Matching problems for unlabelled configurations

John T. Kent*, Kanti V. Mardia & Charles C. Taylor*

University of Leeds

Abstract

In this paper we consider the problem of partial matching for two configurations of landmarks in d dimensions. The objective is to find a suitable subset of landmarks of one configuration and a suitable transformation to take these landmarks onto a corresponding subset of landmarks for the other configuration. Typically the landmarks are unlabelled with possibly different numbers of landmarks in each configuration. The number and selection of matching landmarks are also typically unknown. An important application in three dimensions is to protein matching.

1 Introduction

We start with two configurations of landmarks in d dimensions, $x_j,\ j = 1,\dots,n$ and $\mu_i,\ i = 1,\dots,m,$ repsectively, represented as matrices $\boldsymbol{x}\ (n \times d)$ and $\boldsymbol{\mu}\ (m \times d)$. The objective is to match a subset of landmarks of $\boldsymbol{x}$ to a corresponding set of landmarks of $\boldsymbol{\mu}$. As the notation suggests, we shall tackle this problem by treating $\boldsymbol{x}$ as coming from a suitable distribution with parameters depending on $\boldsymbol{\mu}$.

Two further ingredients are needed to formulate the problem more precisely. The first is a group $\mathcal{G}$ of transformations mapping (some of) the landmarks in $\boldsymbol{\mu}$ to (some of) the landmarks of $\boldsymbol{x}$. We shall use the group $\mathcal{G}_{\text{rigid}}$ containing the rigid body tranformations

$$g(x) = Rx + \alpha,$$

where R is a $d \times d$ rotation matrix and α is a d-dimensional shift vector.

The second ingredient is a set of $\mathcal{P}$ of "permutations" $\boldsymbol{\pi}$ mapping $\{1,\dots,n\}$ to $\{1,\dots,m\} \cup \Delta$. Here Δ denotes a "coffin bin" for those x_j which have no corresponding landmark in $\boldsymbol{\mu}$. In general, we shall suppose the mapping π is injective except for the coffin bin; that is, if $j_1 \neq j_2$ and $\pi(j_1) \neq \Delta,\ \pi(j_2) \neq \Delta$, then

$$\pi(j_1) \neq \pi(j_2). \tag{1.1}$$

This modified permutation set can cope with all three possiblities for landmark numbers, $n < m,\ n = m,\ n > m$.

Given any permutation π define a subset $S = S_\pi \subset \{1,\dots,n\}$ to denote those landmarks not sent to the coffin bin; that is,

$$j \in S_\pi \text{ if and only if } \pi(j) \neq \Delta.$$

The objective of the paper is to find a permutation π and its associated subset S and transformation g such that

$$x_j \approx g(\mu_{\pi(j)}), \quad j \in S. \tag{1.2}$$

This paper updates work last year in Mardia *et al.* (2003) and Taylor *et al.* (2003).

2 Statistical models

A general approach to finding an optimal permutation-like mapping (1.2) is by minimizing some suitable objective function, e.g. Rangajaran *et al.* (1997). In this paper we focus on the objective function given by minus the log likelihood. Two basic statistical models are considered.

Hard model

Treating $\pi \in \mathcal{P}$ and $g \in \mathcal{G}$ as parameters to be estimated, consider the following model for the data, with the landmarks being independent for $j = 1, \ldots, n$:

$$x_j \sim N_d(g(\mu_{\pi(j)}), \sigma^2 I_d), \quad j \in S_\pi, \tag{2.1}$$

$$x_j \sim V, \quad j \notin S_\pi. \tag{2.2}$$

Here V denotes a broad-based distribution with support across the whole domain of the landmarks in $\boldsymbol{x}$, e.g. a uniform distribution or a normal distribution centered at the centroid of $\boldsymbol{x}$ with a large variance $\sigma_0^2 I_d$. The variance σ^2 in (2.1) is a tuning parameter, usually fixed ahead of time, and determines how close a match needs to be between landmarks.

This model is the most intuitive and natural. However, computing MLEs is very difficult computationally due to the combinatorial optimization over π. For this reason, we also consider a simpler model.

Soft model

For this model we use a mixture model, treating the landmarks $\{x_j\}$ as independent and identically distributed observations from the mixture distribution

$$p_0 V + \sum_{i=1}^{m} p_i N_d(g(\mu_i), \sigma^2 I_d), \tag{2.3}$$

with corresponding density

$$f(x) = \sum_{i=0}^{m} p_i f_i(x) \tag{2.4}$$

say. The mixing probabilities are nonnegative and sum to 1, $\sum_{i=0}^{m} p_i = 1$. Note that the component densities $\{f_i\}$ depend implicitly on the various parameters, especially on $g(\cdot)$.

The soft model can almost be viewed as a version of the hard model in which the permutation π is unobserved. The main differences are (a) under the soft model the $\pi(j)$ need not be distinct (even when not in the coffin bin), and (b) there is a "prior" distribution on π in which the the components of π are iid with $Pr(\pi(j) = i) = p_i,\ i = 0, \ldots, n$. Thus the soft model is both less constrained and more structured than the hard model. This type of model has been used by several authors including Cross and Hancock (1998), Luo and Hancock (2001) and Walker (2000).

In this case the EM algorithm (e.g. McLachlan and Krishnan, 1977) can be used to compute the MLEs (at least locally), and it takes the form of a simple explicit iterative updating algorithm. In general it converges quite quickly in this context, though the solution depends heavily on the starting value for g. Hence, to some extent, we are back in the combinatorial jungle of the hard model.

Once g and the prior probabilities $\{p_i\}$ have been estimated, we also obtain estimates of the individual membership probabilities

$$p_{ji} = p_i f_i(x_j)/f(x_j), \tag{2.5}$$

which represent the estimated posterior probabilities that individual j comes from group i. We can can then estimate $\pi(j)$ by "hardening" this soft classification. That is, for each j, we set $\pi(j)$ to be the value of i which maximizes the p_{ji}.

Further, if desired we can modify this allocation rule to impose an injectiveness condition on π as follows. Let $q_j = \max_i(p_{ji})$ and order the indices j in terms of decreasing values of the q_j. Finally, go through this ordered list of indices. At each step estimate $\pi(j)$ to be the value of i maximizing p_{ji}, subject to the proviso that the accepted i, if different from 0, must be distinct from any previously accepted value of i.

3 Algorithms and applications

When $\mathcal{G} = \mathcal{G}_{\text{rigid}}$, it is generally too difficult to make progress with the hard model, so we focus on the soft model.

In this setting the EM algorithm generally coverges very quickly. Note that because $\boldsymbol{\mu}$ is known, it is only the transformation g and the membership probabilities $\{p_i\}$ which are to be estimated. (The more common situation in mixture problems is that there is no group element, but instead the component means μ_i are completely unspecified; this open-ended feature often results in unstable behavior and slow convergence for the EM algorithm.)

Thus the use of the EM algorithm depends on several choices for parameters and initial estimates.

1. Choice of the variances σ^2 and σ_0^2. These choices are very important and determine how good a match has to be in order to be deemed acceptable.

2. Initial esimate of the prior probabilities $\{p_i\}$. This choice is not so important. Often we take them equal, $p_i = 1/(m+1)$.

3. Initial estimate of g. This choice is absolutely crucial and is discussed below.

An important method of obtaining initial estimates of g is through *distance matching*, since rigid body transformations preserve interpoint distances. From the data, construct two distance matrices $D^x = \{d^x(j_1, j_2) = ||x_{j_1} - x_{j_2}||\}$ and $D^\mu = \{d^\mu(i_1, i_2)\}$ of interpoint distances within each of the configurations. By choosing a few distances in D^x which closely match a few distances in D^μ, it is possible to determine (or estimate by least squares) a rigid body motion taking the corresponding landmarks to one another. These choices for g can form the starting point of the EM algorithm.

Distance matching also forms the basis of a deterministic algorithm from graph theory using the maximal connected complete subgraph; see *e.g.* Bron and Kerbosch (1973), and Carraghan and Pardalos (1990). Gold (2003) applied this method to protein matching. The method depends on a threshold below which distances from D^x and D^μ will be deemed a match, and essentially produces a permutation π as in (1.1). Disadvantages of this method are that the estimated permutation can depend dramatically on the choice of threshold and there is no probabilistic assessment of uncertainty.

New work using distance matching to guide the choice of initial estimates and the use of the EM algorithm to refine estimates is currently being developed to tackle these issues.

Two important applications for this methodology are to electrophoretic gel matching in 2 dimensions (*e.g.* Walker, 2000) and to protein matching in 3 dimensions. When applied to protein matching, there is additional information to be incorporated. Namely, each "landmark" comes from one of 24 amino acids, and each matching pair of $\boldsymbol{x}$ and $\boldsymbol{\mu}$ landmarks must have the same (or same class of) amino acid.

Acknowledgements

We would like to thank Dave Westhead and Nicola Gold for useful discussions and help with the data.

References

Bron, C. and Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM* **16**, 575–577.

Carraghan, R. and Pardalos, P.M. (1990). Exact algorithm for the minimal clique problem. *Operations Research Letters* **9**, 375.

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

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

Gold,N,D. (2003) *Computational approaches to similarity searching in a functional site database for protein function prediction*. PhD thesis, University of Leeds.

Luo, B. and Hancock, E.R. (2001). Structural Matching using the EM algorithm and singular value decomposition. *IEEE Trans. PAMI*, **23**, 1120–1136.

Mardia,K. V., Taylor, C. C. and Westhead, D. R. (2003). Structural bioinformatics revisited. *Proceedings in Stochastic Geometry, Biological Structure and Images*, 11–18. Edited by R.G. Aykroyd, K.V. Mardia and M.J. Langdon, Leeds University Press.

McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York, Wiley.

Rangarajan, A., Chui, H. and Bookstein, F. L. (1997). The softassin Procrustes matching algorithm. *Proceedings of the 15th International Conference on Information Processing in Medical Imaging*. 29–42. Edited by J. Duncan and G. Gindi. Lecture Notes in Computer Science. London, Springer-Verlag.

Taylor, C C, Mardia, K V and Kent, J T (2003). Matching unlabelled configurations using the EM algorithm, *Proceedings in Stochastic Geometry, Biological Structure and Images*, 19–21. Edited by R.G. Aykroyd, K.V. Mardia and M.J. Langdon, Leeds University Press.

Walker, G. (2000) *Robust, non-parametric and automatic methods for matching spatial point patterns*. PhD thesis, University of Leeds.