

MATCHING MARKERS AND UNLABELLED CONFIGURATIONS IN PROTEIN GELS

BY KANTI V. MARDIA, EMMA M. PETTY AND CHARLES C. TAYLOR

University of Leeds

Unlabelled shape analysis is a rapidly emerging and challenging area of statistics. This has been driven by various novel applications in bioinformatics. We consider here the situation where two configurations are matched under various constraints, namely the configurations have a subset of manually located ‘markers’ with high probability of matching each other while a larger subset consists of unlabelled points. We consider a plausible model and give an implementation using the EM algorithm. The work is motivated by a real experiment of gels for renal cancer and our approach allows for the possibility of missing and misallocated markers. The methodology is successfully used to automatically locate and remove a grossly misallocated marker within the given dataset.

1. Introduction.

1.1. *Western Blots.* Our motivating application concerns gel techniques used to identify proteins present in human tissue. First, two-dimensional electrophoresis (2-DE) is used to separate all the proteins extracted from a cell. The 2-DE gel is then probed with serum which contains antibodies that will bind to specific proteins. The image of a Western Blot will contain only the location (and intensity) of proteins that have a bound antibody. We can think of Western Blots as containing only a subset of the proteins that are displayed on 2-DE images. The extra step necessary to create a Western Blot allows a further level of variability within the final image. The reproducibility of Western Blots is therefore even more challenging than that of 2-DE images. To help align Western Blots, suitable marker proteins are experimentally determined and are generally expected to be present in all blots under investigation. A stain is applied to each

Keywords and phrases: Primary54C56; secondary 62H35

Keywords and phrases: electrophoresis, shape, Western Blots

blot which will highlight all proteins present, therefore enabling an expert to manually locate the suitable markers. Figure 1 shows an annotated Western Blot image which shows the markers (with the acidity and mass measurements associated with these points) and further points detected by an image analyzer. The markers are used to align the blots by minimizing a sum of squared euclidean distances (usually not the acidity and mass measurements). In some cases, fine adjustments to alignments are made using various heuristic techniques. See, for example, Forgber *et al.* (2009) and Zvelebil & Baum (2007, pp. 613–620) for more details.

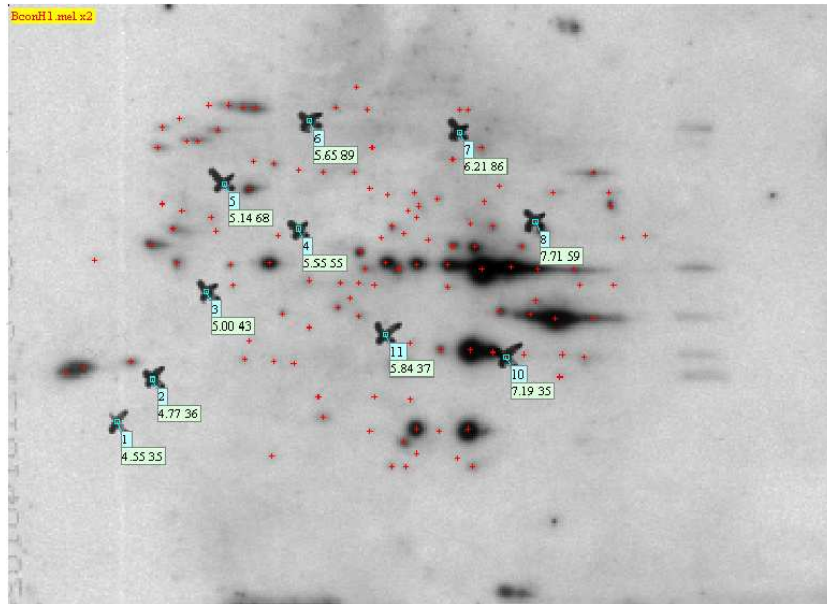


FIG 1. Western Blot image with red crosses depicting the subject-treatment specific non-markers. The larger black crosses indicate the labelled markers, with their acidity and mass measurements (not spatial co-ordinates) highlighted beneath.

Considering the large scope for variation between images and the often vast number of proteins located in a comparatively small area, visual examination to analyse or compare images, although often informative, can be extremely difficult and conclusions unreliable. Visual comparison can also be extremely repetitive and labourious for the expert making the comparisons. Statistical and computational analysis is essential to the *result accuracy* and reduction of expert manual labour. The main aim is to locate a biomarker whose mere presence can be used to measure the progress of disease or treatment effects. In the case of the gel data, a point becomes a biomarker if it is found to have this property. The intensity of a biomarker, indicated by the intensity of the mark on the image, can also provide information about the disease progression or treatment

effect, but this is beyond the scope of this paper.

1.2. *Unlabelled configuration matching.* In the more general setting, the problem is to match two sets — usually of unequal size — of points, in which the correspondence (matching) of the points is unknown. The solution will include the transformation required to align the sets, a list of correspondences which map (some of) the points, and will penalize solutions with many unmatched points, allowing for a trade-off in the goodness-of-fit in the aligned points.

Approaches to closely related problems include the RANSAC algorithm (Fischler & Bolles, 1981), non-rigid point matching using thin-plate splines (Chui & Rangarajan, 2003), a correlation-based approach using kernels (Tsin & Kanade, 2004; Chen, 2011), non-affine matching of distributions (Glaunes *et al.*, 2004) and the Iterative Closest Point Algorithm (Besl & McKay, 1992) for the registration of various representations of shapes. All of these methods avoid making distributional assumptions, with a consequence that probabilistic statements are then difficult to make. By contrast, Czogiel *et al.* (2012), Dryden *et al.* (2007), Kent *et al.* (2010a), Taylor *et al.* (2003) and Green & Mardia (2006) use statistical models to obtain solutions. These latter papers all use examples drawn from protein bioinformatics; a review is given by Green *et al.* (2010).

In this paper, we address a more specific problem in which each configuration contains a subset of points (“markers”) whose labels correspond with high probability, with the remaining points having arbitrary labels (non-markers) as before. Suppose we have two configurations of observed landmarks in d dimensions: markers given by x_j , $j = 1, \dots, K$ and μ_i , $i = 1, \dots, K$, and non-markers μ_i , $i = K + 1, \dots, K + m$ and x_j , $j = K + 1, \dots, K + n$. These are represented as matrices x ($(K + n) \times d$) and μ ($(K + m) \times d$) in which K is usually smaller than m and n . In our model, the markers (the spatial co-ordinates of the large black crosses in Figure 1) μ_i and x_i for $i = 1, \dots, K$ have been identified by an expert to correspond to the same proteins (referred to as a “points” hereafter). However, these are labelled with some uncertainty, so true correspondence is likely but not guaranteed. So it is possible, for example, that markers in μ could correspond to non-markers in x , or have no correspondence at all. For μ_i and x_j with $i, j > K$, (the spatial co-ordinates of the red crosses in Figure 1) we have no prior information

about correspondence probabilities.

1.3. *Statistical Model.* A statistical model in the general setting involves three main components; see Figure 2:

- (a) a group \mathcal{G} , say, on \mathbb{R}^d representing the permitted transformations (g) on (a subset of the landmarks of) μ to bring it close to (a subset of the landmarks of) $x, g \in \mathcal{G}$.
- (b) a matching matrix M , say, identifies which elements of x correspond to which elements of μ for the markers as well as unlabelled points.
- (c) An error model indicating how close the elements of x and μ will be, after the correct transformation and labelling are used.

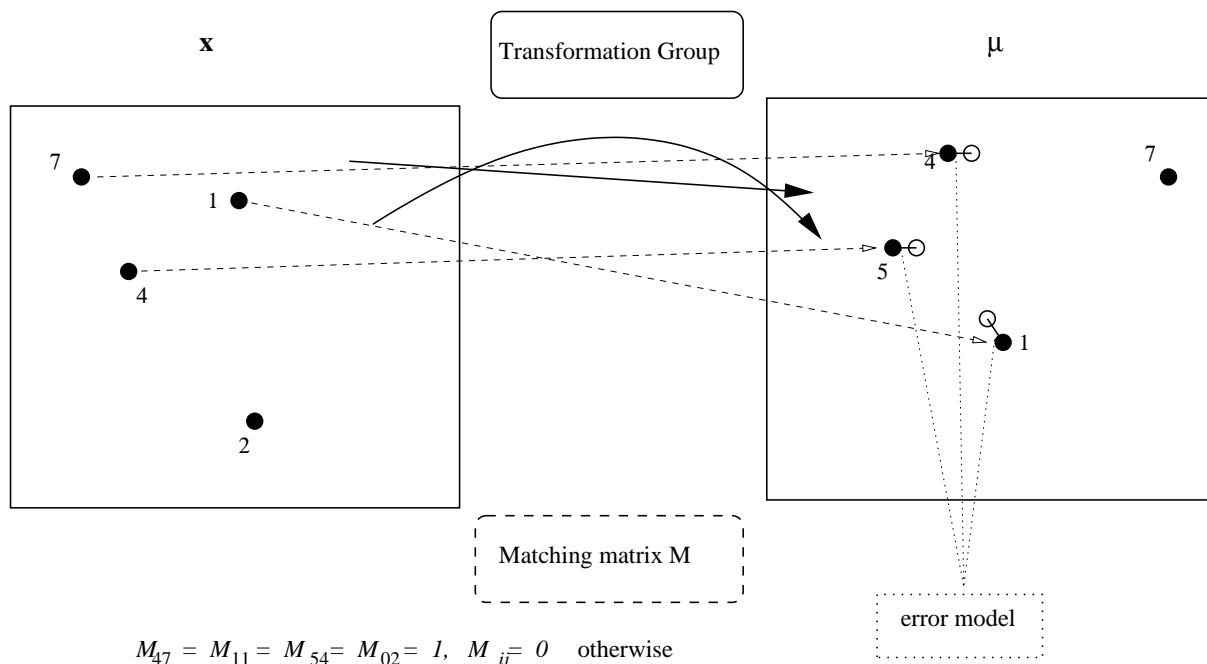


FIG 2. *Illustration of the main ingredients of a statistical model. The labels of the two configurations of points (x and μ) can be considered as arbitrary. Some of the x points are aligned to some of the μ points by a transformation (for example translation, rotation) which belongs to a specified group. An 0/1 matrix M indicates which points match, with unmatched points in x (point 2 in the illustration) assigned to label “0”, and a specific error model assumed for the magnitude of the residual after alignment.*

In Section 2, we introduce our statistical model and emphasize the group of affine transformations belonging to \mathcal{G} which is relevant to our example. The appropriate matching matrix M is estimated under various scenarios, including the use of a matrix Q of prior probabilities, which is introduced to reflect the existence of the markers (labelled points) — an integral part

of the specific problem. In Section 3, we outline likelihood based inference for M , and describe an EM algorithm. In Section 4, we adapt the prior matrix Q when either a marker is missing or a marker is wrongly identified. Two real examples are studied in Section 5 related to renal cancer. In the first example, one marker is grossly misallocated and in the second example, some markers are missing. This procedure has great potential to automate preprocessing of the gels. We conclude with a discussion.

2. Statistical Models.

2.1. *Transformations.* Although the statistical model we later introduce can apply to various types of transformations, we focus on an affine transformation of the form $g(\mu) = \mu A' + B'$, where A is a non-singular $d \times d$ matrix and the $d \times 1$ vector, b , is present in every column of the $d \times (K + m)$ matrix B .

2.2. *Matching matrix.* To estimate the parameters of an appropriate transformation of μ , we can introduce a correspondence system that will indicate whether a point in μ is associated with a point in x , i.e., whether two points *match* across configurations. We can record the correspondence information in a $(K + m + 1) \times (K + n)$ matching matrix, M , where

$$M_{ij} = \begin{cases} 1 & \text{for } i = 0 \text{ if } x_j \text{ does not have a matching point in } \mu \\ 1 & \text{for } i = 1, \dots, K + m \text{ if } x_j \text{ matches } \mu_i \\ 0 & \text{otherwise} \end{cases},$$

for $j = 1, \dots, K + n$. Note that, for simplicity of notation, we use $M_{0j} \equiv M_{K+m+1,j}$, and similarly for other matrices. If $M_{0j} = 1$, then x_j does not have a matching point in μ and we say that x_j is unmatched.

We consider one-to-one or many-to-one matches between points in x and points in μ . We refer to these as *hard* and *soft* matches respectively. Soft matching can be useful in our application since a single protein can produce multiple spots on an image (Banks *et al.*, 2000)

Hard Matches: The matching matrix, M , has the following constraints for the hard model.

$$(1) \quad \sum_{i=0}^{K+m} M_{ij} = 1 \quad \text{for } j = 1, \dots, K + n$$

and

$$(2) \quad \sum_{j=1}^{K+n} M_{ij} \leq 1 \quad \text{for } i = 1, \dots, K + m.$$

So for $i_1 \neq 0$, if $M_{i_1 j_1} = 1$, then $M_{i_1 j_2} = M_{i_2 j_1} = 0$ for all $i_1 \neq i_2$ and $j_1 \neq j_2$. Note that there are no constraints on row $K + m + 1$ in M since each of the $K + n$ points in x is free to remain unmatched. Figure 2 illustrates the case of hard matches in which the point x_2 is unmatched, so $M_{02} = 1$.

Soft Matches: For the soft model, the only constraint is stated in (1). That is, if $M_{i_1 j_1} = 1$ then $M_{i_2 j_1} = 0$ for all $i_1 \neq i_2$, but $M_{i_1 j_2} \in \{0, 1\}$ for $j_1 \neq j_2$. When assigning either hard or soft matches, (1) constrains a point in x to be matched to a single point in μ or, alternatively, to remain unmatched.

2.3. Error distribution. Assuming the transformation parameters, A and b , are known, we can apply a distribution to x_j given the match $M_{ij} = 1$. Given the transformation, we treat the elements of x as conditionally independent with the following densities for $j = 1, \dots, K + n$:

$$(3) \quad p(x_j | M_{ij} = 1) = \begin{cases} \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left\{-\frac{\|x_j - A\mu_i - b\|^2}{2\sigma^2}\right\} & \text{for } i = 1, \dots, K + m \\ \frac{1}{|\Omega|} & \text{for } i = 0 \end{cases}.$$

where Ω is some region in \mathbb{R}^d containing all points in x .

To allow for the possibility of soft matching, we consider points in x to be independent. As we have K markers in each image, we have prior information about the matching across images. Next we introduce notation to deal with prior matching probabilities.

2.4. Prior matching matrix probabilities. Let Q be a $(K + m + 1) \times (K + n)$ matrix with elements $q_{ij} = p(M_{ij} = 1)$. That is, for $j = 1, \dots, K + n$, q_{ij} is the prior probability that μ_i is matched to x_j for $i = 1, \dots, K + m$ and the prior probability that x_j is unmatched for $i = 0$. Again, for simplicity of notation we use q_{0j} in place of $q_{K+m+1,j}$. Note that $\sum_{i=0}^{K+m} q_{ij} = 1$ for $j = 1, \dots, K + n$. We have prior knowledge that corresponding markers, μ_j and x_j for $j = 1, \dots, K$, *should* match. We propose a structure to determine the q_{ij} , which accounts for the possibility of error when allocating markers within a warped image and does not force

corresponding markers to match. In what follows, it will be helpful to note that the matrix Q can be partitioned into sub-matrices of size (rows \times columns) as follows

$$Q \text{ } ((1 + K + m) \times (K + n)) = \left(\begin{array}{c|c} Q^{(0)} \text{ } (1 \times K) & \\ \hline \text{-----} & \\ \hline Q^{(1)} \text{ } ((K + m) \times K) & Q^{(2)} \text{ } ((1 + K + m) \times n) \\ \hline & \end{array} \right)$$

Markers in x : We know that μ_j are the coordinates for marker j in μ , $j = 1, \dots, K$. Let γ_j be the index of the true marker j in μ . If $\gamma_j = j$, then the marker j has been correctly identified. We set the prior probability of a point μ_i being the true marker j , q_{ij} , to be a function of the distance between μ_i and μ_j so that $Q^{(1)}$ has elements

$$(4) \quad q_{ij} = p(\gamma_j = i) = f(d_{ij}) \quad \text{for } i = 1, \dots, K + m, \quad j = 1, \dots, K,$$

where d_{ij} is the Euclidean distance between μ_i and μ_j and choices for f are discussed later.

Next we consider the possibility that a marker within x does not have a corresponding point in μ . Recall that x_j are the coordinates for marker j in x , $j = 1, \dots, K$. To allow for the possibility that x_j remains unmatched, we set the prior probability of $M_{0j} = 1$ to be uniform so that $Q^{(0)}$ has elements

$$(5) \quad q_{0j} = p(\gamma_j = 0) = \frac{1}{|\Omega|}, \quad \text{for } j = 1, \dots, K$$

where Ω is given as in (3).

Non-markers in x : To allow for matching of the non-marker points, we can set the elements of $Q^{(2)}$ as

$$(6) \quad q_{ij} = \frac{1}{K + m + 1}, \quad i = 0, \dots, K + m, \quad j = K + 1, \dots, K + n.$$

So the prior matching probability of a non-marker x_j is uniform.

As an example, we suppose that in Figure 2 only point 1 has been identified as a marker in both x and μ , then we might have $q_{01} = 0.01$ ($= 1/|\Omega|$, say), $q_{11} = 0.89$, $q_{41} = 0.01$, $q_{51} = 0.09$, $q_{71} = 0.00$ (based on the interpoint distances within μ) and $q_{ij} = 1/8$ for the other points shown (taking $m = 6$ in this example).

For ease of reference, the ingredients of the statistical model, together with possible variations are listed in Table 1.

Component of model	Variants	Examples
Configurations x and μ	unlabelled (section 1.2) partially labelled	markers (section 1.1)
Transformation group	rigid-body (section 2.1) affine (section 3.1) non-linear (section 6)	
Matching matrix, M	Hard (section 6) Soft	one-to-one many-to-one (section 6) many-to-many (section 6)
Prior matrix, Q , with $Q_{ij} = P(M_{ij} = 1)$ which depends on		
- markers (section 4)	function of distance (section 3.4.1)	
- non-markers		
Error distribution	isotropic (section 2.3) non-linear (section 6)	

TABLE 1

Main ingredients of the statistical model used for matching of partially labelled configurations of points. Section numbers (e.g. (3.1)) are used to sign-post further details or discussion.

3. EM Algorithms and Inference.

3.1. *EM algorithm.* We use an EM algorithm (McLachlan & Krishnan, 2008) to estimate the transformation parameters, A and b , that will superimpose μ onto x . Throughout this section we assume that σ^2 has been assigned (see Section 3.4.3). In the E-step we calculate the posterior probability that μ_i matches x_j , i.e. the posterior probability that $M_{ij} = 1$. In the M-step the posterior probabilities are input into the expected likelihood of observing M , given the data, x . This enables us to estimate the transformation parameters, A and b .

E-step: We calculate the posterior probability of μ_i matching x_j , given the data, using Bayes' Theorem:

$$(7) \quad p(M_{ij} = 1|x_j) = \frac{p(x_j|M_{ij} = 1)p(M_{ij} = 1)}{p(x_j)},$$

where $p(x_j|M_{ij} = 1)$ is calculated using (3), and $q_{ij} = p(M_{ij} = 1)$ is calculated using (4)–(6).

The denominator of (7) is given by $\sum_{i=0}^{K+m} p(x_j|M_{ij} = 1)p(M_{ij} = 1)$.

M-step: Starting from the multinomial form (McLachlan & Krishnan, 2008, p. 15)

$$l(M|x) = \sum_{i=0}^{K+m} \sum_{j=1}^{K+n} M_{ij} \log p(x_j),$$

we substitute p_{ji} for M_{ij} and $q_{ij}p(x_j|M_{ij} = 1)$ for $p(x_j)$ to obtain the expected log-likelihood of the matching matrix, M , given the data, x :

$$(8) \quad \mathbb{E}[l(M|x)] = \sum_{i=0}^{K+m} \sum_{j=1}^{K+n} p_{ji} [\log q_{ij} + \log p(x_j|M_{ij} = 1)].$$

Here, we suppress the dependence on the parameters A and b .

Both the prior probabilities stored in Q and the conditional distribution of x_j being unmatched are independent of A and b , so, using (8), we estimate the transformation parameters that maximize

$$\sum_{i=1}^{K+m} \sum_{j=1}^{K+n} p_{ji} \log p(x_j|M_{ij} = 1) = \sum_{i=1}^{K+m} \sum_{j=1}^{K+n} p_{ji} \left[-\frac{\|x_j - A\mu_i - b\|^2}{2\sigma^2} - \frac{d}{2} \log(2\pi\sigma^2) \right].$$

Note that the final term is a constant, given that σ is assumed known. Removing further terms independent of A and b , we want to estimate the transformation parameters that minimize

$$\sum_{i=1}^{K+m} \sum_{j=1}^{K+n} p_{ji} \|x_j - A\mu_i - b\|^2.$$

Ignoring the terms independent of b , and noting that $\partial a'x/\partial x = a$ and $\partial x'x/\partial x = 2x$, the maximum likelihood estimates (Walker, 2000) are

$$(9) \quad \hat{b} = \frac{\sum_{i=1}^{K+m} \sum_{j=1}^{K+n} p_{ji} (x_j - A\mu_i)}{\sum_{i=1}^{K+m} \sum_{j=1}^{K+n} p_{ji}}.$$

and

$$(10) \quad \hat{A} = \left[\sum_{i=1}^{K+m} \sum_{j=1}^{K+n} p_{ji} (x_j - \bar{x})(\mu_i - \bar{\mu})' \right] \left[\sum_{i=1}^{K+m} \sum_{j=1}^{K+n} p_{ji} (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})' \right]^{-1}.$$

The algorithm alternates between the E-step and the M-step. At each iteration, the transformation parameters are updated in the M-step to $A^{(r+1)} = \hat{A}^{(r)}$ and $b^{(r+1)} = \hat{b}^{(r)}$, before substitution into the E-step for the next iteration.

We assign convergence to be when r is such that

$$(11) \quad \frac{1}{(K+m+1)(K+n)} \sum_{i=0}^{K+m} \sum_{j=1}^{K+n} [p_{ji}^{(r+1)} - p_{ji}^{(r)}]^2 \leq 10^{-l},$$

where l is chosen and the posterior probability of μ_i matching x_j at the r th and $(r+1)$ st iteration is denoted by $p_{ji}^{(r)}$ and $p_{ji}^{(r+1)}$ respectively, for $i = 0, \dots, K+m$ and $j = 1, \dots, K+n$.

3.2. *Inference for M .* Let P be the $(K + n) \times (K + m + 1)$ matrix containing the final posterior matching probabilities. Let \hat{A} and \hat{b} be the final estimates of the transformation parameters obtained from the EM algorithm.

An obvious route to estimate the matching matrix, M , is to use the posterior matching probabilities, but this will not yield a one-to-one outcome. For one-to-one matches we need to satisfy the constraints in (1) and (2). Given the transformation, the conditional log-likelihood of M is $\sum_{i=0}^{K+m} \sum_{j=1}^{K+n} M_{ij} \log P_{ji}$. We find M that maximises this log-likelihood by mixed integer linear programming. In our implementation we imputed the $2K + m + n$ constraints into `lp_solve` (Berkelaar, 2008), which then yields the estimated one-to-one matching matrix, \hat{M} . We can summarise the steps as follows.

Composite Algorithm

- (i) Assign q_{ij} using (4), (5) and (6) for $i = 0, \dots, K + m$ and $j = 1, \dots, K + n$.
 - (ii) Find initial estimates of the transformation parameters, $A^{(0)}$ and $b^{(0)}$, and assign the variance, σ^2 . Possible choices are discussed in the following subsection.
 - (iii) Run the EM algorithm to get the updated estimates, $p_{ji}^{(1)}$, $A^{(1)}$ and $b^{(1)}$, using (7), (10) and (9) respectively.
 - (iv) Repeat step 3 to find the updated estimates, $p_{ji}^{(r+1)}$, $A^{(r+1)}$ and $b^{(r+1)}$, until convergence (defined in (11)) is reached. Let the final posterior matching probabilities be stored in the $(K + n) \times (K + m + 1)$ matrix P and the final estimated transformation parameters be denoted by \hat{A} and \hat{b} .
 - (v) One-to-one matches are obtained using the hardening algorithm described above.
 - (vi) Treating the matches within the inferred matching matrix, \hat{M} , as known, we can update the transformation parameters using Procrustes methodology (Dryden & Mardia, 1998) to calculate the final estimates, $\hat{\hat{A}}$ and $\hat{\hat{b}}$.
-

3.3. *Assigning the function and parameters within the EM algorithm.* We need to assign the function f stated in (4), as well as starting values for the transformation parameters denoted by $A^{(0)}$ and $b^{(0)}$, and a variance σ^2 . We look at each assignment separately.

3.3.1. *Distance function.* As before, μ_j contains the allocated marker coordinates for marker j in μ , $j = 1, \dots, K$ and γ_j is the index of the true marker j in μ . Let \bar{d}_{ij} denote the expected distance between a point μ_i and μ_j for $i = 1, \dots, K + m$. Due to the freedom for a gel to warp, in reality the distance between μ_i and μ_j in an image is $d_{ij} = \bar{d}_{ij} + \varepsilon$, where ε denotes some error.

Our choice of the function, f , in (4), considers all points in μ as possible true markers. We adopt a multivariate normal distribution for ε , which gives

$$(12) \quad q_{ij} = p(\gamma_j = i) \propto \exp \left\{ -\frac{\|\mu_i - \mu_j\|^2}{2\sigma_*^2} \right\},$$

for $i = 1, \dots, K + m$, where σ_*^2 is the variance between two points in μ (assuming independence across dimensions). So the probability that μ_i is the true marker j will decrease the further it is from μ_j .

3.3.2. *Starting values for transformation parameters.* As we have prior knowledge of allocated corresponding markers in both μ and x , it is sensible that $A^{(0)}$ and $b^{(0)}$ are set as the transformation parameters necessary to best superimpose corresponding markers. Dryden & Mardia (1998) show how these parameters can be estimated from the matrix,

$$(13) \quad R = (\mu_*' \mu_*)^{-1} \mu_*' x^{(m)},$$

where μ_* is the $K \times (d+1)$ matrix $\mu_* = (\mathbf{1}_K, \mu^{(m)})$ and $\mathbf{1}_K$ is a vector of ones of length K . The $K \times d$ matrices, $\mu^{(m)}$ and $x^{(m)}$, contain only the marker coordinates for μ and x respectively.

The first column in R' contains $b^{(0)}$ and the second two columns in R' contain the $d \times d$ matrix $A^{(0)}$.

3.3.3. *Starting values for the variance between images.* We can estimate the variance σ^2 , by considering the mean squared distance between corresponding markers in μ and x after an affine transformation has been applied to superimpose them. That is, set

$$(14) \quad \hat{\sigma}^2 = \frac{1}{\nu} \sum_{j=1}^K \|x_j - A^{(0)}\mu_j - b^{(0)}\|^2,$$

where $\nu = dK - d^2 - d$ and denotes the degrees of freedom. Here dK is the number of error terms in the d components of the K markers. This number is reduced in ν to accommodate the estimates of $A^{(0)}$ and $b^{(0)}$.

4. Grossly misallocated or missing markers. This section describes further refinements to the above Composite Algorithm, which is highly dependent on the transformation parameters input as starting values, $A^{(0)}$ and $b^{(0)}$. We have previously stated that the affine transformation necessary to superimpose corresponding markers in μ and x will provide sensible starting values for the transformation parameters within the EM algorithm. However this would not be the case if gross misallocations occur. The number of missing or grossly misidentified markers are dependent on the quality of the equipment and the expert that creates the images.

Firstly, we provide a method that will highlight grossly misallocated markers across images. Highlighted markers can then be automatically removed or corrected before they are used within the EM algorithm to estimate transformation starting values. Then, in Section 4.2 we deal with the case where some markers are missing from one of the images.

4.1. *Grossly misallocated markers.* Gross misallocations of a marker may occur through human error when inputting marker labels into data spreadsheets. Dryden & Walker (1999) consider procedures based on S estimators, least median of squares and least quartile difference estimators that are highly resistant to outlier points. The RANSAC algorithm (Fischler & Bolles, 1981) uses a similar robust strategy. Here we describe how we can use the EM algorithm previously described.

Here we provide a method that will highlight grossly misallocated markers across images. Highlighted markers can then be automatically removed or corrected before they are used within the EM algorithm to estimate transformation starting values.

Let $\mu^{(m)}$ and $x^{(m)}$ be $K \times d$ coordinate matrices where μ_j and x_j contain the coordinates of marker j in μ and x respectively for $j = 1, \dots, K$. Here we consider the prior matching probabilities to be independent of the distance between a possible marker and the allocated marker so that

$$(15) \quad q_{ij} = \begin{cases} p_M & \text{for } i = j \\ \frac{1-p_M}{K} & \text{for } i \neq j \end{cases},$$

where p_M denotes the probability that the allocated marker μ_j truly corresponds to the allocated marker x_j .

We input $\mu^{(m)}$ and $x^{(m)}$ into steps (i)–(v) of the composite algorithm to estimate the one-to-one matching matrix \hat{M} , replacing (4) and (5) with (15) in stage (i). We use (13) to estimate the starting transformation values, $A^{(0)}$ and $b^{(0)}$. Note that the starting transformation will be distorted by the presence of grossly misallocated markers. There are four possible outcomes for $k = 1, \dots, K$.

- The allocated corresponding markers μ_k and x_k are matched if $\hat{M}_{kk} = 1$. We include both μ_k and x_k in further analyses.
- The marker x_k remains unmatched if $\hat{M}_{0k} = 1$. We exclude both μ_k and x_k from further analyses.
- No point in $x^{(m)}$ is matched to the marker μ_k if $\hat{M}_{kj} = 0$, for all $j = 1, \dots, K$. We exclude both μ_k and x_k from further analyses.
- The marker μ_{k_1} is matched to an allocated non-corresponding marker x_{k_2} if $\hat{M}_{k_1 k_2} = 1$, for $k_1 \neq k_2$. We exclude μ_{k_1} , μ_{k_2} , x_{k_1} and x_{k_2} from further analyses.

See Section 5.1 for an illustration.

4.2. *Missing markers.* It is possible that all K markers are not successfully located in both μ and x . For example, only 10 out of the possible $K = 12$ markers were located in the image displayed in Figure 1.

There are four possible cases we must consider for Marker $k = 1, \dots, K$: (a) located in both μ and x ; (b) located in μ alone; (c) located in x alone; and (d) not located in either μ or x . We first introduce notation to allow for the possibility of missing markers.

Let K_μ and K_x be the total number of markers located in μ and x respectively. As previously noted, let μ be the $(K + m) \times d$ coordinate matrix and x be the $(K + n) \times d$ coordinate matrix.

If marker k is located in μ , then μ_k contains the coordinates of marker k in μ . If marker k is not located in μ , then $\mu_k = \emptyset$. Similarly if marker k is located in x , then x_k contains the coordinates of marker k in x , for $k = 1, \dots, K$. If marker k is not located in x , then $x_k = \emptyset$.

As previously stated, Q is the $(K + m + 1) \times (K + n)$ matrix containing the prior matching probabilities for points in x . We define Q by allowing for the possibility that an allocated marker k is not the true marker k , for $k = 1, \dots, K$.

Markers in x : corresponding to each of the above cases we have

- (a) If $\mu_j \neq \emptyset$ and $x_j \neq \emptyset$, we assign q_{ij} as previously stated in (4) and (5) for $i = 0, \dots, K+m$.
- (b) If $\mu_j \neq \emptyset$ and $x_j = \emptyset$, we treat μ_j as a non-marker.
- (c) If $\mu_j = \emptyset$ and $x_j \neq \emptyset$, we treat x_j as a non-marker.
- (d) If $\mu_j = \emptyset$ and $x_j = \emptyset$, we set $q_{ij} = q_{jk} = \emptyset$ for all i and k .

Non-markers in x : The prior matching probability of a non-marker, x_j , is again set to be uniform over all matching possibilities so that, for $i = 0, \dots, K+m$ and $j = K+1, \dots, K+n$,

$$(16) \quad q_{ij} = \frac{1}{K_\mu + m + 1}.$$

In case 3, when $\mu_j = \emptyset$ and $x_j \neq \emptyset$ for $j = 1, \dots, K$, we treat x_j as a non-marker and use (16) to calculate q_{ij} for $i = 0, \dots, K+m$.

Note that μ contains K_μ markers and m non-markers. There are only $K_\mu + m + 1$ matching possibilities for a point in x , thus producing the denominator in (16). See Section 5.2 for an illustration.

5. Examples. Our full dataset — see Supplementary Material (6) — was collected to represent eight subjects, under two different conditions, treated with two possible treatments.

A replicate image was also produced for each subject-treatment specific case. A typical Western Blot is shown in Figure 1, which is approximately of size 280×220 . In this paper we illustrate the methods on two pairs of images: in the first example, robustness to gross misidentification is explored, and the second example deals with missing markers.

5.1. *Grossly misallocated marker.* Let μ and x represent the co-ordinate sets on Western Blots of a renal cancer cell line cultured under either normoxic or hypoxic conditions. The proteins are then extracted and probed with either patient sera or control sera in a western blot to produce the images generated. All $K = 12$ markers were located in both images.

We input the corresponding markers for μ and x into steps (i)–(v) of the composite algorithm (see Section 3.4) to estimate the one-to-one matching matrix, \hat{M} , found when superimposing $\mu^{(m)}$ onto $x^{(m)}$. That is, we transform the appropriate markers in μ onto the corresponding markers in x . Using only the markers, we estimate the variance in (3) as $\hat{\sigma}^2 = 4.5^2$ and set the

prior matching probability in (15) as $p_M = 0.99$. The starting values for the transformation parameters, $A^{(0)}$ and $b^{(0)}$, are found using (13). We use the final posterior probabilities, P , to estimate M . Marker 1 remains unmatched in both images.

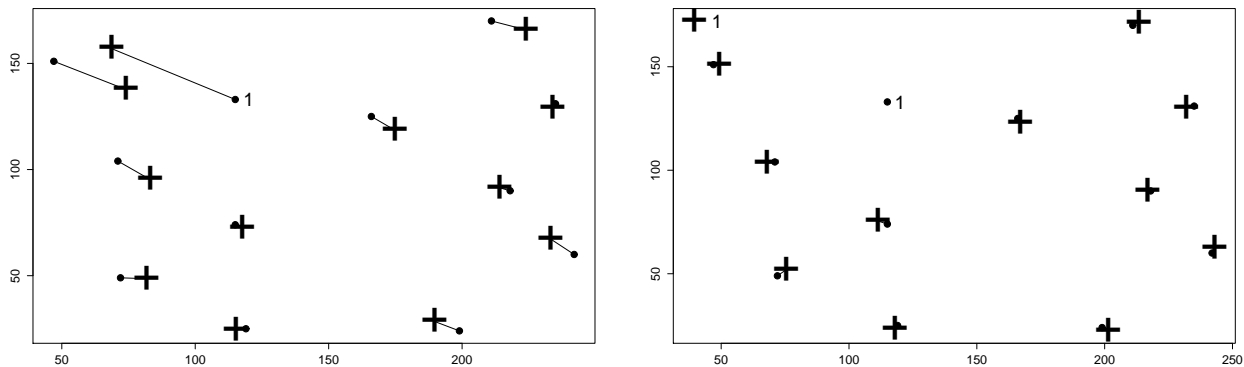


FIG 3. Initial transformation, before (left) and after (right) marker 1 is removed as a marker from both images.

Figure 3 shows the initial transformation of μ onto x before and after marker 1 is removed as a marker (though still displayed) in both images. In this example, the RMSD between the 12 marker pairs before the removal is 19.44. The RMSD between the remaining 11 marker pairs after the removal is 2.96.

Following these discoveries, we were told that marker 1 was incorrectly labelled as spotID 136 when it should have been spotID 153, i.e. the methodology was able to highlight a misidentified marker.

5.2. *Missing markers.* In this example we display the matches made when comparing two *replicate* specimens, representing a cell line cultured under either normoxic conditions, with proteins are extracted and probed with control sera. All 12 markers were located in μ . Markers 9 and 10 were missing in x , so these were treated as non-markers in μ and we set $K = 10$.

We input the images into steps (i)–(v) of the composite algorithm. The starting values for the transformation parameters, $A^{(0)}$ and $b^{(0)}$, are found using (13). We estimate the variance in (3), σ^2 , using (14) with denominator ν . Finally, we set $\hat{\sigma}_*^2 = \hat{\sigma}^2$ in (12). The estimated

transformation parameters are

$$\hat{A} = \begin{pmatrix} 0.980 & -0.047 \\ 0.002 & 1.006 \end{pmatrix},$$

and $\hat{b} = (-1.72, 10.78)'$. We display the matches made in Figure 4 after the final transformation of μ onto x .

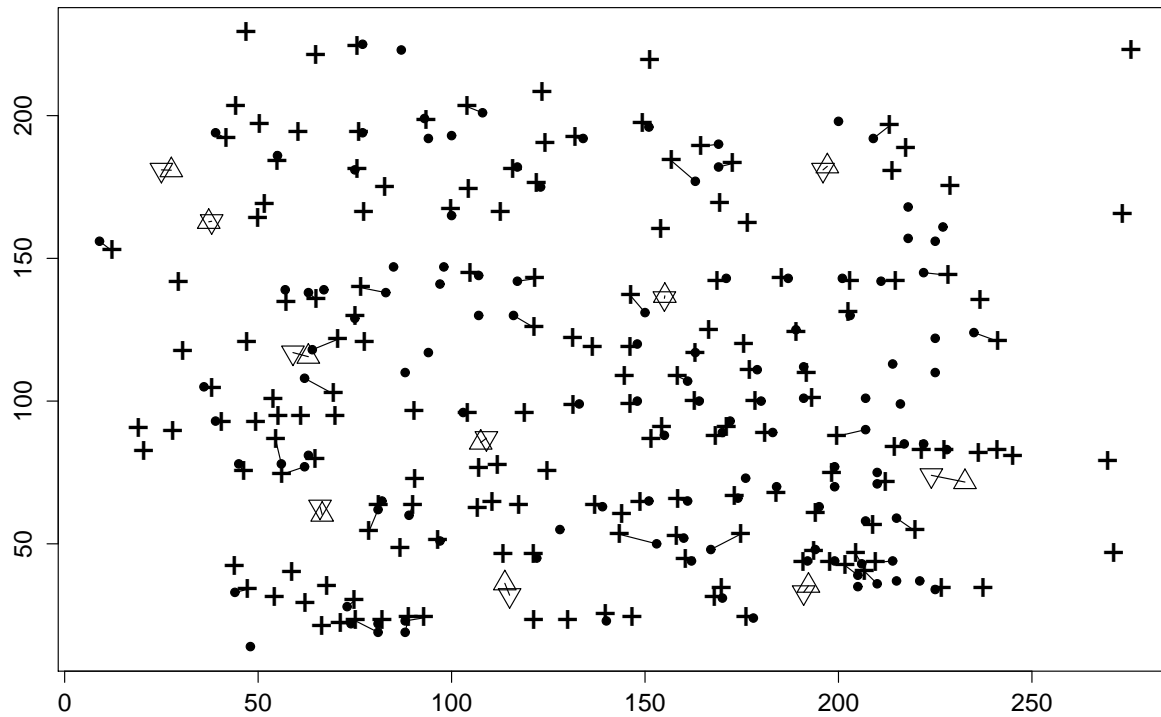


FIG 4. Final transformation of μ onto x and the matches made. Points in x (\bullet), points in transformed μ ($+$), markers in x (∇) and markers in μ (\triangle). The 107 matched points across images are joined by a line.

6. Discussion. Many EM algorithms are known to converge only to a local solution, and this will also apply to the methods considered here. However, the availability of the markers which provide partial information will usually ensure good starting values so this will not be a problem in our application.

Note that it would be possible to adapt the model so that σ could be allowed to vary according to the distance of the point to the edge of the image, which could be used to deal with minor

non-linear deformations. More generally, it should also be possible to adapt our methods to deal with more general transformations, for example using thin-plate splines (Chui & Rangarajan (2003)).

There are situations when clusters occur within a gel which makes it difficult to correctly identify a marker within a cluster of points. We can allow for the increased likelihood that a marker $\mu_j, j = 1, \dots, K$ is misallocated if it exists within a cluster of other points, by using an adaptive choice of f in the prior (4):

$$q_{ij} = p(\gamma_j = i) \propto \begin{cases} \frac{1}{C_j} & \text{if } d_{ij} \leq \varepsilon \\ 0 & \text{if } d_{ij} > \varepsilon \end{cases},$$

where d_{ij} is the Euclidean distance and C_j is the number of points in μ that are within a distance of ε from μ_j , i.e.

$$C_j = \sum_{i=1}^{K+m} I[d_{ij} \leq \varepsilon],$$

where $I[d_{ij} \leq \varepsilon] = 1$ if $d_{ij} \leq \varepsilon$, (0 otherwise) for $i = 1, \dots, K + m$.

A further adaptation of the model, which could be useful in Western blots, would be to incorporate in the priors a measure associated with the grey-scale intensity of the located points in the image (Rohr *et al.*, 2004). Approaches for this, as well as further models for the background noise, are considered in Petty (2009).

Our composite algorithm ensures one-to-one matches, but there are circumstances in which many-to-one or many-to-many matches can be considered. These can be useful when comparing protein images since multiple forms of an individual protein can often be visualised (Banks *et al.*, 2000). That is, a single protein can produce multiple spots on an image.

It should be noted that our model is asymmetric in μ and x . This is not uncommon; for example, the full Procrustes error is not symmetrical (see Dryden and Mardia, 1998). Also the standard RMSD used by bioinformaticians is again not a symmetrical measure. However, there are symmetrical unlabelled shape analyses; see Green and Mardia (2006), for example. However, this method has not been developed for affine transformations and warping as required here. There is also a non-probabilistic method of Rangarajan *et al.* (1997) for similarity shape, but again the extension of the method to affine transformations and warping requires further work;

see Kent *et al.* (2010b) for a statistical framework. For the data considered here, we have verified that reversing the role of μ and x does not change the broad conclusions.

Finally, we note that the methods described in this paper could have applications in other situations in which there are unlabelled points, some of which — possibly with error — have been manually identified. Thus, the method could be used in the preparation of ground truth for training an object recognition system or a pose estimation system; for example, see the survey of Murphy-Chutorian & Trivedi (2008).

Acknowledgements. We would like to thank Roz Banks and Rachel Craven for providing us with gel data and general discussion concerning protein gels. We would also like to thank David Hogg for useful references about further applications. The second author was supported by an CASE studentship funded by the Engineering and Physical Science Research Council and Central Science Laboratories, York, UK.

SUPPLEMENTARY MATERIAL

Western Blot data

(doi: <http://lib.stat.cmu.edu/aoas/???/???>; .tar.gz). The supplementary data contains a zipped file which includes information taken from 28 Western Blots. This represents 8 subjects (four controls and four patients) treated with two possible treatments. A replicate image is also obtained for each subject-treatment combination, though some replicates are missing. Further details are included in the associated README file.

References.

- [1] BANKS, R.E., DUNN, M.J., HOCHSTRASSER, D.F., SANCHEZ, J-C. BLACKSTOCK, W., PAPPIN, D.J. and SELBY, P.J. (2000) Proteomics: new perspectives, new biomedical opportunities, *Lancet*, **356**, 1749–1756.
- [2] BERKELAAR, M. (2008), Interface to lp_solve v. 5.5 to solve linear/integer programs, R package.
- [3] BESL, P.J. & MCKAY, N.D. (1992). A method for registration of 3-D shapes. *IEE trans. PAMI*, **14**, 239–256.
- [4] CHEN, P. (2011). A novel kernel correlation model with the correspondence estimation. *J. Math Imaging Vis.*, **39**, 100–120.
- [5] CHUI, H. & RANGARAJAN, A (2003). A new point matching algorithm for non-rigid registration. *Computer Vision and Understanding*, **89**, 114–141.

- [6] CZOGIEL, I., DRYDEN, I.L. & BRIGNELL, C.J. (2012) Bayesian matching of unlabeled point sets using random fields, with an application to molecular alignment. *Annals of Applied Statistics*, to appear.
- [7] DRYDEN, I.L., HIRST, J.D. & MELVILLE, J.L. (2007) Statistical analysis of unlabelled point sets: comparing molecules in cheminformatics. *Biometrics* **63**, 237–251.
- [8] DRYDEN, I.L. & MARDIA, K.V. (1998) *Statistical Shape Analysis*, J. Wiley & Sons, Chichester.
- [9] DRYDEN, I.L. & WALKER, G. (1999) Highly resistance regression and object matching, *Biometrics*, **55**, 820–825.
- [10] FISCHLER, M.A. and BOLLES, R.C. (1981). Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, **24**, 381–395.
- [11] FORGBER, M., GELLRICH, S., SHARAV, T., STERRY, W. & WALDEN, P. (2009) Proteome-based analysis of serologically defined tumor-associated antigens in cutaneous lymphoma, *PLoS ONE*, **4**, e8376.
- [12] GLAUNES, J., TROUVÉ, A. & YOUNES, L. (2004) Diffeomorphic matching of distributions: a new approach for unlabelled point-sets and sub-manifolds matching. *CVPR (2) 2004*, pp. 712–718.
- [13] GREEN, P.J. & MARDIA, K.V. (2006) Bayesian alignment using hierarchical models, with applications in protein bioinformatics, *Biometrika*, **93**, 235–254.
- [14] GREEN, P.J., MARDIA, K.V., NYIRONGO, V.B. and RUFFIEUX, Y. (2010). Bayesian modeling for matching and alignment of biomolecules. In *The Oxford Handbook of Applied Bayesian Analysis*, O’Hagan, A. and West, M. (Eds.), pp27–50, Oxford University Press.
- [15] KENT, J.T., MARDIA, K.V. & TAYLOR, C.C. (2010a) Matching unlabelled configurations and protein bioinformatics. Research Report STAT10-01, University of Leeds.
- [16] KENT, J.T., MARDIA, K.V. & TAYLOR, C.C. (2010b) An EM interpretation of the SOFTASSIGN algorithm for alignment problems. In *LASR10 — High-throughput sequencing, proteins and statistics*. Eds. Gusnanto, A., Mardia, K.V., Fallaize, C.J. & Voss, J. pp. 29–32.
- [17] McLACHLAN, G.J. & KRISHNAN, T. (2008) *The EM algorithm and extensions*. Wiley, New Jersey.
- [18] MURPHY-CHUTORIAN, E. & TRIVEDI, M.M. (2008) Head pose estimation in computer vision: a survey. *IEEE transactions on pattern analysis and machine intelligence* **31**, 607–626.
- [19] PETTY, E.M. (2009) *Shape analysis in bioinformatics*. PhD thesis, University of Leeds.
- [20] RANGARAJAN, A., CHUI, H., & BOOKSTEIN, F. L. (1997) The SOFTASSIGN Procrustes matching algorithm. In *Conference on Information Processing in Medical Imaging IPMI 97*, 29–42.
- [21] ROHR, K., CATHIER, P., & WÖRZ, S. (2004) Elastic registration of electrophoresis images using intensity information and point landmarks. *Pattern Recognition*, **37**, 1035–1048.
- [22] TAYLOR, C. C., MARDIA, K. V. & KENT, J. T. (2003) Matching unlabelled configurations using the EM algorithm. In Aykroyd, R.G., Mardia, K.V. & Langdon, M.J. (eds.) *LASR Proceedings: Stochastic geometry, biological structure and images*, pp. 19–21.
- [23] TSIN, Y. & KANADE, T. (2004). A Correlation-Based Approach to Robust Point Set Registration. In ECCV, pp. 558–569.

- [24] WALKER, G. (2000) *Robust, non-parametric and automatic methods for matching spatial point patterns*. PhD thesis, University of Leeds.
- [25] ZVELEBIL, M. AND BAUM, J.O. (2007). Understanding Bioinformatics. *Garland Science*, 613–620.

DEPARTMENT OF STATISTICS

UNIVERSITY OF LEEDS

LEEDS LS2 9JT UK

E-MAIL: k.v.mardia@leeds.ac.uk; emma.m.petty@gmail.com; c.c.taylor@leeds.ac.uk