genes in 46 samples (23 humans, 11 bonobos and 12 gorillas). In order to study genes with relevant effects, we select 204 genes corresponding to those which expression scores were greater than 3000.

We consider Euclidean distances between pairs of genes in each species and build a map of the genes in two dimensions for each species, by using an INDSCAL Analysis (see Borg and Groenen (2005)) and we estimate the best affine transformations among points representing genes. It is observed, when applying INDSCAL analysis, that there is apparently more similarity between humans and bonobos.

We estimate the best affine transformations between the resulting points representing the genes of bonobos, gorillas and humans. To carry out posterior inference, we set up a Gibbs sampling scheme following the general method introduced in (1) with 60000 observations (20000 to burn-in). We obtained these results.

Affine transformation between humans and bonobos:

$$A_1 = \begin{pmatrix} 0.9949 & -0.1008 \\ 0.1008 & 0.9949 \end{pmatrix} \qquad \tau_1 = \begin{pmatrix} 0.0015 \\ 0.0013 \end{pmatrix}$$

Rotation: 5.78° (0.10 radians). Translation: zero.

Affine transformation between humans and gorillas:

$$A_2 = \begin{pmatrix} 0.9148 & 0.4039 \\ -0.4039 & 0.9148 \end{pmatrix} \qquad \tau_2 = \begin{pmatrix} 0.0014 \\ 0.0012 \end{pmatrix}$$

Rotation: 23.82° (0.42 radians). Translation: zero.

Estimate of the variance: $\sigma^2 = 0.0446$

We conclude that genes from the three species are fully related, showing that genes of bonobos and humans are more closed related. It may be possible to build a future aid-system to determine relations among different genes based on a reference distance among well known genes.

4 Matching $M = 3$ unlabelled configurations of points in $\mathbb{R}^2$

In this section we consider the case where there are three configurations of points in $\mathbb{R}^2$ with $n_i = m$ points ($i = 1, 2, 3$). All points are matched with triple matchings, that is, $L_3 = m$ and $L_2 = 0$, but we do not know which points are matched. Thence, they are unlabelled or theirs labels are arbitrary. The parameters of the model are A_1, A_2, τ_1, τ_2, σ^2 and $S^{(1,2,3)}$. In this case, all elements of $\mathbf{M}^{(3)}$ are zero, except $M^{(3)}_{123} = m$.

The joint distribution of parameters and observations is

$$P(A_1, A_2, \tau_1, \tau_2, \sigma^2, S^{(1,2,3)}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \propto P(A_1)P(A_2)P(\sigma^2)P(\tau_1)P(\tau_2) \cdot \\ \cdot \left(\frac{1}{\sigma^2\sqrt{3}}\right)^m \exp\left\{-\frac{1}{\sigma^2}\left(\sum_{\substack{\{(i,j,k)| \\ S^{(1,2,3)}_{i,j,k}=1\}}}^{m} \frac{1}{4}\left\|x_{1i} - A_1x_{2j} - \tau_1\right\|^2 + \right.\right. \\ \left.\left. + \frac{1}{3}\sum_{\substack{\{(i,j,k)| \\ S^{(1,2,3)}_{i,j,k}=1\}}}^{m} \left\|\frac{x_{1i} + A_1x_{2j} + \tau_1}{2} - A_2x_{3k} - \tau_2\right\|^2\right)\right\} \qquad (5)$$

4.1 Posterior distribution of $S^{(1,2,3)}$

We consider, as the prior distribution of $S^{(1,2,3)}$, an uniform distribution

$$P(S^{(1,2,3)} \mid M_{123}^{(3)} = m) = \frac{1}{(m!)^2}.$$

In order to simulate from the posterior distribution of $S^{(1,2,3)}$, we consider a Metropolis-Hasting procedure with some possible transitions. We choose at random two points from the configuration $\mathbf{x}_1$, e.g. x_{1i_1} and x_{1i_2}, whose matchings with points of configurations x_2 and x_3 are, respectively, x_{2j_1}, x_{3k_1}, and x_{2j_2}, x_{3k_2}. For simplicity, these matchings will be denoted by (i_1, j_1, k_1) and (i_2, j_2, k_2) and $q(S, S^*)$ denotes the probability of changing from configuration S to S^*. The chain is reversible and the probability of accepting a change is

$$\min\left\{1, r = \frac{P(A_1, A_2, \tau_1, \tau_2, \sigma^2, S^*, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)}{P(A_1, A_2, \tau_1, \tau_2, \sigma^2, S, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)}\right\}.$$

Possible transitions are

(i) With p_1^* probability we substitute matchings (i_1, j_1, k_1) and (i_2, j_2, k_2) by (i_1, j_2, k_1) and (i_2, j_1, k_2). In this case,

$$\begin{aligned} r \quad &= \quad \frac{\exp\left\{-\frac{1}{\sigma^2}\left(\frac{1}{4}\left\|x_{1i_i} - A_1 x_{2j_2} - \tau_1\right\|^2 + \frac{1}{3}\left\|\frac{x_{1i_1} + A_1 x_{2j_2} + \tau_1}{2} - A_2 x_{3k_1} - \tau_2\right\|^2\right.\right.}{\exp\left\{-\frac{1}{\sigma^2}\left(\frac{1}{4}\left\|x_{1i_1} - A_1 x_{2j_1} - \tau_1\right\|^2 + \frac{1}{3}\left\|\frac{x_{1i_1} + A_1 x_{2j_1} + \tau_1}{2} - A_2 x_{3k_1} - \tau_2\right\|^2\right.\right.} \\ & \quad\quad \frac{\left.\left. + \frac{1}{4}\left\|x_{1i_2} - A_1 x_{2j_1} - \tau_1\right\|^2 + \frac{1}{3}\left\|\frac{x_{1i_2} + A_1 x_{2j_1} + \tau_1}{2} - A_2 x_{3k_2} - \tau_2\right\|^2\right)\right\}}{\left.\left. + \frac{1}{4}\left\|x_{1i_2} - A_1 x_{2j_2} - \tau_1\right\|^2 + \frac{1}{3}\left\|\frac{x_{1i_2} + A_1 x_{2j_2} + \tau_1}{2} - A_2 x_{3k_2} - \tau_2\right\|^2\right)\right\}}. \end{aligned}$$

(ii) With p_2^* probability we substitute matchings (i_1, j_1, k_1) and (i_2, j_2, k_2) by (i_1, j_1, k_2) and (i_2, j_2, k_1). In this case,

$$\begin{aligned} r \quad &= \quad \frac{\exp\left\{-\frac{1}{\sigma^2}\left(\frac{1}{4}\left\|x_{1i_1} - A_1 x_{2j_1} - \tau_1\right\|^2 + \frac{1}{3}\left\|\frac{x_{1i_1} + A_1 x_{2j_1} + \tau_1}{2} - A_2 x_{3k_2} - \tau_2\right\|^2\right.\right.}{\exp\left\{-\frac{1}{\sigma^2}\left(\frac{1}{4}\left\|x_{1i_1} - A_1 x_{2j_1} - \tau_1\right\|^2 + \frac{1}{3}\left\|\frac{x_{1i_1} + A_1 x_{2j_1} + \tau_1}{2} - A_2 x_{3k_1} - \tau_2\right\|^2\right.\right.} \\ & \quad\quad \frac{\left.\left. + \frac{1}{4}\left\|x_{1i_2} - A_1 x_{2j_2} - \tau_1\right\|^2 + \frac{1}{3}\left\|\frac{x_{1i_2} + A_1 x_{2j_2} + \tau_1}{2} - A_2 x_{3k_1} - \tau_2\right\|^2\right)\right\}}{\left.\left. + \frac{1}{4}\left\|x_{1i_2} - A_1 x_{2j_2} - \tau_1\right\|^2 + \frac{1}{3}\left\|\frac{x_{1i_2} + A_1 x_{2j_2} + \tau_1}{2} - A_2 x_{3k_2} - \tau_2\right\|^2\right)\right\}}. \end{aligned}$$

(iii) With $1 - p_1^* - p_2^*$ probability we substitute matchings (i_1, j_1, k_1) and (i_2, j_2, k_2) by (i_1, j_2, k_2)

and (i_2, j_1, k_1). In this case,

$$r = \frac{\exp\left\{-\frac{1}{\sigma^2}\left(\frac{1}{4}\left\|x_{1i_1} - A_1x_{2j_2} - \tau_1\right\|^2 + \frac{1}{3}\left\|\frac{x_{1i_1}+A_1x_{2j_2}+\tau_1}{2} - A_2x_{3k_2} - \tau_2\right\|^2 + \frac{1}{4}\left\|x_{1i_2} - A_1x_{2j_1} - \tau_1\right\|^2 + \frac{1}{3}\left\|\frac{x_{1i_2}+A_1x_{2j_1}+\tau_1}{2} - A_2x_{3k_1} - \tau_2\right\|^2\right)\right\}}{\exp\left\{-\frac{1}{\sigma^2}\left(\frac{1}{4}\left\|x_{1i_1} - A_1x_{2j_1} - \tau_1\right\|^2 + \frac{1}{3}\left\|\frac{x_{1i_1}+A_1x_{2j_1}+\tau_1}{2} - A_2x_{3k_1} - \tau_2\right\|^2 + \frac{1}{4}\left\|x_{1i_2} - A_1x_{2j_2} - \tau_1\right\|^2 + \frac{1}{3}\left\|\frac{x_{1i_2}+A_1x_{2j_2}+\tau_1}{2} - A_2x_{3k_2} - \tau_2\right\|^2\right)\right\}}.$$

4.2 An application in Bioinformatics

We consider the same data from section 3.3 and we use an empirical Bayes procedure for estimating the rotation matrices A_1 and A_2 from a known set of 38 labelled genes (those with expression scores greater than 10000). After a Gibbs sampling scheme with 60000 observations (20000 to burn-in) we obtain as estimates of the posterior means of the affine transformations between humans-bonobos and humans-gorillas:

$$A_1 = \begin{pmatrix} 0.9781 & 0.2081 \\ -0.2081 & 0.9781 \end{pmatrix} \qquad A_2 = \begin{pmatrix} 0.9906 & -0.1366 \\ 0.1366 & 0.9906 \end{pmatrix}$$

For angles of rotation matrices, we obtain for humans-bonobos −12.01^o (6.0736 radians) and for humans-gorillas: 7.84^o (0.1370 radians).

Then, we select 23 genes with expression scores between 8000 y 10000 and let consider that points are arbitrarily labelled. We estimate the matching matrix $S^{(1,2,3)}$, by selecting the m most frequent not repeated matchings, and parameters τ_1, τ_2, σ^2 by a MCMC scheme also with 60000 observations (20000 to burn-in). Results are in table 2 that shows how matchings have been estimated correctly.

Humans genes (i)	23	1	2	3	4	9	5	7	6	22
Bonobos genes(j)	23	1	2	3	4	9	5	7	6	22
Gorillas genes (k)	23	1	2	3	4	9	5	7	6	22
Relative Frecuency	0.66	0.49	0.44	0.43	0.43	0.42	0.41	0.40	0.39	0.39

Humans genes (i)	10	21	14	15	16	8	11	18	19	17
Bonobos genes(j)	10	21	14	15	16	8	11	18	19	17
Gorillas genes (k)	10	21	14	15	16	8	11	18	19	17
Relative Frecuency	0.39	0.37	0.37	0.37	0.37	0.39	0.38	0.38	0.38	0.38

Humans genes (i)	20	12	13
Bonobos genes(j)	20	12	13
Gorillas genes (k)	20	12	13
Relative Frecuency	0.38	0.38	0.38

Table 2.

Translations:

$$\tau_1 = \begin{pmatrix} 0.2312 \\ 0.2341 \end{pmatrix} \qquad \tau_2 = \begin{pmatrix} 0.2309 \\ 0.2356 \end{pmatrix}$$

Estimate of the variance: $\sigma^2 = 1.0123$.

5 Conclusions

We have proposed the matchings of $M \geq 2$ configurations, with labelled and with unlabelled points, by generalizing the model presented by Green and Mardia (2006). New class of matching matrices has been defined in order to allow describing all the possible matchings up to order M.

We have considered translation and rigid motion transformations, but a direct generalization may be in terms of scale transformations and non linear transformation among configurations.

As an useful application, we have consider a problem of matching genes among different species, by representing them in $d = 2$ dimensions by a Multidimensional Scaling technique, with $M = 3$ species and labelled genes and with unlabelled genes. However, this model can be used considering different configurations of genes and species and by associating distances to relevant properties of genes compared among species. Moreover, this model can be useful in order to find relations among species, by selecting some critical genes, or to assess diagnostic forecasting by comparing positions of genes in different times of an illness.

Bibliography

I. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling* (2th Edition) (2005). Springer, New York.

B.P. Carlin and T.A. Louis. *Empirical Bayes methods for data analysis* (2th Edition) (2000). Chapman & Hall, Boca Raton.

I.L. Dryden and K.V. Mardia. *Statistical shape analysis* (1998). Wiley, Chichester.

P.J. Green and K.V. Mardia. Bayesian alignment using hierarchical models, with applications in protein bioinformatics. Biometrika (2006), 93(2), p. 235–254.

M.W. Karaman, M.L. Houck, L.G. Chemnick, S. Nagpal, D. Chawannakul, D. Sudano, B.L. Pike, V.V. Ho, O.A. Ryder and J.G. Hacia. Comparative Analysis of Gene-Expression Patterns in Human and African Great Ape Cultured Fibroblasts. Genome Research (2003), 13 p. 1619-1630.

K.V. Mardia and P.E. Jupp. *Directional Statistics* (2000). Wiley, Chichester.