Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics
Benidorm (Alicante, Spain), June 1st–6th, 2006

Fast Bayesian Shape Matching Using Geometric Algorithms

Scott C. Schmidler
Duke University, USA
schmidler@stat.duke.edu

Summary

We present a Bayesian approach to comparison of geometric shapes with applications to classification of the molecular structures of proteins. Our approach involves the use of distributions defined on transformation invariant shape spaces and the specification of prior distributions on bipartite matchings. Here we emphasize the computational aspects of posterior inference arising from such models, and explore computationally efficient approximation algorithms based on a geometric hashing algorithm which is suitable for fully Bayesian shape matching against large databases. We demonstrate this approach on the problems of protein structure alignment, structural database searching, and structure classification. We discuss extensions to flexible shape spaces developed in previous work.

Keywords and Phrases: Bioinformatics; Geometric Hashing; Protein Structure; Shape Analysis.

1. INTRODUCTION

Analysis of data obtained as measurements on geometric objects arises in a number of fields, including image processing and computer vision, biological anthropology and anatomy, molecular biology and chemistry, and mechanical engineering to name but a few. Such data differs from standard high dimensional, multivariate data by the necessity of accounting for natural invariances with respect to geometric transformations such as rigid body motions, viewpoint and projection, and scaling. Treatment of such problems falls under the heading of *statistical shape analysis* and a several different directions of theoretical and methodological work have been developed by various authors (Bookstein, 1991; Small, 1996; Dryden and Mardia, 1998; Kendall et al., 1999; Lele and Richtsmeier, 2001). In the current paper we focus on *landmark* methods for shape analysis, which represent objects by a set of n landmark points in d dimensions. An observed configuration is then given by an $n \times d$ matrix X, with $x_i \in \mathbb{R}^d$ the i^{th} row of X representing the coordinates of the i^{th} landmark.

2. SHAPE AND SHAPE SPACE

For simplicity and concreteness we consider in this paper only *rigid-body shape* (or *size-and-shape*) arising from invariance under special Euclidean transformations.

The *shape* of X denoted by $[X]$ is then defined as the equivalence class

$$[X] = \{XR + \mathbf{1}\mu' \mid R \in \mathbf{SO}(d),\ \mu \in \mathbb{R}^d\}$$

which is the orbit of X under the special Euclidean group $SE(d)$. In general $SE(d)$ may be replaced with a more general class of transformations; common examples include similarity group (including scaling), the affine group, projective transformations, various smooth nonlinear mappings, and piecewise combinations (Dryden, 1999; Schmidler, 2006).

The set of equivalence classes arising from all configurations of n landmarks in d dimensions is called the *(size-and-)shape space* and denoted by $\mathrm{S}\Sigma_n^d$. Shape analysis involves the analysis of observed configuration data in this space, which is invariant under rigid-body transformations. Distances in shape space for two observed configurations with corresponding landmarks may be obtained via the *Procrustes distance*,

$$d_P^2(X,Y) = \min_{\substack{\mu\in\mathbb{R}^3 \\ R\in\mathbf{SO}(3)}} \|Y - (XR + \mathbf{1}\mu')\|_F^2 = \|X_c\|^2 + \|Y_c\|^2 - 2\mathrm{tr}(D)$$

where $\|X\|_F$ denotes Frobenious norm, X_c denotes centering, and $Y'X = UDV'$ is a singular value decomposition. Procrustes distance and related metrics provide one approach to development of distribution theory and multivariate analysis methodology for statistical analysis of shape data (e.g. Dryden and Mardia (1998)).

To date such analyses have largely been predicated on the existence of a one-to-one landmark correspondence between X and Y. However, an important aspect of shape matching in many applied problems is the need to identify such a correspondence, or to determine if one exists. We have previously developed a Bayesian approach to this problem.

3. BAYESIAN SHAPE MATCHING

We have recently developed a Bayesian approach to shape matching motivated by problems in structural proteomics (Schmidler, 2006; Rodriguez and Schmidler, 2006a; Wang and Schmidler, 2006; Rodriguez and Schmidler, 2006b). Let $X_{n\times 3}$ and $Y_{m\times 3}$ be two configuration matrices. We define an *alignment* between X and Y to be a pair $A = [M, \theta]$ where θ is a transformation (here $\theta = (R, \mu) \in SE(3)$) and M is a bipartite matching. Denote the set of matchings by $\mathcal{M}_{n,m}$; a convenient representation for $M \in \mathcal{M}_{n,m}$ is a match matrix $M_{n\times m} = [m_{ij}]$ such that $m_{ij} = 1$ if landmarks X_i and Y_j are matched, and 0 otherwise. Denote by X_M and Y_M the p non-zero rows of MX and $M'Y$ respectively, giving the coordinates of the matched residues.

We have developed a Bayesian approach to shape alignment which defines a prior distribution on alignments $P(M)$ and obtains the posterior distribution:

$$P(M \mid X, Y) = \frac{P(X, Y \mid M)P(M)}{\sum_M P(X, Y \mid M)P(M)}$$

under a likelihood defined on shape space given by the joint density:

$$\begin{aligned} P(X, Y \mid M) &= P(Y_M \mid X_M)P(X_M)P(Y_{\bar{M}})P(X_{\bar{M}}) \\ &= (2\pi\sigma^2)^{-\frac{3p}{2}} \exp -\frac{1}{2\sigma^2} d_P^2(X_M, Y_M) \prod_{y_i \in Y_{\bar{M}}} f(y_i; \lambda) \prod_{x_i \in X_{\bar{M}}} f(x_i; \lambda) \quad (1) \end{aligned}$$

where $\|X\|_F = \mathrm{tr}(X'X)$ is the Frobenious norm and $f(\cdot\,;\lambda)$ is a one-parameter density for unmatched landmarks with λ having an interpretation as a soft threshold for landmark deviations. This density corresponds to an additive model on shape space

$$[Y_M] = [X_M] + \epsilon \qquad\qquad \epsilon \sim N(\mathbf{0}, \sigma^2 I)$$

with ϵ_p a matrix-normal random perturbation; other forms for the error distribution can also be accommodated (Rodriguez and Schmidler, 2006a).

Strictly speaking the density (1) is defined on a tangent-space approximation to shape space, the difference being in the integral required for normalization (see Small (1996); Dryden and Mardia (1998) and Schmidler (2006)); in our case this normalization cannot be ignored.

An alternative approach is to place prior distributions on transformation parameters θ and obtain a posterior distribution over the alignment pair $A = [M, \theta]$, from which the marginal posterior for M may be obtained by integrating out the "nuisance" parameters θ (Schmidler, 2006; Wang and Schmidler, 2006); see also Green and Mardia (2006) for a related approach developed independently. In comparison, the likelihood defined directly on shape space (1) identifies transformations θ directly with matchings M, but can also be viewed as a profile likelihood for A. Schmidler (2006) considers both approaches for protein registration, including the marginalization approach under affine transformations where integration may be done analytically. The marginalization approach has the advantage of accounting for uncertainty in θ given M. In addition, in some applications of image analysis (e.g. tracking), θ is the parameter of interest while M may be the nuisance parameter. In this paper, we restrict attention to models of the form (1) for concreteness; the computational approach described in the next sections applies equally to both approaches.

3.1. *Priors on bipartite matchings and MCMC*

Priors on matchings $P(M)$ may be specified in a variety of ways. Rodriguez and Schmidler (2006a) use a gap-penalty prior:

$$P(M) \propto \exp -(g\, n_g(M) + h \sum_{i=1}^{n_g(M)} l_i(M)) \tag{2}$$

based on affine gap opening and extension penalties used in biological sequence analysis; this is essentially a Markovian process on M. The prior encourages grouping of matches together but requires pre-specification of a landmark ordering, equivalent to assuming a topological equivalence of two proteins.

The advantage of the class of priors (2) is that it enables exact sampling of $P(M \,|\, \theta)$ via an efficient dynamic programming algorithm, thus producing an efficient Gibbs sampling algorithm for iteratively drawing $P(M \,|\, \theta)$ and $P(\theta \,|\, M)$ (Wang and Schmidler, 2006) or efficient Metropolis proposals for the posterior obtained from (1) (Rodriguez and Schmidler, 2006a). However, this MCMC scheme suffers from the combination of facts that $P(M \,|\, \theta)$ may be multimodal and that M and θ are strongly coupled in the posterior $P(M, \theta \,|\, X, Y)$; mixing between multiple modes, when they exist, is therefore unacceptably slow. A solution to this is given by (Rodriguez and Schmidler, 2006a) which constructs a library of θ's to a

proposal distribution used during the Metropolis sampling; this approach is effective but somewhat computationally intensive and is less efficient for multiple shape alignment (Wang and Schmidler, 2006). The approach described in this paper is in a sense a generalization of this library approach, which may potentially replace MCMC sampling entirely.

In this paper we avoid the order-preserving assumptions of (2) and consider priors on unrestricted matchings. Thus the approach described here is applicable to general shape matching when such *a priori* assumptions are not acceptable. More general priors are also considered briefly in (Rodriguez and Schmidler, 2006a), where other MCMC moves are mixed with the dynamic programming steps. Further work is needed to explore the impact of informative priors on bipartite matchings.

3.2. *Posterior summaries*

The model given in the previous section allows for the fully Bayesian comparison of geometric objects. A key advantage of this approach is the ability to accounting for uncertainty in the point correspondences and accurately estimate variability in the resulting matchings, which in many cases may be multimodal. Quantities of interest under such a model include distance measures such as the posterior expectation of Procrustes distance or root-mean-square deviation:

$$E(\text{rmsd} \mid X, Y) = \sum_{M \in \mathcal{M}_{n,m}} |M|^{-\frac{1}{2}} \, d_P(X_M, Y_M) P(M \mid X, Y)$$

or the posterior probability of a match, which may be calculated e.g. as the marginal posterior probability $P(|M|^{-\frac{1}{2}} \, d_P \mid X, Y) \leq \delta$ for some δ or the posterior probability that the fraction of landmarks matched is greater than ϵ. Other important posterior summaries include the posterior mode $\hat{M} = \arg\max_M P(M \mid X, Y)$, or the *marginal alignment matrix* which gives the marginal posterior probability of matching any pair of landmarks:

$$P(m_{ij} \mid X, Y) = \sum_{M \in \mathcal{M}_{n,m}} m_{ij} P(M \mid X, Y) \tag{3}$$

integrating out all other parameters in the model. The space $\mathcal{M}$ grows exponentially in n and m, making exact calculation of these quantities infeasible. Under the gap-penalty based prior, Rodriguez and Schmidler (2006a); Wang and Schmidler (2006) provide Markov chain Monte Carlo (MCMC) sampling algorithms for approximation of such posterior quantities, which take advantage of efficient Gibbs steps involving stochastic dynamic programming recursions. However, the combinatorial nature of the space $\mathcal{M}$ and the existence of multiple posterior modes makes MCMC slow. Thus for applications involving large numbers of shape comparisons such as searching molecular structure or image databases for shapes similar to some query shape, the Bayesian approach remains infeasible. Below we explore deterministic computational approximations based on geometric algorithms to address this issue.

4. BAYESIAN SHAPE CLASSIFICATION

In this paper we also consider the problem of shape *classification*, using a simple Bayesian classifier. Wang and Schmidler (2006) describe Bayesian estimation of

mean shape under the Bayesian framework above, applied to multiple protein structure alignment and analysis of functional conservation. Given a mean shape Z_c for each class $c \in \mathcal{C}$ of interest, we may classify based on the posterior distribution

$$P(Y \in c \,|\, Y) = \frac{P(Y \,|\, Z_c)P(c)}{\sum_{c' \in \mathcal{C}} P(Y \,|\, Z_{c'})P(c')} = \frac{\sum_{M \in \mathcal{M}} P(Y \,|\, Z_c, M)P(M)P(c)}{\sum_{c' \in \mathcal{C}} \sum_{M' \in \mathcal{M}} P(Y \,|\, Z_{c'}, M')P(M)P(c')}$$

This requires calculation of the marginal likelihood $P(Y, Z_c)$ which, as with the posterior summaries above, involves a sum over the exponential space of all possible alignments. More generally, suppose that for each class c of interest we have k_c example shapes $X_1^c, \ldots, X_{k_c}^c$ from class c, and obtain a posterior distribution over Z_c and associated parameters ζ_c represented by a finite set of samples $(Z_c^{(i)}, \zeta_c^{(i)})$ from $P(Z_c, \zeta_c \,|\, X_1^c, \ldots, X_{k_c}^c)$. Denote the set of examples by $\mathbf{X} = \cup_{c \in \mathcal{C}} \{X_1^c, \ldots, X_{k_c}^c\}$. Then we may approximate the posterior classification probability of new observation Y by

$$\begin{aligned} P(Y \in c \,|\, Y, \mathbf{X}) &= \frac{\int\int P(Y \,|\, Z_c, \zeta_c) P(Z_c, \zeta_c \,|\, X_1^c, \ldots, X_{k_1}^c)\, dZ_c\, d\zeta_c P(c)}{\sum_{c' \in \mathcal{C}} \int\int P(Y \,|\, Z_c, \zeta_{c'}) P(Z_c, \zeta_{c'} \,|\, X_1^{c'}, \ldots, X_{k_{c'}}^{c'})\, dZ_{c'}\, d\zeta_{c'} P(c')} \\ &\approx \frac{\sum_i \sum_M P(Y \,|\, Z_c^{(i)}, \zeta_c^{(i)}) P(Z_c^{(i)}, \zeta_c^{(i)} \,|\, X_1^c, \ldots, X_{k_1}^c) P(c)}{\sum_{c' \in \mathcal{C}} \sum_i \sum_M P(Y \,|\, Z_c^{(i)}, \zeta_{c'}^{(i)}) P(Z_c^{(i)}, \zeta_{c'}^{(i)} \,|\, X_1^{c'}, \ldots, X_{k_{c'}}^{c'}) P(c')} \end{aligned}$$

In each case we require calculation of the marginal likelihood $P(Y \,|\, Z_c, \zeta_c)$. In fact this is a closely related problem to that of Bayesian shape matching for database search described above, as can be seen by considering the collected posterior samples $\{(Z_c^{(i)})\}_{c \in \mathcal{C}}$ as a database of shapes. Thus the shared problem is that of marginalizing over the combinatorial space of matchings for a large set of target shapes simultaneously.

5. APPROXIMATE POSTERIORS VIA GEOMETRIC HASHING

As described, computation of posterior quantities by MCMC methods suffers from several drawbacks, including the requirement of order-preserving matchings for efficient Gibbs steps, the difficulty of mixing between multiple posterior modes in the combinatorial space of bipartite matchings, and general infeasibility of Monte Carlo methods for calculations which must be repeated thousands or millions of times to search large databases. In this paper we explore the adaptation of an algorithm from the image processing literature, *geometric hashing*, to compute approximate posterior quantities much more efficiently. Our goal is to approximate the marginal likelihood

$$\begin{aligned} P(Y \,|\, Z_c, \zeta_c) &= \sum_{M \in \mathcal{M}} P(Y \,|\, M, Z_c, \zeta_c) P(M) \\ &\approx \sum_{M \in \mathcal{M}^*} P(Y \,|\, M, Z_c, \zeta_c) P(M) \end{aligned}$$

by consideration of some high posterior density set $\mathcal{M}^*$ which can be calculated efficiently.

5.1. *Geometric hashing*

Geometric hashing is an algorithmic technique developed in the computer vision literature for object recognition and image analysis to rapidly match scenes to a database of models (Wolfson and Rigoutsos, 1997). It has also been applied effectively to alignment and substructure analysis of protein molecules (Nussinov and Wolfson, 1991; Wallace et al., 1996). A key advantage of the approach is the ability to match against a library of models simultaneously in polynomial time. There are several variations on the algorithm; here we describe and implement one of the simplest although not necessarily the most computationally efficient. Here we briefly introduce the algorithm; see above references for more details. In the next section we show how this algorithm may be adapted to perform the Bayesian shape calculations of Section 3 very efficiently.

Geometric hashing begins by representing each object in the database in a hashtable, to which search objects are then compared. All objects are represented as follows: choose three non-collinear landmarks from X denoted by x_a, x_b, and x_c, and define the unique coordinate frame having $s = x_a, x_b, x_c$ lying in the xy-plane with x_a at the origin and x_b on the positive x-axis, and z-axis given by the right-hand rule. Denote by e_1^s, e_2^s, e_3^s the associated orthogonal basis. All other points $X_{[4:n]}$ are represented in this coordinate frame as $x_i - x_0^s = a_i^s e_1^s + b_i^s e_2^s + c_i^s e_3^s$ where x_0^s is the chosen origin; note that the resulting coordinates (a_i^s, b_i^s, c_i^s) are invariant under $SE(3)$. These coordinates are then used to index a location in a table, where the coordinates are stored along with a pointer to the reference set s. This process is then repeated for (a) every ordered subset s of landmark triplets in X, and (b) for every X in the database. This indexing takes $O(n^4)$ processing time per database object, but can be precomputed once offline and then used repeatedly.

To match a newly observed object Y to the database, a landmark triple s is chosen (perhaps randomly) and used to compute coordinates (a_i^s, b_i^s, c_i^s) for the remaining landmarks. These coordinates again serve as indices into the table, where the associated entry contains a list of (point, reference set, object) triplets for matching points in database objects. Each such element of the list is then assigned a "vote" for the associated (object,reference set) pair. The database (object, reference set) pair receiving the most such votes is considered the best match. This voting can be given a probabilistic semantics and the resulting best match considered to be a maximum likelihood match among objects in the database (Wolfson and Rigoutsos, 1997). The key to the approach is that comparison of a new object Y to an entire database may be done rapidly in order $O(nc^*)$ where c^* is a constant giving the average entry list length, related to the density of points.

6. FULLY BAYESIAN SHAPE MATCHING BY GEOMETRIC HASHING

The geometric hashing technique allows rapid large-scale parallel comparison of object configurations against a database. In the paper we explore the use of this technique for approximation of marginal posterior quantities under the models described in Sections 3.2 and 4.

We consider approximation of marginal posterior quantities such as (3) or 95% credible sets:

$$C_{95} = \{M \in \mathcal{M} : P(M \mid X, Y) \geq c_\alpha\} \tag{4}$$

where $c_\alpha = \min$ s.t. $P(C_{95} \mid X, Y) \geq .95$. As mentioned, the original geometric hashing algorithm may be viewed as an approximation of the MAP estimate

$\arg\max_{M\in\mathcal{M}} P(M \mid X, Y)$ under a uniform prior; however, if the posterior $P(M \mid X, Y)$ is multimodal, the MAP estimate may be a poor summary.

6.1. *Priors*

We consider two simple classes of priors. The first are uniform priors on matchings $P(M) = |\mathcal{M}|^{-1}$, where

$$|\mathcal{M}| \;=\; \sum_{k=0}^{\min(n_X, n_Y)} \binom{n_X}{k} \frac{n_Y!}{(n_Y - k)!}$$

The second are *exchangeable* priors:

$$P(M) = f(|M|) = g(k)$$

which depend only on the number of matches but not the precise pattern.

6.2. *Likelihood bound*

In order to approximate quantities such as high posterior density credible sets (4) and marginal posterior probabilities (3) to within a given accuracy, we must be able to identify a subset $\mathcal{M}^* \subset \mathcal{M}$ of elements with high posterior probability. Suppose we wish to find all alignments M such that

$$Z(X,Y)^{-1} g(k) (2\pi\sigma^2)^{-\frac{3k}{2}} \mathrm{e}^{-\frac{1}{2\sigma^2} d_P^2(X_M, Y_M)} P(y_i \in Y_{\bar{M}}, x_i \in X_{\bar{M}}) \geq p^*$$

Where $Z(X, Y)$ denotes the normalizing constant or marginal likelihood. For a match of size $|M| = k$ and conditional on σ^2, we then require

$$d_P(X_M, Y_M) \leq \sqrt{2}\sigma \left[\log\left(\frac{p^*}{g(k)\lambda^{(n+m-2k)} Z(X,Y)^{-1}}\right) + \frac{3k}{2}\log(2\pi\sigma^2) \right]^{\frac{1}{2}} = d_k^*$$

Thus we may construct such a set by

$$\mathcal{M}^* = \bigcup_k \{M \in \mathcal{M} :\; |M| = k,\;\; d_P(X_M, Y_M) \leq d_k^*\}$$

We may adapt the geometric hashing algorithm of Section 5.1 to approximate this set very quickly as described in the following section.

Identification such a set $\mathcal{M}^*$ would allow us to obtain theoretical guarantees on the accuracy of our approximation of the posterior. For example, we would like to guarantee that

$$1 - \alpha \leq \sum_{M\in\mathcal{M}^*} P(M \mid X, Y) = \sum_{M\in\mathcal{M}^*} \frac{L(X, Y \mid M) P(M)}{\sum_{M'} L(X, Y \mid M') P(M')}$$

However, in practice we do not know the marginal likelihood $Z(X, Y)$ needed to compute d_k^*. In order to bound the contributions of individual alignments to the posterior we therefore require a bound on $Z(X, Y)$ as well, so that

$$\sum_{M\in C_\alpha} L(X, Y \mid M) P(M) \geq \beta \quad \text{and} \quad \sum_{M\notin C_\alpha} L(X, Y \mid M') P(M') \leq \gamma$$

where $\beta\gamma^{-1} \geq \alpha$. If we can find $\mathcal{M}^*$ such that it contains all $M \in \mathcal{M}$ with $P(M)L(X, Y \mid M) \geq \delta^*$, a weak bound on $Z(X, Y)$ may be obtained from

$$P(X, Y) \leq \sum_{M \in \mathcal{M}^*} P(X, Y \mid M)P(M) + (|M| - |M^*|)\delta^*$$

It is as yet unclear whether this can lead to a practical algorithm with guaranteed bounds. Nevertheless, the above argument shows that there exists such a d_k^* which will give accurate approximations. Even if we cannot choose d_k^* by theoretical arguments with guarantees, we may be able to obtain accurate approximations with good empirical performance by experimentation; we explore this below.

6.2.1. *Bounding Procrustes distance based on geometric hashing*
The above argument shows that there exists a sequence $d_k^*, k = 0, \ldots, \min(n, m)$ such that the set $\mathcal{M}^* = \{M \in \mathcal{M} : d_P(X_M, Y_M) \leq d_{|M|}^*\}$ yields a high posterior credible set and good approximations to posterior quantities. Even if we cannot determine such a d^* directly to obtain theoretical guarantees, this suggests that choosing a d^* large enough may give good empirical performance. Denote by

$$d_{GH(a,b,c)}^2(X, Y) = \|X - Y\hat{R}_{GH(a,b,c)}\|_F^2 = \sum_{i=2}^{p} \|x_i - \hat{y}_j\|^2$$

the sum-of-squared Euclidean distances of matched landmarks under the geometric hashing transform described above using reference set (a, b, c). For all $(a, b, c) \subset X$ we have

$$d_P(X, Y) \leq d_{GH(a,b,c)}(X, Y)$$

Thus a sufficient condition for $d_P(X_M, Y_M) \leq d_k^*$ for $|M| = k$ is that $\max \|x_i - \hat{y}_j\| \leq d_k^*/\sqrt{k}$ where $\hat{y} = y\hat{R}_{GH(a,b,c)}$. Thus we need only find all matchings with maximum distance between matched points less than $d^* = \max_k d_k^*/\sqrt{k}$. (Alternatively we may build n different hashtables for matches of size $k = 1, \ldots, n$, but we have not implemented this here.)

To obtain all such matching points for a given reference set, we need simply adapt the geometric hashing algorithm as follows: choose the resolution of the hashtable indexing to be d^*. Now when matching an object to the database, each point used as an index into the table at entry (a, b, c) say, must also check the 26 neighboring entries in the surrounding cube: $\{(a-1, b, c), (a, b-1, c), (a, b, c-1), (a-1, b-1, c), \ldots)\}$. In this way every point within d^* of the indexed point may be identified with only a constant factor increase in computational time.

We further adapt the geometric hashing algorithm as follows: we use all $\binom{n}{3}$ unordered triplets of the query object. For each reference set we find all landmark matches within d^*, to construct the maximal matching. Using the corresponding matching we then compute the full least-squares Procrustes distance to obtain the associated likelihood. The resulting matching algorithm has increased computational complexity of $O(n^4)$.

Note that there are multiple approximations being made here. First, the approximation to the posterior by a high density credible set. Second the set itself is approximated: since the above condition is sufficient but not necessary, the hashing algorithm is not guaranteed to obtain all matches with $d_P \leq d^*$. Finally, each match

$M \in \mathcal{M}^*$ is maximal, effectively giving a mode approximation for each region of high posterior probability matches.

Nevertheless, this approach has strong advantages over MCMC-based matching, inheriting the strengths of the original geometric hashing algorithm. First, it is deterministic and non-iterative, making it suitable for rapid, large scale repeated search. Second, it computes the marginal likelihood for the Bayesian match of the query object against an entire set of target objects simultaneously rather than sequentially. Third and related, it is inherently parallelizable. Thus one could imagine a distributed database server which performs shape queries in a fully Bayesian fashion rather than by a standard heuristic optimization criteria.

6.3. *Alternative priors on* $\mathcal{M}$

It is unclear how general $P(M)$ may be and still be amenable to this rapid hashing approach; this merits further exploration. Priors of the form (2) seem difficult to incorporate into the algorithm, but other types of spatial process priors may prove tractable. In addition, the approximating set $\mathcal{M}^*$ obtained under uniform prior as above may be used to obtain approximations to other priors via importance reweighting:

$$P(Y \mid Z_c, \zeta_c) \approx \sum_{M \in \mathcal{M}^*} P(Y \mid M, Z_c, \zeta_c) w(M) / \sum_{M' \in \mathcal{M}^*} w(M')$$

where $w(M) = P(M)/g(|M|)$. Bounds for this case seem to more difficult and rely on the maximum and minimum of $w(M)$ over $M \in \mathcal{M}$.

7. EXAMPLES

We apply this approach to the problem of pairwise protein structure alignment developed previously (Rodriguez and Schmidler, 2006a). Figure 1 shows the results obtained from applying the geometric hashing technique to approximate marginal posteriors as described in Section 6 to match two short protein fragments taken from the N-terminal Helix A region of human deoxyhemoglobin β-chain (4hhb_A) against sperm whale myoglobin (5mbn). The figure shows the marginal posterior match matrices (3) obtained from the hashing approximation, compared with an exact calculation obtained by exhaustive enumeration, which is just feasible for this short $n = 7$ problem.

Figure 3 shows the approximate marginal match matrix obtained from a larger problem which is chosen to exhibit multimodality, where MCMC performs poorly. An N-terminal fragment of the Helix A region of human deoxyhemoglobin β-chain (4hhb_A) is matched against a stretch of 30 N-terminal residues from sperm whale myoglobin spanning both helices A and B. There are expected to be at least two reasonable matches, to each of the two helices, separated by a break for the loop in between. We see this from the hashing-based posterior, as well as lower posterior modes obtained from reversing the N- to C-terminal orientation of the query helix. Exact calculation for this example is infeasible.

8. CONCLUSIONS AND FURTHER WORK

We have described an approximate calculation technique for Bayesian shape matching under the framework described by (Rodriguez and Schmidler, 2006a) using an extension of the geometric hashing technique developed by (Nussinov and Wolfson,

1991). This technique has the advantage of being highly computationally efficient and inherently parallel, allowing particularly efficient simultaneous calculation when shape matching against a large set of shapes is required. As described, this addresses the fundamental computational problem of Bayesian shape matching of a query object against a large database of possible targets, or for calculation of key quantities in Bayesian shape classification when classes are described by mean shapes or posterior distributions summarized by Monte Carlo samples.

This basic approach has been demonstrated for test examples in the alignment of protein structures. Further experimentation and optimization of the implementation is needed to explore the large scale applications which motivate this approach; the algorithm is inherently parallelizable and development of a distributed database server which performs shape queries in a fully Bayesian fashion may be feasible.

There are a number of directions for further work along these lines. It is an interesting question whether more sophisticated bounds for the posterior quantities may be obtained to provide theoretical guarantees on the accuracy of approximation. A simple approach is suggested here but more work is needed to determine if this can be useful in practice. However it is worth pointing out that the MCMC approach, which is also in standard usage for an enormous variety of Bayesian computational problems, also comes with few guarantees in practical problems where the mixing time of the Markov chain cannot be bounded polynomially.

It may be fruitful to view the use of maximal matchings to construct the set $\mathcal{M}^*$ as a finite space analog to standard integral approximation algorithms in continuous spaces based on normal approximations at posterior modes (Tierney et al., 1989; Tanner, 1993). In this sense, the "width" of a mode is clearly missing and it may be helpful to construct a multiplicative factor for the posterior contribution of elements in $\mathcal{M}^*$, e.g. based on the number of subsets of size $k' < k$ for a maximal matching of size k.

The additive error model used here is particularly simple and does not account for spatial covariation between landmark deviations, for example. Addressing this while preserving the efficiency of the hashing calculation requires some care, but increasing d^* and computing the Procrustes distance using appropriate Mahalanobis inner product may be feasible and warrants investigation. Alternatively, as with the discussion of non-factorable priors in Section 6.1, importance reweighting of $\mathcal{M}^*$ may be effective. It remains unclear how practically significant such covariance structure may be for shape matching and classification, and this is likely to be problem specific. It is worth noting that the hashing approach of standardizing the coordinate frame to a single triplet seems reminiscent of Bookstein coordinates (see e.g. Bookstein (1986); Dryden and Mardia (1998)), which is known to induce spurious covariance on the other landmarks. It would be interesting to explore how this affect manifests in the geometric hashing case presented here.

A major advantage of the hashing technique over MCMC sampling is the ability to quickly generate multiple modes which may be well-separated and difficult for a Markov chain algorithm to cross between. One possibility for improving on the hashing approximation described here is to use it to develop a fast approximation which is then refined by MCMC sampling, using the identified modes to generate move proposals in the Markov chain. This has the combined advantage of allowing the MCMC to mix between modes, and of refining the hashing answer in a way which will converge with the usual theoretical guarantees. This is a more powerful generalization of the library-sampling extension to MCMC matching given in Rodriguez and Schmidler (2006a); Wang and Schmidler (2006) without requiring the

order-preserving constraints assumed there. This approach would also allow the incorporation of more general covariance structure and more informative priors. While very promising for improving the quality of the approximation, this approach suffers from the fact that any MCMC technique will be significantly slower than the rapid and simultaneous hashing technique described in this paper, and thus infeasible for large-scale comparison against large databases. More experience with this approach is needed in order to understand the impact of trading off these computational and modeling alternatives.

REFERENCES

Bookstein, F. L. (1986). Size and shape spaces for landmark data in two dimensions. *Stat. Sci.*, 1:181–242.

Bookstein, F. L. (1991). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press.

Dryden, I. L. (1999). *General Shape and Registration Analysis*, volume 80 of *Monographs on Statistics and Applied Probability*, pages 333–364. Chapman & Hall.

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

Green, P. J. and Mardia, K. (2006). Bayesian alignment using hierarchical models, with applications in protein bioinformatics. (*preprint*).

Kendall, D. G., Barden, D., Carne, T. K., and Le, H. (1999). *Shape and Shape Theory*. Wiley.

Lele, S. R. and Richtsmeier, J. T. (2001). *An Invariant Approach to Statistical Analysis of Shapes*. Chapman & Hall.

Nussinov, R. and Wolfson, H. J. (1991). Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc. Natl. Acad. Sci. USA*, 88:10495–10499.

Rodriguez, A. and Schmidler, S. C. (2006a). Bayesian protein structure alignment. (in preparation for *J. Comp. Biol.*).

Rodriguez, A. and Schmidler, S. C. (2006b). Combining sequence and structure information in protein alignments. (submitted to *Intelligent Systems in Molecular Biology 2006*).

Schmidler, S. C. (2006). Bayesian flexible shape matching with applications to structural bioinformatics. (submitted to *Journal of the American Statistical Association*).

Small, C. G. (1996). *The Statistical Theory of Shape*. Springer.

Tanner, M. A. (1993). *Tools for Statistical Inference*. Springer.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc.*, 84(407):710–716.

Wallace, A., Laskowsi, R., and Thornton, J. (1996). Tess: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases: Applications to enzyme active sites. *Prot. Sci.*, 6:2308–2323.

Wang, R. and Schmidler, S. C. (2006). Bayesian multiple protein structure alignment and analysis of protein families. (in preparation).

Wolfson, H. J. and Rigoutsos, I. (1997). Geometric hashing: An overview. *IEEE Comp. Sci. & Eng.*, 4(4):10–21.

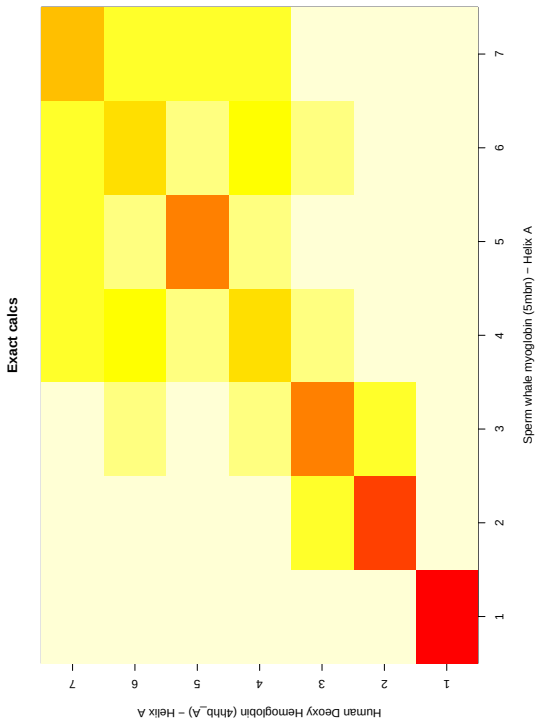

Figure 1: *Marginal posterior match probability matrix obtained for short 7-residue N-terminal fragment including portion of Helix A for human deoxyhemoglobin β-chain (4hhb_A) against sperm whale myoglobin (5mbn). Shown are results from (a) exact calculation, and (b) geometric hashing-based approximation as described in text.*

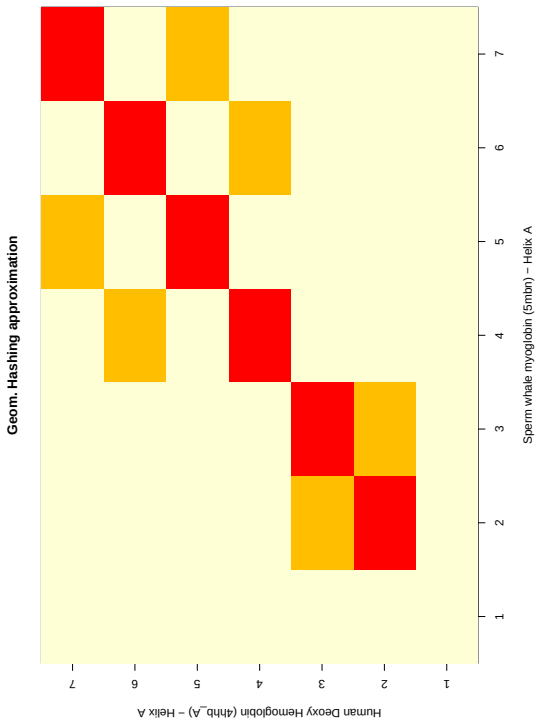

Figure 2: *Marginal posterior match probability matrix obtained for short 7-residue N-terminal fragment including portion of Helix A for human deoxy-hemoglobin β-chain (4hhb_A) against sperm whale myoglobin (5mbn). Shown are results from (a) exact calculation, and (b) geometric hashing-based approximation as described in text.*

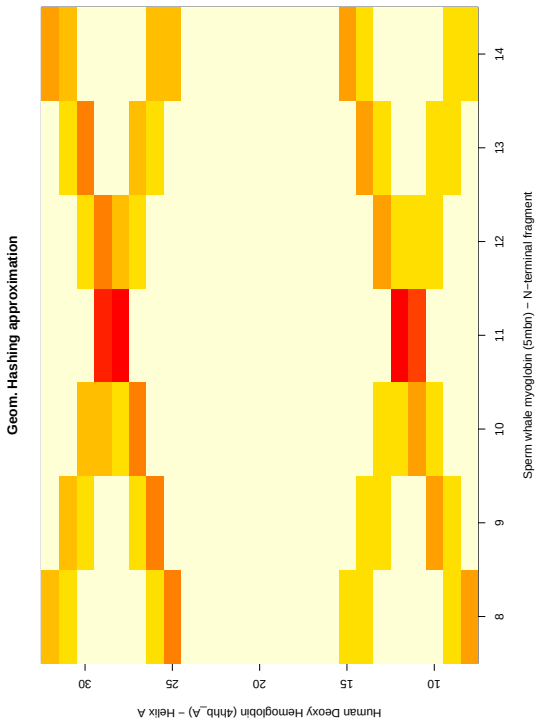

Figure 3: *Application of Bayesian matching algorithm using geometric hashing to larger example exhibiting multimodality: comparison of N-terminal helical fragment of human deoxyhemoglobin β-chain (4hhb_A) against residues 8-32 of sperm whale myoglobin (5mbn) which include helices A and B.*