19

MAD-Bayes matching and alignment for labelled and unlabelled configurations

Peter J Green

University of Bristol, and UTS, Sydney

19.1 Introduction

When Professor Mardia presented a seminar on protein structural bioinformatics in Bristol in February 2003, I was fascinated by one of the problems he described, about matching and alignment, impressed by his visual aids (the things you could do with overlaid acetates!), but rather unsatisfied by the inferential approach he took. The basic problem (which is properly introduced below) involves two key unknown quantities – the matching between unspecified subsets of two data clouds and the geometrical transformations the clouds had each been subjected to – and it seemed to me essential to treat these two things simultaneously, not sequentially: if that is accepted, then it is natural to use a Bayesian treatment. I think I said something to this effect in discussion, and followed it up later with a proposed model framework, which Professor Mardia and I investigated, with the results eventually becoming a *Biometrika* paper, Green and Mardia (2006).

Some subsequent developments of this idea appear in Mardia et al. (2007) (using the formal Bayesian fitting algorithm as a numerical technique for refining a non-Bayesian solution), Ruffieux and Green (2009) (extending the idea to alignment of multiple configurations), Green et al. (2010) (largely a review article, but describing broader classes of biomolecular matching and alignment problems, and anticipating extensions to the modelling) and Fallaize et al. (2014) (employing a 'gap prior' to use sequence information when it is available).

I have also enjoyed robust, but friendly, conversations about the approach with both of the Editors of this volume, each of whom has also made significant contributions to understanding and addressing the problem, including Kent et al. (2004) and Kenobi and Dryden (2012).

This paper revisits inference based on the models like those in Green and Mardia (2006) and Fallaize et al. (2014), using MAD-Bayes, a new perspective on fast approximate inference due to Broderick et al. (2013). This view might help to reconcile rival paradigms applied to this problem: it turns out to nicely bridge the gap between Bayesian and optimisation approaches to inferring matching and alignment.

This is a Book Title Name of the Author/Editor

19.2 Modelling protein matching and alignment

A mathematical abstraction of a certain problem in protein alignment involves a form of unlabelled shape analysis: we observe two point configurations $x = \{x_j : j = 1, 2, \ldots, m\}$ and $y = \{y_k : k = 1, 2, \ldots, n\}$ in $\mathcal{R}^d$ (typically $d = 2$ or 3); unknown subsets of each configuration are assumed to be matched, apart from noise, but the two configurations have been subject to different unknown geometrical transformations. These transformations are assumed to lie in prescribed families, e.g. translations, rotations, rigid-body or affine transformations, or perhaps there has been some nonlinear warping. The problem is to make simultaneous inference about the alignment and the (relative) transformations. In turn this abstraction can be set up in various ways: to preserve symmetry in the treatment of x and y, Green and Mardia (2006) supposed both configurations to be transformed from some latent configuration in another space, after being subject to both thinning and the addition of noise.

For the case of affine transformations, Green and Mardia (2006) assumed that the x configuration lies in the same d-dimensional space as the latent points, while the y configuration needs transforming to $Ay + \tau$ to lie in this space. The noise is assumed zero-mean spherical gaussian with variance σ^2, independently for each point. The alignment between the configurations is represented by the binary (0/1) matrix M, where $M_{jk} = 1$ if and only if x_j and y_k are matched. Each point can be matched at most once, so there is at most one non-zero entry in each row and each column of M. We will write $\{j \underset{M}{\sim} k\}$ for the set of (j, k) pairs matched according to M, that is $\{(j, k) : M_{jk} = 1\}$.

In Green and Mardia (2006), a stochastic model for the point configurations and their alignment is derived, leading to a posterior distribution of the form

$$p(M, A, \tau|\sigma, x, y) \propto |A|^n p(A)p(\tau) \prod_{j \underset{M}{\sim} k} \left(\frac{\rho\phi\{(x_j - Ay_k - \tau)/\sigma\sqrt{2}\}}{\lambda(\sigma\sqrt{2})^d} \right) \tag{19.1}$$

over the unknown parameters A, τ and M, assuming here that σ is fixed, where ϕ is the standard normal density.

In the modelling, the distribution of the alignment M arises indirectly through a thinned-hidden-point formulation, and the induced prior for M has the form

$$p(M) \propto \left(\frac{\rho}{\lambda v} \right)^L \tag{19.2}$$

where $L = \sum_{jk} M_{jk}$ is the number of matches. It follows that all feasible alignment matrices M with the same value for L have the same prior probability: $M|L$ is uniformly distributed. Of course the number of different M with the same value of L varies greatly with the value of L – in fact it is $m!n!/[L!(m - L)!(n - L)!]$ (Green and Mardia 2006).

Expressions similar to (19.1) can arise from other underlying formulations by other authors, perhaps with $\sigma\sqrt{2}$ replaced by σ, and perhaps with $\rho/\lambda v$ expressed as a single parameter.

Green and Mardia (2006) build a methodology using the posterior distribution (19.1), concentrating primarily on the case of a rigid-body motion in 2- or 3-dimensions, where A is a rotation matrix, modelled a priori by a matrix Fisher distribution. Posterior sampling can be accomplished with a relatively straightforward Markov chain Monte Carlo (MCMC) sampler. This uses Gibbs updates for σ^2 and τ, Metropolis–Hastings updates for M (in which addition, deletion, or switching of matches are proposed), and, in the 3-D case, a novel Metropolis sampler for the matrix Fisher distribution for updating A.

For Bayesian point estimation of the alignment, we can take a decision theory approach based on a loss function that is additive over (j, k) pairs and exchangeable with respect to indexing. This turns out to require only the pairwise posterior match probabilities $P\{M_{jk} = 1|x, y\}$, which are readily estimated by direct enumeration from a MCMC sample. The resulting optimisation computation is equivalent to a mathematical programming assignment problem, and standard methods can be used to solve it.

These methodologies were illustrated by application to alignment of 2-D protein gels, and of 3-D configurations of active sites. The MCMC methodology is in principle vulnerable to mixing problems caused by multi-modality in the posterior distribution, although such problems are not apparent in the examples shown.

19.3 Gap priors and related models

When sequence information is available, it is appealing to consider using it, and an attractive approach is to use a 'gap prior' of the form

$$p(M) \propto \exp(-U(M))$$

using the so-called gap penalty $U(M)$ given by

$$U(M) = gS(M) + h \sum_{i=1}^{S(M)} (l_i - 1), \tag{19.3}$$

where $S(M)$ is the number of instances where a new gap in the alignment is opened, l_i is the length of the ith gap, and g, h are positive hyperparameters, with commonly, $g > h$. See Rodriguez and Schmidler (2010) and Fallaize et al. (2014). Informally, the effect of using this prior with $g > h > 0$ compared to $g = h = 0$ is that among alignments with the same likelihood, preference is given to those where consecutively numbered atoms are matched, and where this fails, preference goes to those where the unmatched atoms are consecutive.

Using this prior in place of that used by Green and Mardia (2006), with other modelling details unchanged, leads to the posterior

$$p(M, A, \tau|\sigma, x, y) \propto |A|^n p(A) p(\tau) v^L \exp(-U(M)) \prod_{j \underset{M}{\sim} k} \left(\frac{\phi\{(x_j - Ay_k - \tau)/\sigma\sqrt{2}\}}{(\sigma\sqrt{2})^d} \right). \tag{19.4}$$

Although the gap penalty is commonly expressed in the form (19.3), this form is arguably ambiguous, and it can helpful to express it more explicitly (Fallaize et al. 2014). Let M be a binary $m \times n$ matrix with L 1s, located in entries $(j_i, k_i), i = 1, 2, \ldots, L$, where the js and ks are consistently ordered: $j_1 < j_2 < \ldots < j_L$ and $k_1 < k_2 < \ldots < k_L$. This represents, of course the matching of x_{j_i} and y_{k_i}, for $i = 1, 2, \ldots, L$. Then the gap penalty can be written

$$U(M) = \sum_{i=1}^{L+1} [f(j_i - j_{i-1}) + f(k_i - k_{i-1})] \tag{19.5}$$

where $f(1) = 0$ and for $r \geq 2$, $f(r) = g + (r-2)h$. Here, we write $j_0 = k_0 = 0$ and $j_{L+1} = m + 1$, $k_{L+1} = n + 1$. We take $U(M) = +\infty$ if the js and ks cannot be consistently ordered, that is, if the alignment M is inconsistent with the sequence ordering; such M have zero prior probability under this model.

In Fallaize et al. (2014), the MCMC algorithm of Green and Mardia (2006) is adapted to sampling for the posterior distribution for the gap prior model. The resulting algorithm relies upon proposing stepwise updates to M corresponding to adding, removing, or switching a match. These are particularly easy to implement for the gap prior. If we insert a new match $(j^\star, k^\star)$ between (j_i, k_i) and (j_{i+1}, k_{i+1}), then the reduction in total gap penalty is the sum of two terms, one from the js and one from the ks. The term from the js is equal to

$$\begin{cases} g & \text{if } j_{i+1} - j_i = 2, \\ h & \text{if } j_{i+1} - j_i > 2 \text{ and } j^\star = j_i + 1 \text{ or } j_{i+1} - 1, \text{ and} \\ 2h - g & \text{otherwise.} \end{cases}$$

These three possibilities correspond to filling (and so eliminating) a gap, shortening a gap, or splitting a gap into two. The term from the ks has the same form.

A feature of this gap model that some might feel unappealing intuitively is that, conditional on the number of matches and the number of gaps, the indices of the x and y points forming those matches are a priori independent. In fact, the penalty $U(M)$, and hence the probability $p(M)$, depends only on L and S where S is the total number of gaps in the two sequences combined; S is the number of blocks of consecutive all-zero rows or columns in M. To be explicit,

$$U(M) = (g-h)S + h(m+n-2L). \tag{19.6}$$

Thus, e.g., if there are 3 matches and the x indices are $(4, 5, 9)$, then under this model the y indices $(7, 8, 12)$ are exactly as probable as $(7, 11, 12)$. Indeed, if $m = 9$ and $n = 15$, this probability is also the same as that the x indices $(1, 2, 3)$ match y indices $(2, j, 14)$, for any $j = 4, 5, \ldots, 12$ as all of these situations give $L = 3$ and $S = 5$. The penalty (19.3) should therefore more accurately termed a ‘gap-count’ penalty!

Changing the specification of $U(M)$ to better match intuition, or different scientific judgement, about likely patterns of insertion and deletion would often still yield a distribution amenable to posterior sampling using an appropriately modified MCMC algorithm. This would be especially straightforward if the penalty remained a sum over the individual gaps, but all that is really needed is that the change to the penalty when a match is deleted, added or switched is cheaply computed, meaning in practice that it uses only information that is local to the revision in M. Two possibilities that come immediately to mind are to use (19.5) but with a function f that is strictly concave but still increasing for positive gap lengths, or to use a form where the penalty is a decreasing function of the correlation between the matched (j, k) indices – with the effect that in the first example above, the x indices $(4, 5, 9)$ are less likely to be matched to the y indices $(7, 11, 12)$ than to $(7, 8, 12)$.

19.4 MAD-Bayes

MAD-Bayes (MAP-based Asymptotic Derivations from Bayes) is a novel methodology for fitting complex stochastic models due to Broderick et al. (2013). It was devised to meet the sometimes contradictory desiderata of complying with the Bayesian paradigm and delivering practical methodology that can be executed very quickly even on large data-sets.

MAD-Bayes is essentially a simple framework for delivering small-variance asymptotic approximations to MAP (*maximum a posteriori*) estimation, yielding results that, while not usually of closed form, are neverthess typically amenable to solution using fast optimisation techniques. It exploits the fact that in many statistical models, when the likelihood is taken to a ‘small-variance’ limit, a non-trivial limit is obtained for the MAP estimator, provided that hyperparameters in the prior are also taken to appropriate limits. Except in the simplest of cases, there may be more than one way to do this, giving different non-trivial limits, so some judgement is needed.

Although MAD-Bayes was conceived as a perspective to take in the presence of nonparametric priors and models with discrete allocation structures such as mixtures and clustering, the idea can be more simply illustrated and understood with a toy example from parametric Bayes. Suppose $y \sim N(X\beta, \sigma^2)$ with a normal prior: $\beta \sim N(\beta_0, \tau^2 I)$. Then of course the posterior is

$$\beta | y \sim N\left(\{\sigma^{-2}X^TX + \tau^{-2}I\}^{-1}\{\sigma^{-2}X^Ty + \tau^{-2}\beta_0\}, \{\sigma^{-2}X^TX + \tau^{-2}I\}^{-1}\right) \tag{19.7}$$

The posterior mean and mode are both $\{X^TX + \alpha I\}^{-1}\{X^Ty + \alpha\beta_0\}$, the value minimising $||y - X\beta||^2 + \alpha||\beta - \beta_0||^2$ over β, where $\alpha = \sigma^2/\tau^2$. This is a nontrivial combination of data and prior information, providing $0 < \sigma^2/\tau^2 < \infty$ strictly. Unlike the other applications of the MAD-Bayes principle for approximating the posterior mode and later the posterior distribution, discussed later in this chapter, these results hold exactly for any positive σ^2.

The canonical example of MAD-Bayes presented by Broderick et al. (2013) provides an extension to the classical K-means clustering algorithm that they call *DP*-means. They propose clustering multivariate data $(x_1, x_2, \ldots, x_n)$ by partitioning the index set $\{1, 2, \ldots, n\}$ as a disjoint union $\bigcup_{j=1}^{K} C_j$, where K, $\{C_j\}$ and cluster means $\{\mu_j\}$ are chosen to minimise

$$\sum_{j=1}^{K} \sum_{i \in C_j} ||x_i - \mu_j||^2 + (K-1)\lambda^2, \tag{19.8}$$

λ being a regularisation constant. This approach, intuitively reasonable in itself, can be derived by a MAD-Bayes argument approximating the MAP estimate of the clustering under a Dirichlet/Chinese restaurant process mixture model (Lo 1984). As with α in the normal linear model example above, the constant λ^2 is the ratio of the variance σ^2 to a function of a hyperparameter in the prior, so the asymptotic framework again demands that the prior concentrates as the variance decreases. Broderick et al. (2013) further illustrate the idea applied to feature learning, particularly exploiting other Bayesian nonparametric prior models such as the Indian buffet process, and various extensions. The idea has more recently been used in feature learning for studying tumour heterogeneity by Xu et al. (2014).

A different kind of recent application is to image segmentation. Pereyra and McLaughlin (2014) apply a MAD-Bayes argument to the posterior arising from an image model based on a hidden Potts–Markov random field. Computing the MAP estimate in this problem is NP-hard, but a convex relaxation is possible, leading ultimately to an objective function of the form

$$\sum_{j=1}^{K} \sum_{i \in C_j} \{||y_i - x_i||^2 + ||x_i - \mu_j||^2\} + \beta ||\nabla x||_1, \tag{19.9}$$

to be minimised over $x, \mu, \{C_j\}$ and K, given a data image y. Here $||\nabla x||_1$ is the ℓ_1 norm of the 1st order discrete gradient of the hidden image x, a convexification of the $||\nabla x||_0$ arising formally from the model. The minimisation over x is equivalent to a total-variation denoising problem, of a kind which has been extensively studied in the recent optimisation literature and that can be solved very efficiently even in very high-dimensional scenarios using parallel proximal splitting methods. The minimisation over the other variables involves K-means clustering.

19.5 MAD-Bayes for unlabelled matching and alignment

To develop a MAD-Bayes method for matching and alignment, we use (19.1) to obtain, ignoring additive constants in the log-posterior,

$$\begin{aligned} & -4\sigma^2 \log p(M, A, \tau | \sigma, x, y) = -4\sigma^2 \log\{|A|^n p(A) p(\tau)\} \\ & -4\sigma^2 L \log(\rho/\lambda) + 4\sigma^2 d L \log(\sigma/\sqrt{2}) + 2\sigma^2 \log 2\pi + \sum_{j \underset{M}{\sim} k} ||x_j - A y_k - \tau||^2. \end{aligned} \tag{19.10}$$

According to the MAD-Bayes approximation paradigm of Broderick et al. (2013), we should examine this function in the small-variance limit, as $\sigma^2 \to 0$. For a non-degenerate limit in this asymptotic analysis, the prior cannot be held fixed. Suppose $\rho/\lambda = \exp(\alpha/4\sigma^2)$ for some real constant α. Then as $\sigma \to 0$ in (19.10) we obtain

$$-4\sigma^2 \log p(M, A, \tau | \sigma, x, y) \to -\alpha L + \sum_{j \underset{M}{\sim} k} ||x_j - A y_k - \tau||^2.$$

Thus finding the MAP estimate of M, A, τ, the values maximising the log posterior, is asymptotically equivalent to minimising the penalised sum-of-squares

$$-\alpha L + \sum_{j \underset{M}{\sim} k} ||x_j - Ay_k - \tau||^2. \tag{19.11}$$

The similarity in general form between (19.11), and (19.8) or (19.9) is clear.

When $\alpha > 0$ there is a non-trivial solution, and the optimisation serves to limit the number of matches L; informally, with A and τ held fixed for simplicity of the argument, including an additional match (j', k') will decrease the penalised sum-of-squares if and only if $||x_{j'} - Ay_{k'} - \tau||^2 < \alpha$.

The parameter α controls the behaviour of the prior parameter ρ/λ in the small variance limit: positive α implies that $\rho/\lambda \to \infty$ as $\sigma^2 \to 0$, at a particular rate. This is easy to understand qualitatively: if the noise variance is reduced so that matches become harder to find, that must be compensated by concentrating the prior for M on higher numbers of matches L.

In summary, this simple analysis of MAP inference in our Bayesian model has reduced to an optimisation problem, penalised least-squares, one with a fairly simple structure by the standards of problems addressable by modern optimisation techniques. For fixed A, τ, optimisation over M is an instance of a weighted matching problem for a bipartite graph, for which the Hungarian algorithm (Jacobi 1890; Munkres 1957) provides a solution; this is usually posed as a maximisation problem and the weight on edge (j, k) to be used would be simply $\max\{0, \alpha - ||x_j - Ay_k - \tau||^2\}$. For fixed M, optimisation over A and τ (say, in the case of rigid body transformation) is an example of Procrustes analysis. It is easy to see (since each step reduces the value of the criterion (19.11) and because the set of possible alignments is finite) that alternating between these two steps defines an algorithm that converges to a possibly local optimum in a finite number of iterations. We stress that this may not be a global optimum as complex models often lead to multi-modal posteriors; we comment futher on multi-modality in Section 19.11.

This simple idea could no doubt be improved using techniques from modern optimisation methodology. But even without such improvement, this algorithm runs very quickly. Without making any attempt to optimise coding of the outer loop, an implementation in `R`, using function `solve_LSAP` from package `clue` and function `procOPA` from package `shapes` provides an algorithm that runs in 0.03 seconds on a 3.20GHz processor for the small problem in Section 4.2 of Green and Mardia (2006), to be compared to 10.85 seconds for 10^6 sweeps of the MCMC sampler on the same problem (but which of course provides a much richer inference).

Note that use of the Hungarian algorithm, or other code for the assignment problem, guarantees that the inferred alignment is feasible in the sense that no point is simultaneously matched to more than one point in the other configuration, in contrast to the formally somewhat similar method using the EM algorithm to compute the maximum likelihood estimate of the alignment (see for example Kent et al. (2004)).

There is a related approach called ‘Softassign Procrustes’ to this problem due to Rangarajan et al. (1997). This proceeds by first relaxing the constraint that M is a binary matrix to set up an iterative deterministic annealing algorithm using Lagrange multipliers that alternates between updating the geometrical parameters and updating M; the method appeals to a theorem of Sinkhorn (1964) to deliver a solution in which M is in fact binary. The Softassign Procrustes algorithm has been given an EM-like interpretation by Kent et al. (2010).

19.6 Omniparametric optimisation of the objective function

An interesting perspective on the optimisation of (19.11) allows simultaneous consideration of all $\alpha \in (0, \infty)$, delivering what is often called a ‘regularisation path’ (for example in the context of the Lasso (Efron et al. 2004)). Picture a two-dimensional scatter plot of points, each representing a possible alignment M, with horizontal coordinate $L(M)$ and vertical coordinate $\sum_{j \underset{M}{\sim} k} ||x_j - Ay_k - \tau||^2$.

The optimal M according to (19.11) corresponds to the point where a line of slope α is a lower tangent to the scatter of points, and the set of all M that are optimal for some α is represented by the lower convex hull of the configuration. Because there are only finitely many possible values of M, this lower convex hull is a polygonal line, so there exists a finite grid of values of α, say $\alpha_0 > \alpha_1 > \alpha_2 > \cdots$, such that for all $\alpha \in (\alpha_{i+1}, \alpha_i)$, $i = 0, 1, \ldots$, the optimal M is constant, say $\hat{M}_i$. Note that $L(\hat{M}_i)$ will decrease with i. One approach to constructing the α_i and $\hat{M}_i$, following a suggestion of the referees, is to proceed sequentially for $i = 0, 1, \ldots$, using each α_i as a starting point for determining α_{i+1}.

The set-up also invites comparison with that of Lau and Green (2007), who discussed optimal Bayesian point estimation of a clustering (of gene expression profiles) based on a pairwise-coincidence loss function. ‘Omniparametric’ optimisation of the expected loss over all values of the parameter in the loss function was implemented in a fast heuristic algorithm, which might be used to inspire a similar approach to the present problem. Following that paradigm would suggest iteratively refining the grid (α_i), starting with an initial pair of (low, high) values; the recursive step to split an interval (α_{i+1}, α_i) would search for a alignment M whose representative point in this diagram lies outside the line segment determined by the interval endpoints.

19.7 MAD-Bayes in the sequence-labelled case

In the sequence-labelled case, the points in each configuration are numbered in sequential order (along a protein, in typical application) and we can use this numbering in specifying a prior on the alignment matrix M. This leads to the posterior (19.4) instead of (19.1). The ‘energy function’ $U(M)$ in the prior for M may take the gap penalty form (19.3) or something more general, either with the same intention of promoting or insisting upon sequence order being maintained, or with some other purpose.

For such a posterior, we obtain

$$\begin{aligned} -4\sigma^2 \log p(M, A, \tau|\sigma, x, y) &= -4\sigma^2 \log\{|A|^n p(A) p(\tau) v^L\} \\ &+ 4\sigma^2 U(M) + 4\sigma^2 dL \log(\sigma/\sqrt{2}) + 2\sigma^2 \log 2\pi + \sum_{j \underset{M}{\sim} k} ||x_j - Ay_k - \tau||^2. \end{aligned} \quad (19.12)$$

Since all of the other terms vanish as $\sigma^2 \to 0$, we need for a non-trivial limit that $U(M)$ or its parameters scale in such a way that $4\sigma^2 U(M)$ has a non-trivial limit. For example, in the case of the gap penalty (19.3), if $8\sigma^2 h \to \alpha$ and $4\sigma^2(g - h) \to \beta$, then according to (19.6) the resulting optimisation problem is to minimise

$$-\alpha L + \beta S + \sum_{j \underset{M}{\sim} k} ||x_j - Ay_k - \tau||^2. \quad (19.13)$$

Intuitive interpretation of this objective function is less straightforward: adding a match always increases L by 1, but the associated change in S may be $+2, +1, 0, -1$ or -2. Optimisation over the alignment for fixed A and τ is no longer a weighted matching problem, taking this set-up out of reach of the Hungarian algorithm; as suggested by the referees, there may be a role here for dynamic programming.

19.8 Other kinds of labelling

In their section 3.6, Green and Mardia (2006) propose a way to extend the model leading to (19.1) to allow simultaneous model-based inference about alignment when the points in the observed configurations are recorded as belonging to different clusters, or 'colours', and pairs of points where both belong to the same cluster are more likely to be matched. An example in protein bioinformatics arises when the amino acids characterising the observed points are categorised as hydrophobic or hydrophilic (possibly subdivided into charged, polar and glycine). The model extension achieving this amounts to modifying the prior on the alignment matrix M to favour like-coloured matches, so provides a general mechanism for handling 'partially labelled' configurations, where labels are not unique.

The modified prior on M that was proposed has the form

$$p(M) \propto \left(\frac{\rho}{\lambda v}\right)^L \prod_{j \underset{M}{\sim} k} \exp(\gamma I[r_j = s_k] + \delta I[r_j \neq s_k])$$

where x_j is coloured r_j and y_k coloured s_k, instead of (19.2). This modification needs only trivial changes to the Metropolis–Hastings updating of M in the posterior simulation.

It is easy to see that such modified priors lead also to a simply-modified MAD-Bayes objective function. The penalised sum-of-squares (19.11) is replaced by

$$-\alpha L + \sum_{j \underset{M}{\sim} k} \left\{||x_j - Ay_k - \tau||^2 + \gamma' I[r_j = s_k] + \delta' I[r_j \neq s_k]\right\}, \qquad (19.14)$$

where $\gamma' = 4\sigma^2\gamma$ and $\delta' = 4\sigma^2\delta$.

Numerical optimisation of (19.14) can again in principle be addressed by alternating between optimising over M and over A and τ, and again the former step is an instance of a weighted matching problem, since the objective function can be expressed as a sum over $\{(j,k) : M_{jk} = 1\}$.

The extensions in this section and the previous one can readily be combined, simultaneously penalising gaps and favouring like-coloured matches, and giving the objective function

$$-\alpha L + \beta S + \sum_{j \underset{M}{\sim} k} \left\{||x_j - Ay_k - \tau||^2 + \gamma' I[r_j = s_k] + \delta' I[r_j \neq s_k]\right\}.$$

19.9 Simultaneous alignment of multiple configurations

Ruffieux and Green (2009) generalised the two-configuration methodology of Green and Mardia (2006) to handle the case of multiple configurations. They argue that information is lost by treating the configurations pairwise; the truth of this is most easily seen in the kind of latent-true-configuration model they use (since we should want to use all information at once in the implicit inference about the positions of the latent points), but the point will be generally true. Kenobi and Dryden (2012) match multiple configurations using a model that considers them only two at a time. The ideas illustrated in this chapter will continue to apply *mutatis mutandis* to the multiple-configuration case, although I do not know whether the discrete optimisation algorithms that would be needed for implementation are still instances of standard optimisation theory problems.

19.10 Beyond MAD-Bayes to posterior approximation?

The motivating example in the Gaussian case delivered the whole posterior (19.7) not only the posterior mode. Could we extend the MAD-Bayes perspective to deliver at least an approximation to the posterior, by slightly refining the asymptotic argument? In this section I attempt only a preliminary, speculative answer to this question, which seems a promising subject for further investigation.

For the unlabelled case, leaving aside technicalities for the moment, the argument leading to the penalised least-squares objective function (19.11) equally well delivers the formal approximation, valid as $\sigma^2 \to 0$,

$$p(M, A, \tau|\sigma, x, y) \approx e^{\alpha L/4\sigma^2} \exp\{(-1/4\sigma^2) \sum_{j \underset{M}{\sim} k} ||x_j - Ay_k - \tau||^2\}. \tag{19.15}$$

Our focus will be to investigate the form of the density on the right hand side. For definiteness, we take the case of rigid-body transformations, so that A is special orthogonal.

It is possible to make some progress interpreting the approximate joint posterior (19.15) by considering the full conditionals for each of A, τ and M in turn.

For A,

$$\begin{aligned} \sum_{j \underset{M}{\sim} k} ||x_j - Ay_k - \tau||^2) &= \sum_{j \underset{M}{\sim} k} \{||x_j - \tau||^2 + ||y_k||^2 - 2(Ay_k)^T(x_j - \tau)\} \\ &= \sum_{j \underset{M}{\sim} k} ||x_j - \tau||^2 + \sum_{j \underset{M}{\sim} k} ||y_k||^2 - 2\text{tr}\{A^T \sum_{j \underset{M}{\sim} k} (x_j - \tau)y_k^T\} \end{aligned} \tag{19.16}$$

This reveals that under the approximate distribution (19.15), A given τ and M (and x, y, σ) has a matrix Fisher distribution (Mardia and Jupp 2000, p.289), as shown in Green and Mardia (2006). The normalising constant of this distribution is known, so that A can be integrated out to give

$$p(M, \tau|\sigma, x, y) \approx e^{\alpha L/4\sigma^2} \exp\{(-1/4\sigma^2) \sum_{j \underset{M}{\sim} k} ||x_j - \tau||^2 + ||y_k||^2\} {}_0F_1(p/2, (1/16\sigma^4)F^T F), \tag{19.17}$$

where $F = \sum_{j \underset{M}{\sim} k} (x_j - \tau)y_k^T$ depends on both τ and M, as well as the data. This does not seem amenable to further analytic simplification.

Similarly, we can evidently extract from the right hand side of (19.15) the approximate conditional for τ given M and A as

$$\tau|M, A, x, y, \sigma \sim N(L^{-1} \sum_{j \underset{M}{\sim} k} (x_j - Ay_k), 2\sigma^2/L),$$

while the approximate conditional for M given τ and A is also explicit but hardly tractable.

In an effort to gain more insight into the form of the approximate posterior, we could consider one of the approximations to the matrix Fisher distribution developed by Khatri and Mardia (1977) and Bingham et al. (1992). However, these seem too intricate to use for practical statistical analysis.

So let us consider further approximation: we could try to use a Normal approximation for $p(A|M, \tau, x, y, \sigma)$. Suppose that $A \sim \text{MatrixFisher}(F)$ with F non-singular; note that this demands that the M in question matches sufficiently many (x, y) pairs with coordinates in general position. Now let $K = (F^T F)^{1/2}$ be the elliptical part of F and $N = FK^{-1}$ its polar part (Mardia and Jupp 2000, p. 286). Let $V\Delta V^T$ with $\Delta = \text{diag}(\delta_1, \delta_2, \ldots, \delta_d)$ be the spectral decomposition of K. In the concentrated case, where all δ_i become large (many matches), we have (Peter Jupp, *personal communication*)

$$(A - N) \approx NVSV^T$$

where S is a skew-symmetrix matrix with $(\delta_i + \delta_j)^{1/2} s_{ij} \sim N(0,1)$, independently.

It seems probable that the argument leading to this can be refined to yield a joint Normal approximation for $p(A, \tau | M, x, y, \sigma)$, although I have not attempted to verify the details. Under such an approximation, the approximate joint posterior (19.15) becomes a Normal mixture distribution, and this seems to be the analysis of (19.15) most likely to be useful for numerical implementation. More work is needed here.

Returning to the mathematical basis for the approximation, a rigorous analysis would need to establish that the approximation of densities that we have investigated really does imply convergence of the probability measures (say, in the sense of total variation norm) under suitable regularity conditions.

19.11 Practical uses of MAD-Bayes approximations

It is hoped that the optimisation-based techniques suggested in this chapter could be developed to make a practically-useful contribution to methodology. They seem to offer to supply some of the advantages of the Bayesian approach – notably treating uncertainty about the alignment and the geometrical transformation symmetrically and simultaneously – without having to pay the price of relying on Monte Carlo computation.

However, even neglecting the fact that the Bayesian set-up has to be approximated to allow delivery of these optimisation solutions, there are other caveats. In particular, they are not a panacea for the problems of multi-modality that can bedevil MCMC methods. The MAD-Bayes perspective is really blind to the possibly existence of modes other that the one under consideration, and numerical optimisation methods need to be special and carefully chosen to deliver optima of multi-modal objective functions reliably, just as MCMC methods have to be specifically designed to handle multi-modal target distributions.

It may be useful to regard optimisation approaches as complementary to posterior sampling – for example, MAD-Bayes might provide a starting point for a MCMC simulation, from which perhaps a rather short MCMC run might be used to assess variability; again this would demand some guarantee about unimodality for reliable inference. This is very much in the spirit of the work of Mardia et al. (2007).

Acknowledgements

I am indebted to Peter Jupp for kindly deriving the high-concentration Normal approximation to the matrix Fisher distribution for me, to the referees for very helpful comments and pointers, to my collaborators on these problems, and above all to Kanti Mardia for his friendship and many stimulating discussions over the years.

References

Bingham C, Chang T and Richards D 1992 Approximating the matrix Fisher and Bingham distributions: applications to spherical regression and Procrustes analysis. *Journal of Multivariate Analysis* **41**, 314–337.

Broderick T, Kulis B and Jordan MI 2013 MAD-Bayes: MAP-based asymptotic derivations from Bayes *Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA. JMLR: W&CP volume 28.* See also arXiv:1212.2126.

Efron B, Hastie T, Johnstone I and Tibshirani R 2004 Least angle regression. *The Annals of Statistics* **32**(2), 407–499.

Fallaize CJ, Green PJ, Mardia KV and Barber S 2014 Bayesian protein sequence and structure alignment Current version at arXiv:1404.1556.

Green PJ and Mardia KV 2006 Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika* **93**, 235–254. doi:10.1093/biomet/93.2.235.

Green PJ, Mardia KV, Nyirongo VB and Ruffieux Y 2010 Bayesian modelling for matching and alignment of biomolecules *The Oxford Handbook of Applied Bayesian Analysis* Oxford University Press pp. 27–50.

Jacobi CGJ 1890 De aequationum differentialum systemate non normali ad formam normalem revocando *C.G.J. Jacobi's gesammelte Werke, fünfter Band* K. Weierstrass, Berlin, Bruck und Verlag von Georg Reimer pp. 485–513.

Kenobi K and Dryden IL 2012 Bayesian matching of unlabeled point sets using Procrustes and configuration models. *Bayesian Analysis* **7**(3), 547–566.

Kent JT, Mardia KV and Taylor CC 2004 Matching problems for unlabelled configurations In *Bioinformatics, Images, and Wavelets* (ed. Aykroyd RG, Barber S and Mardia KV), pp. 33–36.

Kent JT, Mardia KV and Taylor CC 2010 An EM interpretation of the Softassign algorithm for alignment problems In *High-throughput Sequencing, Proteins and Statistics* (ed. Gusnanto A, Mardia KV, J FC and Voss J), pp. 29–32. Leeds University Press.

Khatri CG and Mardia KV 1977 The von Mises–Fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society, B* **39**, 95–106.

Lau JW and Green PJ 2007 Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics* **16**, 526–558.

Lo AY 1984 On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics* **12**(1), 351–357.

Mardia KV and Jupp PE 2000 *Directional statistics*. Wiley, Chichester.

Mardia KV, Nyirongo VB, Green PJ, Gold ND and Westhead DR 2007 Bayesian refinement of protein functional site matching. *BMC Bioinformatics* **8**, 257.

Munkres J 1957 Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics* **5**(1), 32–38.

Pereyra M and McLaughlin S 2014 Small-variance asymptotics of hidden Potts-MRFs: Application to fast Bayesian image segmentation *Proc. European Signal Processing Conf. (EUSIPCO)*, Lisbon, Portugal.

Rangarajan A, Chui H and Bookstein FL 1997 The Softassign Procrustes matching algorithm *Information Processing in Medical Imaging, Lecture Notes in Computer Science Volume 1230*, pp. 29–42.

Rodriguez A and Schmidler S 2010 Bayesian protein structural alignment Submitted to Annals of Applied Statistics. Current version at http://www.stat.duke.edu/~scs/Papers/BayesStructAlignAAS.pdf.

Ruffieux Y and Green PJ 2009 Alignment of multiple configurations using hierarchical models. *Journal of Computational and Graphical Statistics* **18**, 756–773. doi:10.1198/jcgs.2009.07048.

Sinkhorn R 1964 A relationship between arbitrary positive matrices and doubly stochastic matrices. *Annals of Mathematical Statistics* **35**, 876–879.

Xu Y, Mueller P, Yuan Y, Gulukota K and Ji Y 2014 MAD Bayes for tumor heterogeneity – feature allocation with exponential family sampling. Technical report, Department of Mathematics, University of Texas Austin. arXiv:1402.5090.

INDEX