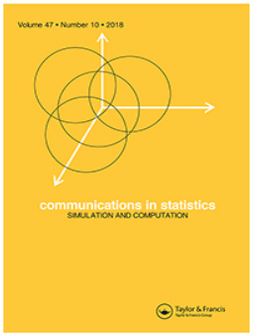


Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Parallel tempering strategies for model-based landmark detection on shapes

Justin Strait, Oksana Chkrebtii & Sebastian Kurtek

To cite this article: Justin Strait, Oksana Chkrebtii & Sebastian Kurtek (2019): Parallel tempering strategies for model-based landmark detection on shapes, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2019.1670843

To link to this article: https://doi.org/10.1080/03610918.2019.1670843

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2019.1670843

Parallel tempering strategies for model-based landmark detection on shapes

Justin Strait[a], Oksana Chkrebtii[b], and Sebastian Kurtek[b]

[a]Department of Statistics, University of Georgia, Athens, Georgia, USA; [b]Department of Statistics, The Ohio State University, Columbus, Ohio, USA

ABSTRACT

In the field of shape analysis, landmarks are defined as a low-dimensional, representative set of important features of an object's shape that can be used to identify regions of interest along its outline. An important problem is to infer the number and arrangement of landmarks, given a set of shapes drawn from a population. One proposed approach defines a posterior distribution over landmark locations by associating each landmark configuration with a linear reconstruction of the shape. In practice, sampling from the resulting posterior density is challenging using standard Markov chain Monte Carlo (MCMC) methods because multiple configurations of landmarks can describe a complex shape similarly well, manifesting in a multi-modal posterior with well-separated modes. Standard MCMC methods traverse multi-modal posteriors poorly and, even when multiple modes are identified, the relative amount of time spent in each one can be misleading. We apply new advances in the parallel tempering literature to the problem of landmark detection, providing guidance on implementation generalized to other applications within shape analysis. Proposal adaptation is used during burn-in to ensure efficient traversal of the parameter space while maintaining computational efficiency. We demonstrate this algorithm on simulated data and common shapes obtained from computer vision scenes.

ARTICLE HISTORY
Received 19 April 2019
Accepted 17 September 2019

KEYWORDS
shape analysis; parallel tempering; landmarks; Markov chain Monte Carlo; elastic metric

1. Introduction

A challenging problem in shape analysis is the automatic identification of important shape features (known as *landmarks*). This is motivated by applications in medical imaging: doctors often use expertise to mark (or annotate) anatomically-relevant points on shapes extracted from magnetic resonance image (MRI) scans, for instance. With a large number of patients, this process is cumbersome and error-prone. Previous work has focused on landmark inference either for pre-specified images (e.g., X-rays in Chen et al. (2014)) or shape classes (e.g., human faces in Tie and Guan 2013; Segundo et al. 2010; Gilani, Shafait, and Mian 2015). The methods proposed by these authors are deterministic, and do not readily quantify uncertainty in the obtained estimates. Some work has attempted to generalize this problem to arbitrary shape classes (Domijan and

CONTACT Justin Strait jds49205@uga.edu 310 Herty Dr., University of Georgia, Athens, GA 30602.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.
Supplemental data for this article can be accessed here.

Wilson 2005; Strait, Chkrebtii, and Kurtek 2018). The former takes an image analysis perspective, inferring landmarks on images using a hierarchical model. The latter also takes a Bayesian approach, but focuses on curve-level data (rather than images). A general issue with the model proposed in Strait, Chkrebtii, and Kurtek (2018) is the difficulty of obtaining reliable posterior estimates for landmarks. There are two reasons for this. First, shape data is typically quite high-dimensional, resulting in a model likelihood which is highly-peaked. Second, landmark posteriors can be multi-modal in some cases, i.e., there may exist several sets of landmarks which can be deemed "important." Traditional Markov chain Monte Carlo samplers often get trapped in local modes, resulting in biased estimates of shape landmarks.

In this paper, we discuss modern strategies for improved sampling from the posterior distribution specified in Strait, Chkrebtii, and Kurtek (2018), allowing for multi-modalities to be efficiently explored. We should note that while the main motivation for this work is inference for the landmark detection model, we believe that the discussed sampler is relevant to numerous other problems which rely on infinite-dimensional models. While infinite-dimensional models do not have tractable normalizing constants, finite-dimensional approximations of this model are less restrictive than their counterparts which introduce dimension-reduction at an early stage. For instance, a hierarchical model for estimating the registration function in the functional data setting was explored by Cheng, Dryden, and Huang (2016) under a similar Gaussian process model to Strait, Chkrebtii, and Kurtek (2018), which is reliant on a squared-distance term in the likelihood – in fact, the authors of this paper discuss a simulated tempering algorithm to handle posterior multi-modality. Other problems in shape analysis where multi-modalities may be present include model-based image segmentation under an elastic shape prior (Joshi and Srivastava 2009; Bryner, Srivastava, and Huynh 2013), imputation of missing segments for functions and planar shapes (He, Yucel, and Raghunathan 2011), estimation of landmark-constrained registration functions (Strait et al. 2017), among others. Consequently, we have tried to keep the actual sampler discussions as general as possible. The improved sampler is based on work by Lacki and Miasojedow (2016); Miasojedow, Moulines, and Vihola (2013), which has generally been applied to low-dimensional models. Our work applies their adaptive procedure to the high-dimensional data setting (e.g., shapes), and we discuss complications which may arise under its implementation.

The rest of this paper is organized as follows. Section 2 provides a brief overview of elastic shape analysis; the model proposed in Strait, Chkrebtii, and Kurtek (2018) is summarized in Sec. 3, along with the sampling method used. Section 4 describes an adaptive parallel tempering sampler (in the spirit of Miasojedow, Moulines, and Vihola 2013; Lacki and Miasojedow 2016), which improves on the existing sampling procedure. Finally, we conclude with results and closing remarks. A supplementary materials document is available online – this includes supporting plots that were not included within the main text.

2. Statistical shape analysis

Kendall (1984) defined shape as a property that remains after certain shape-preserving transformations (rigid motion and scaling) are filtered out from an object. Early work

in shape analysis used a finite-dimensional set of important points, known as *landmarks* (Kendall 1984; Good 1994; Dryden and Mardia 2016), to represent shape. While numerous statistical tools have been developed using landmark-based representations, these require the researcher to manually specify appropriate landmark locations – a difficult task in itself, especially for large datasets. In addition, these landmarks must be selected to be in correspondence across the shape population, so that they represent the same feature on each individual shape. For instance, if the researcher selects the first landmark at the tip of the ring finger on the contour of a human hand, then the first landmark should be selected at the tip of the ring finger on all other human hand shapes being analyzed. Thus, landmark selection is quite tedious and subject to human error. A more modern approach, called *elastic* shape analysis, is motivated by treating shape as an infinite-dimensional object. Elastic methods allow one to be landmark-free in a sense, by reducing the dependence of statistical analyses on the pre-specified landmarks. A benefit of using elastic representations is that the optimal correspondence of points is the solution to a minimization problem, ensuring that prominent features are matched appropriately. Recently, Strait et al. (2017) combined the elastic framework with landmarks, allowing for prior knowledge to be introduced into the solution for optimally registering shapes. In this section, we present a summary of necessary topics from elastic shape analysis. For further discussion, consult Srivastava et al. (2011), Kurtek et al. (2012), and Srivastava and Klassen (2016).

2.1. The square-root velocity function

Assume $\beta : \mathcal{D} \to \mathbb{R}^2$ is an absolutely continuous, planar curve defining the contour of an object. The curve domain $\mathcal{D}$ is assumed to be $[0, 1]$ for open curves, and $\mathbb{S}^1$ for closed curves (where there is no well-defined start or end point). The object which underlies elastic shape analysis is the *square root velocity function (SRVF)*, defined as $q(t) = \frac{\dot{\beta}(t)}{\sqrt{|\dot{\beta}(t)|}}$, where $\dot{\beta}$ is the time-derivative of β, $|\cdot|$ is the Euclidean norm, and $q(t) := 0$ at points where $\dot{\beta}$ vanishes or is non-differentiable. The SRVF has numerous benefits. First, it encodes the instantaneous velocity of β. In addition, the original curve function β can be recovered from q via $\beta(t) = \beta(0) + \int_0^t q(s)|q(s)|ds$. These two points show that use of SRVF results in no loss of information about the original curve, while also being automatically invariant to translation of curves (a shape-preserving transformation): if $\beta_2 = \beta_1 + c$, then the corresponding SRVFs are equal, i.e., $q_1 = q_2$. The next section describes the advantages of using an elastic approach in the context of shape analysis.

2.2. Elastic metric

The greatest benefit of the SRVF pertains to the choice of an appropriate shape metric. It has been shown (Joshi et al. 2007; Srivastava et al. 2011) that the $\mathbb{L}^2$ metric between two curves β_1 and β_2 depends on the parameterization function $\gamma \in \Gamma$, i.e., $||\beta_1 - \beta_2|| \neq ||\beta_1 \circ \gamma - \beta_2 \circ \gamma||$, where $||\cdot||$ denotes the $\mathbb{L}^2$ function norm, and $\Gamma = \{\gamma : \mathcal{D} \to \mathcal{D} | \gamma$ is an orientation$-$preserving diffeomorphism$\}$ is the group of re-parameterization functions. This property is not desirable; ultimately, we want to remove variation

associated with shape-preserving transformations. When a curve is represented by a function, shape should also be invariant to any re-parameterization of the function. The parameterization γ controls the rate at which an individual curve is traversed. In the context of a pair of curves, parameterizations dictate the correspondence of points between them. Simultaneous re-parameterization of β_1 and β_2 by the same γ does not change the pointwise correspondence, and thus, the distance between the two shapes should also remain the same – the $\mathbb{L}^2$ metric between β_1 and β_2 does not satisfy this property.

However, the $\mathbb{L}^2$ distance between the SRVFs q_1 and q_2 corresponding to β_1 and β_2 *is* preserved under re-parameterization. Furthermore, a key result states that the $\mathbb{L}^2$ distance between SRVFs is equivalent to an *elastic metric* between the original curve functions β_1 and β_2 (see Srivastava et al. (2011) for its explicit form). The elastic metric is a Riemannian metric which quantifies the amount of bending and stretching necessary to deform β_1 into β_2. Use of this Riemannian metric can be computationally demanding for methods which rely on repeated metric calculation. However, Joshi et al. (2007) mention that by converting β_1 and β_2 to SRVFs q_1 and q_2, the elastic metric simplifies to the $\mathbb{L}^2$ metric. Thus, we can use the relationship between β and q to compute SRVFs, use the $\mathbb{L}^2$ metric with SRVFs, and map back to the original curve function for visualization. We state the simplified form of the elastic metric under the SRVF transformation below:

$$d_{\text{Elastic}}(\beta_1, \beta_2) = ||q_1 - q_2|| = \sqrt{\int_{\mathcal{D}} |q_1(t) - q_2(t)|^2 dt}. \tag{1}$$

This metric also plays an important role in solving the *registration problem*, where the goal is to find the optimal correspondence of points between two curves. This is one of the strengths of elastic shape analysis – simply using landmarks automatically imposes a correspondence which may potentially be sub-optimal. The registration problem is detailed extensively in Srivastava and Klassen (2016), but is not of primary concern here.

3. Landmark detection model

We now return to the original problem at hand: model-based inference of landmark locations on a collection of shapes. Two models are proposed in Strait, Chkrebtii, and Kurtek (2018): the first assumes the number of landmarks is known (and fixed), which is then extended to a second model, where the number of landmarks is allowed to vary. Our focus is on landmark location inference under the first model, i.e., when the number of landmarks is assumed to be fixed. We will also assume that the population of shapes is homogeneous, meaning shapes are already registered to each other. In practice, this means that an arc-length parameterization is sufficient for all curves.

Formally, let $\beta_1, ..., \beta_M : \mathcal{D} \to \mathbb{R}^2$ be a sample of curve outlines from a homogeneous shape population. The goal is to infer k landmark locations $\boldsymbol{\theta} = (\theta_1, ..., \theta_k) \in \mathcal{D}^k$, subject to the constraint $\theta_1 < ... < \theta_k$ (ensuring landmarks are ordered appropriately with respect to curve parameterizations). With the assumption that curve registration is not

necessary, only variability associated with translation and scale is to be removed. Translation-invariance is achieved through transformation of curves to their SRVFs, denoted $q_{\beta_1}, ..., q_{\beta_M}$. Scale is removed by pre-processing curves to be of unit length, preventing shapes with larger size from dominating inference. In practice, the previously mentioned functions must be discretized. The issue of discretization is discussed extensively in Strait, Chkrebtii, and Kurtek (2018), and as such, we will assume all curves are sampled to N points, where N reflects the resolution of the data. The discretization of a function f to N points is denoted $f^{(N)}$.

3.1. Evaluation of landmark configurations

In order to infer landmarks based on a statistical model, a discrepancy between the model and the data must be defined through the likelihood function. This is equivalent to evaluating landmark arrangements over the sample of shapes to reflect the two important goals of landmark selection: reconstruction and low-dimensional representation. Strait, Chkrebtii, and Kurtek (2018) base the likelihood on the linear reconstruction error as follows. Let $\boldsymbol{\theta}$ be a candidate landmark configuration for the curve β_m with SRVF q_{β_m}. Consider construction of the linearization of β_m, with knot points at $\beta_m(\boldsymbol{\theta})$ (i.e., the curve landmark locations); call this function $L_m(t;\boldsymbol{\theta})$. This piecewise-linear curve is formed simply by joining straight lines between landmarks – i.e., the segment of L_m connecting the points $\beta_m(\theta_i)$ and $\beta_m(\theta_{i+1})$ for $i = 1, ..., k-1$ is given by $L_m(t;\boldsymbol{\theta}) = \left(1-\frac{t-\theta_i}{\theta_{i+1}-\theta_i}\right)\beta_m(\theta_i) + \left(\frac{t-\theta_i}{\theta_{i+1}-\theta_i}\right)\beta_m(\theta_{i+1})$, for $\theta_i \leq t < \theta_{i+1}$. We suppress the t argument for the remainder of the paper, thus referencing this curve by $L_m(\boldsymbol{\theta})$. For open curves, the starting point is additionally connected to the first landmark, and the last landmark is connected to the ending point. For closed curves, the first and final landmarks are connected so that the resulting piecewise-linear function is a closed curve as well.

It is clear that the degree to which the linear reconstruction $L_m(\boldsymbol{\theta})$ resembles the original curve β_m will depend on the landmark configuration. For instance, Figure 1 shows three landmark configurations (under the assumption of $k = 4$ landmarks): the one on the left yields a worse linear reconstruction than the two on the right. At the same time, the two figures on the right that are more desirable in terms of linear reconstruction error appear to select features that are more important to the understanding of the shape of interest. The discrepancy between $L_m(\boldsymbol{\theta})$ and β_m can be measured using the elastic metric given by Eq. (1). This is referred to as the *reconstruction error* associated with a landmark set $\boldsymbol{\theta}$. After discretization to N points, this quantity is given by:

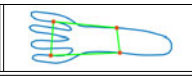
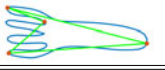
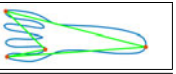
			
(1)	$(0.20, 0.40, 0.60, 0.80)$	$(0.17, 0.25, 0.51, 0.76)$	$(0.25, 0.51, 0.76, 0.84)$
(2)	0.713	0.402	0.399

Figure 1. Linear reconstructions (green) of curve β (blue) through three different landmark configurations (red). (1) Landmark values $\boldsymbol{\theta}$ and (2) $d^2_{\text{Elastic}}(\beta^{(N)}, L^{(N)}(\boldsymbol{\theta}))$ are listed below.

$$d_{\text{Elastic}}(\beta_m^{(N)}, L_m^{(N)}(\boldsymbol{\theta})) = \left|\text{vec}\left(q_{\beta_m}^{(N)} - q_{L_m(\boldsymbol{\theta})}^{(N)}\right)\right|, \tag{2}$$

where the vectorization operator $\text{vec}(\cdot)$ converts the input, a $2 \times N$–dimensional curve, into a $2N$-dimensional vector by vertical concatenation of the 2 rows. A landmark configuration $\boldsymbol{\theta}$ yielding a low value of Eq. (2) "approximates" the full shape well (and thus represents a desirable set of landmarks), while a high value indicates landmarks which do not capture important features of the original shape well. This is reflected in the values computed for the configurations of Figure 1: the squared reconstruction error is much larger for the landmark placements on the left, while smaller (and virtually the same) for the right two.

3.2. Model

A model for the curves which allows for inference on a landmark configuration $\boldsymbol{\theta}$ can be specified conditionally using the linear reconstructions described in the previous section as:

$$\text{vec}\left(q_{\beta_1}^{(N)} - q_{L_1(\boldsymbol{\theta})}^{(N)}\right), \ldots, \text{vec}\left(q_{\beta_M}^{(N)} - q_{L_M(\boldsymbol{\theta})}^{(N)}\right) \Big| \boldsymbol{\theta}, \kappa \overset{\text{iid}}{\sim} \mathcal{N}\left(0_{2N}, \frac{1}{2\kappa} I_{2N}\right), \tag{3}$$

where κ is a precision parameter. For a given landmark configuration, this discretized Gaussian process model proposed in Strait, Chkrebtii, and Kurtek (2018) compares the *SRVFs* of all M curves in the sample with their linear reconstruction. The resulting likelihood function,

$$\mathcal{L}(\boldsymbol{\theta}, \kappa) = \pi^{-NM} \kappa^{NM} \exp\left(-\kappa \sum_{m=1}^{M} d_{\text{Elastic}}^2\left(\beta_m^{(N)}, L_m^{(N)}(\boldsymbol{\theta})\right)\right), \tag{4}$$

is inversely proportional to the cumulative reconstruction error, defined by Eq. (2). This likelihood rewards landmark sets which yield a low reconstruction error cumulatively over all curves, as desired by the argument presented in Sec. 3.1.

The Bayesian paradigm is well-suited to the problem of inference on $\boldsymbol{\theta}$ in this setting, due to the ease of representing highly structured uncertainty, the availability of prior information about the number of landmarks present, and the ease of updating the posterior as new data becomes available. By updating prior models on $\boldsymbol{\theta}$ and κ given the data, uncertainty estimates for these quantities can be obtained from approximate posterior samples. Prior probability models on these parameters should be chosen to reflect the analyst's belief about a particular group of shapes before the data is collected. In Strait, Chkrebtii, and Kurtek (2018), prior models which generalize well to a variety of applications are provided, and are therefore less informative. A natural prior for the precision parameter κ is a Gamma(a, b) distribution. Using this, one can obtain the marginal likelihood $\mathcal{L}(\boldsymbol{\theta})$ by integrating with respect to the prior measure for κ, i.e., $\mathcal{L}(\boldsymbol{\theta}) = \int \mathcal{L}(\boldsymbol{\theta}, \kappa)\pi(\kappa)d\kappa$. Selecting hyperparameters $a = 1, b = 0.01$ results in a diffuse prior, which is appropriate when little is known about this parameter a-priori. The prior on $\boldsymbol{\theta}$ is specified indirectly on the *consecutive differences*, $\boldsymbol{s}$, between its components – the precise expression for components of $\boldsymbol{s}$ is given in Strait, Chkrebtii, and Kurtek (2018). A common prior in this situation is a Dirichlet distribution, i.e., $\boldsymbol{s} \sim \text{Dir}(\alpha 1)$.

This construction ensures that the ordering constraint of $\boldsymbol{\theta}$ is satisfied. For example, in each panel of Figure 1, this ensures that the reconstructions correspond to a single landmark arrangement by preventing permutation of the labels of $\boldsymbol{\theta}$ that leave the values of its elements unchanged. In a general setting, selecting $\alpha = 1$ adds little subjective information about the spacing between elements of $\boldsymbol{\theta}$.

3.3. Posterior sampling

Let $\mathcal{L}(\boldsymbol{\theta})$ be the marginal likelihood described in the previous section, and $\pi(\boldsymbol{\theta})$ be the prior density for $\boldsymbol{\theta}$ induced by the Dirichlet prior on component-wise landmark differences $\boldsymbol{s}$ from Sec. 3.2. As is often the case in practice, summaries of interest for the posterior,

$$\pi(\boldsymbol{\theta}|\boldsymbol{\beta}^{(N)}) \propto \left(b + \sum_{m=1}^{M} d^2_{\text{Elastic}}\left(\beta_m^{(N)}, L_m^{(N)}(\boldsymbol{\theta})\right)\right)^{-(a+NM)} \pi(\boldsymbol{\theta}), \tag{5}$$

are not available in closed form. Indeed, the posterior is only known up to a normalizing constant, as integration is intractable. Posterior summaries are instead typically estimated from a Monte Carlo sample (Robert and Casella 1999). A general random walk Metropolis sampler (Metropolis et al. 1953) is as follows. First, draw initial value $\boldsymbol{\theta}^{[0]}$ from $\pi(\boldsymbol{\theta})$. Then, form a Markov chain, where at step t, proposal $\boldsymbol{\theta}^*$ is drawn from a symmetric kernel G with density $g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{[t-1]})$. For a random walk algorithm, the kernel is centered at $\boldsymbol{\theta}^{[t-1]}$, with covariance K. Many choices of G are suitable; most standard is to let $g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{[t-1]})$ be a $\mathcal{N}(\boldsymbol{\theta}^{[t-1]}, K)$ density, which is our assumption for the remainder of this paper. The proposal $\boldsymbol{\theta}^*$ is accepted with probability $\min\{1, \alpha^{[t]}_{\text{within}}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{[t-1]})\}$, where,

$$\alpha^{[t]}_{\text{within}}\left(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{[t-1]}\right) = \frac{\mathcal{L}(\boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)}{\mathcal{L}\left(\boldsymbol{\theta}^{[t-1]}\right)\pi\left(\boldsymbol{\theta}^{[t-1]}\right)}, \tag{6}$$

is the Metropolis acceptance ratio. This construction results in a Markov chain whose stationary distribution is the target posterior $\pi(\boldsymbol{\theta}|\boldsymbol{\beta}^{(N)})$. In other words, under some mild assumptions (Robert and Casella 1999), there exists c such that the sequence $\{\boldsymbol{\theta}^{[c]}, \boldsymbol{\theta}^{[c+1]}, \ldots\}$ is a sample from the target posterior $\pi(\boldsymbol{\theta}|\boldsymbol{\beta}^{(N)})$. In practice, c is unknown and convergence diagnosis is based on heuristics, such as visual analysis of trace plots and autocorrelation plots. Once convergence is suspected, the actual sample used to approximate posterior summaries is $\{\boldsymbol{\theta}^{[\hat{c}]}, \boldsymbol{\theta}^{[\hat{c}+1]}, \ldots, \boldsymbol{\theta}^{[\hat{c}+R-1]}\}$, where R is the size of the MCMC sample, and all samples before $\hat{c}$, the user-specified burn-in period, are discarded. If it is suspected that $\hat{c} < c$, the sample produced may also be thinned to reduce autocorrelation in the sequence of draws. For smooth and unimodal posteriors, the choice of $\hat{c}$ may well be reliable. However, when multiple posterior modes are present, and especially if they are separated by large regions of low posterior probability, one can typically be sure that c would be extremely large (Geyer 1991). Indeed, the Markov chain described above, is known to become trapped in local modes in the sense that the time spent in that mode is not proportional to the mode's relative probability mass,

unless an extremely long chain is considered, which may well be beyond the user's computational resources.

3.4. Algorithm discussion

There are several points of discussion regarding this MCMC algorithm in the context of model-based landmark selection. First, it should be noted that Strait, Chkrebtii, and Kurtek (2018) use an even simpler proposal kernel G: instead of a multivariate normal random walk proposal, they simply select one of the k components of $\boldsymbol{\theta}^{[t-1]}$ at random to update using a univariate random walk, keeping the other $k - 1$ components fixed. While updates are slightly faster, this scheme mixes very slowly. In addition, this slow mixing can lead to misleading conclusions due to improper diagnosis of convergence. Components of $\boldsymbol{\theta}$ can get trapped in certain regions of the state space due to the posterior dependence between components. The multivariate proposal scheme described in Sec. 3.3 partially resolves this issue. A related important choice is that of the proposal covariance matrix K. There is a tradeoff between the dispersion of the proposal kernel and the acceptance rate of the MCMC algorithm. Balancing this tradeoff becomes especially difficult in the presence of constraints on the parameter space (Golchi and Campbell 2016). Chosen optimally, K would match the posterior covariance. However, without knowing this, we can make it as close as possible via adaptation. Indeed, a chain-dependent adaptation during burn-in of the proposal to target an expected acceptance rate has been shown to be effective in its selection (for example, Roberts, Gelman, and Gilks (1997), Roberts and Rosenthal (2001) 0.234 is optimal for random walk algorithms under certain assumptions).

Another challenge arises with the presence of posterior multi-modality. For certain shapes, there may be several landmark configurations which are similarly important, as defined by the reconstruction error criterion of Sec. 3.1. This is evident in the right two panels of Figure 1 where both choices of $\boldsymbol{\theta}$ yield a relatively similar value of $d^2_{\text{Elastic}}(\beta_m^{(N)}, L_m^{(N)}(\boldsymbol{\theta}))$. Since the quality of landmark locations is incorporated into the likelihood only through this reconstruction error, several values of $\boldsymbol{\theta}$ could map to the same or similar likelihood values, i.e., the mapping $\boldsymbol{\theta} \mapsto d^2_{\text{Elastic}}(\beta_m^{(N)}, L_m^{(N)}(\boldsymbol{\theta}))$ is not injective. This is especially likely for complex shapes, where it is unclear how many landmarks should be selected, like the fork of Figure 1. Random walk Metropolis algorithms are notoriously poor at efficient exploration of multi-modal densities, as they have a tendency to get trapped in local modes. This makes single-chain MCMC inappropriate for models where parameters are only incorporated in the likelihood through non-injective functions (e.g., distances).

4 Adaptive parallel tempering MCMC

4.1. Motivation

In this section, we outline a strategy for sampling posterior landmark arrangements in the presence of multi-modality. Recall that the posterior of interest is $\pi(\boldsymbol{\theta}|\boldsymbol{\beta}^{(N)})$, as specified in Eq. (5), with state space $\mathcal{X} = \{\boldsymbol{\theta} \in \mathcal{D}^k : \theta_1 < \theta_2 < \ldots < \theta_k\}$. As described in

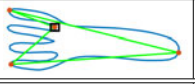
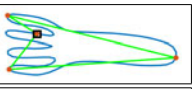
		
(1)	$(0.17, 0.25, 0.51, 0.76)$	$(0.13, 0.25, 0.51, 0.76)$
(2)	0.402	0.471
(3)	1.91×10^{39}	2.64×10^{32}

Figure 2. Linear reconstructions (green) of curve β (blue) through two different landmark configurations (red). Perturbed landmark 1 is marked in black. (1) Landmark values $\boldsymbol{\theta}$, (2) $d^2_{\text{Elastic}}(\beta^{(N)}, L^{(N)}(\boldsymbol{\theta}))$, and (3) unnormalized posterior density values are listed below.

Sec. 3.3, applying a standard random walk Metropolis-Hastings algorithm means proposing $\boldsymbol{\theta}^*$ from a symmetric proposal kernel G, which is centered at the previous value, $\boldsymbol{\theta}^{[t-1]}$, and accepting it with probability given by the ratio of unnormalized posterior densities, as seen in Eq. (6). However, even when adapting the proposal covariance matrix, proposals that are much larger than the local posterior variance are very unlikely. In our motivating application, posterior modes over landmark arrangements tend to be well-separated and highly-peaked, as a result of the posterior's exponential dependence on the number of discretization points N (which is typically between 50 and 200).

As a simple illustration, again consider the fork in Figure 2, which has been discretized to 101 points ($N = 101$, $M = 1$), and set $\alpha = 1$ for the Dirichlet prior on $\boldsymbol{s}$. This implies $\pi(\boldsymbol{\theta}) \propto I(\theta_1 < \ldots < \theta_k)$, where $I(\cdot)$ is the indicator function. For a weakly informative choice of $a = 1, b = 0.01$ on the κ prior, the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{\beta}^{(N)})$ is approximately proportional to $(d^2_{\text{Elastic}}(\beta^{(N)}, L^{(N)}(\boldsymbol{\theta})))^{-NM}$. Consider $\boldsymbol{\theta}^{[t-1]} = (0.17, 0.25, 0.51, 0.76)$ (left panel), and proposal $\boldsymbol{\theta}^* = (0.13, 0.25, 0.51, 0.76)$ (right panel) which only perturbs the first component – this acts to move the red point between the upper two prongs of the fork closer to the tip of the third prong. This small move in the state space $\mathcal{X}$ results in a squared reconstruction error increase of 0.069, resulting in a configuration which is on the order of 10^{-6} less likely. For a perturbation of all components, this type of proposal occurs quite often, reflecting the large negative curvature of the posterior modes.

4.2. Parallel tempering

One way to traverse the low-density region between highly-peaked modes of a target distribution is to *temper* its density, which has the effect of "flattening" out the posterior of interest. Given a specified temperature $T > 1$, the tempered target posterior is given by $\pi_T(\boldsymbol{\theta}|\boldsymbol{\beta}^{(N)}) \propto (\mathcal{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta}))^{1/T}$. The acceptance probability in Eq. (6) for a Markov chain targeting a tempered posterior is therefore larger compared to that of the corresponding non-tempered case. This allows for faster traversal of lower-probability regions between modes. Of course, the goal is ultimately to construct a Markov chain that targets the actual posterior, while exploiting the flexibility of tempering. This can be achieved by running a number $z > 1$ of MCMC chains in parallel, one targeting the true posterior and the rest targeting progressively more highly tempered posteriors, according to a temperature schedule $T_1 = 1 < T_2 < \ldots < T_z$ chosen by the user. Between-chain swaps are proposed with a fixed probability, and can be accepted or rejected, which

allows for the exchange of information between chains. Crucially, the single non-tempered chain targets the desired posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{\beta}^{(N)}) = \pi_{T_1}(\boldsymbol{\theta}|\boldsymbol{\beta}^{(N)})$. This technique belongs to the class of population MCMC methods, and is known as *parallel tempering* (as earliest introduced in Swendsen and Wang 1986; Geyer 1991).

At time t, let $\boldsymbol{\theta}_1^{[t]}, ..., \boldsymbol{\theta}_z^{[t]}$ be the current states of the z chains, corresponding to tempered posteriors $\pi_{T_1}(\boldsymbol{\theta}|\boldsymbol{\beta}^{(N)}), ..., \pi_{T_z}(\boldsymbol{\theta}|\boldsymbol{\beta}^{(N)})$, respectively. Once the within-chain proposal step is completed, a swapping of the current states between chains can be proposed. Any two chains $i, j, i \neq j$ can be swapped. However, in practice, most restrict to swaps of adjacent chains – i.e., randomly select $i \in \{1, ..., z-1\}$ and then propose swap between chains i and $i+1$. This swap is accepted with probability $\min\{1, \alpha^{[t]}_{\text{swap},\, i}(\boldsymbol{\theta}_i^{[t]}, \boldsymbol{\theta}_{i+1}^{[t]})\}$, where,

$$\alpha^{[t]}_{\text{swap},\, i}\left(\boldsymbol{\theta}_i^{[t]}, \boldsymbol{\theta}_{i+1}^{[t]}\right) = \left(\frac{\mathcal{L}\left(\boldsymbol{\theta}_i^{[t]}\right)\pi\left(\boldsymbol{\theta}_i^{[t]}\right)}{\mathcal{L}\left(\boldsymbol{\theta}_{i+1}^{[t]}\right)\pi\left(\boldsymbol{\theta}_{i+1}^{[t]}\right)}\right)^{\frac{1}{T_{i+1}} - \frac{1}{T_i}}, \tag{7}$$

is the Metropolis acceptance ratio for between-chain swaps. This ratio takes a similar form to the within-chain proposal because one can view swap proposals as switching adjacent components of one product Markov chain, where the state space is the product space $\mathcal{X}^z$. Once convergence is suspected for all of the chains (diagnosis is done for each chain individually), inference is performed based on the sample generated from the non-tempered chain, i.e., the one corresponding to the target posterior $\pi_{T_1}(\boldsymbol{\theta}|\boldsymbol{\beta}^{(N)})$.

4.3. Adapting parallel tempering

Parallel tempering MCMC can be challenging to implement, due to the many user-specified settings. Choosing the temperature scale efficiently allows for some of the chains to easily propose values of $\boldsymbol{\theta}$ corresponding to the various local modes, while the total number of the parallel chains controls how quickly any between-mode moves will make their way to the target chain via swaps. Temperature gaps between adjacent chains that are too large may result in a low number of swap moves. Moreover, running numerous chains is computationally burdensome, especially for high-dimensional parameter spaces, as parallelization is only partial (since communication between chains is critical). We attempt to address these issues in the spirit of Lacki and Miasojedow (2016).

The basic idea is to let the temperatures be updated at each step of a sweep through all z chains (i.e., the temperature of chain i, $T_i^{[t]}$, is now time-dependent), with $T_1^{[t]} = 1$ for all t. Swaps are proposed and accepted using the Metropolis ratio in Eq. (7). Once all proposal steps have beem completed, temperatures are adapted so that swap acceptance probabilities target a theoretically optimal value of 0.234 (suggested by Atchade, Roberts, and Rosenthal 2011; Kone and Kofke 2005), ensuring that the gap between adjacent temperatures is not too large. The number of chains is also adapted at each step, i.e., $z^{[t]}$ is a function of time t. Chains are eliminated if a function of the variance scaling factor exceeds the optimal proposal covariance for that particular chain. Since it

may take time to properly tune individual chains, we only begin to prune chains after enough full sweeps through all chains have been performed. The full adaptive parallel tempering MCMC algorithm is shown below.

Algorithm 1. Adaptive PT

I. *Initialization*
   1. Determine number of landmarks k to infer.
   2. Set number of chains $z^{[0]}$, and temperatures $1 = T_1^{[0]} < T_2^{[0]} < \ldots < T_{z^{[0]}}^{[0]}$. Also, set N_0 to be the desired iteration before chains can be in consideration for deletion.
   3. Compute $\rho_i^{[0]} = \log(T_{i+1}^{[0]} - T_i^{[0]})$ for $i = 1, \ldots, z^{[0]} - 1$.
   4. For chains $i = 1, \ldots, z^{[0]}$:
      a. Initialize landmarks $\boldsymbol{\theta}_i^{[0]}$ and set $\boldsymbol{\mu}_i^{[0]} = \boldsymbol{\theta}_i^{[0]}$ (running mean).
      b. Set $\Sigma_i^{[0]} = I_k$ (running covariance).
      c. Set $\phi_i^{[0]} = 0$ (scale factor), and let $K_i^{[0]} = \exp(\phi_i^{[0]})\Sigma_i^{[0]}$ (proposal covariance).

II. *Within-chain adaptive random walk:* at time t, for chains $i = 1, \ldots, z^{[t-1]}$:
   1. Update chain i:
      a. Sample $\boldsymbol{\theta}_i^* \sim \mathcal{N}(\boldsymbol{\theta}_i^{[t-1]}, K_i^{[t-1]})$.
      b. Compute $\alpha_{\text{within},i}^{[t]}(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}_i^{[t-1]}) = (\alpha_{\text{within}}^{[t]}(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}_i^{[t-1]}))^{1/T_i^{[t-1]}}$ (latter is Equation 6 adjusted by temperature of chain i).
      c. Set $\tilde{\boldsymbol{\theta}}_i^{[t]} = \boldsymbol{\theta}_i^*$ w.p. $\min\{1, \alpha_{\text{within},i}^{[t]}(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}_i^{[t-1]})\}$; else, set $\tilde{\boldsymbol{\theta}}_i^{[t]} = \boldsymbol{\theta}_i^{[t-1]}$.
   2. Adapt proposal covariance of chain i:
      a. Compute $\tau_1^{[t]} = \min\{0.9, k(t+1)^{-0.6}\}, \tau_2^{[t]} = (t+1)^{-0.6}$.
      b. Update $\boldsymbol{\mu}_i^{[t]} = (1-\tau_1^{[t]})\boldsymbol{\mu}_i^{[t-1]} + \tau_1^{[t]}\tilde{\boldsymbol{\theta}}_i^{[t]}$.
      c. Update $\Sigma_i^{[t]} = (1-\tau_1^{[t]})\Sigma_i^{[t-1]} + \tau_1^{[t]}(\tilde{\boldsymbol{\theta}}_i^{[t]} - \boldsymbol{\mu}_i^{[t-1]})(\tilde{\boldsymbol{\theta}}_i^{[t]} - \boldsymbol{\mu}_i^{[t-1]})^\top$.
      d. Update $\phi_i^{[t]} = \phi_i^{[t-1]} + \tau_2^{[t]}(\min\{1, \alpha_{\text{within},i}^{[t]}(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}_i^{[t-1]})\} - 0.234)$, and set $K_i^{[t]} = \exp(\phi_i^{[t]})\Sigma_i^{[t]}$.

III. *Between-chain adaptive state swap:* at time t,
   1. Randomly select $\eta \in \{1, \ldots, z^{[t-1]} - 1\}$.
   2. Compute $\alpha_{\text{swap},\eta}^{[t]}(\tilde{\boldsymbol{\theta}}_\eta^{[t]}, \tilde{\boldsymbol{\theta}}_{\eta+1}^{[t]})$ (via Equation 7).
   3. Set $\boldsymbol{\theta}_{\eta+1}^{[t]} = \tilde{\boldsymbol{\theta}}_\eta^{[t]}$ and $\boldsymbol{\theta}_\eta^{[t]} = \tilde{\boldsymbol{\theta}}_{\eta+1}^{[t]}$ w.p. $\min\{1, \alpha_{\text{swap},\eta}^{[t]}(\tilde{\boldsymbol{\theta}}_\eta^{[t]}, \tilde{\boldsymbol{\theta}}_{\eta+1}^{[t]})\}$; else, set $\boldsymbol{\theta}_{\eta+1}^{[t]} = \tilde{\boldsymbol{\theta}}_{\eta+1}^{[t]}$ and $\boldsymbol{\theta}_\eta^{[t]} = \tilde{\boldsymbol{\theta}}_\eta^{[t]}$. For all other chains $i \neq \eta, i \neq \eta + 1$, set $\boldsymbol{\theta}_i^{[t]} = \tilde{\boldsymbol{\theta}}_i^{[t]}$.

IV. *Adapt temperature scheme:* at time t,
   1. Set $T_1^{[t]} = 1$. For $i = 1, \ldots, z^{[t-1]} - 1$:
      a. Compute $\alpha_{\text{swap},i}^{[t]}(\boldsymbol{\theta}_i^{[t]}, \boldsymbol{\theta}_{i+1}^{[t]})$ (via Equation 7).
      b. Update $\rho_i^{[t]} = \rho_i^{[t-1]} + \tau_2^{[t]}(\min\{1, \alpha_{\text{swap},i}^{[t]}(\boldsymbol{\theta}_i^{[t]}, \boldsymbol{\theta}_{i+1}^{[t]})\} - 0.234)$.
      c. Set $T_{i+1}^{[t]} = \exp(\rho_i^{[t]}) + T_{i+1}^{[t-1]}$.

2. If $t > N_0$, set $z^{[t]} = \min\left\{j \in \{1, \ldots, z^{[t-1]}\} : \exp(\phi_j^{[t]}) \geq \frac{2.38}{\sqrt{k}}\right\}$; else, set $z^{[t]} = z^{[t-1]}$. Delete all chains indexed after $z^{[t]}$.

4.4. Algorithm discussion

There are several additional parameters in this algorithm, which should be briefly discussed. First, in order to adapt the within-chain proposal scheme to ensure proper mixing, Lacki and Miasojedow (2016) use a combination of two adaptive procedures described in the following papers: Haario, Saksman, and Tamminen (2001), Andrieu and Thoms (2008), Roberts and Rosenthal (2009), and Atchade and Fort (2010). For a given chain i, a running estimate of the covariance matrix for $\boldsymbol{\theta}_i$ is computed ($\Sigma_i^{[t]}$), which requires a running estimate of the mean ($\boldsymbol{\mu}_i^{[t]}$) as well. This covariance is then adjusted by a scale factor, $\exp(\phi_i^{[t]})$, which is also updated iteratively. Step sizes for all of these updates are $\tau_1^{[t]}$ and $\tau_2^{[t]}$, which decrease as the number of chain iterations increases; this results in smaller updates as the chain progresses in time (with the hope that the target stationary distributions have been reached). The quantities $\tau_1^{[t]} = \min\{0.9, k(t+1)^{-0.6}\}$ and $\tau_2^{[t]} = (t+1)^{-0.6}$ are chosen by Miasojedow, Moulines, and Vihola (2013) (denoted γ_2 and γ_3, respectively, in their work) for multiple reasons: they satisfy required ergodicity conditions for this adaptive parallel tempering algorithm, and agree with common choices in the iterative stochastic algorithm literature.

It should also be pointed out that temperatures evolve over time through updates of a vector comprised of log temperature-differences between adjacent components. Miasojedow, Moulines, and Vihola (2013) prove the existence and uniqueness of a choice for this vector which achieves a targeted swap acceptance rate (0.234 has been shown to be optimal in both the physics and statistics literature for swap moves by Kone and Kofke 2005; Atchade, Roberts, and Rosenthal 2011). This rate is also used to decouple the choice of temperatures with the number of temperature levels. A simplification to this update can be made by assuming geometric spacing between temperatures, meaning only one parameter for the log temperature-difference is updated at each iteration; however, with our particular implementation, we found that this was not necessary, as we still maintain computational efficiency with the original update. Lastly, the final step in an iteration of this algorithm selects the lowest-temperature chain which exceeds the optimal proposal covariance for that particular chain, and prunes the remaining higher-temperature chains. Elimination of chains should only be considered once each chain's proposal covariance has stabilized. Thus, N_0 should be selected large enough to ensure this has been satisfied. In our implementation, $N_0 = 40,000$ seems to work empirically.

5 Results

5.1 Simulated curve

First, we illustrate the advantages of adaptive parallel tempering from Sec. 4.3 for a simple simulated curve, defined by $\beta(t) = (t, |\sin(2\pi t)|)^\top, t \in [0, 1]$. Suppose one wants to

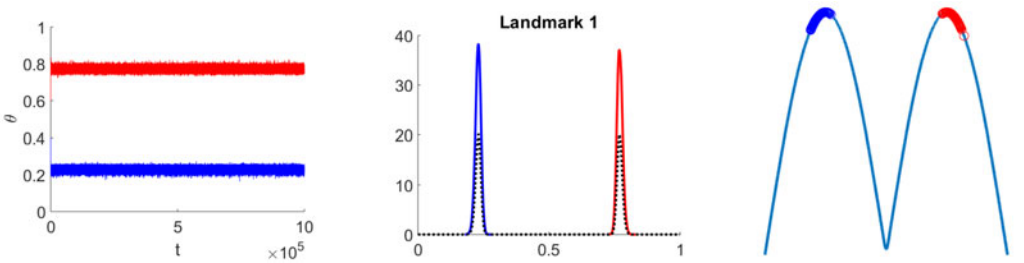


Figure 3. Posterior inference by way of the adaptive random walk Metropolis algorithm for the simulated curve using initializations (1) and (2), in blue and red, respectively. Left: Trace plots for θ. Middle: Computed posterior for θ, with the true posterior in black. Right: Posterior samples superimposed on curve β.

estimate $k = 1$ landmark θ on this curve. In this simplified setting, we can numerically evaluate the posterior in Eq. (5) over a grid (selecting 200 equally-spaced values of θ) as a baseline to compare the performance of different sampling schemes. For this curve (blue in the right plot of Figure 3), each peak appears to represent an equally important feature of the shape due to symmetry. Thus, we would expect the posterior of θ to be bimodal, with each mode concentrated around one of the peaks of β.

We begin by implementing the adaptive random walk Metropolis algorithm of Sec. 3.3 and Strait, Chkrebtii, and Kurtek (2018). Two independent chains are initialized at (1) $\theta^{[0]} = 0.4$ and (2) $\theta^{[0]} = 0.6$. Each chain is run for 10^6 iterations. To form the final posterior sample, the first ten percent of iterations are discarded as samples prior to burn-in, and the remaining sample is thinned by every 100 steps to reduce chain autocorrelation. The results are shown in Figure 3 for initializations (1) and (2), in blue and red, respectively. We first note the trace plots in the left panel appear to suggest that each chain has converged. However, initializing the chain at two different values shows convergence to two different regions of the parameter space! This is well illustrated in the middle panel, which shows the approximated posterior based on random walk Metropolis samples for these two initializations (in red and blue); (1) concentrates around 0.25, while (2) concentrates around 0.75. The true posterior (as evaluated on a grid) is shown as the black dashed line; since the algorithm only remains in one of the modes, it overestimates the true density at the corresponding values of θ. The posterior samples can be superimposed on the original shape, as shown in the right panel; as expected, each one targets a separate peak on the curve, but neither chain is able to move to the opposite peak, due to the low-density region of the posterior which separates the two modes. We note that acceptance rates for (1) and (2) are 0.2340 and 0.2338, respectively, which are near the 0.234 target in the adaptive random walk algorithm.

In order to explore both modes, we now implement the adaptive parallel tempering algorithm described in Sec. 4.3. For the initialization step, we set the initial number of chains $z^{[0]} = 10$, and set initial temperatures $T_i^{[0]} = i$ for $i = 1, \ldots, 10$. The algorithm is run for 10^6 sweeps; we resort to pruning of chains after $N_0 = 40,000$ iterations, in order to ensure that proposal covariances have stabilized for each chain. After the sampling algorithm is completed, the posterior is approximated from the samples of the first (non-tempered) chain, and is post-processed to remove burn-in and autocorrelations in

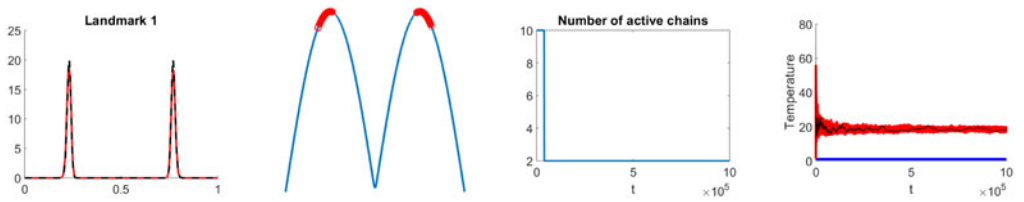


Figure 4. Posterior inference by way of the adaptive parallel tempering sampler for the simulated curve. From left to right: computed posterior for θ, with the true posterior in black.; posterior samples superimposed on curve β; number of active chains as a function of number of iterations; temperature evolution of the final two active chains under 25 different initializations (black line represents the temperature course corresponding to the second chain for the original initialization).

the same way as above. The left-most plot of Figure 4 shows the resulting posterior density, which is bimodal and very similar to the true posterior. When superimposed on the curve β (second panel), it is clear that this adaptive method is able to traverse between the two modes efficiently, unlike the random walk Metropolis algorithm. The third plot of Figure 4 shows a plot of the number of active chains, as a function of each chain's time step (using the original settings). As specified above, we begin with ten chains; however, right after iteration number $N_0 = 40,000$, the algorithm immediately prunes the number of active chains to two, and remains at two for the duration of the algorithm. The temperature evolution of the two active chains is plotted on the far right: note that the first chain stays fixed at one (since this is the target posterior). Within-chain acceptance rates for the first and second chains were 0.2336 and 0.2337, respectively, while the swap acceptance rate between the two chains overall was 0.2263 (including iterations where all ten initial chains were included); this rate increases to the targeted acceptance rate of 0.2347 when only focusing on iterations after the pruning of chains.

Diagnosis of convergence via trace plots is more challenging for multi-modal targets, due to the large number of jumps between modes; see Figure 1 of the supplementary materials online. We can check a few different items to assess convergence. First, one can confirm that the final posterior density is independent of initialization $\boldsymbol{\theta}_i^{[0]}$ for each chain $i = 1, \ldots, 10$; this is true for this example. In addition, the pruning of chains should also be independent of the initialization. To test this, we repeat the adaptive parallel tempering algorithm 25 independent times, with random initialization of the landmark vector. We specify five active chains for each run to start, since our initial run suggested only two chains will be kept. In all 25 runs, five chains are immediately pruned to two after iteration N_0, illustrating independence to initialization. Finally, we would like the temperatures to adapt to roughly the same value for each independent run of the algorithm. In the right panel of Figure 4, temperature evolution for the two active chains are shown. Note that the first chain is fixed at one (as this is the target posterior). For the second chain, the plot shows all 25 temperature courses simultaneously, approximately forming a band which seems to stabilize around a temperature of 20. As a check, we have also superimposed the temperature profile of the original run of the algorithm, which also falls within this band. We also note that for the 25 different initializations, both the within-chain and between-chain swap acceptance rates are near the targeted 0.234.

Lastly, we compare computation time for the different samplers. It is clear that parallel tempering is more expensive, due to the necessity for running multiple chains simultaneously; however, we have found all algorithm times to be within reason for the applications studied throughout this paper. For the simulated curve, our implementation of adaptive random walk Metropolis required 989.3 and 949.5 seconds for the two initializations, respectively; compare this to our original adaptive parallel tempering scheme with ten initial chains, which required 1614.9 seconds. If we only initialized with five chains (which was acceptable since all 25 replicates of the sampler converged to two active chains), this requires a similar 1590.7 seconds: this is understandable, as all but two of the chains are pruned fairly early during the course of the algorithm.

5.2. MPEG-7 shapes

Parallel tempering is almost necessary for shapes with complex features, including the ones in the MPEG-7 shape dataset (http://www.dabi.temple.edu/~shape/MPEG7/dataset.html). This is a collection of shapes extracted from images which are popularly used in computer vision research. The curves representing the outlines of objects in these images are closed, and thus we will resort to posterior inference on the curve domain $\mathcal{D} = \mathbb{S}^1$. Many of these shapes present modeling challenges, as the number of landmarks k is not easily specified. Consider the fork from Figure 2, with the goal of inferring the $k = 4$-dimensional landmark vector $\boldsymbol{\theta}$. Here, the curve β is sampled using $N = 101$ points. To compare performance of the adaptive parallel tempering algorithm, we first proceed by running 10^6 iterations of random walk Metropolis with adaptive covariance. Due to the complexity of the shapes, we remove the first thirty percent of the chain, and further thin the sample by 100 to reduce autocorrelations. Figure 5 shows the marginal posterior densities for each component θ_i for $i = 1, 2, 3, 4$, as plotted on $\mathcal{D}$; the obtained marginals all appear unimodal. However, when the posterior samples are superimposed on the curve in the right panel, the fourth landmark (in yellow) only captures a region between the first two prongs. Since the fork is relatively symmetric, there

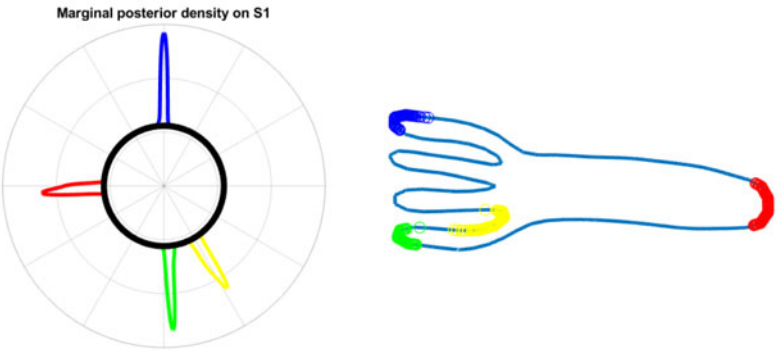


Figure 5. Posterior inference by way of the adaptive random walk Metropolis algorithm for the simulated curve. Left: Marginal posterior densities for components of $\boldsymbol{\theta}$, plotted on the curve domain $\mathbb{S}^1$ (in black, where the horizontal grid line on the right half denotes the zero radian line, and angles are traversed counterclockwise around the circle). Right: Posterior samples superimposed on curve β. Colors denote the four landmarks: θ_1 = blue, θ_2 = red, θ_3 = green, θ_4 = yellow.

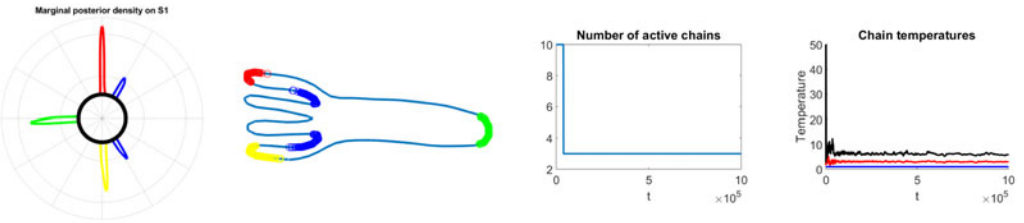


Figure 6. Posterior inference by way of the adaptive parallel tempering sampler for the fork shape. From left to right: marginal posterior densities for $\boldsymbol{\theta}$, plotted on $\mathcal{S}^1$ (in black); posterior samples superimposed on fork β; number of active chains as a function of number of iterations; temperature evolution of the final three active chains. Colors denote the four landmarks: $\theta_1 =$ blue, $\theta_2 =$ red, $\theta_3 =$ green, $\theta_4 =$ yellow.

may be a second mode for this particular landmark occurring between the last two prongs (similar to the middle panel of Figure 1); however, random walk Metropolis is not able to explore this region of the parameter space in 10^6 iterations.

Since we suspect there is a second mode for one of the landmarks (due to the relative symmetry of the fork), adaptive parallel tempering is implemented. As in the simulated curve of Sec. 5.1, we begin with $z^{[0]} = 10$ chains and consider pruning chains after $N_0 = 40,000$ iterations. The initial temperature of chain i is set to i for $i = 1, ..., 10$. The final posterior sample is generated exactly as above, using values from the first chain. The left plot in Figure 6 shows the marginal posterior densities once again; this time, the portion corresponding to regions between prongs displays two equal-sized modes. Note that the colors change compared to the plot in Figure 5; this is the result of a label-switching on components of $\boldsymbol{\theta}$ compared to the random walk Metropolis output. Post-processing is done to remove the identifiability issue associated with ordering of these components (as detailed in Strait, Chkrebtii, and Kurtek (2018)). This means that in this particular figure, landmark labels get shifted by one (i.e., the first landmark here represents the same feature as the fourth landmark in Figure 5). This can be more clearly seen when viewing the posterior samples superimposed on the shape: θ_1 now represents the bimodal feature that was missing from θ_4 under the adaptive random walk method. For this particular shape, three chains were selected to be active after N_0, and temperatures stabilized throughout the duration of the algorithm (see right two panels of Figure 6). We note that the temperature of the third chain actually spikes toward a value of 600, before eventually settling down – this instability seems to occur often during our implementation of the adaptive parallel tempering sampler. Acceptance rates within the three chains were 0.2267, 0.2308, and 0.2240, with swap rates around 0.1137 and 0.1128 for the two possible moves (chain 1 to/from 2, chain 2 to/from 3). We also have verified that these results are independent of chain initialization. A comparison of marginal posterior densities under the random walk sampler and adaptive parallel tempering is included as Figure 2 of the supplementary materials online.

With this example, the bimodal posterior for the landmark corresponding to the between-prong region could suggest that this particular feature is less important in aiding the linear reconstruction of β. This could also potentially indicate a misspecification of the number of landmarks, if the ultimate goal is to approximate the curve using just

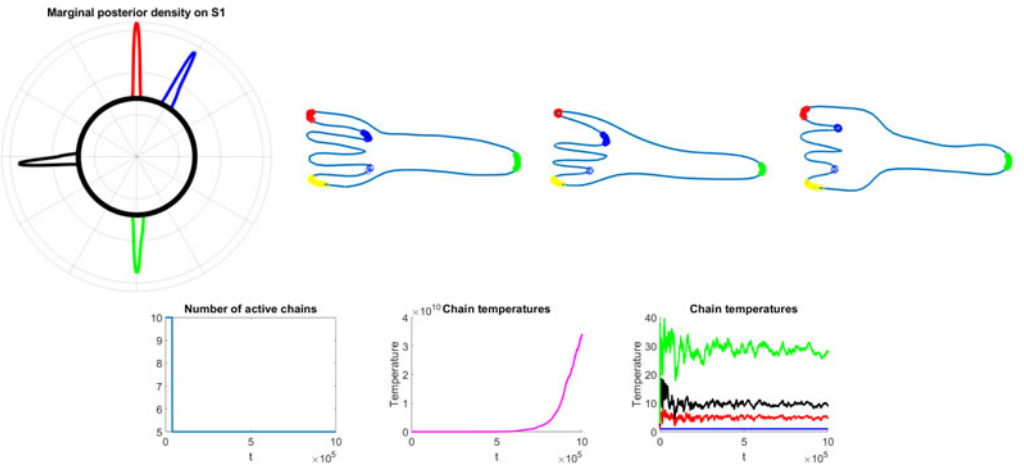


Figure 7. Posterior inference by way of the adaptive parallel tempering sampler for the sample of three forks. Top (from left to right): Marginal posterior densities for $\boldsymbol{\theta}$, plotted on $\mathcal{S}^1$ (in black); posterior samples superimposed on forks $\beta_1, \beta_2, \beta_3$. Colors denote the four landmarks: $\theta_1 =$ blue, $\theta_2 =$ red, $\theta_3 =$ green, $\theta_4 =$ yellow. Bottom (from left to right): Evolution of number of active chains. The right two plots show the temperature evolutions: one for all five chains, and the other for just the first four chains. Note the fifth chain's temperature continues to increase without bound.

the set of inferred landmarks. However, if one solely wants to learn the four most important features, this procedure allows us to see that this landmark can be placed in two different locations with equal value towards the linear reconstruction. Computing time for the random walk algorithm is 1829.7 seconds, while adaptive parallel tempering required 6018.1 seconds – longer as a result of the number of landmarks and number of active chains, but certainly not prohibitively expensive.

The parallel tempering procedure can also be applied to the model when inference is desired for multiple homogeneous shapes simultaneously. Consider $M = 3$ similarly-shaped forks, with the goal of still estimating $k = 4$ landmarks. In order to ensure a proper correspondence between the three shapes, we perform simple pairwise registrations between β_2 and β_3 to a baseline β_1 (this ensures that, for instance, the fork prongs occur at the same time as each of the three curves are traversed simultaneously). We implement adaptive parallel tempering under the same settings as the one fork sample; marginal posterior densities for components of $\boldsymbol{\theta}$, as well as posterior samples superimposed on $\beta_1, \beta_2, \beta_3$ are shown in Figure 7. Note the somewhat similar locations of the posterior densities to Figure 6, along with less variability in the regions associated with $\theta_2, \theta_3, \theta_4$. However, the marginal posterior density plot for θ_1 appears unimodal, while the posterior samples superimposed on the forks preserve the bi-modality. The combination of asymmetry over the sample of three forks leads the sampler to prefer the upper prong region to the lower one, since this has smaller reconstruction error.

This example is also noteworthy due to an observation about the adaptive parallel tempering algorithm. Lacki and Miasojedow (2016) choose to prune any chain *after* the one which begins to exceed the optimal proposal covariance. For the three forks sample, this means selection of five active chains after the $N_0 = 40,000^{\text{th}}$ iteration – more than the three necessary for just one fork. Intuitively, this makes sense, as the posterior will

become even more highly-peaked with more data, and thus require more active chains in order to properly temper the target density while ensuring the desired swap rate. However, from the bottom, middle plot of Figure 7, it appears that the fifth chain's temperature increases without bound; the right plot shows that the other four chains do converge to a particular set of temperatures. This suggests that this final chain is not useful to the tempering procedure, and raises a potential question about computational cost with regards to the discussed criteria. Lacki and Miasojedow (2016) errs on the side of caution, preferring to keep chains which are not useful in order to ensure correct sampling, particularly in an example they discuss with a Gaussian mixture model, where 50 chains can be initialized with relative ease. However, for the shape data used here, an additional chain can add a costly amount of time to inference. For our purposes, we prefer to keep this chain, as not all examples exhibit this same temperature instability for the final active chain. For other high-dimensional applications, this phenomenon may be worth studying further.

Finally, we return to analysis of $M = 1$ shape and compare results from random walk Metropolis to the proposed adaptive parallel tempering method for two other shapes in MPEG-7. For the butterfly, the goal is to identify $k = 5$ landmarks; the same is true for the camel. The supplementary materials contain plots for the number of active chains and temperature evolution for both of these examples. The left box of Figure 8 show posterior samples superimposed on a butterfly (left) and camel (right) from the traditional sampler using two different initializations (with identical settings to previous examples in this section). Posterior samples obtained via parallel tempering are displayed in the right box. Once again, we observe that the original sampler converges to two different posteriors depending on the initialization, whereas adaptive parallel tempering is able to detect multi-modalities. Specifically, the butterfly exhibits bi-modality

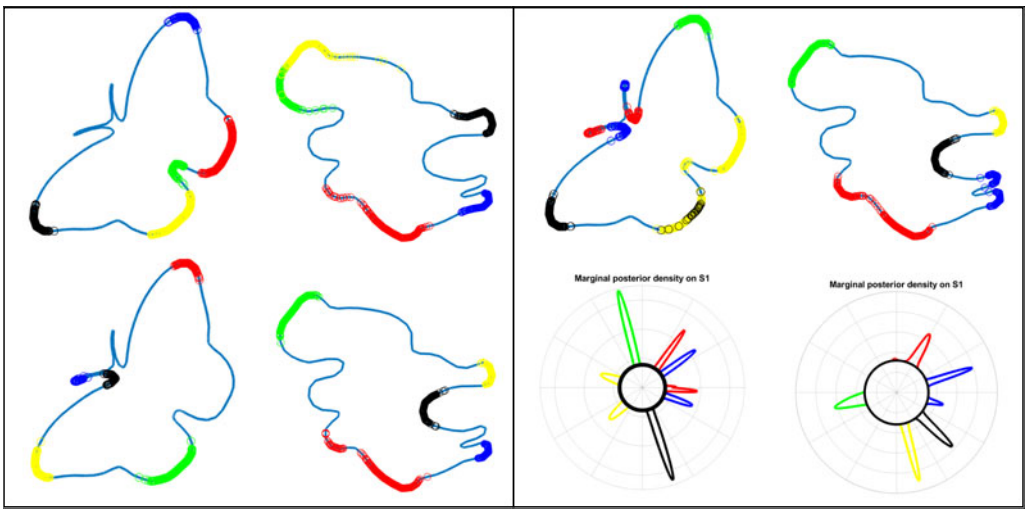


Figure 8. Posterior samples superimposed on two different shapes, obtained via (left box) random walk Metropolis for two different initializations, and (right box) adaptive parallel tempering. The bottom right shows the marginal posterior densities under the parallel tempering sampler. Colors denote labeled landmarks: θ_1 = blue, θ_2 = red, θ_3 = green, θ_4 = yellow, θ_5 = black. Note that relabeling of landmarks is due to post-processing of the posterior sample, as in the fork example.

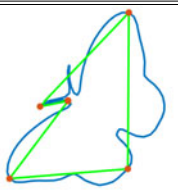
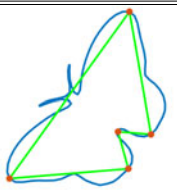
	"Usual" mode	"Unusual" mode
		
(1)	$(0.95, 0.99, 0.29, 0.61, 0.80)$	$(0.80, 0.29, 0.48, 0.54, 0.61)$
(2)	0.372	0.392
(3)	4.26×10^{42}	2.34×10^{40}

Figure 9. Linear reconstructions (green) of curve β (blue) through frequently-occurring (left) and infrequently-occurring (right) posterior landmark configurations (red). (1) Landmark values $\boldsymbol{\theta}$, (2) $d^2_{\text{Elastic}}(\beta^{(N)}, L^{(N)}(\boldsymbol{\theta}))$, and (3) unnormalized posterior density values are listed below.

around the antennae as well as the butterfly's base, while the camel's rear two legs form a bimodal landmark for its posterior.

In particular, the butterfly example is interesting, as numerous landmarks are multi-modal: the two which are immediately obvious capture either the left or right antenna (due to the symmetry), corresponding to θ_1 and θ_2. The region around the base of the butterfly, θ_3, appears bimodal in order to capture one of the two parts of this region. It also looks like θ_5, which concentrates around the left wing and the bottom left part of the base, is bimodal; however, its marginal posterior does not even seem to indicate this second mode. Since individual posterior samples are plotted as circles, the vast majority of the black landmarks are actually on the upper left wing, with very few on the base. This distribution is in fact bimodal, but the second mode is extremely small: in particular, we picked out a "usual" posterior sample and compared it to one corresponding to this "unusual" mode. For each one, Figure 9 shows the resulting linearization of the butterfly β, along with the corresponding squared reconstruction error. Notice that these values are somewhat similar, but due to the high-dimensionality of the curve being studied, the ratio of unnormalized posterior densities at these two values indicates that the "usual" mode is 182 times more likely than the "unusual" mode. Furthermore, almost all of the "unusual" modes corresponded to a similar landmark vector. This is another benefit of parallel tempering: discovery of multiple modes for landmarks that may necessarily be known to exist a priori.

6. Summary

We have proposed an automated parallel tempering scheme for exploring multi-modal posterior distributions arising from landmark detection models in shape analysis. Our approach has the benefit of both identifying multiple landmark configurations which are deemed equally "important" according to the model-based criterion specified (i.e., reconstruction error), which represents an improvement over single-chain alternatives. The scheme we use is adaptive in three ways: (1) modification of proposal covariances within-chain to ensure proper mixing; (2) time-dependent temperature settings for each parallel chain; and (3) adaptation of the number of parallel chains run. Item (3) ensures that this procedure for posterior sampling is also as computationally efficient as possible, given the high-dimensionality of the data. Benefits of implementation have been

demonstrated on a simulated curve, as well as shapes from the MPEG-7 dataset. This sampler is able to detect multi-modal landmark posteriors for many complex shapes, particularly when the number of landmarks k is not easy to specify prior to inference. As discussed in Sec. 1, we feel that the aforementioned adaptive parallel tempering sampler is well-suited to other inferential problems in elastic shape and functional data analysis. Code will be made available on the lead author's webpage, as well as on other repositories.

Funding

This work was partially supported by the following grants (SK): NSF (DMS 1613054), NSF (CCF 1740761), NSF (CCF 1839252) and NIH (R37 CA214955).

Reference

Andrieu, C., and J. Thoms. 2008. A tutorial on adaptive MCMC. *Statistics and Computing* 18 (4): 343–73. doi:10.1007/s11222-008-9110-y.

Atchade, Y., and G. Fort. 2010. Limit theorems for some adaptive MCMC algorithms. *Bernoulli* 16:116–54. doi:10.3150/09-BEJ199.

Atchade, Y., G. Roberts, and J. Rosenthal. 2011. Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Statistics and Computing* 21:555–68. doi:10.1007/s11222-010-9192-1.

Bryner, D., A. Srivastava, and Q. Huynh. 2013. Elastic shapes models for improving segmentation of object boundaries in synthetic aperture sonar images. *Computer Vision and Image Understanding* 117 (12):1695–710. doi:10.1016/j.cviu.2013.07.001.

Chen, C.,. W. Xie, J. Franke, P. Grutzner, L. Nolte, and G. Zheng. 2014. Automatic X-ray landmark detection and shape segmentation via data-driven joint estimation of image displacements. *Medical Image Analysis* 18 (3):487–99. doi:10.1016/j.media.2014.01.002.

Cheng, W., I. L. Dryden, and X. Huang. 2016. Bayesian registration of functions and curves. *Bayesian Analysis* 11 (2):447–75. doi:10.1214/15-BA957.

Domijan, K., and S. P. Wilson. 2005. A Bayesian method for automatic landmark detection in segmented images. In Proceedings of the International Conference on Machine Learning, Bonn, Germany.

Dryden, I. L., and K. V. Mardia. 2016. *Statistical shape analysis: With applications in R*. 2nd ed. New York: Wiley.

Geyer, C. J. 1991. Markov chain Monte Carlo maximum likelihood. Computing Science and Statistics Proceedings of the 23rd Symposium on the Interface, Seattle, WA, 156–63.

Gilani, S. Z., F. Shafait, and A. Mian. 2015. Shape-based automatic detection of a large number of 3D facial landmarks. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA.

Golchi, S., and D. A. Campbell. 2016. Sequentially constrained Monte Carlo. *Computational Statistics & Data Analysis* 97:98–113. doi:10.1016/j.csda.2015.11.013.

Good, P. 1994. *Permutation tests*. New York: Springer-Verlag.

Haario, H.,. E. Saksman, and J. Tamminen. 2001. An adaptive Metropolis algorithm. *Bernoulli* 7 (2):223–42. doi:10.2307/3318737.

He, Y., R. Yucel, and T. Raghunathan. 2011. A functional multiple imputation approach to incomplete longitudinal data. *Statistics in Medicine* 30 (10):1137–56. doi:10.1002/sim.4201.

Joshi, S., and A. Srivastava. 2009. Intrinsic Bayesian active contours for extraction of object boundaries in images. *International Journal of Computer Vision* 81 (3):331–55. doi:10.1007/s11263-008-0179-8.

Joshi, S. H., E. Klassen, A. Srivastava, and I. H. Jermyn. 2007. A novel representation for Riemannian analysis of elastic curves in $\mathbb{R}^n$. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, 1–7.

Kendall, D. G. 1984. Shape manifolds, procrustean metrics, and complex projective shapes. *Bulletin of the London Mathematical Society* 16 (2):81–121. doi:10.1112/blms/16.2.81.

Kone, A., and D. Kofke. 2005. Selection of temperature intervals for parallel-tempering simulations. *Journal of Chemical Physics* 122:206101–206101-2.

Kurtek, S., A. Srivastava, E. Klassen, and Z. Ding. 2012. Statistical modeling of curves using shapes and related features. *Journal of the American Statistical Association* 107 (499):1152–65. doi:10.1080/01621459.2012.699770.

Lacki, M., and B. Miasojedow. 2016. State-dependent swap strategies and automatic reduction of number of temperatures in adaptive parallel tempering algorithm. *Statistics and Computing* 26 (5):951–64. doi:10.1007/s11222-015-9579-0.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. *The Journal of Chemical Physics* 21 (6):1087–92. doi:10.1063/1.1699114.

Miasojedow, Błażej, Eric Moulines, and Matti Vihola. 2013. An adaptive parallel tempering algorithm. *Journal of Computational and Graphical Statistics* 22 (3):649–64. doi:10.1080/10618600.2013.778779.

Robert, C., and G. Casella. 1999. *Monte Carlo statistical methods*. New York: Springer-Verlag.

Roberts, G., A. Gelman, and W. Gilks. 1997. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability* 7:110–20. doi:10.1214/aoap/1034625254.

Roberts, G., and J. Rosenthal. 2001. Optimal scaling for various Metropolis–Hastings algorithms. *Statistical Science* 16 (4):351–67. doi:10.1214/ss/1015346320.

Roberts, G., and J. Rosenthal. 2009. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* 18 (2):349–67. doi:10.1198/jcgs.2009.06134.

Segundo, M. P., L. Silva, O. R. P. Bellon, and C. C. Queirolo. 2010. Automatic face segmentation and facial landmark detection in range images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 40 (5):1319–30. doi:10.1109/TSMCB.2009.2038233.

Srivastava, A., E. Klassen, S. H. Joshi, and I. H. Jermyn. 2011. Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (7):1415–28. doi:10.1109/TPAMI.2010.184.

Srivastava, A., and E. P. Klassen. 2016. *Functional and shape data analysis* Vol. 1, 1–416. New York: Springer-Verlag.

Strait, J., O. Chkrebtii, and S. Kurtek. 2018. Automatic detection and uncertainty quantification of landmarks on elastic curves. *Journal of the American Statistical Association* 1.

Strait, J., S. Kurtek, E. Bartha, and S. MacEachern. 2017. Landmark-constrained elastic shape analysis of planar curves. *Journal of the American Statistical Association* 112 (518):521–33. doi:10.1080/01621459.2016.1236726.

Swendsen, R., and J. Wang. 1986. Replica Monte Carlo simulation of spin glasses. *Physical Review Letters* 57 (21):2607–9. doi:10.1103/PhysRevLett.57.2607.

Tie, Y., and L. Guan. 2013. Automatic landmark point detection and tracking for human facial expressions. *EURASIP Journal on Image and Video Processing* 8 (1):1–15. doi:10.1186/1687-5281-2013-8.