Protein gels matching

Kanti V. Mardia[1], Vic Patrangenaru[2], and Samanmalee Sugathadasa*[2]

[1] Department of Statistics, University of Leeds
[2] Department of Mathematics and Statistics,
Texas Tech University

1 Introduction

2D-gel electrophoresis (2DGE) is a widely used technique for separating protein molecules for identification purposes. Different proteins have different molecular weights and electrical charges, hence behave differentially in an applied electric field. In 2DGE protein molecules are allowed to move in a gel medium under an X-Y electric field, and a snapshot is taken once molecules have had sufficient time to move away from the initial point. Due to the variability of gels, electric fields, and environmental conditions, the configuration of protein molecules is generally considered to be determined only up an affine transformation. Problem of matching electrophoresis data amounts to matching locations of the same protein in different 2DGE images.

Here we discuss the problem of matching proteins from two 2DGE images under the assumption that the image points are matched onto each other via a linear transformation. A data set in Horgan *et al.* (1992), consisting two images from the same protein sample consisting of 35 data points is used to illustrate the methodology discussed. In the images the first 10 invariant points have already been matched. The 2DGE images can be seen in Dryden and Mardia (1998, p.20).

Rather than using planar Cartesian coordinates, in our approach we regard the linear shape of a configuration of k points, as a point on an linear shape space. The space of planar linear shapes of such configurations is in a one to one correspondence with the Grassmann manifold $\mathrm{G}_2(k,\mathbb{R})$. We embed this manifold in the space of k by k symmetric matrices (Dimitric, 1996; Patrangenaru and Mardia, 2003) and consider on the linear shape space the metric induced by the Euclidean distance in the space of matrices via this embedding. The matching procedure is thus an extrinsic algorithm.

2 An extrinsic matching algorithm on linear shape spaces and on affine shape spaces

Let us suppose that two images of 2DGE are to be matched, and the images have been annotated with coordinates of k data pairs $\{x_i = (x_{i,1}, x_{i,2})\}_{i=1}^k$ and $\{y_i = (y_{i,1}, y_{i,2})\}_{i=1}^k$ respectively. Let us assume that the first m data points have been matched. We are seeking a permutation $\pi \in S_{k-m}$ which yields a match for the remaining $k-m$ coordinates.

Let us denote by G either the group of linear transformations or the group of affine transformations of $\mathbb{R}^2$. Let S_{k-m} denote the group of permutations on last $k-m$ elements of an ordered set of k elements. Let us consider the group action, $\Pi \times G \times (\mathbb{R}^2)^k \to (\mathbb{R}^2)^k$,

$$(\pi, g).[x_1^T, \cdots, x_k^T]^T = [(x_{\pi(1)}g)^T, \cdots, (x_{\pi(k)}g)^T]^T.$$

In the case when $G = GL(2, \mathbb{R})$ we may identify the G orbit of a point in $\mathbb{R}^{2k}$ with the two dimensional subspace in $\mathbb{R}^k$ spanned by the vectors $[x_{1,1}, \cdots, x_{k,1}]^T$ and $[x_{1,2}, \cdots, x_{k,2}]^T$. Hence the space of orbits under the $GL(2, \mathbb{R})$ action is the Grassmann manifold of two dimensional subspaces in $\mathbb{R}^k$, $\mathrm{G}_2(k, \mathbb{R})$. In the case when G is the affine group $GL(2) \times \mathbb{R}^2$, we may first subtract the mean from the data to account for the translation part of the action, hence the orbit space may be treated as the Grassmann manifold $\mathrm{G}_2(k-1, \mathbb{R})$. Now the permutation part of the action remains a permutation action on the Grassmannian. Thus the matching problem in 2DGE amounts to finding a permutation so that the distance between the two affine shapes of configurations of ordered points in the two images is minimized. Each such affine shape is a point on the Grassmannian. Let us recall that the distance $\rho(V, W)$ between points V and W of a Grassmannian is naturally defined as the distance between the two linear operators corresponding to orthogonal reflections of $\mathbb{R}^k$ about V and W respectively (Patrangenaru and Mardia, 2003). Thus we have the following

Problem: Given two subspaces V and W in $\mathrm{G}_2(k, \mathbb{R})$, and an integer $m < k$, find the permutation π on the last $k - m$ elements of an ordered set of k elements, such that $\rho(\pi(V), W)$ is minimized.

Here we describe a numerical scheme which is guaranteed to converge to the permutation that solves our problem in the case of the matching under the linear group. This algorithm appears to scale reasonably well with the configuration size k of the problem. We haven't done a careful analysis of the algorithm yet. Our numerical experiments have thus far shown that the algorithm is rapidly convergent.

Let us begin with an orthogonal bases $\{v_1, v_2\}$ of V and $\{w_1, w_2\}$ of W respectively. When π is a permutation let us denote $\pi(v_i)$ by v_i^π. Let P_V and P_W denotes orthogonal reflection operators from $\mathbb{R}^k$ about V and W respectively. Then we have,

$$\begin{aligned}(\rho(V, W))^2 &= \|P_V - P_W\|^2 \\ &= \|P_V\|^2 + \|P_W\|^2 - 2 < P_V, P_W > \\ &= 4 + 4 - 2\mathrm{trace}(\mathrm{P_V P_W}).\end{aligned}$$

Thus we are seeking for a permutation $\pi*$ such that $\mathrm{trace}(\mathrm{P}_{\pi*(\mathrm{V})}\mathrm{P_W}) \geq \mathrm{trace}(\mathrm{P}_\pi(\mathrm{V})\mathrm{P_W})$ for all allowable permutations π. Since $P_V = [v_1, v_2][v_1, v_2]^T$ we may state,

$$\mathrm{trace}\mathrm{P}_{\pi(\mathrm{V})}\mathrm{P_W} = \sum_{\mathrm{i,j=1,2}} [(\mathrm{v_i^\pi})^\mathrm{T}\mathrm{w_j}]^2. \tag{2.1}$$

A straightforward search over all permutations isn't possible since it will involve examining $(k - m)!$ rearrangements. Our proposed algorithm for $\mathrm{G}_2(k, \mathbb{R})$ involves three phases:

Step 1: for $i = m + 1$ to $k - 1$ set π to be the permutation which transposes $m + 1$ and i. Use formula (2.1) to find i that maximize $\mathrm{trace}\mathrm{P}_{\pi(\mathrm{V})}\mathrm{P_W} - \mathrm{trace}\mathrm{P_V P_W}$. Interchange rows $m + 1$ and i of v_1 and v_2 vectors.

Step 2: repeat step 1 until no i can be found.

Step 3: Repeat steps 1 and 2 with $m + 1$ replaced by $m + 2, m + 3, \cdots, k - 1$.

Observe that each step terminate with a larger value of $\mathrm{trace}\mathrm{P}_{\pi(\mathrm{V})}\mathrm{P_W}$ than before, hence the algorithm proceeds in the right direction. However, due the quadratic nature of the equation (2.1), it may run into a situation in which no transposition will improve the error, yet there are still acceptable permutations that will improve the error. The way out of this is to use a three element permutation to get started again. This follows from the following

Lemma 1. *Suppose* $\pi \in \mathrm{id} \times S_{k-m}$ *is a permutation such that* $\mathrm{trace} \mathrm{P}_{\pi(\mathrm{V})} \mathrm{P}_{\mathrm{W}} - \mathrm{trace} \mathrm{P}_{\mathrm{V}} \mathrm{P}_{\mathrm{W}} > 0$*. Then, for generic subspaces* V *and* W*,* π *may be found so that it only permutes three elements.*

The algorithm is applied to the gel data in Horgan *et al.* (1992) (see also Dryden and Mardia, 1998). Numerical results will be presented in the poster with comparison to the EM algorithm of Kent *et al.* (2004).

Acknowledgements

We would like to thank Ananda Bandulasiri for useful conversations. VP's research is supported by NSF Grant DMS-0406151.

References

Dimitric, I. (1996). A note on equivariant embeddings of Grassmannians. *Publ. Inst. Math. (Beograd) (N.S.)* **59**, 131-137.

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

Horgan G. W., Creasey, A. and Fenton, B. (1992). Superimposing two dimensional gels to study genetic variation in malaria parasites, *Electrophoresis* **13** 871-875.

Kent, J.T., Mardia, K.V. and Taylor, C.C. (2004). Matching problems for unlabelled configurations. *Proceedings in Bioinformatics, Images and Wavelets*. 33-36. Edited by Aykroyd, R.G., Barber, S. and Mardia, K.V. Leeds, Leeds University Press.

Patrangenaru, V. and Mardia, K.V. (2003). Affine Shape Analysis and Image Analysis. *Proceedings in Stochastic Geometry, Biological Structure and Images*, 57-62. Edited by Aykroyd, R.G., Mardia, K.V. and Langdon, M.J. Leeds, Leeds University Press.