2016

Registration and Clustering of Functional Observations

Zizhen Wu
University of South Carolina

Registration and Clustering of Functional Observations

by

Zizhen Wu

Bachelor of Economics
Wuhan University, China, 2008
Master of Public Administration
Iowa State University, 2010
Master of Science
University of Minnesota, 2012

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Statistics

College of Arts and Sciences

University of South Carolina

2016

Accepted by:

David B. Hitchcock, Major Professor

Timothy E. Hanson, Committee Member

Lianming Wang, Committee Member

Gabriel A. Terejanu, Committee Member

Lacy Ford, Senior Vice Provost and Dean of Graduate Studies

Dedication

To my parents, Yiyi Wu and Jianhua Fan.

Acknowledgments

I would like to express the deepest appreciation to my Ph.D. advisor Dr. David Hitchcock for his motivation, patience, and knowledge. Without his advice, inspiration, and support, it would not have been possible to write this doctoral dissertation. I would also like to thank the members of my dissertation committee: Dr. Timothy Hanson, Dr. Lianming Wang, and Dr. Gabriel Terejanu for their valuable comments and suggestions.

Above all, I would like to thank my fiancée Lin for her encouragement and support of my study. I also wish to acknowledge the Department of Statistics at University of South Carolina for creating a supportive environment during my graduate study here.

Finally, I would like to thank all my friends in USC for leaving me such good memories in Columbia.

Abstract

As an important exploratory analysis, curves of similar shape are often classified into groups, which we call clustering of functional data. Phase variations or time distortions are often encountered in the biological processes, such as growth patterns or gene profiles. As a result of time distortion, curves of similar shape may not be aligned. Regular clustering methods for functional data usually ignore the presence of phase variations, which may result in low clustering accuracy. However, it is difficult to account for phase variation without knowing the cluster structure.

In this dissertation, we first propose a Bayesian method that simultaneously clusters and registers functional data. We model a warping function with a discrete approximation generated from the family of Dirichlet distributions, which allows great flexibility and computational simplicity. Then, we modify our Bayesian algorithm to obtain a fast registration method, which does not require any template curve. We propose a distance-based clustering method that uses a "derivative sign" to measure the dissimilarity between two curves after potential phase variations are removed. Finally, we derive a modified variational approximation for our Bayesian method for simultaneous registration and clustering, which produces a faster alternative for the full Markov chain Monte Carlo (MCMC) sampling.

We demonstrate our proposed methods on simulated data as well as the famous Berkeley growth data, a set of yeast gene profile data, and a set of response of human fibroblasts to serum data.

Table of Contents

Dedication iii

Acknowledgments iv

Abstract v

List of Tables viii

List of Figures ix

Chapter 1 Introduction 1

Chapter 2 Literature Review 4

Chapter 3 A Bayesian Method for Simultaneous Registration and Clustering of Functional Observations 10

3.1 Model Assumption 10

3.2 Likelihood and Bayesian Analysis 12

3.3 Choosing the Number of Clusters 20

3.4 Simulation Study 22

3.5 Real Data Analysis 29

3.6 Discussion 34

Chapter 4 Clustering Functional Observations with Time Warpings via Derivative-Shape Measure 37

4.1 Model Assumption 37

4.2 Pairwise Derivative-Shape Dissimilarity Measure Algorithm 38

4.3 Simulation Study 47

4.4 Real Data Analysis 50

4.5 Discussion 52

Chapter 5 Adapted Variational Bayes Method 57

5.1 Adapted Variational Bayes Algorithm 59

5.2 Choosing Initial Values 66

5.3 Choosing the Number of Clusters 67

5.4 Simulation Study 68

5.5 Real Data Analysis 70

5.6 Discussion 74

Bibliography 78

Appendix A Choosing the Number of Clusters C 82

Appendix B Bayesian Registration for One Cluster 83

B.1 Likelihood and Bayesian Analysis 83

B.2 Sampling Algorithm 84

Appendix C Derivation of the Gradient for Optimizing γ_i 88

Appendix D Convergence Criterion 92

List of Tables

Table 3.1 5 sets of coefficients for the B-spline basis functions 23

Table 3.2 Sensitivity analysis for simulated data 29

Table 3.3 Clustering results for Berkeley acceleration curves. 31

Table 3.4 Clustering results for cell cycle when $C = 5$ 33

Table 3.5 Clustering results for cell cycle when $C = 4$ 34

Table 4.1 Parameter choices for growth data 50

Table 4.2 Clustering results for Berkeley acceleration curves. 51

Table 4.3 Clustering results for Berkeley acceleration curves based on Euclidean distance. 51

Table 4.4 Parameter choices for HFS data 53

Table 5.1 Clustering results for Berkeley acceleration curves by AVB. 72

List of Figures

Figure 1.1 Examples of warping functions 2

Figure 3.1 Left: Warping functions from $Dir(0.8, 0.8, \ldots, 0.8)$. Right: Warping functions from $Dir(5, 5, \ldots, 5)$. 13

Figure 3.2 Left: Simulated data with phase variations. Right: Simulated data with additional vertical shifts generated from $Unif\,(-0.5, 0.5)$. 14

Figure 3.3 (a) A set of 56 simulated observations with 5 clusters. (b) Simulated data with phase variation removed, with superimposed posterior estimated mean curves (solid black). (c) True mean curves (gray) and estimated mean curves (black). (d) Estimated warping functions for all 5 clusters. 26

Figure 3.4 Solid curves represent estimated warpings, and dashed curves represents true warpings. Curves with the same color are warpings for the same observation. 27

Figure 3.5 ψ values for different choices of M. 28

Figure 3.6 Left: original growth acceleration; Right: growth acceleration without first 5 measures. 30

Figure 3.7 (a)-(b) Unregistered growth acceleration data for 39 boys (blue dashed) and 54 girls (pink dashed) with cross-sectional mean superimposed. (c) Registered cluster 1 with 37 boys (blue dashed) and 3 girls (pink solid). (d) Registered cluster 2 with 51 girls (pink dashed) and 3 boys (blue solid). (e)-(f) Estimated warping functions for cluster 1 and cluster 2, respectively. 32

Figure 3.8 (a) Raw gene expression with cluster structure determined by the biologists. (b) Registered curves with 4 clusters. (c)-(f) Registered four clusters with their estimated mean curves superimposed. 36

Figure 4.1 Left: raw curves. Right: registered curves. 44

Figure 4.2 Left: first-order derivative curves. Right: registered curves. . . . 44

Figure 4.3 Left: 2nd order derivative curves. Right: registered curves. . . . 45

Figure 4.4 Left: raw curves and their derivatives. Right: registered curves. . 46

Figure 4.5 (a) A set of 24 simulated observations with 5 clusters. (b) Simulated data with phase variation removed, with superimposed posterior estimated mean curves (solid black). (c) True mean curves (gray) and estimated mean curves (black). (d) Estimated warping functions for all 5 clusters. 48

Figure 4.6 Left: cRate using DSM metric. Right: cRate using Euclidean metric. 49

Figure 4.7 (a)-(b) Unregistered growth acceleration data for 39 boys (blue dashed) and 54 girls (pink dashed) with cross-sectional mean superimposed. (c) Registered cluster 1 with 36 boys (blue dashed) and 3 girls (pink solid). (d) Registered cluster 2 with 46 girls (pink dashed) and 8 boys (blue dashed). (e)-(f) Estimated warping functions for cluster 1 and cluster 2, respectively. 54

Figure 4.8 Raw data of the response of human fibroblasts to serum. 55

Figure 4.9 The number of clusters versus the average silhouette width. . . . 55

Figure 4.10 (a) Raw data with cluster structure determined by the algorithm. (b) Registered curves with vertical shifts removed. (c)-(f) Registered four clusters with their estimated mean curves superimposed. 56

Figure 5.1 Lower bounds of the simulation study 69

Figure 5.2 (a) A set of 23 simulated observations with 4 clusters. (b) Simulated data with phase variation removed, with superimposed posterior estimated mean curves (solid black). (c) True mean curves (gray) and estimated mean curves (black). (d) Estimated warping functions for all 4 clusters. 70

Figure 5.3 Lower bound of AVB for Berkeley acceleration data 71

Figure 5.4 (a)-(b) Unregistered growth acceleration data for 39 boys (blue dashed) and 54 girls (pink dashed) with cross-sectional mean superimposed. (c) Registered cluster 1 with 36 boys (blue dashed) and 9 girls (pink solid). (d) Registered cluster 2 with 45 girls (pink dashed) and 3 boys (blue dashed). (e)-(f) Estimated warping functions for cluster 1 and cluster 2, respectively. 73

Figure 5.5 Left: Raw HFS data. Right: Raw HFS data with estimated membership. 74

Figure 5.6 Lower bound of AVB for HFS data 75

Figure 5.7 (a) Registered curves with vertical shifts removed. (b)-(f) Registered five clusters with their estimated mean curves superimposed. 76

Figure 5.8 Estimated warping functions for HFS data 77

Figure C.1 An example how to calculate $\frac{\partial}{\partial \gamma_{i5}} \gamma_i(t_j)$: the four cases are illustrated in the white, gray, purple, and red segments, respectively. The dotted vertical lines represent time $\mathbf{t}$. we approximate the warping function with $M = 20$. 91

Chapter 1

Introduction

Functional data analysis (FDA) extends existing statistical tools to data represented by curves or surfaces over time, space, or another domain. FDA techniques are widely used in the studies of gene expression, handwriting, and image data, among other applications. One advantage of functional data analysis over traditional multivariate analysis is the ability to examine higher-order derivatives of fitted functions. For example, the first-order derivative of a fitted monotone smoothing function [Ramsay and Silverman, 2005] measuring children's height over a given period represents the estimated growth velocity, and the second-order derivative is the estimated growth acceleration, etc. This dissertation focuses on the curves over a time domain, which are usually fitted by some basis function expansion. Throughout this paper, we use the B-spline basis [De Boor, 2001].

In some data sets, curves present similar patterns within subgroups, which requires a cluster analysis assigning observations that share similar characteristics into the same subgroup. After identifying the cluster structure, follow-up analysis usually focuses on two major variations among curves within a cluster: amplitude and phase variations [Ramsay and Silverman, 2005]. The amplitude variations characterize variations along the vertical direction over time, which consist of measurement error and departures from the underlying mean function. The phase variations are caused by the misalignment between the unobserved biological/mechanical clock and the chronological clock. A classic example is the Berkeley growth data [Tuddenham and Snyder, 1953]. The growth accelerations among girls and boys display similar

patterns, however, the growth peaks and valleys happen at very different ages. In the presence of phase variation, even a simple summary, such as the mean curve, may fail to capture the pattern of any individual curve. Thus, it is desirable to remove the phase variations for better statistical analysis. The process of eliminating phase variation is called registration in the literature [Ramsay and Silverman, 2005]. The phase variation is usually modeled by a warping function $h(\cdot)$ [Ramsay and Silverman, 2005], which is a non-decreasing continuous function defined on the time domain $\mathcal{T}$ satisfying the endpoint conditions $h(a) = a$, and $h(b) = b$, where a and b are two endpoints of the time domain. Figure 1.1 shows eight warping functions, while the bold dashed line is the 45° reference line representing an identity warp.

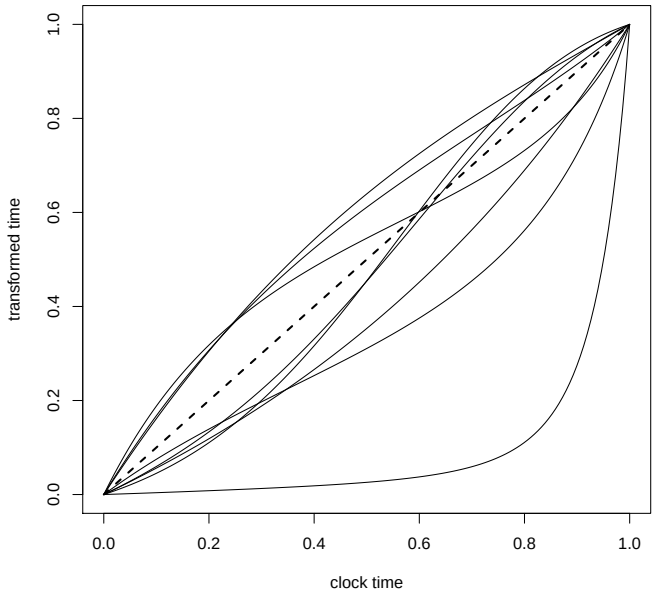

Figure 1.1 Examples of warping functions

In a clustering problem, the cluster structure is blurred by the effect of the time distortions, which should be eliminated for the purpose of clustering. However, when the phase variation in a curve depends on which cluster the curve belongs to, it is not feasible to estimate the warping functions without knowing the cluster memberships.

On the other hand, it is challenging to obtain a high clustering accuracy due to phase variation. The main focus of this dissertation is clustering functional observations in the presence of phase variation. In Section 2, clustering methods, registration methods and recent methods for joint clustering and registration are reviewed. In Chapter 3, we propose a Bayesian method for simultaneous registration and clustering of functional observations. In Chapter 4, we propose a distance-based method that takes advantage of our fast registration procedure in the previous chapter. In Chapter 5, we obtain a faster approach to inference for the model proposed in Chapter 3 by applying the variational Bayes method.

Chapter 2

Literature Review

Clustering functional observations involves grouping curves that share similar characteristics. Classic multivariate clustering methods, including the hierarchical agglomerative method, K-means method, and model-based method [Everitt et al., 2011] could be applied to the functional data by viewing each observation as a vector. The hierarchical agglomerative method starts with each observation belonging to its own cluster. At each step, two clusters with the closest distance are merged together, and the total number of clusters is reduced by 1. We stop the algorithm once the desired number of clusters is achieved. Typical distance measures between two clusters are single linkage, complete linkage, and average linkage. The K-means method starts with K centroids, which are the multidimensional means of the K clusters. Each observation is assigned to the cluster that has the shortest distance between its centroid and that observation. Then, the centroids are updated based on the current clusters. The algorithm iterates between last two steps until some convergence criterion is satisfied. Fraley and Raftery [2002] popularized model-based clustering. They assumed observations are sampled from a mixture of normal distributions, i.e., $x_i \sim \sum_{j=1}^{k} w_j N_p(\boldsymbol{\mu}_j, \Sigma_j)$. This model is also called a finite mixture model. The parameter estimation is carried out via the EM algorithm. However, these methods designed for multivariate data fail to capture the time dependency of the mean functions, which may result in poor clustering accuracy for functional data.

Recently, several methods have been developed for clustering functional data. Luan and Li [2003] used mixed-effect models for time-course gene expression. For the

i-th gene expression measured at time t_{ij}, the proposed model is

$$Y_i(t_{ij}) = \sum_{l=1}^{p} \beta_l^{(c)} \bar{B}_l(t_{ij}) + \sum_{l=1}^{q} \gamma_{il} B_l(t_{ij}) + \epsilon_{ij},$$

where the first term models the mean curve of cluster c, and the second term is the random effect. Both fixed and random effects are fitted via B-spline basis function expressions. The parameter estimation and the posterior cluster probabilities are calculated via the EM algorithm. This mixed-effect model is a special case of the model proposed by James and Sugar [2003]. They proposed a finite mixture model of the form

$$\mathbf{Y}_i = S_i(\boldsymbol{\lambda}_0 + \Lambda \boldsymbol{\alpha}_{\mathbf{Z}_i} + \boldsymbol{\gamma}_i) + \boldsymbol{\epsilon}_i,$$

where the error term $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, R)$, and random effect $\boldsymbol{\gamma}_i \sim N(\mathbf{0}, \Gamma)$. The parameter estimation is obtained via the EM algorithm.

Motivated by an application in epidemiology, Dunson and Herring [2006] proposed a semiparametric model using a finite Dirichlet process mixture of the form

$$y_i = \eta_i + \epsilon_i,$$

where $\eta \sim G(\cdot) = \sum_{h=1}^{k} p_h \delta_{\Theta_h}(\cdot)$, and the mean trajectory $\Theta_h \sim GP(\mathcal{C}_{\kappa_h})$ follows a Gaussian process prior. The parameter estimates are obtained via Markov chain Monte Carlo (MCMC).

As an important preprocessing step, registration eliminates the phase variation. A simple and intuitive way of registering data is shift registration [Ramsay and Silverman, 2005]. Assuming that each sample function x_i is defined beyond the interval $[T_1, T_2]$ on which the sample functions are taken, we shift variable t horizontally, i.e., the registration is of the form

$$x_i^*(t) = x_i(t + \delta_i),$$

where δ_i is a parameter that aligns function x_i.

The shifting parameter δ_i is estimated by an iterative method called the Procrustes method. We define a global measurement of the goodness of registration by the sum of squared error as

$$\texttt{REGSSE} = \sum_{i=1}^{N} \int_{\mathcal{T}} [x_i(t+\delta_i) - \hat{\mu}(t)]^2 ds,$$

where $\mathcal{T}$ is the interval of registration, and $\hat{\mu}(t)$ is the overall mean function which can be evaluated by a smoothing method, possibly including a roughness penalty. The iterative procedure is described as follows. Starting with the original data, in each step, we calculate δ_i to minimize `REGSSE` by the Newton-Raphson algorithm and update $\hat{\mu}(t)$, $x_i(t+\delta_i)$, and `REGSSE`, repeating the procedure until some convergence criterion is satisfied.

Ramsay and Li [1998] proposed a warping function of the form

$$h(t) = C_1\{D^{-1}\exp(D^{-1}w)\}(t),$$

where $D^{-1}\exp$ is the monotonization operator, which guarantees the monotonicity of the warping function. The function w is formed by B-spline basis functions for flexibility. The parameter estimation is carried out via minimizing the penalized squared error criterion

$$F_\lambda(y, x|h) = \int ||y(t) - x[h(t)]||^2 \, \mathrm{d}t + \lambda \int w^2(t) \, \mathrm{d}t.$$

The magnitude of parameter λ determines the smoothness of the warping function.

Ramsay and Silverman [2005] developed a similar model, in which the warping function is

$$h(t) = C_0 + C_1 \int_0^t \exp W(u) \, \mathrm{d}u,$$

where W is an unconstrained function, which can be expressed by a set of a B-spline functions for example. The monotonicity is achieved by an integral over an exponential function. To estimate W, a continuous fitting criterion is used. They

define

$$\mathbf{T}(h) = \begin{bmatrix} \int \{x_0(t)\}^2 \, \mathrm{d}t & \int x_0(t)x[h(t)] \, \mathrm{d}t \\ \int x_0(t)x[h(t)] \, \mathrm{d}t & \int \{x[h(t)]\}^2 \, \mathrm{d}t \end{bmatrix}.$$

Using principal component analysis, the registration is complete if the smaller eigenvalue, which measures departures from unidimensionality, is 0. A roughness penalty term is used in conjunction with the minimal eigenvalue criterion to ensure the smoothness.

Recently, Cheng et al. [2015] developed a Bayesian method for 1D curve and 2D image registration. The key idea is to model the warping functions by a discrete approximation generated from the family of Dirichlet distributions. Let $(\gamma_1, \ldots, \gamma_M)$ be a vector of realizations from a Dirichlet distribution; this vector satisfies $\gamma_i > 0$ for $i = 1, \ldots, M$ and $\sum_i \gamma = 1$. It follows that the linear interpolation of the cumulative sum is strictly monotone increasing in $[0, 1]$. Without loss of generality, we can map the original time domain into $[0, 1]$, and register the curves on $[0, 1]$. The advantage of such a discrete approximation is its simple formulation and great flexibility.

Other recently developed Bayesian registration methods include Earls and Hooker [2015]. They model both the mean curves and warping functions by Gaussian processes. The warping function $h_i(t) = t_i + \int_{t_1}^{t} e^{w_i(s)}$, where $w_i(t)$ follows a Gaussian process distribution. The proposed model is

$$X_i(h_i(t))|z_{0i}, z_{1i}, f(t) \sim GP(z_{0i} + z_{1i}f(t), \gamma_R^{-1}\Sigma(s,t)) \qquad s, t \in \mathcal{T}.$$

The details of how to model the covariance matrix are given in Earls and Hooker [2014]. They also proposed fast inference via a modified variational Bayes method [Ormerod and Wand, 2010], which they refer to as adapted variational Bayes. When the prior of w_i is non-conjugate, they directly maximize the log-likelihood function with respect to the variables determining the warping function to obtain the optimal value in the current iteration without deriving the approximated distribution.

Recently, several papers have tackled the problem of joint clustering and registration. Liu and Yang [2009] proposed the SACK model, which is capable of clustering functional data when a simple time translation is presented. The proposed model is $y_{ij} = d_i + \sum_{l=1}^{L} \beta_l B_l(b_i + t_{ij}) + \epsilon_{ij}$, where d_i is the amplitude variable and b_i is the time translation variable. They translate the shift in the time domain into variation in the measurement space by a first-order Taylor expansion on the B-spline basis functions. The transformed model is $y_{ij} = d_i + \sum_{l=1}^{L} \beta_l (B_l(t_{ij}) + b_i B_l'(t_{ij})) + \epsilon_{ij}$, which is a mixture model. The conditional cluster probabilities are calculated via the EM algorithm. They use pBIC, a modified BIC, for model selection, since the regularity conditions of BIC do not hold for the mixture model.

Also assuming a simple time translation, Sangalli et al. [2010] proposed an iterative method based on a dissimilarity measure called k-mean alignment, which iterates among a template identification step, an alignment and cluster step, and a normalization step until convergence.

To handle more realistic scenarios under arbitrary time warpings, Tang and Müller [2009] propose a method based on pairwise warping. For observation i and k, the pairwise warping function is a composition of two individual warping functions defined as $g_{ik}(t) = h_i(h_k^{-1}(t))$. The pairwise warping function is estimated via minimizing

$$C_\lambda(Y_i, Y_k, g) = E\left\{\int_{\mathcal{T}} (Y_i(g(t)) - Y_k(t))^2 + \lambda(g(t) - t)^2 \, \mathrm{d}t\right\},$$

which essentially minimizes the L^2 distance between unaligned observation k and time-transformed observation i with curve k as its template. To avoid extreme time distortion and solve the identifiability issue, the warping function is estimated by

$$\hat{h}_i^{-1} = \frac{1}{\sum_{i=1}^{n} \mathbf{1}_{\{d_{pw}(i,k) < d_0\}}} \sum_{i=1}^{n} \tilde{g}_{ik}(t) \mathbf{1}_{\{d_{pw}(i,k) < d_0\}},$$

which is based on the assumption that $E(h_i(t)) = t$ and the L^2 distance between curves from two clusters is relatively large. Note that d_0 is a threshold that determines the pairs used in estimating $h_i(\cdot)$. However, this method assumes the mean curves in

different clusters are well separated vertically to some degree, a condition potentially too strong for some applications.

Zhang and Telesca [2014] proposed a model with flexible warping functions of the form

$$y_i(t) = c_i + a_i m_i\{\mu_i(t, \boldsymbol{\phi}_i), \theta_i\} + \epsilon_i(t),$$

where $\mu_i(\cdot)$ is the curve-specific warping function modeled by a B-spline basis expansion with restricted coefficients to guarantee the monotonicity. The parameter c_i accounts for vertical shift, and a_i serves as a stretching/shrinking factor. They model the B-spline coefficients of the i-th mean curve $\boldsymbol{\theta}_i$ by a Dirichlet process mixture. Consequently, the number of clusters K is determined implicitly.

Chapter 3

A Bayesian Method for Simultaneous Registration and Clustering of Functional Observations

We develop a Bayesian method that simultaneously registers and clusters functional data of interest. Unlike other existing methods, which often assume a simple translation in the time domain, our method uses a discrete approximation generated from the family of Dirichlet distributions to allow warping functions of great flexibility. Under this Bayesian framework, a MCMC algorithm is proposed for posterior sampling. We demonstrate this method via simulation studies and applications to growth curve data and cell cycle regulated yeast genes.

3.1 Model Assumption

In a functional dataset, we assume that there are N objects, on which we take K measurements over time. Given a certain number of repeated measurements, we may model the response trajectory as a function of time using some basis (such as splines) in the context of functional data analysis.

We assume that each observation is composed of a signal function and random error terms, that is,

$$\mathbf{Y} = af(\mathbf{t}) + \boldsymbol{\epsilon},$$

where $a \in \mathbb{R}^+$ is a stretching/shrinking factor [Zhang and Telesca, 2014], $f(\mathbf{t})$ is the set of underlying responses at the vector of time points $\mathbf{t}$, and $\boldsymbol{\epsilon}$ is an i.i.d. $N(0, \sigma^2)$

error vector.

When our observed data must be aligned, we model the effect of the warping function associated with $\mathbf{Y}$ as $\mathbf{Y} = f[h(\mathbf{t})] + \boldsymbol{\epsilon}$, where h is the underlying warping function, and therefore,

$$\mathbf{Y}|\boldsymbol{\beta}, \gamma, \sigma^2, a \sim \text{MVN}(af[h(\mathbf{t})], \sigma^2\mathbf{I}).$$

For the purpose of clustering, we introduce notation for different groups. For a fixed number of clusters C, we use the vector $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iC})$ to denote the cluster membership for the i-th observation. Note that only one element of $\boldsymbol{z}_i$ equals 1 and the rest all equal 0. Throughout this paper, we will use B-splines with q basis functions to model the signal curve. It follows that for a K-dimensional observation $\mathbf{Y}$, we have $f(\mathbf{t}) \approx \boldsymbol{\phi}(\mathbf{t})\boldsymbol{\beta}$, where $\boldsymbol{\phi}$ is a $K \times q$ matrix of coefficients of the B-spline basis evaluated at each time point. To be more specific,

$$\boldsymbol{\phi}(\mathbf{t}) = \begin{pmatrix} \phi_1(t_1) & \phi_2(t_1) & \ldots & \phi_q(t_1) \\ \phi_1(t_2) & \phi_2(t_2) & \ldots & \phi_q(t_2) \\ \ldots & \ldots & \ldots & \ldots \\ \phi_1(t_K) & \phi_2(t_K) & \ldots & \phi_q(t_K) \end{pmatrix},$$

where $\phi_i(\cdot)$ denotes the i-th B-spline basis function, and $\boldsymbol{\beta}$ is a vector of B-spline coefficients. We use the same basis functions and assume the same variance σ^2 across all groups. Let $\boldsymbol{\beta}_i$ denote the spline coefficients for the i-th group, $i = 1, 2, \ldots, C$. The discretized mean curve for the i-th cluster is represented as $\boldsymbol{\mu}_i \approx \boldsymbol{\phi}[\gamma(\mathbf{t})]\boldsymbol{\beta}_i$, where $\gamma(\cdot)$ is the discrete approximation of the corresponding warping function h, which will be discussed in the next section.

Prior Distributions on Parameters

To estimate the warping function h_i for the i-th observation, a discrete approximation generated by a Dirichlet distribution is utilized [Cheng et al., 2015]. Without loss of generality, let us assume that the time domain $\mathcal{T} = [0, 1]$. Any general time domain $[T_1, T_2]$ may be converted into $[0, 1]$ by the transformation $g(t) = (t - T_1)/(T_2 - T_1)$. Let $\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iM} \sim Dir(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is a M-vector of positive parameters.

For the Dirichlet distribution, we have $\sum_j \gamma_{ij} = 1$, which suggests that the linear interpolation of the cumulative sum over γ_{ij} can serve as a discrete approximation of the continuous warping h_i. The parameter M controls the smoothness of the approximation. A large M results in a smoother approximation, but more computational burden.

The hyperparameter $\boldsymbol{\alpha}$ can be chosen to affect the "concentration" of the warping functions relative to the 45° reference line, which corresponds to no warping. Small values in $\boldsymbol{\alpha}$ allow more variability in each step of the approximation, and vice versa. Figure 3.1 shows two sets of discrete warping functions, each with 20 jumps, generated from $Dir(0.8, 0.8, \ldots, 0.8)$, and $Dir(5, 5, \ldots, 5)$, respectively.

If observation i is assigned to cluster j, then the cluster membership indicator $\boldsymbol{z}_i$ is a vector of size C containing a 1 in the j-th position and 0 elsewhere. We model $\boldsymbol{z}_i$ with a multinomial distribution, i.e., $\boldsymbol{z}_i \sim Multi\ (1, (p_1, \ldots, p_C))$, where $p_1, \ldots, p_C$ are the membership probabilities satisfying $\sum_j p_j = 1$. We choose a conjugate Dirichlet prior for those probabilities; i.e., $p_1, \ldots, p_C \sim Dir(\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is a vector of hyperparameters.

For the i-th cluster, we assume that $\boldsymbol{\beta}_i \sim \text{MVN}(\boldsymbol{\beta}_{0i}, \Gamma)$. It will be seen later that the full conditional distribution of $\boldsymbol{\beta}_i$ is still multivariate normal. We model the precision parameter $\tau = 1/\sigma^2$ with a (conjugate) gamma prior, i.e., $\tau \sim Gamma(\kappa, \theta)$.

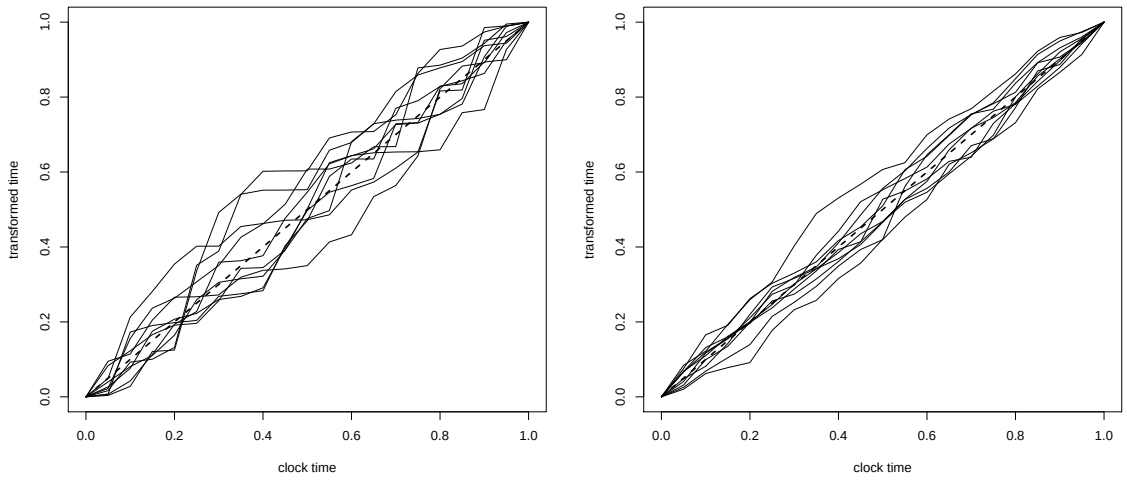

Figure 3.1 Left: Warping functions from $Dir(0.8, 0.8, \ldots, 0.8)$. Right: Warping functions from $Dir(5, 5, \ldots, 5)$.

For functional observations, one possible source of amplitude variation is composed of vertical shifts among observations in the same cluster. The left panel in Figure 3.2 shows a set of simulated observations from the same cluster with phase variations; the right panel shows the same observations with additional vertical shifts following $Unif\,(-0.5, 0.5)$. The bold curve is the true signal function generating the observations.

Our prior model assumes the vertical shift S_i for the i-th observation is $Unif\,(-\phi, \phi)$ for some positive ϕ. On the stretching/shrinking factors a_i, we place independent $N(1, \sigma_a^2)$ priors, $i = 1, 2, \ldots, N$.

Likelihood and Posterior of the Model

Under the preceding model assumptions, for a vector of measurements taken on the same functional observation, we have

$$\mathbf{Y} = a\boldsymbol{\phi}[\gamma(\mathbf{t})]\boldsymbol{\beta} + \mathbf{S} + \boldsymbol{\epsilon},$$

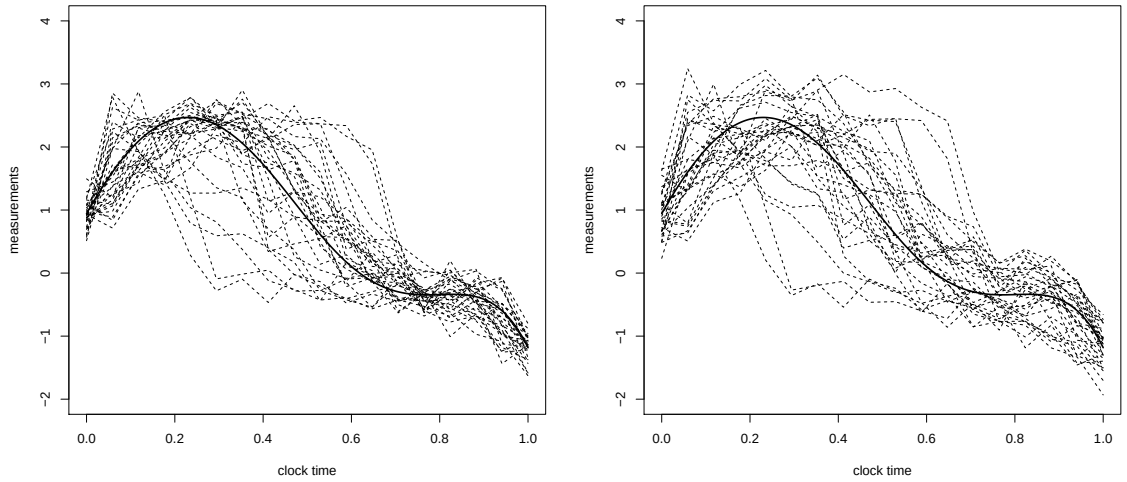

Figure 3.2 Left: Simulated data with phase variations. Right: Simulated data with additional vertical shifts generated from *Unif* $(-0.5, 0.5)$.

where $\mathbf{S} = S \otimes \mathbf{1}$ ($\otimes$ is the Kronecker product) is a vector of size K containing the same vertical shifts. Hence, the distribution of the i-th observation $\boldsymbol{y}_i$ belonging to a specific cluster in the presence of phase variation is given by

$$\mathbf{Y}_i|\boldsymbol{\beta}, \gamma_i, \boldsymbol{z}_i, \tau, s \sim \text{MVN}\left(a_i \boldsymbol{\phi}[\gamma_i(\mathbf{t})] \prod_{c=1}^{C} \boldsymbol{\beta}_c^{z_{ic}} + \mathbf{s}_i, \tau^{-1}\mathbf{I}\right).$$

With the above prior distributions on the parameters, the joint distribution of the data and parameters is

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_C,\gamma_1,\ldots,\gamma_N,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_N,p_1,\ldots,p_C,\tau,s_1,\ldots,s_N,a_1,\ldots,a_N,\\
&\quad \boldsymbol{y}_1,\ldots,\boldsymbol{y}_N)\\
=\;& \prod_{i=1}^{N}\mathcal{P}(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{z}_i,\gamma_i,p_1,\ldots,p_C,\tau,s_1,\ldots,s_N,a_1,\ldots,a_N)\prod_{c=1}^{C}\mathcal{P}(\boldsymbol{\beta}_c|\boldsymbol{\beta}_0^c,\Gamma)\prod_{i=1}^{N}\mathcal{P}(\gamma_i|\boldsymbol{\alpha})\\
&\prod_{i=1}^{N}\mathcal{P}(\boldsymbol{z}_i|p_1,\ldots,p_C)\mathcal{P}(p_1,\ldots,p_C|\boldsymbol{\eta})\mathcal{P}(\tau|\kappa,\theta)\prod_{i=1}^{N}\mathcal{P}(s_i|\phi)\prod_{i=1}^{N}\mathcal{P}(a_i|\sigma_a^2)\\
\propto\;& \prod_{i=1}^{N}\tau^{K/2}\exp\left\{-\frac{1}{2}\tau\left[\boldsymbol{y}_i-a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\prod_{c=1}^{C}\boldsymbol{\beta}_c^{z_{ic}}-\mathbf{s}_i\right]'\left[\boldsymbol{y}_i-a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\prod_{c=1}^{C}\boldsymbol{\beta}_c^{z_{ic}}-\mathbf{s}_i\right]\right\}\\
&\prod_{c=1}^{C}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_c-\boldsymbol{\beta}_0^c)'\Gamma^{-1}(\boldsymbol{\beta}_c-\boldsymbol{\beta}_0^c)\right\}\prod_{i=1}^{N}\prod_{m=1}^{M}\gamma_{im}^{\alpha_m-1}\prod_{i=1}^{N}\prod_{c=1}^{C}p_c^{z_{ic}}\prod_{c=1}^{C}p_c^{\eta_c-1}\\
&\tau^{\kappa+1}\exp\{-\tau\theta\}\prod_{i=1}^{N}\mathbf{1}_{\{-\phi<s_i<\phi\}}\prod_{c=1}^{N}\exp\left\{-\frac{1}{2}(a_i-1)^2\right\}\\
\propto\;& \tau^{KN/2}\exp\left\{-\frac{1}{2}\tau\sum_{i=1}^{N}\left\|\boldsymbol{y}_i-a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\prod_{c=1}^{C}\boldsymbol{\beta}_c^{z_{ic}}-\mathbf{s}_i\right\|^2\right\}\\
&\prod_{c=1}^{C}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_c-\boldsymbol{\beta}_0^c)'\Gamma^{-1}(\boldsymbol{\beta}_c-\boldsymbol{\beta}_0^c)\right\}\\
&\prod_{i=1}^{N}\prod_{m=1}^{M}\gamma_{im}^{\alpha_m-1}\prod_{c=1}^{C}p_c^{\sum_{i=1}^{N}z_{ic}+\eta_c-1}\tau^{\kappa-1}\exp\{-\tau\theta\}\\
&\prod_{i=1}^{N}\mathbf{1}_{\{-\phi<s_i<\phi\}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{N}(a_i-1)^2\right\}.
\end{aligned}
$$

This joint distribution will be used to obtain the relevant full conditional distributions for the MCMC algorithm.

Due to the complexity of the proposed model, an analytical posterior derivation is intractable, so our inference is based on MCMC sampling of the posterior distribution. At iteration t, the MCMC algorithm is as follows:

- **Gibbs Sampling for Cluster Membership $\boldsymbol{z}_i$**

  The full conditional distribution of $\boldsymbol{z}_i$ is

$$
\mathcal{P}(\boldsymbol{z}_i|\text{rest}) \;\propto\; \exp\left\{-\frac{1}{2}\tau^{[t-1]}\left\|\boldsymbol{y}_i-a_i^{[t-1]}\boldsymbol{\phi}[\gamma_i^{[t-1]}(\mathbf{t})]\prod_{c=1}^{C}(\boldsymbol{\beta}_c^{[t-1]})^{z_{ic}}-\mathbf{s}_i^{[t-1]}\right\|^2\right\}
\prod_{c=1}^{C}(p_c^{[t-1]})^{z_{ic}}.
$$

The cluster membership indicator vector is discrete and follows a multinomial distribution. The probability of belonging to the j-th cluster is proportional to

$$\exp\left\{-\frac{1}{2}\tau^{[t-1]}\left\|\boldsymbol{y}_i - a_i^{[t-1]}\boldsymbol{\phi}[\gamma_i^{[t-1]}(\mathbf{t})]\ \boldsymbol{\beta}_j^{[t-1]} - \mathbf{s}_i^{[t-1]}\right\|^2\right\} p_j^{[t-1]}.$$

Let us denote the above quantity by q_j. We have

$$\boldsymbol{z}_i|\text{rest} \sim multi\left(\frac{q_1}{\sum_j q_j}, \ldots, \frac{q_C}{\sum_j q_j}\right).$$

- **Gibbs Sampling for Cluster Probabilities** $p_1, \ldots, p_C$

  After updating the cluster membership, the full conditional distribution of the probabilities is

$$\begin{aligned}\mathcal{P}(p_1, \ldots, p_C|\text{rest}) &\propto \prod_{i=1}^{N}\prod_{c=1}^{C} p_c^{z_{ic}^{[t]}} \prod_{c=1}^{C} p_c^{\eta_c - 1} \\ &\propto \prod_{c=1}^{C} p_c^{\sum_{i=1}^{N} z_{ic}^{[t]} + \eta_c - 1}.\end{aligned}$$

  It follows that

$$p_1, \ldots, p_C|\text{rest} \sim Dir\left(\sum_{i=1}^{N} z_{i1}^{[t]} + \eta_1, \ldots, \sum_{i=1}^{N} z_{iC}^{[t]} + \eta_C\right).$$

- **Metropolis-Hastings Algorithm for Sampling Warping** γ_i

  We update $\gamma_{i1}, \ldots, \gamma_{iM-1}$. The two endpoints satisfy the conditions $\gamma_{i0} = 0$, and $\gamma_{iM} = 1 - \sum_{j=1}^{M-1}\gamma_{ij}$, because of the constraints of the warping function, and hence are not involved in the updating procedure. After updating the $\boldsymbol{z}_i$, we propose a value of γ_{ij}^* from a truncated normal with mean $\gamma_{ij}^{[t-1]}$ and variance σ_γ^2 on $[0, \gamma_{iM} + \gamma_{ij}]$ to guarantee a positive γ_{ij}^* and γ_{iM}^*. We accept the proposed value with probability

$$\lambda = \min\left\{1, \frac{\exp\left\{-\frac{1}{2}\tau^{[t-1]}\left\|\boldsymbol{y}_i - a_i^{[t-1]}\boldsymbol{\phi}[\gamma_i^{*(j)}(\mathbf{t})]\prod_{c=1}^{C}\boldsymbol{\beta}_c^{[t-1]z_{ic}^{[t]}} - \mathbf{s}_i^{[t-1]}\right\|^2\right\}}{\exp\left\{-\frac{1}{2}\tau^{[t-1]}\left\|\boldsymbol{y}_i - a_i^{[t-1]}\boldsymbol{\phi}[\gamma_i^{(j-1)}(\mathbf{t})]\prod_{c=1}^{C}\boldsymbol{\beta}_c^{[t-1]z_{ic}^{[t]}} - \mathbf{s}_i^{[t-1]}\right\|^2\right\}} \times \right.$$
$$\left. \frac{(\gamma_{ij}^*)^{\alpha_j-1}(\gamma_{iM}^*)^{\alpha_M-1}\left[\Phi\left(\frac{r_{ij}^{[t]}-\gamma_{ij}^*}{\sigma_\gamma}\right) - \Phi\left(\frac{-\gamma_{ij}^*}{\sigma_\gamma}\right)\right]}{(\gamma_{ij}^{[t-1]})^{\alpha_j-1}(\gamma_{iM}^{[t-1]})^{\alpha_M-1}\left[\Phi\left(\frac{r_{ij}^{[t]}-\gamma_{ij}^{[t-1]}}{\sigma_\gamma}\right) - \Phi\left(\frac{-\gamma_{ij}^{[t-1]}}{\sigma_\gamma}\right)\right]}\right\},$$

where $\gamma_i^{(j)}$ is the warping function with the jump updated through the j-th element, and Φ is the standard normal CDF.

- **Gibbs Sampling for Spline Coefficients $\boldsymbol{\beta}_k$**

  After updating the γ_i's and $\boldsymbol{z}_i$'s, we use a superscript as the updated membership indicator. For example, $\boldsymbol{y}_i^{(k)}$ signifies that we classify observation $\boldsymbol{y}_i$ into group k. Furthermore, let $n_k^{[t]}$ denote the size of group k at the current iteration. The full conditional of $\boldsymbol{\beta}_k$ is given by

$$\begin{aligned}
\mathcal{P}(\boldsymbol{\beta}_k|\text{rest}) \propto\ & \exp\left\{-\frac{1}{2}\tau^{[t-1]}\sum_{l=1}^{n_k^{[t]}}\left\|\boldsymbol{y}_l^{(k)} - a_i^{[t-1]}\boldsymbol{\phi}[\gamma_l^{[t]}(\mathbf{t})]\boldsymbol{\beta}_k - \mathbf{s}_l^{[t-1]}\right\|^2\right\} \\
& \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{0k})'\Gamma^{-1}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{0k})\right\} \\
\propto\ & \exp\left\{-\frac{1}{2}\boldsymbol{\beta}_k'\underbrace{\left(\tau^{[t-1]}\sum_{l=1}^{n_k^{[t]}}\left[\left(a_l^{[t-1]}\right)^2\boldsymbol{\phi}'[\gamma_l^{[t]}(\mathbf{t})]\boldsymbol{\phi}[\gamma_l^{[t]}(\mathbf{t})]\right] + \Gamma^{-1}\right)}_{\text{call it } \mathbf{A}_k}\boldsymbol{\beta}_k - \right. \\
& \left.\boldsymbol{\beta}_k'\underbrace{\left(\tau^{[t-1]}\sum_{l=1}^{n_k^{[t]}}a_l^{[t-1]}\boldsymbol{\phi}'[\gamma_l^{[t]}(\mathbf{t})](\boldsymbol{y}_l^{(k)} - \mathbf{s}_l^{[t-1]}) + \Gamma^{-1}\boldsymbol{\beta}_{0k}\right)}_{\text{call it } \mathbf{c}_k}\right\} \\
\propto\ & \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_k - \mathbf{A}_k^{-1}\mathbf{c}_k)'\mathbf{A}_k(\boldsymbol{\beta}_k - \mathbf{A}_k^{-1}\mathbf{c}_k)\right\}.
\end{aligned}$$

  Therefore,

$$\boldsymbol{\beta}_k|\text{rest} \sim \text{MVN}(\mathbf{A}_k^{-1}\mathbf{c}_k, \mathbf{A}_k^{-1}).$$

- **Gibbs Sampling for Precision τ**

  After updating the γ_i's, $\boldsymbol{z}_i$, and $\boldsymbol{\beta}_k$'s, the full conditional distribution of τ is given by

$$\begin{aligned}
& \mathcal{P}(\tau|\text{rest}) \\
\propto\ & \tau^{KN/2} \exp\left\{-\frac{1}{2}\tau \sum_{i=1}^{N} \left\| \boldsymbol{y}_i - \prod_{c=1}^{C} \left[a_i^{[t-1]} \boldsymbol{\phi}(\gamma_i^{[t]}(\mathbf{t}))\boldsymbol{\beta}_c^{[t]}\right]^{z_{ic}^{[t]}} - \mathbf{s}_i^{[t-1]} \right\|^2 \right\} \tau^{\kappa-1} \\
& \exp\{-\tau\theta\} \\
\propto\ & \tau^{\frac{KN}{2}+\kappa-1} \exp\left\{-\tau\left(\frac{1}{2}\sum_{i=1}^{N} \left\| \boldsymbol{y}_i - \prod_{c=1}^{C} \left[a_i^{[t-1]} \boldsymbol{\phi}(\gamma_i^{[t]}(\mathbf{t}))\boldsymbol{\beta}_c^{[t]}\right]^{z_{ic}^{[t]}} - \mathbf{s}_i^{[t-1]} \right\|^2 + \theta\right)\right\}.
\end{aligned}$$

  It follows that

$$\tau|\text{rest} \sim Gamma\left(\frac{KN}{2}+\kappa, \frac{1}{2}\sum_{i=1}^{N} \left\| \boldsymbol{y}_i - \prod_{c=1}^{C} \left[a_i^{[t-1]} \boldsymbol{\phi}(\gamma_i^{[t]}(\mathbf{t}))\boldsymbol{\beta}_c^{[t]}\right]^{z_{ic}^{[t]}} - \mathbf{s}_i^{[t-1]} \right\|^2 + \theta\right).$$

- **Gibbs Sampling for Vertical Shift** S_i

  After updating the γ_i's, $\boldsymbol{z}_i$, $\boldsymbol{\beta}_k$'s, and τ, the full conditional distribution of S_i is given by

$$\begin{aligned}
& \mathcal{P}(s_i|\text{rest}) \\
\propto\ & \exp\left\{-\frac{1}{2}\tau^{[t]} \left\| \boldsymbol{y}_i - \prod_{c=1}^{C} \left[a_i^{[t-1]} \boldsymbol{\phi}(\gamma_i^{[t]}(\mathbf{t}))\boldsymbol{\beta}_c^{[t]}\right]^{z_{ic}^{[t]}} - \mathbf{s}_i \right\|^2 \right\} \mathbf{1}_{\{-\phi<s_i<\phi\}}.
\end{aligned}$$

  To simplify the notation, let us define d_l as the l-th element of the vector $\boldsymbol{y}_i - \prod_{c=1}^{C} \left[a_i^{[t-1]} \boldsymbol{\phi}(\gamma_i^{[t]}(\mathbf{t}))\boldsymbol{\beta}_c^{[t]}\right]^{z_{ic}^{[t]}}$. The posterior then is

$$\begin{aligned}
\mathcal{P}(s_i|\text{rest}) &\propto \exp\left\{-\frac{1}{2}\tau^{[t]} \sum_{l=1}^{K} (s_i - d_l)^2\right\} \mathbf{1}_{\{-\phi<s_i<\phi\}} \\
&\propto \exp\left\{-\frac{1}{2}\tau^{[t]} \sum_{l=1}^{K} (s_i^2 - 2d_l s_i)\right\} \mathbf{1}_{\{-\phi<s_i<\phi\}} \\
&\propto \exp\left\{-\frac{1}{2}\tau^{[t]} K (s_i - \sum_{l=1}^{K} d_l/K)^2\right\} \mathbf{1}_{\{-\phi<s_i<\phi\}}
\end{aligned}$$

The normal kernel indicates that the posterior distribution of the vertical shift S_i is a truncated normal with mean $\sum_{l=1}^{K} d_l/K$, and variance $1/(\tau^{[t]}K)$, i.e.,

$$S_i|\text{rest} \sim N\left(\frac{\sum_{l=1}^{K} d_l}{K}, \frac{1}{\tau^{[t]}K}\right)\mathbf{1}_{\{-\phi<s_i<\phi\}}.$$

Note that for a point estimate of these shifts, we simply require $\sum_i s_i = 0$ to ensure identifiability.

- **Gibbs Sampling for Stretching/Shrinking Factor** a_i

  After updating the γ_i's, $\boldsymbol{z}_i$, $\boldsymbol{\beta}_k$'s, τ, and s_i's, the full conditional distribution of a_i is given by

$$\begin{aligned}
&\mathcal{P}(a_i|\text{ rest}) \\
\propto\ &\exp\left\{-\frac{1}{2}\tau^{[t]}\left\|\boldsymbol{y}_i - \prod_{c=1}^{C}\left[a_i\boldsymbol{\phi}(\gamma_i^{[t]}(\mathbf{t}))\boldsymbol{\beta}_c^{[t]}\right]^{z_{ic}^{[t]}} - \mathbf{s}_i^{[t]}\right\|^2\right\}\exp\left\{-\frac{1}{2\sigma_a^2}(a_i-1)^2\right\}.
\end{aligned}$$

  For economy of notation, we denote the l-th element of $\prod_{c=1}^{C}[\boldsymbol{\phi}(\gamma_i^{[t]}(\mathbf{t}))\boldsymbol{\beta}_c^{[t]}]^{z_{ic}^{[t]}}$ and $\boldsymbol{y}_i$ by $\mu_{il}^{[t]}$ and y_{il}, respectively. The posterior becomes

$$\begin{aligned}
&\mathcal{P}(a_i|\text{ rest}) \\
\propto\ &\exp\left\{-\frac{1}{2}\tau^{[t]}\left[\sum_{l=1}^{K} a_i^2(\mu_{il}^{[t]})^2 - \sum_{l=1}^{K} 2a_i\mu_{il}^{[t]}(y_{il}-s_i^{[t]})\right]\right\}\exp\left\{-\frac{1}{2\sigma_a^2}a_i^2 + \frac{1}{\sigma_a^2}a_i\right\} \\
\propto\ &\exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma_a^2} + \tau^{[t]}\sum_{l=1}^{K}(\mu_{il}^{[t]})^2\right]a_i^2 + \left[\tau^{[t]}\sum_{l=1}^{K}\mu_{il}^{[t]}(y_{il}-s_i^{[t]}) + \frac{1}{\sigma_a^2}\right]a_i\right\}.
\end{aligned}$$

  By completing the square, we have

$$a_i|\text{ rest} \sim N\left(\frac{\tau^{[t]}\sum_{l=1}^{K}\mu_{il}^{[t]}(y_{il}-s_i^{[t]}) + 1/\sigma_a^2}{1/\sigma_a^2 + \tau^{[t]}\sum_{l=1}^{K}(\mu_{il}^{[t]})^2}, \frac{1}{1/\sigma_a^2 + \tau^{[t]}\sum_{l=1}^{K}(\mu_{il}^{[t]})^2}\right).$$

From experimentation using various simulated data, we note two concerns: (1) The posterior cluster memberships converge quickly usually after several hundred iterations, and barely change afterwards; (2) the "converged" cluster memberships depend on the initial values of the Markov chain. These phenomena are partially due to the fact that the misclassified observations affect the posterior sampling of

coefficients $\boldsymbol{\beta}$, and cluster memberships are in turn influenced by those coefficients in the next iteration.

We need to "force" the individual curves to accept new group membership from time to time to avoid the vicious circle described above. We adjust the sampling algorithm in the following way: In the burn-in stage, every I iterations, $p\%$ of the curves in each group switch clusters at random (for practical purposes, we recommend $I = 10$ to 100, $p = 3$ to 15). We make these switches only in the burn-in stage, and thus we use an ordinary MCMC algorithm afterward with the initial values obtained from the burn-in stage. This switch reduces the influence of initial values. Should the switch result in a poorer clustering, we note based on experimentation that the chain can adjust itself and is likely to recover individual classifications of the previous partitions that were correct.

Our proposed method can cluster observations under nonlinear time distortion and vertical shifting and does not require or estimate any template for the purpose of registration.

3.3 Choosing the Number of Clusters

Determining the number of clusters is a common problem in cluster analysis. A wide variety of solutions have been proposed. The "elbow criterion" examines the percentage of variation explained as a function of the number of clusters, with the number of clusters chosen where when the plot levels off. The variance ratio criterion [Caliński and Harabasz, 1974] chooses the number of clusters which maximizes the ratio of the between-cluster and the within-cluster sum-of-squares. For model-based clustering methods, information criteria such as AIC [Akaike, 1974] and BIC [Schwarz et al., 1978] are frequently employed as a measure of clustering quality. These information-theoretic approaches are based essentially on the log-likelihood and penalize the number of parameters in the model.

For our method, it is simple to calculate the log-likelihood for a given cluster number C^* at each iteration. Recall that

$$\mathbf{Y}_i|\boldsymbol{\beta}, \gamma_i, \boldsymbol{z}_i, \tau, s_i, a_i \sim \text{MVN}\left(a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\prod_{c=1}^{C}\boldsymbol{\beta}_c^{z_{ic}} + \mathbf{s}_i, \tau^{-1}\mathbf{I}\right).$$

The likelihood is given by

$$\begin{aligned}
& \mathcal{L}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N|\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{C^*}, \gamma_1, \ldots, \gamma_N, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N, s_1, \ldots, s_N, a_1, \ldots, a_N) \\
= & \prod_{i=1}^{C^*}\prod_{j=1}^{n_i}(2\pi)^{K/2}|\tau^{-1}\mathbf{I}|^{-1/2}\exp\left\{-1/2\tau||\boldsymbol{y}_j - a_j\boldsymbol{\phi}[\gamma_j(\mathbf{t})]\prod_{c=1}^{C}\boldsymbol{\beta}_c^{z_{jc}} - \mathbf{s}_j)||^2\right\} \\
= & (2\pi)^{-K/2}\prod_{i=1}^{C^*}\prod_{j=1}^{n_i}\tau^{K/2}\exp\left\{-1/2\tau||\boldsymbol{y}_j - a_j\boldsymbol{\phi}[\gamma_j(\mathbf{t})]\prod_{c=1}^{C}\boldsymbol{\beta}_c^{z_{jc}} - \mathbf{s}_j)||^2\right\}.
\end{aligned}$$

The log-likelihood follows as

$$\begin{aligned}
& \log\mathcal{L}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N|\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{C^*}, \gamma_1, \ldots, \gamma_N, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N, s_1, \ldots, s_N, a_1, \ldots, a_N) \\
= & \text{constant} + \frac{KN}{2}\tau - \frac{1}{2}\tau\sum_{i=1}^{C^*}\sum_{j=1}^{n_i}||\boldsymbol{y}_j - a_j\boldsymbol{\phi}[\gamma_j(\mathbf{t})]\prod_{c=1}^{C}\boldsymbol{\beta}_c^{z_{jc}} - \mathbf{s}_j)||^2.
\end{aligned}$$

To be conservative, we start our algorithm with an excessive initial number of clusters (at least 1/4 of the total number of observations) and allow the number of non-empty clusters to decrease across iterations. Such a decrease occurs when at iteration t, based on the last sampled parameter values, no objects are assigned to some cluster in the Metropolis-Hastings cluster membership step.

We apply the following procedure to select the number of clusters during the initial (burn-in) stage of the algorithm, in conjunction with the cluster membership-switching procedure described at the end of Section 4. After this initial stage, we fix the number of clusters and proceed with ordinary MCMC, using only the Gibbs step to assign cluster membership to each observation.

When the total number of non-empty clusters decreases from C^* to $C^* - 1$, we calculate the average log-likelihood for the most recent block of iterations with C^* clusters (denoted by avg $\log\mathcal{L}_{C^*}$) and compare it to the average log-likelihood for

the most recent block of iterations with $C^* + 1$ clusters (denoted by avg $\log \mathcal{L}_{C^*+1}$). If avg $\log \mathcal{L}_{C^*} >$ avg $\log \mathcal{L}_{C^*+1}$, we accept the decrease. Otherwise, we reset the cluster membership to the first iteration in the most recent block of iterations where the number of clusters is $C^* + 1$. The pseudo code is given in Appendix A.

If the number of clusters remains constant for a long period of time, it either achieves the optimal number of clusters in terms of average log-likelihood, or the algorithm is trapped at the current number of clusters. Let M_{C^*} be number of consecutive iterations that the Markov chain stays at the current number of clusters. If M_{C^*} is larger than some predetermined threshold, we compare the average log-likelihood of the current block of iterations where $C = C^*$ to the average log-likelihood of the most recent block of iterations with $C = C^* + 1$. If the average log-likelihood is smaller for the current C^*, we reset $C = C^* + 1$, and reset the cluster membership to the first iteration in the most recent block of iterations where the number of clusters is $C^* + 1$. The pseudo code is given in Appendix A.

3.4 Simulation Study

To illustrate our algorithm's ability to estimate warping functions and cluster structure, we generate a simulated dataset and apply our method to it.

On the domain $\mathcal{T} = [0, 1]$, we choose 6 B-spline basis functions of order 5 using an equally-spaced knots sequence. We specify 5 clusters, and thus generate 5 sets of B-spline coefficients of size 6 distributed as MVN$(\mathbf{0}, 2 \times \mathbf{I})$, which are shown in Table 3.1. We assign 10, 12, 11, 10, and 13 observations (56 total observations) to each cluster, respectively, and generate 56 warping functions with 20 steps distributed as $Dir(\boldsymbol{\alpha} = (1, \ldots, 1))$. We assume that 30 equally spaced measurements on $\mathcal{T}$ are taken from each curve. The simulated warping functions are applied to the clock time and the underlying process times are obtained for each observation. For the i-th observation, we evaluate the B-spline function at its corresponding process times. A

set of stretching/shrinking factors of size 56 is generated as independent $N(1, 0.05^2)$ and multiplied to the mean values of the corresponding observations. Finally, we add white noise with $\sigma^2 = 0.01$ to each observation at each time point. A vertical shift generated from $Unif\,(-1, 1)$ is added to each observation. A plot of the simulated dataset is shown in the top left panel of Figure 3.3.

Table 3.1 5 sets of coefficients for the B-spline basis functions

	β_1	β_2	β_3	β_4	β_5	β_6
coef 1	2.38	0.46	-2.18	0.39	-2.56	1.81
coef 2	0.38	-2.89	-0.65	3.24	-1.38	4.08
coef 3	0.94	2.50	4.06	-2.79	0.59	-1.18
coef 4	-0.48	1.35	-4.88	2.48	-0.65	0.31
coef 5	-1.12	-0.00	0.83	2.62	4.48	0.70

To analyze the simulated data, we use a B-spline representation with 9 basis functions of order 6 with equally spaced knots. Our simulation experimentation indicates the clustering results are insensitive to the choice of spline basis having reasonable number and order, which is also noted by Liu and Yang [2009] and James and Sugar [2003]. The means $\boldsymbol{\beta}_0$ of the B-spline are taken to be $\mathbf{0}$, and we assume those coefficients are independent with variance $\mathbf{1}$, i.e., $\boldsymbol{\beta}|\boldsymbol{\beta}_0, \Gamma \sim N(\mathbf{0}, \mathbf{I})$. Based on Appendix A, the posterior samples for those coefficients are dominated by the data unless we have very strong prior knowledge. For the hyperparameters, we choose $\kappa = 100, \theta = 1$ for the precision, $\phi = 1$ for the vertical shifts, and $\alpha = 1$ for the warping functions. Following our algorithm for choosing the number of clusters, we start with $C = 30$ clusters having equal prior cluster probabilities.

We perform 20000 iterations, with the first 10000 discarded as burn-in. There are 5859 iterations in all whose number of non-empty clusters is 5, indicating that $C = 5$ is the most appropriate choice for this simulated dataset. To find a good set of starting values, we run another chain with $C = 5$ for 20000 iterations, with the first 10000 discarded as burn-in. We switch 15% of the observations in each cluster

every 20 iterations. Finally, a regular MCMC is performed using the initial values obtained from the last step. The correct classification rate (cRate) [Liu and Yang, 2009], defined as the maximum proportion of agreements between estimated and true cluster memberships (among all labeling permutations), is a measure of clustering quality. The cRate of our simulation study is 100%. We compare the result from joint registration and clustering to other existing methods using these simulated data. Of methods involving only clustering, the K-means method [Hothorn and Everitt, 2014], Ward's hierarchical agglomerative method [Hothorn and Everitt, 2014], and a model-based clustering method [Fraley and Raftery, 2002] produce a cRate of 83.93%, 85.71%, and 89.28%, respectively. To compare our result to a stepwise registration and clustering approach, we apply the registration method of Ramsay and Silverman [Ramsay and Silverman, 2005] implemented with the `register.fd` function in the `fda` package in `R` [Ramsay et al., 2013] to smooth and register the curves. Applied to the resulting registered curves, the cRate of the above three methods are 62.50%,75%, and 82.14%, respectively.

The lower left panel of Figure 3.3 displays the true signal curves (gray) and our posterior estimated signal curves (black). We use means of the posterior samples having 5 clusters as the point estimates of the B-spline coefficients. The estimated signal curves basically capture the characteristics of the true signal curve. The lower right panel of Figure 3.3 shows the estimated warping functions.

Notice that in Figure 3.3, there appears to be some phase variation that causes a slight discrepancy between the estimated mean curves (black) and the true mean curves (gray), which is due to a type of identifiability issue. The observed curves are the composition of the underlying mean curves, the subject-specific stretching/shrinking factors, and the subject-specific warping functions. Consequently, a slightly different set of mean curves, along with slightly different warping functions, could produce identical observed curves. Figure 3.4 displays three sets of estimated

and true warping functions selected from three clusters. The discrepancies we see between the true and estimated warping functions are what produce the phase variation in Figure 3.3(c). This issue is not closely related to clustering accuracy, however. The warpings are designed to align the curves within their identified cluster, in order to better measure the distances between curves in a cluster. Thus our clustering accuracy should still be good despite the discrepancies in Figure 3.3(c).

To test the convergence of the chain, we use the Heidelberg-Welch stationarity test [Heidelberger and Welch, 1981]. One advantage of this method is that it does not require multiple chains with different initial values, since our chain starts with the initial values determined by a preliminary run. For our simulation study, the sample for τ passes the test; 85% of the spline coefficient samples pass the test; 85% of the stretching/shrinking factor samples pass the test; 95% of the vertical shift samples pass the test; 93% of the warping function jumps pass the test. Overall, the vast majority of the posterior samples are considered to be drawn from their stationary distributions.

The goals of our study are estimating cluster membership and the warping functions associated with each observation. For a given observation, each step of the discrete warping function is estimated via the mean posterior jump at that step. The phase variation can be removed by applying the estimated warping function to the clock time for each observation. For our simulated dataset, the curves with phase variation removed are shown in the top right panel of Figure 3.3, from which we see a clear cluster structure.

The user-chosen value of M determines the degree of discretization of the warping function. Our philosophy is to achieve a balance between a reasonable approximation and affordable computational time. As a guide for the choice of M, we proposed the criterion

$$\psi^{M,\alpha} = \sum_{i=1}^{N} \int_0^1 |\gamma_i^{M,\alpha}(t) - t| dt$$

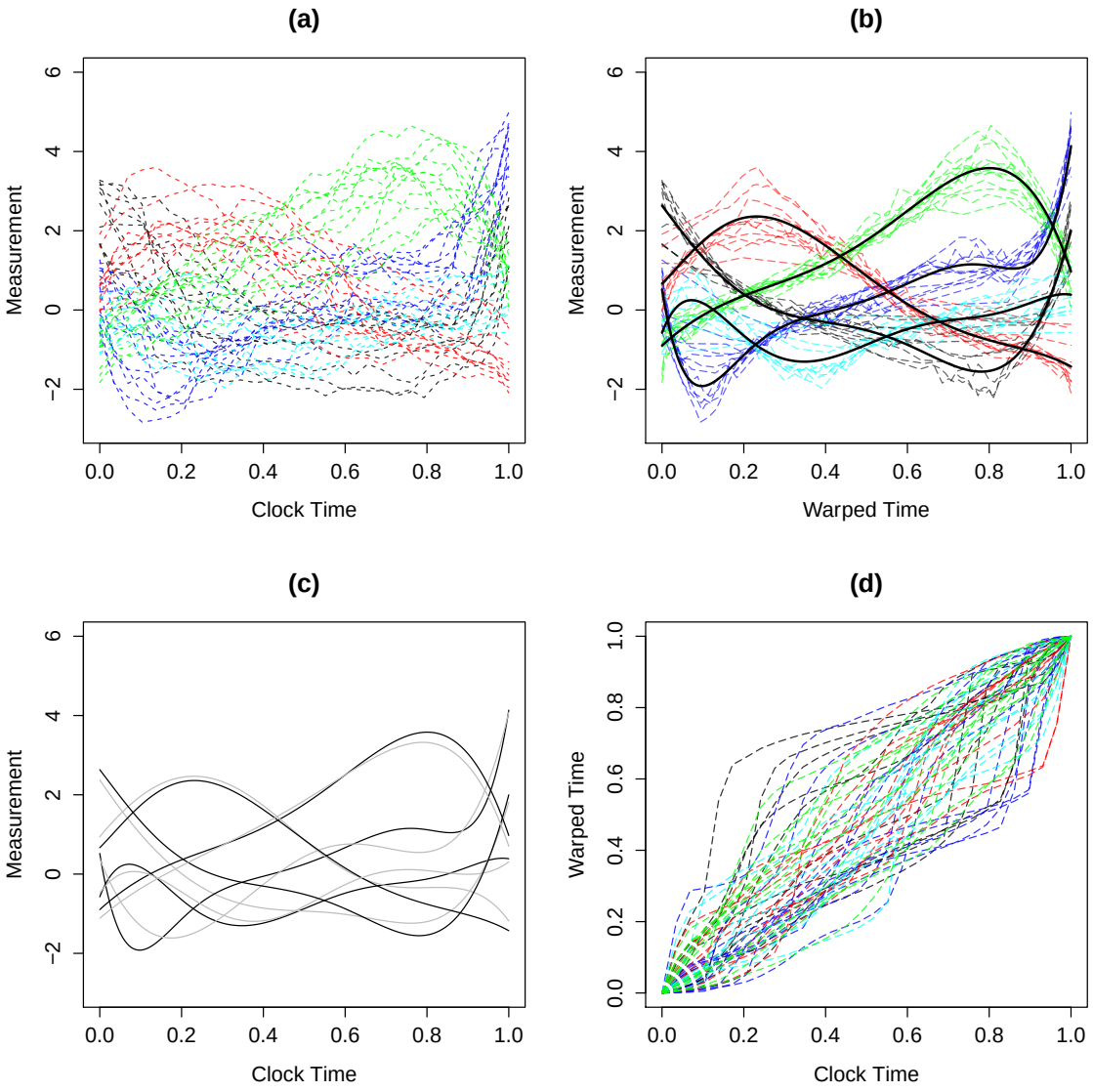

Figure 3.3 (a) A set of 56 simulated observations with 5 clusters. (b) Simulated data with phase variation removed, with superimposed posterior estimated mean curves (solid black). (c) True mean curves (gray) and estimated mean curves (black). (d) Estimated warping functions for all 5 clusters.

to measure the concentration of the warping functions (as a function of the dimension M and concentration parameter α) around the 45° reference line. If we change M, we need to adjust α simultaneously so that the variabilities among the warping functions remain roughly the same across different choices of M and α. We may obtain a positive real K by specifying a base Dirichlet distribution with $M = M_0$ and $\alpha = \alpha_0$,

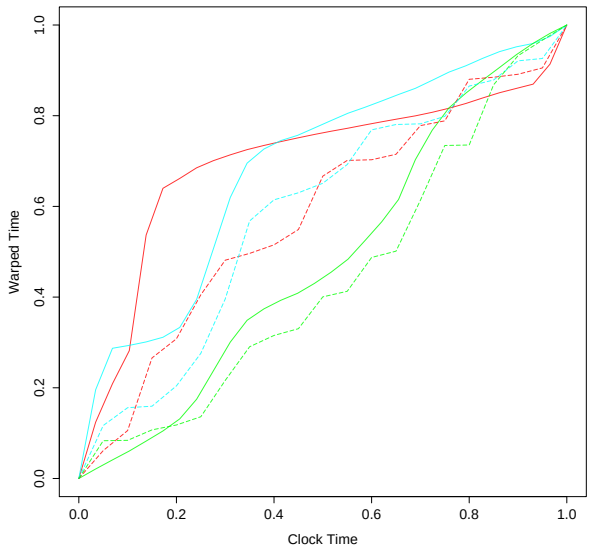

Figure 3.4 Solid curves represent estimated warpings, and dashed curves represents true warpings. Curves with the same color are warpings for the same observation.

and then letting $K = \alpha M$.

To inform the choice of M, we run 5 preliminary chains with 5000 iterations on our simulated data. We hold all parameters and hyperparameters constant except M and α, which we vary. We choose $M = 20, \alpha = 1$ as the base distribution and thus $K = 20$. We examined the cases of $M = 5, 10, 20, 30, 40$, and 50. Figure 3.5 shows a scatter plot of ψ against M. A value of M around the "elbow" of this plot should be sufficiently large to represent well the true nature of the distribution of warpings. We see that values of $M \geq 10$ are acceptable, since the elbow of Figure 3.5 is at $M = 10$. We still prefer using $M = 20$ due to more precise approximation and a still reasonable computing time. Note that the classification rate (cRate) for $M = 5$ is only 68%, while all other cases have cRate around 95% even for such a preliminary run.

We conduct a sensitivity analysis by examining the specifications of several hy-

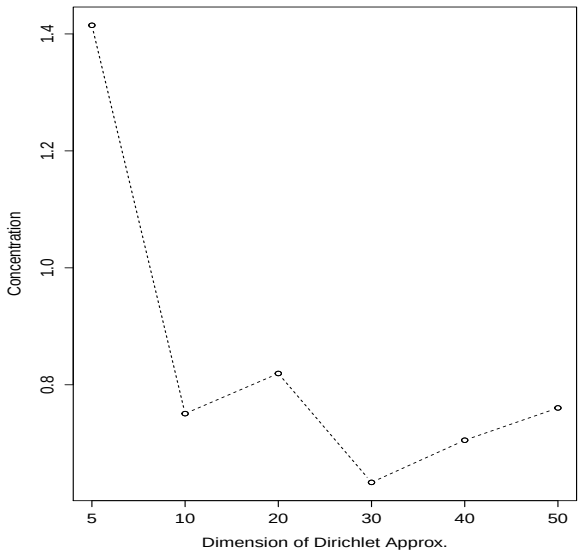

Figure 3.5 ψ values for different choices of M.

perparameters. We investigate the effect of various choices of α, ϕ, and σ_a^2. We vary the hyperparameters one at a time, separately multiplying each by 10, then by 0.1. The original values for α, ϕ, and σ_a^2 are 1, 1, and 0.05^2, respectively.

Table 3.2 shows the cRate for different altered choices of hyperparameters. The alteration of α only results in 1 and 3 incorrect curves, respectively. Using a large shift parameter $\phi = 10$ misclassifies 3 curves, while the small shift misclassifies 5 curves. This makes sense since the conditional posterior distribution of ϕ is a truncated normal bounded at $-\phi$ and ϕ. The small choice of σ_a^2 results in a much better cRate.

Based on our simulation study, our method seems to be insensitive to the specification of α. One exception is for data like the Berkeley accelerations that we present in Section 6, for which all the curves are similar and the phase variation contributes significantly to the cluster structure. In such a case, α must be chosen with caution. We would recommend choosing ϕ fairly large rather than small, since a small ϕ may be too restrictive to sample a proper shift. Finally, we would recommend choosing

σ_a^2 relatively small when uncertain.

Table 3.2 Sensitivity analysis for simulated data

Parameter	Value	cRate
α_1	0.1	94.64%
α_2	10	98.21%
ϕ_1	0.1	91.07%
ϕ_2	10	94.64%
$\sigma_{a_1}^2$	0.025	87.50%
$\sigma_{a_2}^2$	0.00025	100%

We perform another simulation study based on the previous setup but with only 10 evenly spaced measurement points. The cRate is 100%, which suggests our method performs well for sparsely sampled data.

3.5 Real Data Analysis

Berkeley Growth Curves

The Berkeley growth data [Tuddenham and Snyder, 1953] measured 54 girls and 39 boys at 31 time points from age 1 to age 18. In the literature, this dataset often serves as a benchmark to test clustering accuracy. A monotone smoothing spline [Ramsay and Silverman, 2005] can be applied to the original height data. If we evaluate the corresponding second order derivatives at these 31 measurement time points, there exists obvious phase variation as shown in Figure 3.6. The left panel shows the acceleration data; the right panel shows the acceleration values without first 5 timepoints excluded due to the bias of the function estimation near the boundary [Cheng et al., 1997]. Based on Figure 3.6, we assume that there are small vertical shifts with $\phi = 1.2$ and the variation among observations is caused by both phase variation and random error $\boldsymbol{\epsilon}$. We choose $\kappa = 50$ and $\theta = 10$ to accommodate possible amplitude variation, and we choose $\alpha = 4$ for the Dirichlet approximation. We model

the signal functions with 8 B-spline basis functions of order 6 defined on a equally spaced knot sequence. The prior means of the spline coefficients are generated as $N(1, 4)$, and the spline coefficients are assumed independent with variance 1. We switch 10 percent of the observations from each cluster every 10 iterations in the burn-in stage. The number of clusters is fixed at 2 throughout the entire MCMC. The prior cluster probabilities are both 0.5 for males and females. We perform 20000 iterations, with the first 10000 discarded as burn-in.

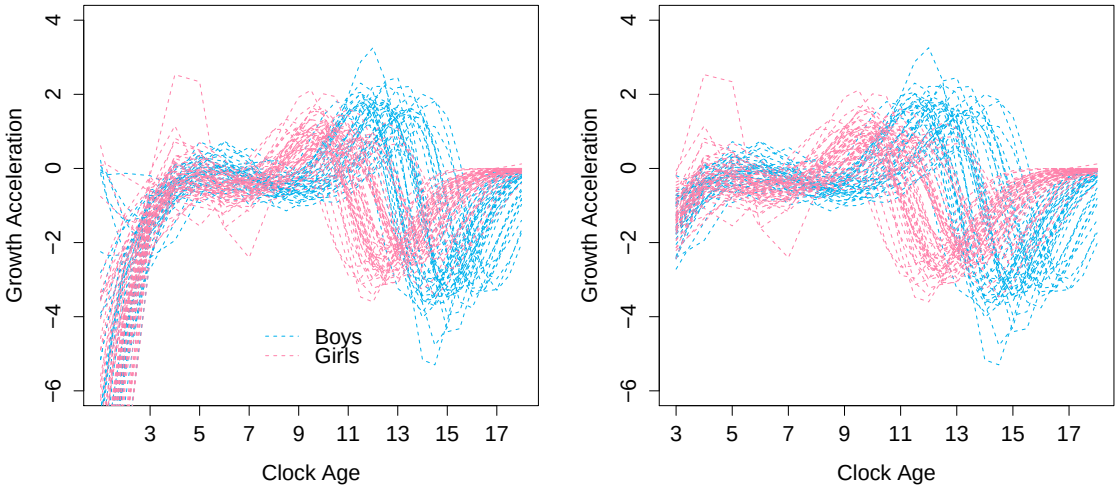

Figure 3.6 Left: original growth acceleration; Right: growth acceleration without first 5 measures.

The clustering results are shown in Table 4.2; only 3 females and 2 males are misclassified to the opposite gender, yielding overall cRate 94.6%. The clustering results are plotted in the second row of Figure 3.7; the bold solid curves represent those boys who are misclassified as girls, and the bold dashed curves represent the misclassified girls. The right panel shows the curves after registration. For comparison, we apply Ward's hierarchical clustering on the unregistered data [Hothorn and Everitt, 2014], which produces a cRate of 75.26% with 23 girls misclassified as boys. A model-based method [Fraley and Raftery, 2002] produces a 73.08% cRate with 23 girls misclas-

sified as boys. After registration, the Ward's method and the model-based method yields a 63.44% and 68.82% cRate, respectively.

Table 3.3 Clustering results for Berkeley acceleration curves.

	True cluster	
	Male	Female
Cluster I	37	3
Cluster II	2	51

We also apply the proposed method to the original height data and velocity data. For the original height curves, we set $\alpha = 100$ and $\sigma_a^2 = 10^{-3}$, since there is no strong evidence of time distortion and the vertical shifts constitute the majority of the variation. We put a strong precision-related hyperparameter with $\kappa = 5 \times 10^4$ and $\theta = 1$ due to the highly precise height measurements. Our corresponding cRate is 91.4%, while the SACK model (Liu and Yang, 2009) reports a 86% accuracy rate, and KCFC (Chiou and Li, 2007) reports a 93.35% accuracy rate. For the velocity curves, we apply our method with $\alpha = 10, \kappa = 50, \theta = 10, \phi = 5$ and $\sigma_a^2 = 0.1^2$. The cRate produced is 84.9%, while Zhang and Telesca [2014] reported a cRate of 83%.

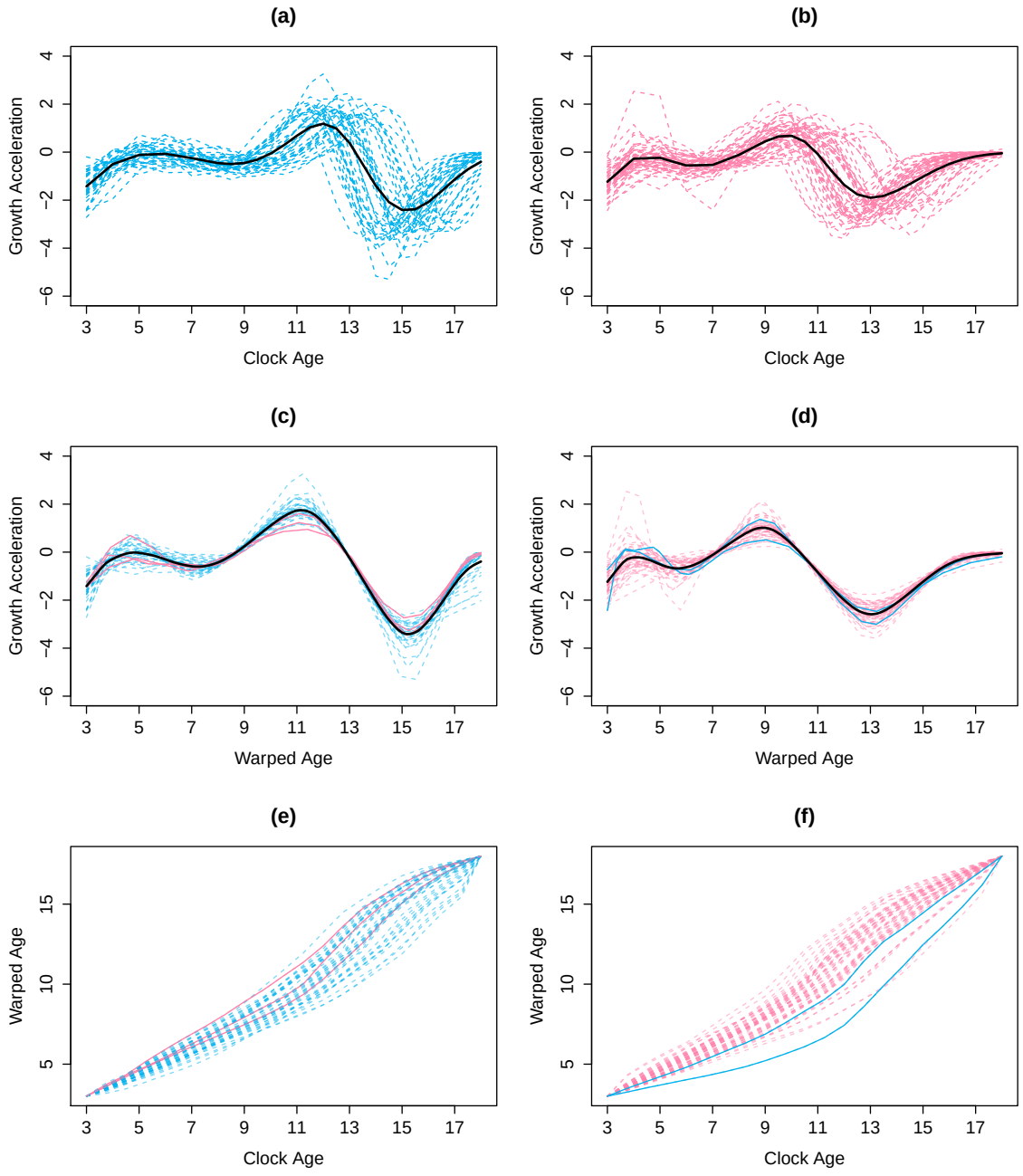

Figure 3.7 (a)-(b) Unregistered growth acceleration data for 39 boys (blue dashed) and 54 girls (pink dashed) with cross-sectional mean superimposed. (c) Registered cluster 1 with 37 boys (blue dashed) and 3 girls (pink solid). (d) Registered cluster 2 with 51 girls (pink dashed) and 3 boys (blue solid). (e)-(f) Estimated warping functions for cluster 1 and cluster 2, respectively.

Elutriation-Synchronized Cell Cycle

The elutriation dataset, collected by Alter et al. [2000], measures ratios of gene expression levels in log-scale 18 times, at 7-minute intervals. We apply our proposed Bayes method to a subset of 78 gene expressions. According to Spellman et al. [1998], this dataset is classified into five cell-cycle subgroups: M/G_1, G_1, S, S/G_2 and G_1/M. Among these 78 gene expressions, genes 1 to 13, genes 14 to 52, genes 53 to 60, gene 61 to 67, and gene 68 to 78 are classified into these five respective phases. Note that these different cycle phases are based on biologists' beliefs, and therefore are not absolutely true cluster structure. The trajectories of the dataset are shown in the left panel of Figure 3.8.

First, we apply our method with five clusters to examine whether the clustering results agree with the underlying biological process. Table 3.4 shows that 39 out of 78 genes are classified in their corresponding cycle phases, highlighted by bold numbers. The gene expressions adjacent to each other should behave similarly due to adjacent-phase correlation. Therefore, we also highlight in italics cells adjacent to the diagonal elements. Note that 67 out of 78 gene expression profiles are clustered on the tridiagonal positions including 5 on the left lower corner and 1 on the right upper corner, since cluster I and cluster V are considered to be adjacent phases by the circular property of the data.

Table 3.4 Clustering results for cell cycle when $C = 5$

Cluster	M/G_1	G_1	S	S/G_2	G_1/M
I (9)	**5**	*1*	0	2	*1*
II (26)	*2*	**19**	*1*	3	1
III (24)	1	*15*	**7**	*1*	0
IV (5)	0	4	0	0	*1*
V(14)	*5*	0	0	*1*	**8**
total	13	39	8	7	11

We next allow the algorithm to choose the number of clusters, initially using 20

clusters. The mode of the number of non-empty clusters is 4, indicating 4 clusters. The clustered gene expression profiles are shown in Figure 3.8 (right panel). The aligned curves show a clear cluster structure, and all curves in the same cluster display roughly the same pattern.

Table 3.5 Clustering results for cell cycle when $C = 4$

Cluster	M/G_1	G_1	S	S/G_2	G_1/M
I (9)	5	1	0	2	1
II (38)	0	27	8	2	1
III (13)	1	11	0	1	0
IV (18)	7	0	0	2	9
total	13	39	8	7	11

3.6 Discussion

We have developed a Bayesian clustering method for functional observations that works especially well for data having phase variations. If one believes the phase variations are important characteristics in distinguishing different clusters or that there is no phase variation, one may specify large values of $\boldsymbol{\alpha}$ to discourage the warping functions from departing from a 45° straight line. In this case, our method approximates a Bayesian clustering of functional data without registration.

We demonstrate our algorithm's ability to capture cluster structure and estimate warping functions through simulation studies and real data analyses. Based on our simulation, we observe that one should pick hyperparameters $\boldsymbol{\alpha}$ carefully when phase variations contribute significantly to the clustering structure. We recommend large ϕ and small σ_a^2 when uncertain.

By using the Dirichlet warping approach, our method allows fairly arbitrary warping functions and places no assumptions on the vertical separation among clusters. Thus, the scope of application of our method may exceed that of existing methods,

which make more restrictive assumptions. Our simultaneous registration and clustering approach simplifies the analysis procedure and should benefit researchers who cluster functional data.

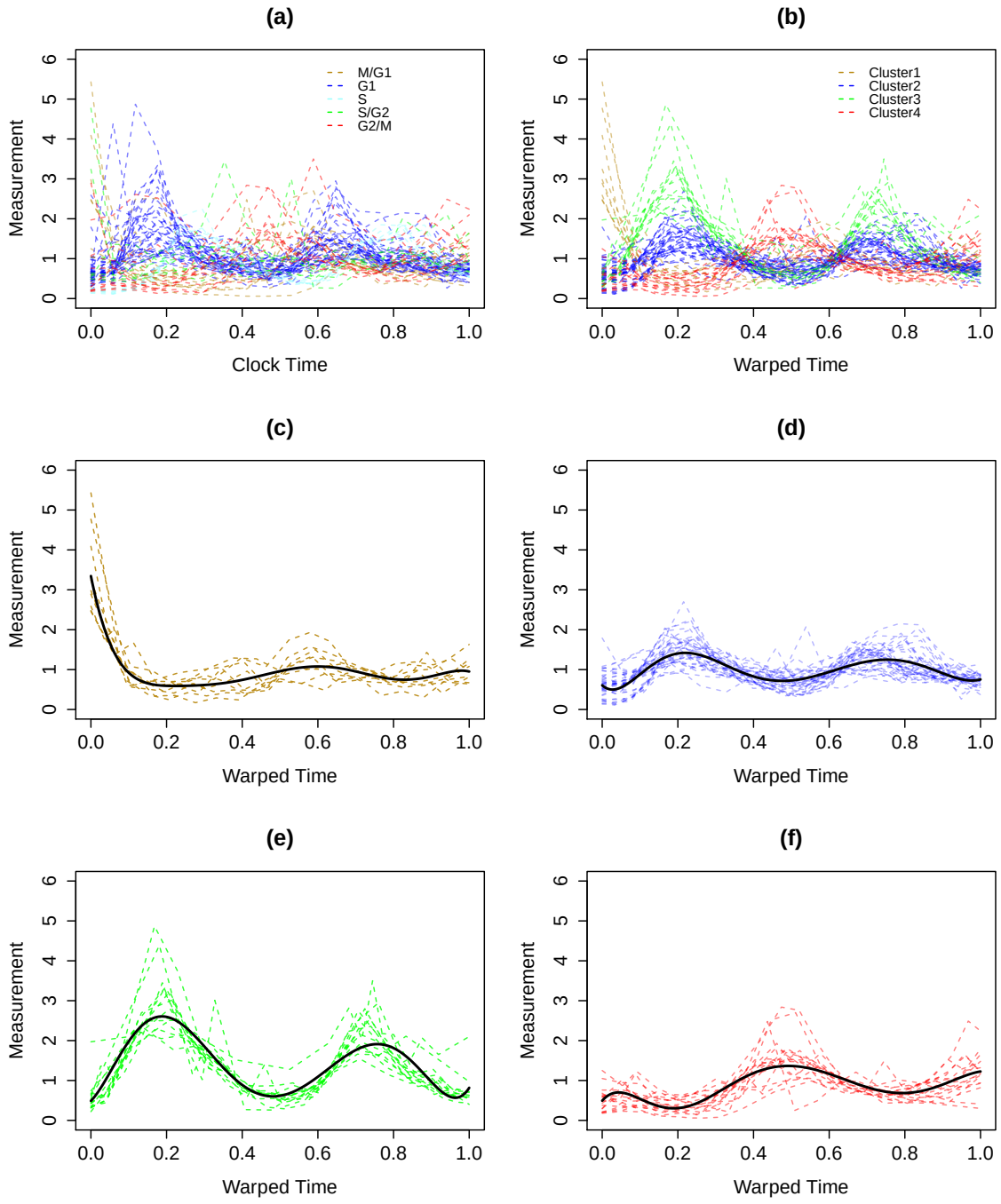

Figure 3.8 (a) Raw gene expression with cluster structure determined by the biologists. (b) Registered curves with 4 clusters. (c)-(f) Registered four clusters with their estimated mean curves superimposed.

Chapter 4

Clustering Functional Observations with Time Warpings via Derivative-Shape Measure

We apply a Bayesian method to register pairs of curves potentially belonging to the same cluster. By employing a discrete approximation generated from the Dirichlet distribution, our Bayesian method is capable of detecting arbitrary warping functions. After registration, we develop a "derivative sign" method to measure the dissimilarity between two functional data based on their shapes, which serves as a"distance" for clustering purposes. The clustering result can then be obtained via any desired distance-based method afterwards.

4.1 Model Assumption

Following the Bayesian model proposed in chapter 3, we assume that there are N objects, which belong to C clusters, with K measurements taken on each. For a discretized functional observation in a given cluster, the response vector is modeled by

$$\mathbf{Y} = af(\mathbf{t}) + \boldsymbol{\epsilon},$$

where a is a stretching/shrinking factor accounting for amplitude variation, and $f(\mathbf{t})$ is a $K \times 1$ vector of responses measured at a vector of time points $\mathbf{t}$. The random errors $\boldsymbol{\epsilon}$ are generated from $N(0, \sigma^2)$.

If our observed data exhibit both amplitude and phase variability, the associated warping function is denoted by $h(\cdot)$. The response now is $\mathbf{Y} = af[h(\mathbf{t})] + \boldsymbol{\epsilon}$. The

mean function $f(\cdot)$ is represented by a B-spline expansion with q basis functions so that

$$f(t) = \sum_{j=1}^{q} \phi_j(t)\beta_j$$

throughout this paper, and the warping function is approximated by a piecewise linear function $\gamma(\cdot)$ [Cheng et al., 2015]. See the appendix for more details.

Another possible mode of amplitude variation that the data may exhibit besides the stretching/shrinking factor is composed of vertical shifts among observations in the same cluster. Denote the $K \times 1$ vertical shifts by $\mathbf{S} = S \otimes \mathbf{1}$. The left panel in Figure 3.2 shows a set of simulated observations from the same cluster with phase variations; the right panel shows the same observations with additional vertical shifts following $Unif\,(-0.5, 0.5)$. The bold curve is the true signal function generating the observations.

Our observed response becomes

$$\mathbf{Y} = a\boldsymbol{\phi}[\gamma(\mathbf{t})]\boldsymbol{\beta} + \mathbf{S} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\phi}$ is a $K \times q$ matrix consisting of basis functions evaluated at $\mathbf{t}$, and $\boldsymbol{\beta}$ is the vector of basis function coefficients. Thus,

$$\mathbf{Y}|\boldsymbol{\beta}, \gamma, \tau, s \sim \text{MVN}\left(a\boldsymbol{\phi}[\gamma(\mathbf{t})]\boldsymbol{\beta} + \mathbf{s}_i, \tau^{-1}\mathbf{I}\right).$$

We parameterize the variance using the precision τ, which is convenient for our Bayesian registration described algorithm in the appendix.

4.2 Pairwise Derivative-Shape Dissimilarity Measure Algorithm

Given a pair of curves belonging to the same cluster that differ only based upon phase, such two curves should be similar in terms of shape if the phase variations are removed. On the other hand, if they belong to different clusters, these two curves should be significantly different in shape no matter what registration method we apply to them.

Consequently, we need to address the following two issues: (1) how to efficiently remove phase variations for a pair of observations belonging to the same cluster; (2) how to measure the dissimilarity between two curves after the phase variations are removed.

Pairwise Registration

Following Cheng et al. [2015], we model the warping function by the cumulative sum of realizations generated from a Dirichlet distribution, and we propose a Bayesian registration method. The MCMC sampling algorithm is given in the Appendix B; it is essentially a variation of the MCMC algorithm proposed in Chapter 3 by assuming the number of clusters $C = 1$. Our proposed method can efficiently remove nonlinear time distortion and vertical shifting. Compared to the popular `register.fd` function in the `R` package `fda` [Ramsay et al., 2013], our proposed registration method is about 10 times faster in achieving a reasonably good alignment with code written in `C++`, and can handle the case when vertical shifts exist. Furthermore, our proposed Bayesian algorithm does not require any template curve for registration. Figure 4.1 shows a pair of raw curves belonging to the same cluster (left panel), and the registered curves (right panel) with vertical shifts removed. Notice that our registration algorithm requires that the common domain $\mathcal{T} = [0, 1]$.

In section 4.2, a method of measuring the dissimilarity between two curves is introduced and the result serves as a "distance" for the clustering purpose.

Derivative-Shape Dissimilarity Measure

For a curve $y(t)$, define

$$\psi_{ij}(t) = \begin{cases} 0.5 & \text{if } y'(t) \geq 0 \\ -0.5 & \text{if } y'(t) < 0, \end{cases}$$

where y' represents the derivative. Given a pair of curves y_i and y_j we assume they belong to the same "group" if these two functions are generated by two similar (up to phase variation) underlying mechanisms. For example, consider the growth accelerations of two boys with similar growth patterns. If there is no phase variation, then we will expect that the derivatives of those two curves have the same sign. So, a dissimilarity measure of y_i and y_j is given by

$$der(y_i, y_j) \equiv \int_{\mathcal{T}} |\psi_i - \psi_j| \, \mathrm{d}t,$$

which is near zero for two curves that are "similar" in the sense just described. However, this measure is sensitive to phase variations. For example, the landmarks (local maximum, minimum, and inflection points, etc.) of the observed curves $y_i(t)$ and $y_j(t)$ are likely to appear at different times. Thus, $der(y_i, y_j)$ may be much greater than 0 even if the underlying generating mechanisms (except phase variations) are exactly the same.

Suppose curves y_i and y_j belong to the same cluster but have phase variation. Denote $h_i : \mathcal{T} \to \mathcal{T}$ and $h_j : \mathcal{T} \to \mathcal{T}$ as the warping functions for y_i and y_j, respectively. We define a derivative-shape measure to be

$$der\{y_i[h_i(t)], y_j[h_j(t)]\} = \int_{\mathcal{T}} |\psi_i[h_i(t)] - \psi_j[h_j(t)]| \, dt$$

which will be near 0 for a pair of curves belonging to the same cluster.

The Bayesian registration algorithm introduced in the section 4.2 provides an effective way of estimating the warping functions $h_i(t)$ and $h_j(t)$. Combining the fact that our registration algorithm requires the domain $\mathcal{T}$ to be $[0, 1]$, the range of possible values of the derivative-shape measure (DSM) henceforth is also $[0, 1]$. Note that DSM $= 0$ means two functions increase, decrease, or remain flat simultaneously throughout the entire $\mathcal{T}$; DSM $= 1$ means they always behave in an opposite fashion (one curve increases when the other one decreases, and *vice versa*).

Adjustment for Monotonicity

For a pair of curves, the derivative-shape measure described in section 4.2 is always 0 if both curves are monotone increasing or decreasing. However, the true underlying mean curves could be quite different. For example, consider $5\sqrt{t}$ and t^2 on the interval of $[0, 1]$. Under such a situation, the DSM would not provide any useful information to distinguish these curves, and hence should be ignored. As discussed in section 4.2, we will apply the DSM to the higher-order derivatives. The final distance between two curves is a weighted average of all DSMs taken on the original curves and their derivatives. If we apply the DSM up to the K-th derivative, we should place weights $(w_0, w_1, \ldots, w_K)$ on the K DSM values.

To address the monotonicity issue just described, we make the following adjustment. For each of a pair of curves, define ξ_{ij} as the ratio of the length of combined intervals on which the j-th derivative of the i-th curve is monotone increasing to the total length of the i-th curve's domain, $i = 1, 2$ and $j = 0, 1, \ldots, K$. Let $l_j = |\xi_{1j} + \xi_{2j} - 1|$, then define $m_j = w_j(1 - l_j)$. If both curves are monotone increasing or decreasing then $m_j = 0$, which means we ignore the information on the j-th derivative. In contrast, suppose one curve is monotone increasing and another decreasing, then the DSM provides the complete information for clustering and $m_j = w_j$. Finally, we normalize the weights based on the monotonicity adjustment as

$$w_j^* = \frac{m_j}{\sum_{l=1}^{k} m_l}.$$

Full Algorithm

To start, we presmooth all curves using a B-spline basis expansion with curvature penalized using smoothing parameters chosen by the generalized cross validation (GCV) [Golub et al., 1979], and we use the smoothed function evaluated at the original measurement points as our input. The first-order DSM is not sufficient to measure

the shape dissimilarity between two functions. For example, $der(r, s) = 0$ for any monotonic increasing (decreasing) functions r and t. It is necessary to apply this derivative-shape measure to higher-order derivative(s).

Let $\mathbf{Y}_i$ be the vector of values evaluated at $\mathbf{t}$ on the presmoothed curve for the i-th observation, and $\mathbf{t}$ be the vector of points where the measurements are taken. The full algorithm is as follows:

1. For a pair of observed curves y_i and y_j, initially assume that they belong to the same cluster regardless of the true cluster memberships. Register $(\mathbf{Y}_i, \mathbf{t})$ and $(\mathbf{Y}_j, \mathbf{t})$ by our Bayesian registration method.

2. Let γ_i and γ_j be the discrete approximation of the warping function h_i and h_j, respectively. Fit smoothing spline curves with smoothing penalty λ_1 on $(\mathbf{Y}_i, \gamma_i(\mathbf{t}))$ and $(\mathbf{Y}_j, \gamma_j(\mathbf{t}))$, and denote the fitted function by $\hat{y}_i^*$ and $\hat{y}_j^*$, respectively, with the asterisk indicating the functions are registered. Note that λ_1 should be small or even 0 since we have already presmoothed all curves. Applying the DSM defined above, we obtain $der(\hat{y}_i^*, \hat{y}_j^*)$ as a measure of dissimilarity between these two functions after removing the phase variation.

3. We may apply the same derivative measure procedure to the derivatives of the original functions, since the shape of a function is determined by these derivatives. For the 1st-order derivative, we take the following steps:

   - We evaluate $\hat{y}_i'[\alpha_i(t)] = \frac{d}{dt}\hat{y}_i[\alpha_i(t)]$ at another set of points $\mathbf{t}^*$. Note that $\mathbf{t}^*$ could be finer than $\mathbf{t}$, since we consider the pair of discrete realizations

   $$(\hat{y}_i'[\alpha_i(\mathbf{t}^*)], \mathbf{t}^*) \text{ and } (\hat{y}_j'[\alpha_j(\mathbf{t}^*)], \mathbf{t}^*)$$

   as the input of the Bayesian registration algorithm. Let us denote the estimated warping functions for these two derivatives as β_i and β_j (they are

presumably "small" if y_i and y_j belong to the same cluster, since the phase variations of the raw curves have been removed in step 2), respectively.

- Then we evaluate

$$\hat{y}_i'[\alpha_i(\beta_i(\mathbf{t}^*))],$$

and fit another smoothing spline $\hat{y}_i^{*\prime}$ with smoothing parameter λ_2, which should also be small since the derivative is taken on a smoothed curve. Apply the same procedure on $\hat{y}_j'[\alpha_j(t)]$ as well.

Now, we calculate the DSM $der(\hat{y}_i^{*\prime}, \hat{y}_j^{*\prime})$ on the registered first-derivative function.

4. Apply step 3 to higher-order derivatives if necessary, where $der(\hat{y}_i^{*(k)}, \hat{y}_j^{*(k)})$ denotes the DSM of the pair of k-th order derivative functions.

5. Choose a set of weights $(w_1, w_2, \ldots, w_K)$, such that $\sum_i w_i = 1$, and calculate the pairwise distance between y_i and y_j as

$$w_1^* der(\hat{y}_i^*, \hat{y}_j^*) + w_2^* der(\hat{y}_i^{*\prime}, \hat{y}_j^{*\prime}) + \cdots + w_K^* der(\hat{y}_i^{*(K)}, \hat{y}_j^{*(K)}),$$

where the weights $w_1^*, \ldots, w_K^*$ are defined in the end of section 4.2.

6. perform steps 1-5 on every pair of curves y_i, y_j to obtain a distance matrix containing all pairwise distances.

7. Apply any preferred dissimilarity-based clustering methods, such as hierarchical clustering or the K-medoids method, on the distance matrix calculated in step 6.

To illustrate the algorithm, let us consider the data shown in Figure 4.1. The dissimilarity measure of the curves on the right panel is 0.02. However, the dissimilarity measure of the raw curves is 0.21, which confirms the fact that the DSM is sensitive to the phase variation. Figure 4.2 shows the first-order derivatives and their

registered counterpart. The DSM of the curves in the left panel is 0.106, and the DSM of those in right one is 0.072.

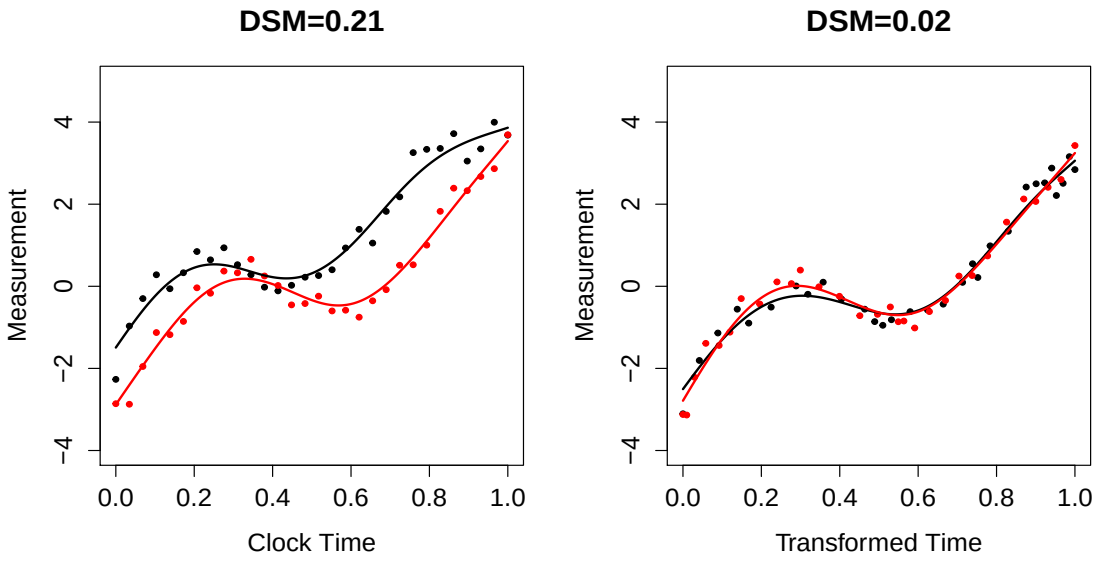

Figure 4.1 Left: raw curves. Right: registered curves.

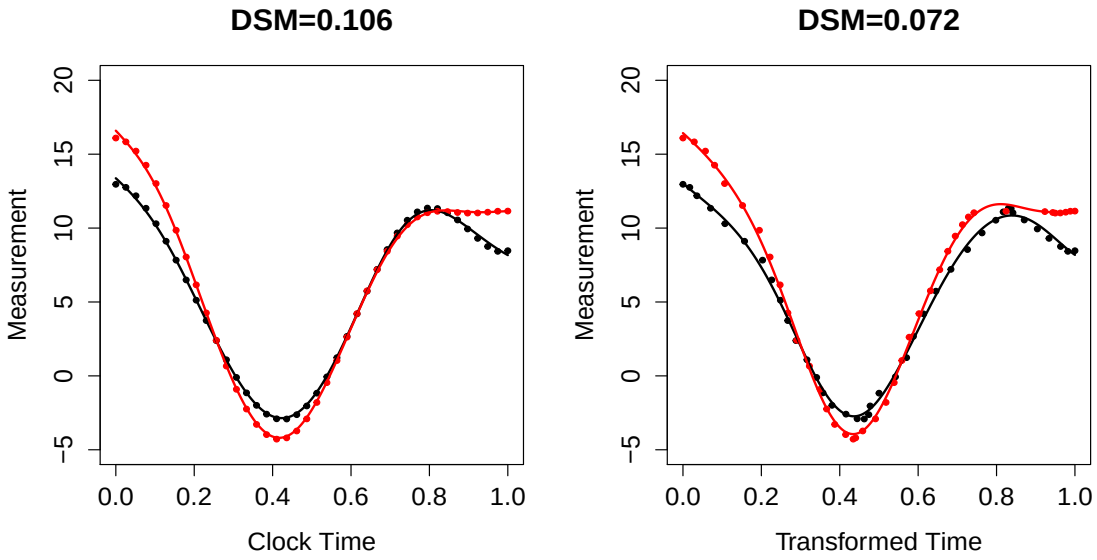

Figure 4.2 Left: first-order derivative curves. Right: registered curves.

Next, let us examine the second-order derivatives. Figure 4.3 shows the second-

order derivatives and their registered counterparts. The DSM of the curves in left panel is 0.04, and the DSM of those in the right one is 0.028.

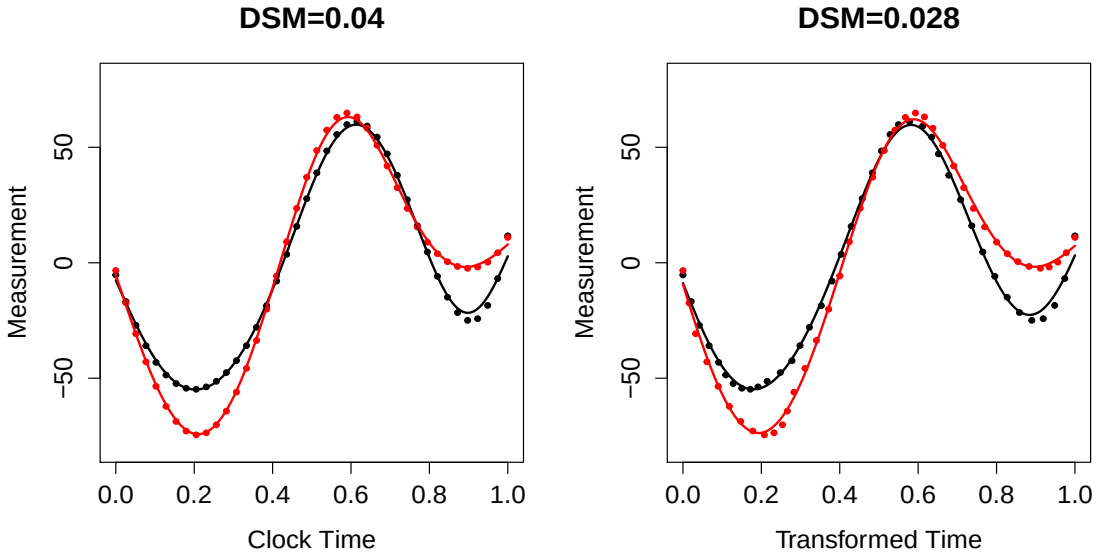

Figure 4.3 Left: 2nd order derivative curves. Right: registered curves.

Figure 4.4 illustrates the case when we apply our procedure to curves belonging to different clusters. The curves on left panel are raw curves and their derivatives, while the curves in the right panel are the registered counterparts. The DSMs of the curves in the right panels are 0.838, 0.476, and 0.368 for the original functions, first derivative, and second derivatives, respectively; the DSMs of those in the right panels are 0.838, 0.356, and 0.236, respectively. The DSMs are indeed reduced after the registrations as we expected; however, the values are still large compared to the DSMs for the pair of curves belonging to the same cluster that was shown in the previous example. This numerical example illustrates the idea that the DSMs could be significantly greater than 0 even after registration for curves belonging to the different clusters.

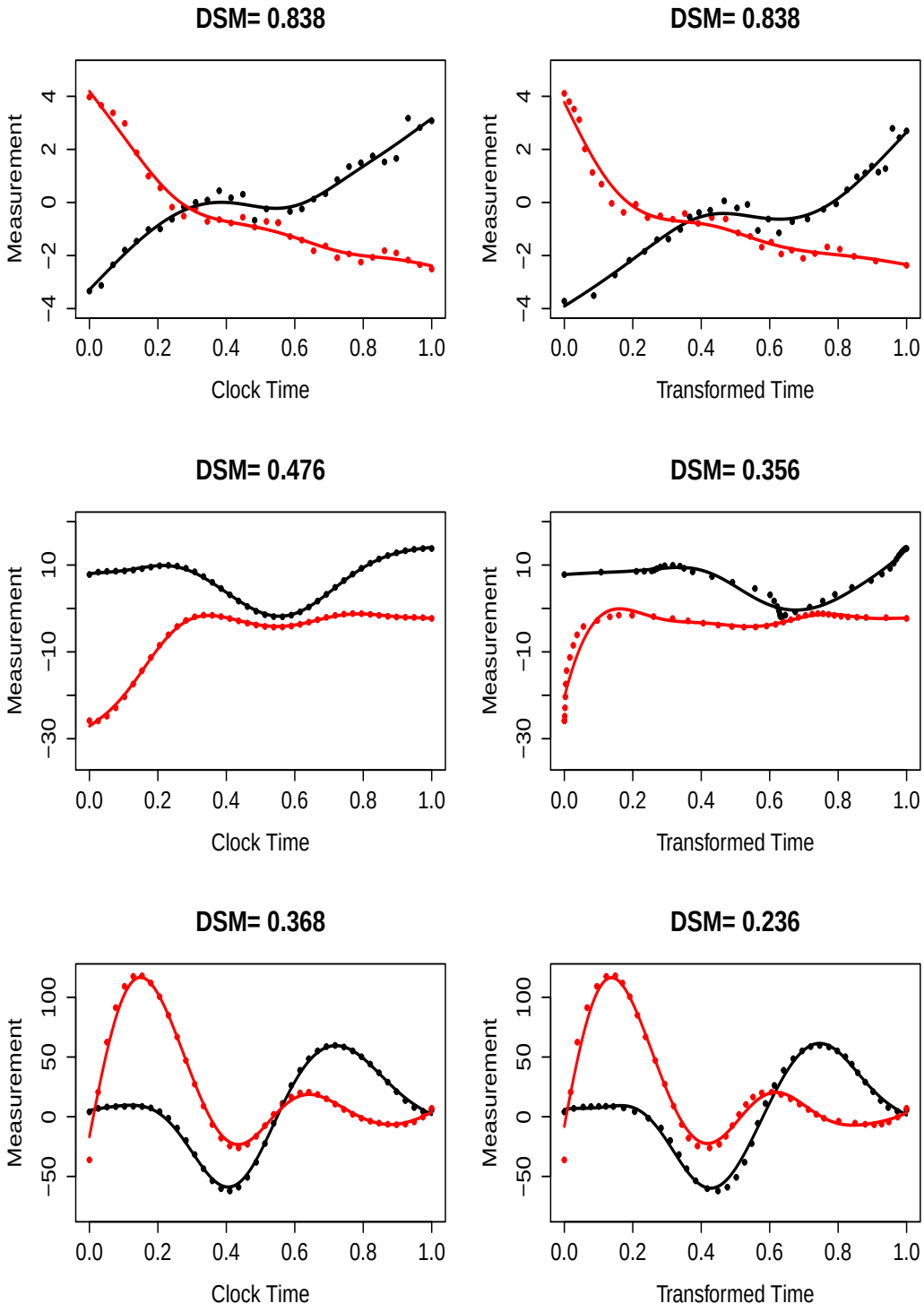

Figure 4.4 Left: raw curves and their derivatives. Right: registered curves.

4.3 Simulation Study

We apply our proposed method to a set of simulated data to demonstrate its clustering accuracy.

We generate simulated curves having domain $\mathcal{T} = [0, 1]$. We specify 5 clusters, with each cluster associated with its mean function generated from a B-spline expansion. We set 8 B-spline basis functions of order 5 with an equally-spaced knot sequence for each of 5 clusters, and we generate 8 sets of coefficients of size 6 from $N(2, 9)$ independently.

We assign 5, 4, 6, 5, and 4 observations to each cluster, respectively. We evaluate the mean function associated with each observation at 30 equally spaced points on $\mathcal{T}$. The phase variation of each observation is introduced via applying an warping function, approximated by the cumulative sum over 20 steps distributed as $Dir(\boldsymbol{\alpha} = (0.8, \ldots, 0.8))$, to the clock time, with the result serving as the unobserved system time.

For one source of amplitude variation within a cluster, we generate a set of stretching/shrinking factors from $N(1, 0.04)$ independently. Vertical shifts, generated from independent $Unif(-2, 2)$, serve as another source of amplitude variation. We add normal random errors with mean 0 and variance 0.04. The top left panel of Figure 4.5 shows one set of simulated data.

For a pair of observations, we perform 200 iterations of our Bayesian registration algorithm with the first half as burn-in. To smoothly represent the generated data, we use a B-spline representation with 10 basis functions of order 5 with equally spaced knots. The Bayesian registration requires the specification of several hyperparameters. For registration, we choose $\alpha_1 = 1$ for the warping function, $\beta_0 = \mathbf{0}$, $\Gamma = \mathbf{I}_{10}$ for the spline coefficient, $\kappa = 50, \theta = 1$ for the precision, $\sigma^2_{a_1} = 0.02^2, \sigma^2_{a_2} = 0.03^2$ for the stretching/shrinking factors, and $\phi_1 = 1, \phi_2 = 10$ for vertical shifts. **?** gave recommendations about how to set these parameters. For smoothing penalty parameters,

we choose $\lambda_1 = \lambda_2 = 10^{-8}$ for the raw curve smoothing and its derivatives. We set the raw weights $w_1 = 0.7$, $w_2 = 0.2$, and $w_3 = 0.1$ for the original curves, first-order, and second-order derivatives, respectively, to calculate the distance matrix. Finally, we use Ward's method in the `R` function `hclust` to determine the cluster membership.

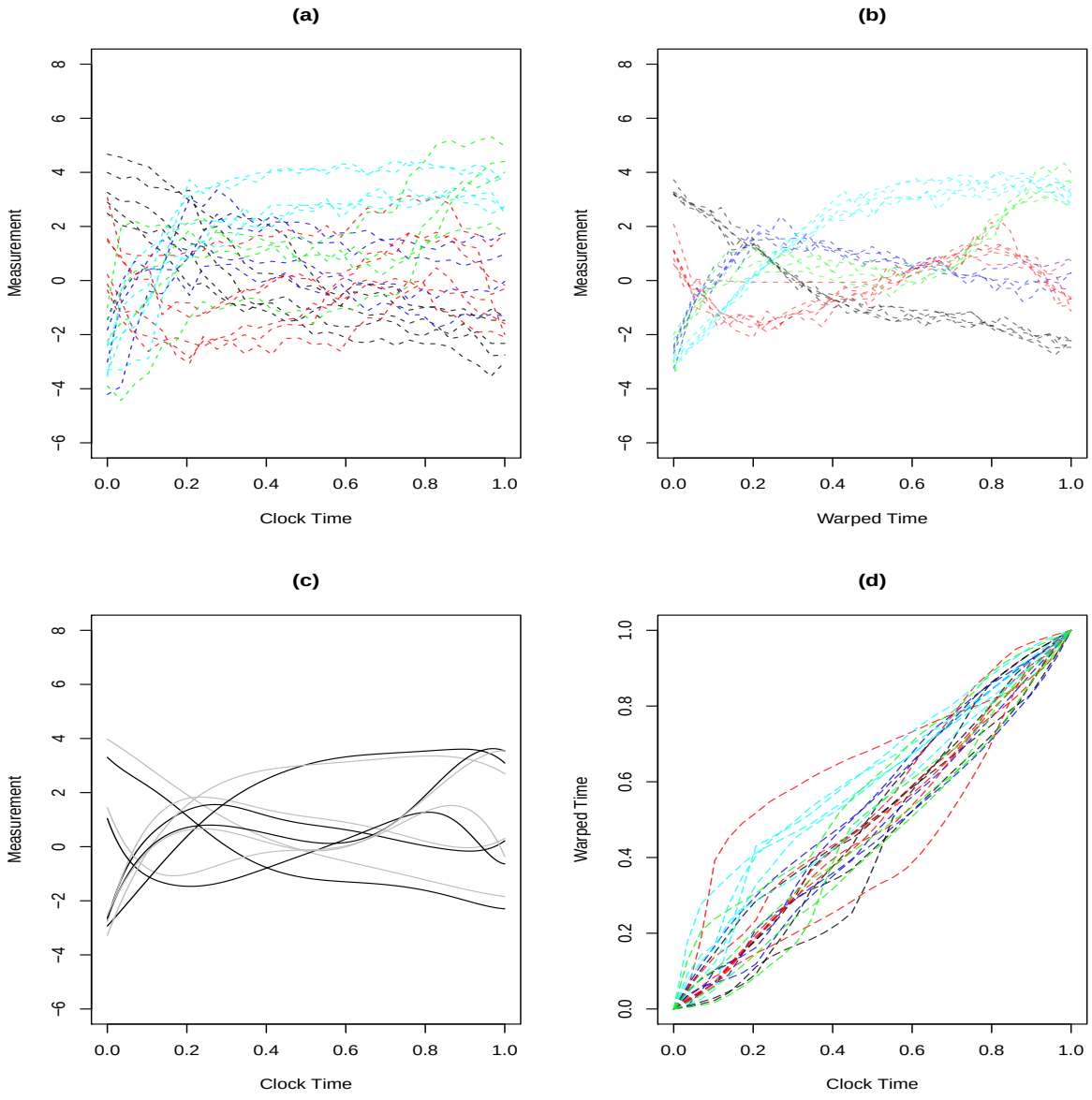

Figure 4.5 (a) A set of 24 simulated observations with 5 clusters. (b) Simulated data with phase variation removed, with superimposed posterior estimated mean curves (solid black). (c) True mean curves (gray) and estimated mean curves (black). (d) Estimated warping functions for all 5 clusters.

To examine the overall performance of our proposed algorithm, we implement this data generation and clustering procedure 50 times. The correct classification rate (cRate) [Liu and Yang, 2009], defined as the maximum proportion of agreements between estimated and true cluster memberships (among all labeling permutations), is a measure of clustering quality. The average cRate over these 50 repetitions is 92%.

After the cluster memberships are obtained via the DSM method, we could register clustered curves within the same group by the Bayesian clustering method. For one set of simulated data, our method classifies all curves in Figure 4.5(a) correctly. The registered curves are shown in Figure 4.5(b).

To compare our method with another common dissimilarity measure, we repeat the same simulation and clustering procedure 50 times using Euclidean distance as the dissimilarity measure. The average cRate is 67.67% using Euclidean distance. The side-by-side boxplots of the classification rates using the DSM metric and using the Euclidean metric are given in Figure 4.6. Note that 9 out of 50 repetitions using the DSM metric have 100% cRate.

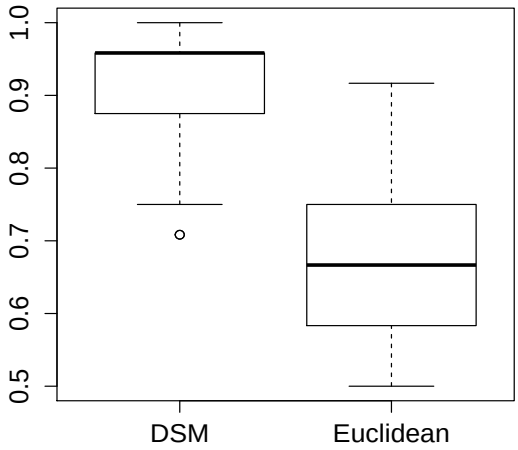

Figure 4.6 Left: cRate using DSM metric.
Right: cRate using Euclidean metric.

Berkeley Growth Curves

We applied our proposed method on the Berkeley growth acceleration data described in Section 3.5. Assuming that the number of clusters $C = 2$, we perform 200 iterations with first 80 as burn-in for a pair of curves with the tuning parameters shown in Table 4.1, and use Ward's method in the `R` function `hclust` to determine the cluster membership. We model the signal functions with 10 B-spline basis functions of order 5 defined on a equally spaced knot sequence. The prior mean of each coefficient is set to be 0, and the variance is $\mathbf{I}_{10}$.

Table 4.1 Parameter choices for growth data

Parameter	Description	Value
M	# of jumps of Warping Approx.	20
ϕ_1	vert. shift for raw curve	0.5
ϕ_2	vert. shift for deriv.	0.1
α_1	conc. param. for raw curve	30
α_2	conc. param. for deriv.	60
λ_1	smoothing. param. for raw curve	10^{-9}
λ_2	smoothing param. for deriv.	10^{-5}
$\sigma^2_{a_1}$	var. of stretching factor for raw curves	0.02^2
$\sigma^2_{a_2}$	var. of stretching factor for deriv.	0.03^2
T	# of points taken for deriv.	40
(w_1, w_2, w_3)	weights for clustering	$(0.7, 0.2, 0.1)$

The clustering results are shown in Table 4.2; only 8 females and 3 males are misclassified as the opposite gender, yielding an overall cRate of 90.32%. The clustering results are plotted in Panel (c) and (d) of Figure 4.7; the bold solid curves represent those boys who are misclassified as girls, and the bold dashed curves represent the misclassified girls. The right panel shows the curves after registration.

To compare our method with a classic distance-based clustering method, we cluster the growth acceleration data based on Euclidean distance. The results are shown

Table 4.2 Clustering results for Berkeley acceleration curves.

	True cluster	
	Male	Female
Cluster I	36	8
Cluster II	3	46

in Table 4.3; all boys are classified correctly but 23 girls are misclassified, yielding a 75.87% cRate.

Table 4.3 Clustering results for Berkeley acceleration curves based on Euclidean distance.

	True cluster	
	Male	Female
Cluster I	39	0
Cluster II	23	31

Response of Human Fibroblasts to Serum

Iyer et al. [1999] measured the response of fibroblasts to serum of 8613 time-course gene expressions using cDNA microarrays. Normal human fibroblasts require growth factors for proliferation, which is usually provided by fetal bovine serum (FBS). The authors stimulated the fibroblasts of serum deprivation by addition of medium containing 10% FBS. The responses were measured at 12 times, ranging from 15 minutes to 24 hours after serum stimulation. They applied a cluster analysis on a subset of 517 genes whose expression changed substantially in response to serum. We analyze a subset of 80 gene expressions. The raw data are shown in Figure 4.8.

We perform 200 iterations (with first 80 as burn-in) for a pair of curves with the tuning parameters shown in Table 4.4. We model the signal functions with 10 B-spline basis functions of order 5 defined on a equally spaced knot sequence. The prior mean

of each coefficient is set to be 0, and the variance is $\mathbf{I}_{10}$. We use Ward's method in the `R` function `hclust` to determine the cluster membership. To determine the number of clusters, **?** proposed a graphical method called silhouettes to validate the clustering quality and determine the proper number of clusters. The plot of the number of clusters versus the average silhouette width is shown in Figure 4.9. The silhouette width is between -1 and 1 with a larger number indicating a better clustering quality. Figure 4.9 suggests the proper number of clusters is between 3 and 6. After a careful graphical examination of the registered curves in each cluster, we decide choose the number of clusters $C = 4$.

The clustering result is given in Figure 4.10. The registered curves shows a clearer pattern after the vertical shifts are removed as shown in Figure 4.10 (b). All curves in the same cluster are roughly follow the same pattern as shown in panel (c)-(f) of Figure 4.10. Note that four curves separate from the majority in panel (b) due to the vertical shifts. One advantage of our method over the classic registration method proposed by Ramsay and Silverman [2005] is the ability of handling vertical shifts. The mean (bold) curves of panel (c) and (f) seem to follow a similar shape. However, the left portion in (c) is concave down while the counterpart in (f) is concave up. Subtle differences in terms of shape like that are successfully captured by the DSM method. For an explanation of the connection between the clustering result and the functionalities of each gene, for example, see Zhang and Telesca [2014], who previously analyzed this data set.

4.5 Discussion

We have developed a derivative-based method to measure the dissimilarity between a pair of curves with their possible phase variation removed by our Bayesian registration method. If one believes the important difference among clusters is subtle shape variation like different concavity on the same increasing or decreasing interval, our

Table 4.4 Parameter choices for HFS data

Parameter	Description	Value
M	# of jumps of Warping Approx.	20
ϕ_1	vert. shift for raw curve	3
ϕ_2	vert. shift for deriv.	1
α_1	conc. param. for raw curve	5
α_2	conc. param. for deriv.	10
λ_1	smoothing. param. for raw curve	10^{-9}
λ_2	smoothing param. for deriv.	10^{-5}
$\sigma^2_{a_1}$	var. of stretching factor for raw curves	0.05^2
$\sigma^2_{a_2}$	var. of stretching factor for deriv.	0.05^2
T	# of points taken for deriv.	30
(w_1, w_2, w_3)	weights for clustering	$(0.7, 0.2, 0.1)$

proposed method is effectively capture such difference by examining the higher order derivatives. Our Bayesian registration scheme provides a flexible yet computational efficient way to register a pair of curves. Compared to the classic registration methods, the ability of handling vertical shifts of our Bayesian approach is necessarily for applications like RHFS in section 4.4.

We demonstrate clustering accuracy our algorithm via simulation studies and real data analyses. For the choices of various parameters, see the discussion in section 3.4. They recommended to chose steps of discrete approximation M based on a scree plot, and use large shift parameter ϕ and small stretching/shrinking parameter σ_a^2 when uncertain.

In the spirit of nonparametric method, our algorithm imposes fewer assumptions on the curves. By using the derivative-shape measure as a distance proxy, our method is robust against vertical shifts, stretches, and shrinkages among curves. Our Bayesian registration method allows a fairly flexible approximation to the warping functions, which greatly extends the scope of applications compared to the existing methods.

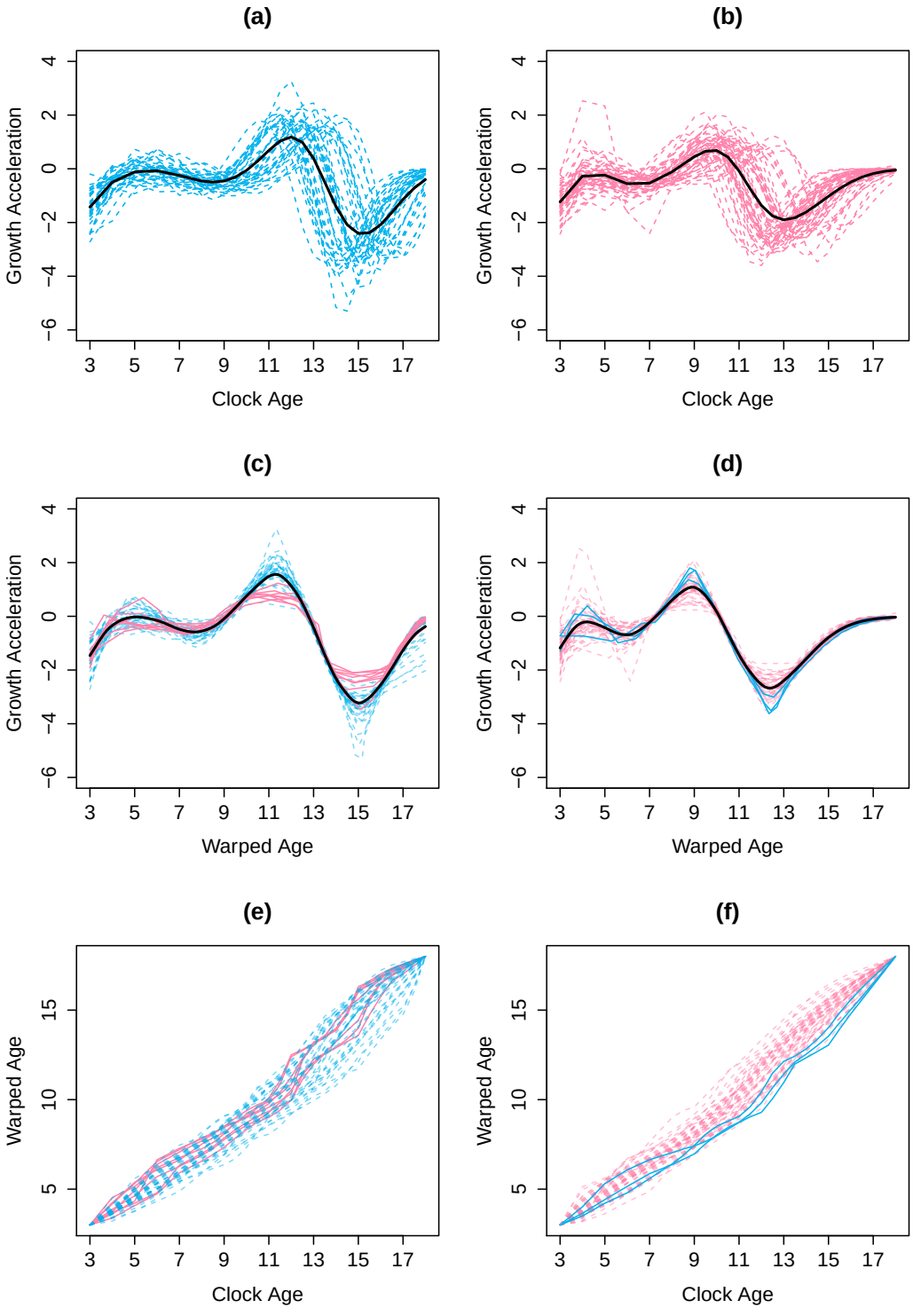

Figure 4.7 (a)-(b) Unregistered growth acceleration data for 39 boys (blue dashed) and 54 girls (pink dashed) with cross-sectional mean superimposed. (c) Registered cluster 1 with 36 boys (blue dashed) and 3 girls (pink solid). (d) Registered cluster 2 with 46 girls (pink dashed) and 8 boys (blue dashed). (e)-(f) Estimated warping functions for cluster 1 and cluster 2, respectively.

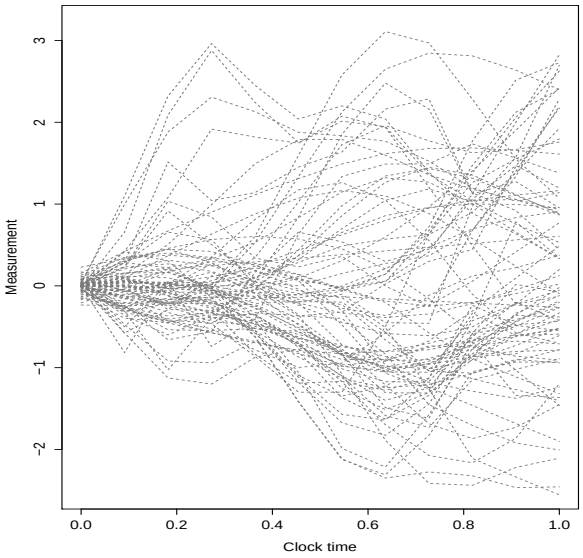

Figure 4.8 Raw data of the response of human fibroblasts to serum.

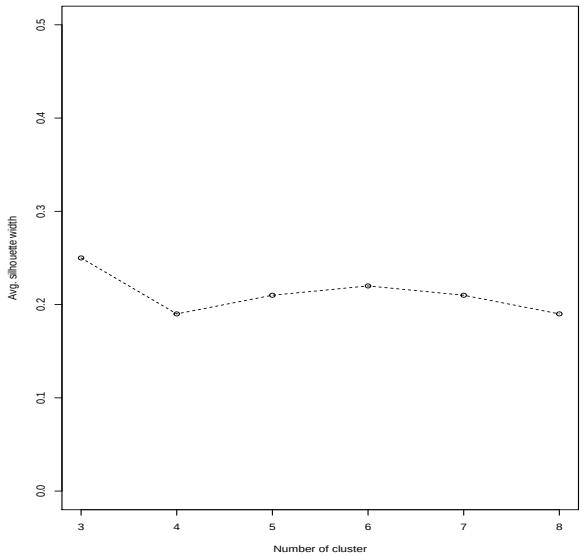

Figure 4.9 The number of clusters versus the average silhouette width.

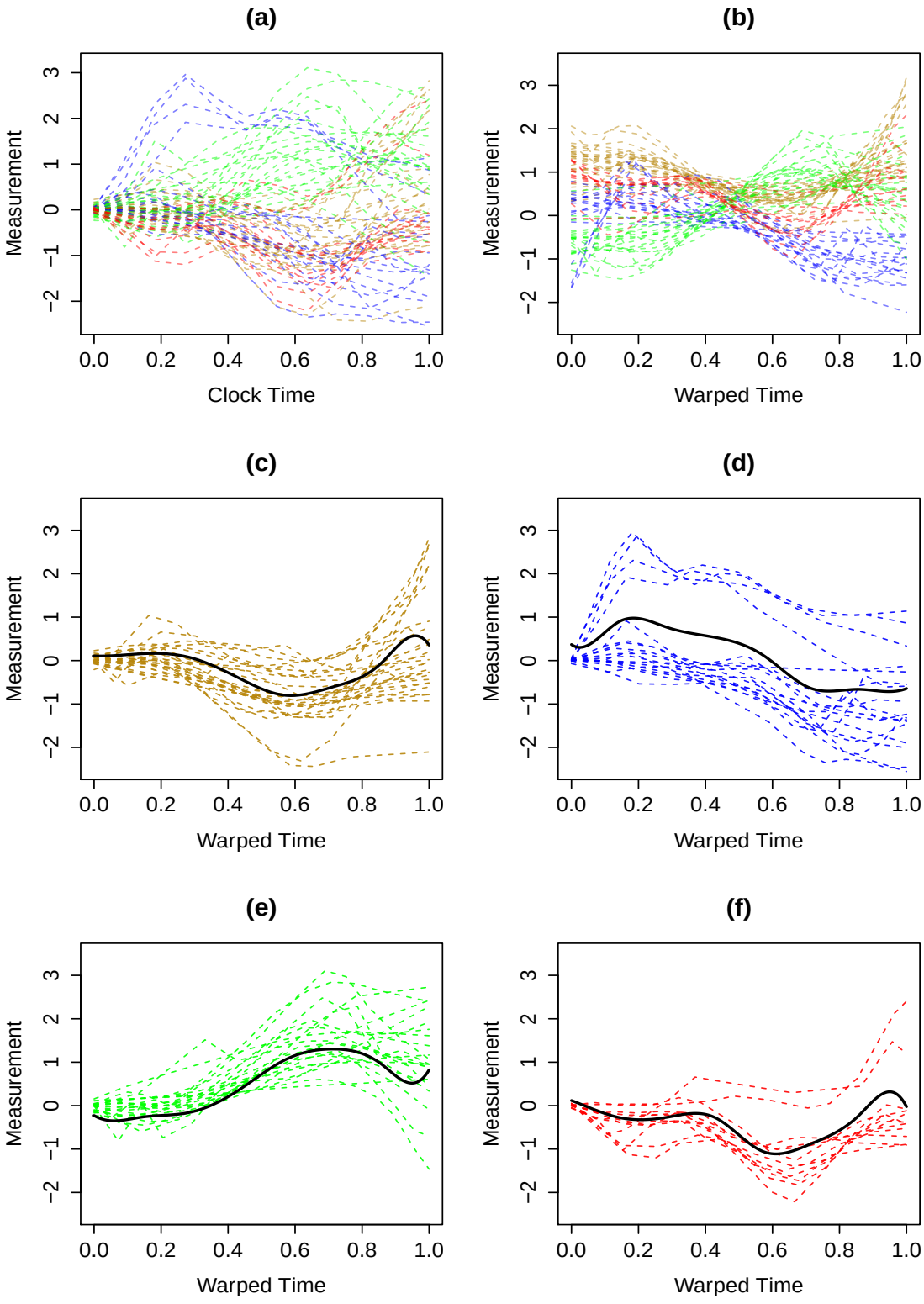

Figure 4.10 (a) Raw data with cluster structure determined by the algorithm. (b) Registered curves with vertical shifts removed. (c)-(f) Registered four clusters with their estimated mean curves superimposed.

Chapter 5

Adapted Variational Bayes Method

In this chapter, we derive an adapted variational Bayes method [Earls and Hooker, 2015] for our model proposed in Chapter 2. As an inference method that is fast compared to Markov chain Monte Carlo, the application of variational approximations are popular in the computer science community and gaining more attention from the statistics community [Ormerod and Wand, 2010]. Here, we focus on the approximation under the product density transformation [Bishop, 2006, Ormerod and Wand, 2010]. Let $\boldsymbol{y}$ and $\boldsymbol{\theta}$ denote observations and parameters, respectively. Then it can be shown that

$$\log p(\boldsymbol{y}) = \int q(\boldsymbol{\theta}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \left\{ \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})} \right\} \mathrm{d}\boldsymbol{\theta}, \tag{5.1}$$

where $q(\cdot)$ is an arbitrary density function of the parameter space Θ, and $p(\boldsymbol{y})$ is the marginal likelihood function. Note that the second term in the RHS of equation (5.1) is the Kullback-Leibler divergence [Kullback and Leibler, 1951] between $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\boldsymbol{y})$, satisfying $\int q(\boldsymbol{\theta}) \log \left\{ \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})} \right\} \mathrm{d}\boldsymbol{\theta} \geq 0$. It follows that

$$\log p(\boldsymbol{y}) \geq \int q(\boldsymbol{\theta}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\theta} \tag{5.2}$$

with equality attained when $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{y})$ almost everywhere. We define the lower bound of the variational approximation as $\exp \int q(\boldsymbol{\theta}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\theta}$. The goal is to maximize this lower bound, so that the K-L divergence between the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ and the approximation $q(\boldsymbol{\theta})$ is minimized. In the machine learning literature, $p(\boldsymbol{y})$ is called the model evidence [Bishop, 2006] which provides a foundation for performing Bayesian model selection.

The product density transforms assume the approximated posterior distribution $q(\boldsymbol{\theta}) = \prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i)$. It is important to notice that this is the only assumption imposed in order to make the approximation. It takes an iterative approach to maximize the lower bound. The current update for the i-th parameter $\boldsymbol{\theta}_i$, $i = 1, 2, \ldots, M$, involves updating

$$q_i^*(\boldsymbol{\theta}_i) \leftarrow \frac{\exp\{\mathbb{E}_{-\boldsymbol{\theta}_i} \log p(\boldsymbol{y}, \boldsymbol{\theta})\}}{\int \exp\{\mathbb{E}_{-\boldsymbol{\theta}_i} \log p(\boldsymbol{y}, \boldsymbol{\theta})\} \, \mathrm{d}\boldsymbol{\theta}_i}, \tag{5.3}$$

where the expectations are taken with respect to all updated parameters but $\boldsymbol{\theta}_i$, and the asterisk indicating the updated optimal approximation of $p(\boldsymbol{\theta}_i|\boldsymbol{y})$ at the current iteration. This updating scheme increases the lower bound $\exp \int q(\boldsymbol{\theta}) \log \left\{ \frac{p(\boldsymbol{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\theta}$ at each iteration, and therefore the local optimum is guaranteed [Ormerod and Wand, 2010]. In practice, we monitor the lower bound until a certain convergence criterion is satisfied. It can be shown that the expectation-maximization (EM) [Dempster et al., 1977] algorithm could be viewed as a special case of this variational approximation algorithm [Tzikas et al., 2008].

If the prior of $\boldsymbol{\theta}_i$ is in the conjugate family, we obtain the closed-form update of $\boldsymbol{\theta}_i$ similar to the MCMC counterpart without integrating the denominator in the expression (5.3). If some parameter $\boldsymbol{\theta}_k$ does not have a conjugate prior, Earls and Hooker [2015] suggest updating the estimate of $\boldsymbol{\theta}_k$ by maximizing $q_k(\boldsymbol{\theta}_k)$ with respect to $\boldsymbol{\theta}_k$, which is equivalent to updating

$$\boldsymbol{\theta}_k^* = \arg\sup_{\boldsymbol{\theta}_k} \left\{ \exp\{\mathbb{E}_{-\boldsymbol{\theta}_i} \log p(\boldsymbol{y}, \boldsymbol{\theta})\} \right\}. \tag{5.4}$$

Earls and Hooker [2015] refer to this modified approach as the adapted variational Bayes (AVB). It is straightforward to show that the AVB algorithm increases the lower bound at each iteration. Following the derivation of the posterior sampling in section 3.2 for our Bayesian model proposed in Chapter 3, there is no closed-form update for the Dirichlet jumps $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_N$. We will update these jumps by directly maximizing the corresponding $q()$ function, which is defined in (5.7), via a

constrained optimization. The details of the constrained optimization are discussed in Appendix C; The details of the AVB algorithm are given in Section 5.1, and the convergence criterion is given in Appendix D.

5.1 Adapted Variational Bayes Algorithm

Following the full Bayesian method described in section 3.1, the joint distribution of the data and parameters is

$$
\begin{aligned}
& \mathcal{L}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_C, \gamma_1, \ldots, \gamma_N, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N, p_1, \ldots, p_C, \tau, a_1, \ldots, a_N, s_1, \ldots, s_N, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) \\
\propto\ & \prod_{i=1}^{N} \tau^{K/2} \exp\left\{ -\frac{1}{2}\tau \sum_{i=1}^{N}\sum_{c=1}^{C} \left(\|\boldsymbol{y}_i - a_i \boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta}_c - \mathbf{s}_i\|^2 \right) z_{ic} \right\} \\
& \prod_{c=1}^{C} \exp\left\{ -\frac{1}{2}(\boldsymbol{\beta}_c - \boldsymbol{\beta}_0^c)^T \Gamma^{-1} (\boldsymbol{\beta}_c - \boldsymbol{\beta}_0^c) \right\} \\
& \prod_{i=1}^{N}\prod_{m=1}^{M} \gamma_{im}^{\alpha_0 - 1} \prod_{c=1}^{C} p_c^{\sum_{i=1}^{N} z_{ic} + \eta_c - 1} \tau^{\kappa - 1} \exp\{-\tau\theta\} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{N}(a_i - 1)^2 \right\} \prod_{i=1}^{N} \mathbf{1}_{\{-\phi < s_i < \phi\}} \\
\propto\ & \exp\left\{ \frac{KN}{2}\ln\tau - \frac{1}{2}\tau \sum_{i=1}^{N}\sum_{c=1}^{C} \left(\|\boldsymbol{y}_i - a_i \boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta}_c - \mathbf{s}_i\|^2 \right) z_{ic} \right\} \\
& \prod_{c=1}^{C} \exp\left\{ -\frac{1}{2}(\boldsymbol{\beta}_c - \boldsymbol{\beta}_0^c)^T \Gamma^{-1} (\boldsymbol{\beta}_c - \boldsymbol{\beta}_0^c) \right\} \\
& \prod_{i=1}^{N}\prod_{m=1}^{M} \gamma_{im}^{\alpha_0 - 1} \prod_{c=1}^{C} p_c^{\sum_{i=1}^{N} z_{ic} + \eta_c - 1} \tau^{\kappa - 1} \exp\{-\tau\theta\} \exp\left\{ -\frac{1}{2\sigma_a^2}\sum_{i=1}^{N}(a_i - 1)^2 \right\} \prod_{i=1}^{N} \mathbf{1}_{\{-\phi < s_i < \phi\}}
\end{aligned}
$$

The log-likelihood function is given by

$$
\begin{aligned}
& \ln \mathcal{L}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_C, \gamma_1, \ldots, \gamma_N, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N, p_1, \ldots, p_C, \tau, a_1, \ldots, a_N, s_1, \ldots, s_N, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) \\
=\ & \frac{KN}{2}\ln\tau - \frac{1}{2}\tau \sum_{i=1}^{N}\sum_{c=1}^{C} \left(\|\boldsymbol{y}_i - a_i \boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta}_c - \mathbf{s}_i\|^2 \right) z_{ic} + \sum_{c=1}^{C} -\frac{1}{2}(\boldsymbol{\beta}_c - \boldsymbol{\beta}_0^c)^T \Gamma^{-1} (\boldsymbol{\beta}_c - \boldsymbol{\beta}_0^c) \\
& + (\alpha_0 - 1)\sum_{i=1}^{N}\sum_{m=1}^{M} \ln\gamma_{im} + \sum_{i=1}^{C}\sum_{i=1}^{N} (z_{ic} + \eta_c - 1)\ln p_c + (\kappa - 1)\ln\tau - \tau\theta - \\
& \frac{1}{2\sigma_a^2}\sum_{i=1}^{N}(a_i - 1)^2 + \sum_{i=1}^{N} \ln \mathbf{1}_{\{-\phi < s_i < \phi\}} + \text{const.}
\end{aligned}
$$

Under the product decomposition assumption, we have

$$
q(\boldsymbol{z}, \boldsymbol{p}, \gamma, \boldsymbol{\beta}, \tau, \boldsymbol{a}, \boldsymbol{s}) = q(\boldsymbol{z})q(\boldsymbol{p})q(\boldsymbol{\gamma})q(\boldsymbol{\beta})q(\tau)q(\boldsymbol{a})q(\boldsymbol{s})
$$

For each iteration, we need to calculate

$$\begin{aligned}
&\ln q^*(\mathbf{Z}) = \mathbb{E}_{-\mathbf{Z}}[\ln \mathcal{P}(\mathbf{Y}, \mathbf{Z}, \boldsymbol{p}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{s}, \tau)] \\
= \;& \mathbb{E}_{\boldsymbol{\beta},\tau,\boldsymbol{a},\boldsymbol{s}}[\ln \mathcal{P}(\mathbf{Y}, \boldsymbol{\gamma}|\mathbf{Z}, \boldsymbol{a}, \boldsymbol{s})] + \mathbb{E}_{\boldsymbol{p}}[\ln \mathcal{P}(\boldsymbol{z}|\boldsymbol{p})] \\
= \;& \mathbb{E}_{\boldsymbol{\beta},\tau,\boldsymbol{a},\boldsymbol{s}}\left[K/2 \ln \tau - \frac{1}{2}\tau \sum_{i=1}^{N}\sum_{c=1}^{C}\left(\|\boldsymbol{y}_i - a_i \boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta}_c - \mathbf{s}_i\|^2\right) z_{ic}\right] + \\
& \mathbb{E}_{\boldsymbol{p}}\left[\sum_{c=1}^{C}\sum_{i=1}^{N}(z_{ic} + \eta_c - 1)\ln p_c\right] \\
= \;& \sum_{i=1}^{N}\sum_{c=1}^{C} z_{ic}\left\{\frac{K}{2}\mathbb{E}_\tau(\ln \tau) - \frac{1}{2}\mathbb{E}_\tau(\tau)\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}}\|\boldsymbol{y}_i - a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta}_c - \mathbf{s}_i\|^2 + \mathbb{E}_{\boldsymbol{p}}(\ln p_c)\right\} + \\
& \text{constant not involving } \boldsymbol{z}.
\end{aligned}$$

Further, we have

$$\begin{aligned}
& \mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}}\|\boldsymbol{y}_i - a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta}_c - \mathbf{s}_i\|^2 \\
= \;& \mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}}\left(\sum_{j=1}^{K}(y_{ij} - a_i\phi[\gamma_i(t_j)]\boldsymbol{\beta}_c - s_i)^2\right) \\
= \;& \sum_{j=1}^{K}\Big(y_{ij}^2 + \mathbb{E}_{\boldsymbol{a}}(a_i^2)\mathbb{E}_{\boldsymbol{\beta}}(\phi[\gamma_i(t_j)]\boldsymbol{\beta}_c)^2 + \mathbb{E}_{\boldsymbol{s}}(s_i^2) - 2\mathbb{E}_{\boldsymbol{a}}(a_i)\phi[\gamma_i(t_j)]\mathbb{E}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_c)y_{ij} - 2\mathbb{E}_{\boldsymbol{s}}(s_i)y_{ij} \\
& +2\mathbb{E}_{\boldsymbol{a}}(a_i)\mathbb{E}_{\boldsymbol{s}}(s_i)\phi[\gamma_i(t_j)]\mathbb{E}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_c)\Big) \\
= \;& \sum_{j=1}^{K}\Big(y_{ij}^2 + \left[(\sigma_{a_i}^2)^* + (\mu_{a_i}^*)^2\right]\phi[\gamma_i(t_j)](\Sigma_{\boldsymbol{\beta}_c}^* + \boldsymbol{\mu}_{\boldsymbol{\beta}_c}^*\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^{*T})\phi^T[\gamma_i(t_j)] + (\sigma_{s_i}^2)^* + (\mu_{s_i}^*)^2 \\
& -2\mu_{a_i}^*\phi[\gamma_i(t_j)]\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^* y_{ij} - 2\mu_{s_i}^* y_{ij} + 2\mu_{a_i}^*\mu_{s_i}^*\phi[\gamma_i(t_j)]\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^*\Big). \qquad (5.5)
\end{aligned}$$

Define

$$\begin{aligned}
& \ln \rho_{ic} \\
= \;& \frac{K}{2}\mathbb{E}_\tau(\ln \tau) - \frac{1}{2}\mathbb{E}_\tau(\tau)\left\{\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}}\|\boldsymbol{y}_i - a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta}_c - \mathbf{s}_i\|^2\right\} + \mathbb{E}_{\boldsymbol{p}}(\ln p_c).
\end{aligned}$$

We have

$$q^*(\mathbf{Z}) \;\propto\; \prod_{i=1}^{N}\prod_{c=1}^{C}\rho_{ic}^{z_{ic}}.$$

By normalizing the above distribution, we obtain

$$q^*(\mathbf{Z}) = \prod_{i=1}^{N}\prod_{c=1}^{C} r_{ic}^{z_{ic}},$$

where we have defined

$$r_{ic} = \frac{\rho_{ic}}{\sum_{l=1}^{C} \rho_{il}}. \tag{5.6}$$

The quantity r_{ic} is also called "responsibility" in the machine learning literature.

For the warping function, we need to maximize $\ln q^*(\gamma_i)$ with respect to γ_i. For the i-th observation, the log-likelihood function is given by

$$\begin{aligned}
& \ln q^*(\gamma_i) \\
= \; & -\frac{1}{2}\mathbb{E}_{\tau,\boldsymbol{\beta},\boldsymbol{z}}\left\{\tau\left[\sum_{c=1}^{C}\left(\|\boldsymbol{y}_i - a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta}_c - \mathbf{s}_i\|^2\right) z_{ic}\right]\right\} + (\alpha_0 - 1)\sum_{m=1}^{M}\ln\gamma_{im} + \text{const.} \\
= \; & -\frac{1}{2}\sum_{c=1}^{C}\left\{r_{ic}\mu_\tau^*\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}}\|\boldsymbol{y}_i - a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta}_c - \mathbf{s}_i\|^2\right\} + (\alpha_0 - 1)\sum_{m=1}^{M}\ln\gamma_{im} + \text{const} \\
= \; & -\frac{1}{2}\sum_{c=1}^{C}\Bigg\{r_{ic}\mu_\tau^*\sum_{j=1}^{K}\Bigg(\left[(\sigma_{a_i}^2)^* + (\mu_{a_i}^*)^2\right]\phi[\gamma_i(t_j)](\Sigma_{\boldsymbol{\beta}_c}^* + \boldsymbol{\mu}_{\boldsymbol{\beta}_c}^*\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^{*T})\phi^T[\gamma_i(t_j)] \\
& -2\mu_{a_i}^*\phi[\gamma_i(t_j)]\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^* y_{ij} + 2\mu_{a_i}^*\mu_{s_i}^*\phi[\gamma_i(t_j)]\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^*\Bigg)\Bigg\} \\
& +(\alpha_0 - 1)\sum_{m=1}^{M}\ln\gamma_{im} + \text{const}
\end{aligned} \tag{5.7}$$

under the constraint $\sum_{m=1}^{M}\gamma_{im} = 1$ and $\gamma_{im} > 0$. We maximize (5.7) with respect to $\gamma_{i1}, \ldots, \gamma_{iM}$ to obtain the optimal $\boldsymbol{\gamma}_i^*$. For an efficient and accurate optimization, it is necessary to derive the gradient of (5.7). Due to the simplicity of the linear approximation, it is possible to derive the closed-form expression. The formula of the derivative is given by (C.5) in Appendix C.

For cluster probability $\boldsymbol{p}$, we have

$$\begin{aligned}
\ln q^*(\boldsymbol{p}) &= \mathbb{E}_{-\boldsymbol{p}}\left\{\sum_{c=1}^{C}\left(\sum_{i=1}^{N} z_{ic} + \eta - 1\right)\ln p_c\right\} + \text{ const} \\
&= \sum_{c=1}^{C}\left(\sum_{i=1}^{N} r_{ic} + \eta - 1\right)\ln p_c + \text{ const}
\end{aligned}$$

Therefore,

$$q^*(\boldsymbol{p}) \sim Dir\left(\sum_{i=1}^{N} r_{i1} + \eta, \ldots, \sum_{i=1}^{N} r_{iC} + \eta\right).$$

To ease the notation in deriving the lower bound of the algorithm, let us denote

$$\alpha_c = \sum_{i=1}^{N} r_{ic} + \eta, \tag{5.8}$$

and

$$\hat{\alpha} = \sum_{c=1}^{C}\sum_{i=1}^{N} r_{ic} + C\eta = N + c\eta. \tag{5.9}$$

For the spline coefficient $\boldsymbol{\beta}_k$,

$$\begin{aligned}
& \ln q^*(\boldsymbol{\beta}_k) \\
= \;& \mathbb{E}_{-\beta_k}\left\{-\frac{1}{2}\tau\sum_{i=1}^{N}\left(\|\boldsymbol{y}_i - a_i\boldsymbol{\phi}[\gamma_i^*(\mathbf{t})]\boldsymbol{\beta}_k - \mathbf{s}_i\|^2\right) z_{ik} - \frac{1}{2}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{0k})^T\Gamma^{-1}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{0k})\right\} \\
= \;& -\frac{1}{2}\sum_{i=1}^{N}\left\{r_{ik}\mathbb{E}_\tau(\tau)\mathbb{E}_{\boldsymbol{a},\boldsymbol{s}}\left(\|\boldsymbol{y}_i - a_i\boldsymbol{\phi}[\gamma_i^*(\mathbf{t})]\boldsymbol{\beta}_k - \mathbf{s}_i\|^2\right)\right\} - \frac{1}{2}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{0k})^T\Gamma^{-1}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{0k}) \\
& + \text{const} \\
= \;& -\frac{1}{2}\sum_{i=1}^{N}\left(r_{ik}\mathbb{E}_\tau(\tau)\mathbb{E}_{\boldsymbol{a}}(a_i^2)\boldsymbol{\beta}_k^T\boldsymbol{\phi}^T[\gamma_i^*(\mathbf{t})]\boldsymbol{\phi}[\gamma_i^*(\mathbf{t})]\boldsymbol{\beta}_k\right) - \frac{1}{2}\boldsymbol{\beta}_k^T\Gamma^{-1}\boldsymbol{\beta}_k + \\
& \sum_{i=1}^{N}(\boldsymbol{\beta}_k^T r_{ik}\mathbb{E}_\tau(\tau)\mathbb{E}_{\boldsymbol{a}}(a_i)\boldsymbol{\phi}^T[\gamma_i^*(\mathbf{t})])(\boldsymbol{y}_i - \boldsymbol{\mu}_{s_i}^*) - \boldsymbol{\beta}_k^T\Gamma^{-1}\boldsymbol{\beta}_{0k} + \text{ const} \\
= \;& -\frac{1}{2}\boldsymbol{\beta}_k^T\underbrace{\left(\mu_\tau^*\sum_{i=1}^{N} r_{ik}((\sigma_{a_i}^2)^* + \mu_{a_i}^2)\boldsymbol{\phi}^T[\gamma_i^*(\mathbf{t})]\boldsymbol{\phi}[\gamma_i^*(\mathbf{t})] + \Gamma^{-1}\right)}_{\text{call it } \mathbf{A}_k}\boldsymbol{\beta}_k - \\
& \qquad \boldsymbol{\beta}_k^T\underbrace{\left(\mu_\tau^*\sum_{i=1}^{N} r_{ik}\mu_{a_i}^*\boldsymbol{\phi}^T[\gamma_i^*(\mathbf{t})](\boldsymbol{y}_i - \mu_{s_i}) + \Gamma^{-1}\boldsymbol{\beta}_{0k}\right)}_{\text{call it } \mathbf{c}_k} + \text{const.}
\end{aligned}$$

It follows that $\boldsymbol{\beta}_k^* \sim \text{MVN}(\mathbf{A}_k^{-1}\mathbf{c}_k, \mathbf{A}_k^{-1})$.

For precision parameter τ,

$$\begin{aligned}
\ln q^*(\tau) \;=\; & \frac{KN}{2}\ln\tau - \frac{1}{2}\tau\sum_{i=1}^{N}\sum_{c=1}^{C} r_{ic}\left(\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}}\|\boldsymbol{y}_i - a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta}_c - \mathbf{s}_i\|^2\right) + \\
& (\kappa - 1)\ln\tau - \theta\tau + \text{ const.}
\end{aligned}$$

It follows that

$$\tau^* \sim Gamma\left(\frac{KN}{2} + \kappa, \frac{1}{2}\sum_{i=1}^{N}\sum_{c=1}^{C} r_{ic}\left(\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}}\|\boldsymbol{y}_i - a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta}_c - \mathbf{s}_i\|^2\right) + \theta\right),$$

where $\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}}(\|\boldsymbol{y}_i - a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta}_c - \mathbf{s}_i\|^2)$ is given by (5.5). Let us denote the updated shape and rate parameters by

$$\kappa^* = \frac{KN}{2} + \kappa, \tag{5.10}$$

and

$$\theta^* = \frac{1}{2}\sum_{i=1}^{N}\sum_{c=1}^{C} r_{ic}\left(\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}}\left\|\boldsymbol{y}_i - a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta}_c - \mathbf{s}_i\right\|^2\right) + \theta. \tag{5.11}$$

For stretching/shrinking factor a_i, we have

$$\begin{aligned}
& \ln q^*(a_i) \\
= \;& \mathbb{E}_{-\boldsymbol{a}}\left\{-\frac{1}{2}\tau\sum_{c=1}^{C} z_{ic}\left\|\boldsymbol{y}_i - a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta}_c - \mathbf{s}_i\right\|^2\right\} - \frac{1}{2\sigma_a^2}(a_i-1)^2 + \text{const} \\
= \;& \mathbb{E}_{-\boldsymbol{a}}\left\{-\frac{1}{2}\tau\sum_{c=1}^{C} z_{ic}\left[\sum_{j=1}^{K} a_i^2(\phi[\gamma_i(t_j)]\boldsymbol{\beta}_c\boldsymbol{\beta}_c^T\phi^T[\gamma_i(t_j)]) - \sum_{j=1}^{K} 2a_i\phi[\gamma_i(t_j)]\boldsymbol{\beta}_c(y_{ij}-s_i)\right]\right\} \\
& -\frac{1}{2\sigma_a^2}a_i^2 + \frac{1}{\sigma_a^2}a_i + \text{const} \\
= \;& -\frac{1}{2}\left\{\frac{1}{\sigma_a^2} + \mu_\tau^*\sum_{c=1}^{C} r_{ic}\left[\sum_{j=1}^{K}\phi[\gamma_i(t_j)])(\Sigma_{\boldsymbol{\beta}_c}^* + \boldsymbol{\mu}_{\boldsymbol{\beta}_c}^*(\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^*)^T)\phi^T[\gamma_i(t_j)])\right]\right\} a_i^2 \\
& + \left\{\mu_\tau^*\sum_{c=1}^{C} r_{ic}\left[\sum_{j=1}^{K}\mu_{icj}^*(y_{ij}-\mu_{s_i})\right] + \frac{1}{\sigma_a^2}\right\} a_i + \text{const},
\end{aligned}$$

where we have defined the j-th element of $\boldsymbol{\phi}(\gamma_i^*(\mathbf{t}))\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^*$ and $\boldsymbol{y}_i$ by μ_{icj}^* and y_{ij}, respectively.

By completing the square, we have the optimal $q^*(a_i)$ follows a normal distribution with

$$\mu_{a_i}^* = \frac{\mu_\tau^*\sum_{c=1}^{C} r_{ic}\left[\sum_{j=1}^{K}\mu_{icj}^*(y_{ij}-\mu_{s_i})\right] + 1/\sigma_a^2}{1/\sigma_a^2 + \mu_\tau^*\sum_{c=1}^{C} r_{ic}\left[\sum_{j=1}^{K}\phi[\gamma_i^*(t_j)])(\Sigma_{\boldsymbol{\beta}_c}^* + \boldsymbol{\mu}_{\boldsymbol{\beta}_c}^*(\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^*)^T)\phi^T[\gamma_i(t_j)])\right]}, \tag{5.12}$$

and

$$(\sigma_{a_i}^2)^* = \frac{1}{1/\sigma_a^2 + \mu_\tau^*\sum_{c=1}^{C} r_{ic}\left[\sum_{j=1}^{K}\phi[\gamma_i^*(t_j)])(\Sigma_{\boldsymbol{\beta}_c}^* + \boldsymbol{\mu}_{\boldsymbol{\beta}_c}^*(\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^*)^T)\phi^T[\gamma_i(t_j)])\right]}. \tag{5.13}$$

For the i-th shift parameter s_i, we have

$$\begin{aligned}
& \ln q^*(s_i) \\
= \; & \mathbb{E}_{-\boldsymbol{a}} \left\{ -\frac{1}{2}\tau \sum_{c=1}^{C} z_{ic} \left\| \boldsymbol{y}_i - a_i \boldsymbol{\phi}[\gamma_i^*(\mathbf{t})]\boldsymbol{\beta}_c - \mathbf{s}_i \right\|^2 \right\} + \ln \mathbf{1}_{\{-\phi<s_i<\phi\}} + \text{const} \\
= \; & \mathbb{E}_{-\boldsymbol{a}} \left\{ -\frac{1}{2}\tau \sum_{c=1}^{C} z_{ic} \left[\sum_{j=1}^{K} (s_i^2 - 2(y_{ij} - a_i \phi[\gamma_i^*(t_j)]\boldsymbol{\beta}_c) s_i \right] \right\} + \ln \mathbf{1}_{\{-\phi<s_i<\phi\}} + \text{const} \\
= \; & -\frac{1}{2}\mathbb{E}_\tau(\tau) K \left(s_i - \frac{\sum_{c=1}^{C} \left(r_{ic} \sum_{j=1}^{K} \mu_{a_i}^* d_{icj} \right)}{K} \right)^2 + \ln \mathbf{1}_{\{-\phi<s_i<\phi\}} + \text{const},
\end{aligned}$$

where we have defined $d_{icj} = y_{ij} - a_i \phi[\gamma_i^*(t_j)]\boldsymbol{\beta}_c$.

It follows that

$$q^*(s_i) \sim N\left(\frac{\sum_{c=1}^{C} \left(r_{ic} \sum_{j=1}^{K} \mu_{a_i}^* d_{icj} \right)}{K}, \frac{1}{\mu_\tau^* K} \right) \mathbf{1}_{\{-\phi<s_i<\phi\}}.$$

Let us denote

$$\tilde{\mu}_{s_i} = \frac{\sum_{c=1}^{C} \left(r_{ic} \sum_{j=1}^{K} \mu_{a_i}^* d_{icj} \right)}{K}, \tag{5.14}$$

and

$$\tilde{\sigma}_{s_i}^2 = \frac{1}{\mu_\tau^* K}. \tag{5.15}$$

By a property of the truncated normal distribution, the approximated mean and variance of s_i at current iteration are given by

$$\mu_{s_i}^* = \tilde{\mu}_{s_i} + \frac{\phi(\frac{-\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}}) - \phi(\frac{\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}})}{\Phi(\frac{-\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}}) - \Phi(\frac{\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}})} \tilde{\sigma}_{s_i}, \tag{5.16}$$

and

$$(\sigma_{s_i}^2)^* = \tilde{\sigma}_{s_i}^2 \left[1 + \frac{\frac{-\phi-\tilde{\mu}_{s_i}}{\sigma_{s_i}} \phi(\frac{-\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}}) - \frac{\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}} \phi(\frac{\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}})}{\Phi(\frac{\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}}) - \Phi(\frac{-\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}})} - \left(\frac{\phi(\frac{-\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}}) - \phi(\frac{\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}})}{\Phi(\frac{\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}}) - \Phi(\frac{-\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}})} \right)^2 \right] \tag{5.17}$$

where $\phi(\cdot)$ is the probability density function of the standard normal distribution and $\Phi(\cdot)$ is its cumulative distribution function.

A summary of the AVB algorithm is given in Algorithm 1.

Algorithm 1: Adapted variational Bayes algorithm for simultaneous clustering and registration of functional data

Initialize:

For $i = 1, \ldots, N$,

$\gamma_{ik} > 0$ for $k = 1, \ldots, M$, such that $\sum_{k=1}^{M} \gamma_{ik} = 1$;

$r_{ic} > 0$ for $c = 1, \ldots, C$, such that $\sum_{c=1}^{C} r_{ic} = 1$;

$\mu_{s_i} \in \mathbb{R}$ and $\sigma_{s_i}^2 > 0$; $\mu_{a_i} \in \mathbb{R}$ and $\sigma_{a_i}^2 > 0$.

For $c = 1, \ldots, C$,

$\boldsymbol{\mu}_{\beta_c} \in \mathbb{R}^q$ and $\Sigma_{\boldsymbol{\beta}} > 0$;

$\mu_{p_c} > 0$ such that $\sum_{c=1}^{C} \mu_{p_c} = 1$; $\mu_{\ln p_c} < 0$.

$\mu_\tau > 0$; $\mu_{\ln \tau} \in \mathbb{R}$.

Cycle:

For $i = 1, \ldots, N$,

$\boldsymbol{\gamma}_i \leftarrow$

$\arg\sup_{\gamma_i} \left\{ -\frac{1}{2} \sum_{c=1}^{C} \left\{ r_{ic} \mu_\tau \mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}} \left\| \boldsymbol{y}_i - a_i \boldsymbol{\phi}[\gamma_i(\mathbf{t})] \boldsymbol{\beta}_c - \mathbf{s}_i \right\|^2 \right\} + (\alpha_0 - 1) \sum_{m=1}^{M} \ln \gamma_{im} \right\}$,

update $\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}} \left\| \boldsymbol{y}_i - a_i \boldsymbol{\phi}[\gamma_i(\mathbf{t})] \boldsymbol{\beta}_c - \mathbf{s}_i \right\|^2$ with $\boldsymbol{\gamma}_i$ updated,

$r_{ic} \leftarrow \frac{\exp\left\{ \frac{K}{2} \mu_{\ln \tau} - \frac{1}{2} \mu_\tau \left\{ \mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}} \| \boldsymbol{y}_i - a_i \boldsymbol{\phi}[\gamma_i(\mathbf{t})] \boldsymbol{\beta}_c - \mathbf{s}_i \|^2 \right\} + \mu_{\ln p_c} \right\}}{\sum_{l=1}^{C} \exp\left\{ \frac{K}{2} \mu_{\ln \tau} - \frac{1}{2} \mu_\tau \left\{ \mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}} \| \boldsymbol{y}_i - a_i \boldsymbol{\phi}[\gamma_i(\mathbf{t})] \boldsymbol{\beta}_l - \mathbf{s}_i \|^2 \right\} + \mu_{\ln p_l} \right\}}$,

$\mu_{a_i} \leftarrow \frac{\mu_\tau \sum_{c=1}^{C} r_{ic} \left[\sum_{j=1}^{K} \mu_{icj} (y_{ij} - \mu_{s_i}) \right] + 1/\sigma_a^2}{1/\sigma_a^2 + \mu_\tau \sum_{c=1}^{C} r_{ic} \left[\sum_{j=1}^{K} \phi[\gamma_i(t_j)]) (\Sigma_{\boldsymbol{\beta}_c} + \boldsymbol{\mu}_{\boldsymbol{\beta}_c} (\boldsymbol{\mu}_{\boldsymbol{\beta}_c})^T) \phi'[\gamma_i(t_j)]) \right]}$,

$\sigma_{a_i}^2 \leftarrow \frac{1}{1/\sigma_a^2 + \mu_\tau \sum_{c=1}^{C} r_{ic} \left[\sum_{j=1}^{K} \phi[\gamma_i(t_j)]) (\Sigma_{\boldsymbol{\beta}_c} + \boldsymbol{\mu}_{\boldsymbol{\beta}_c} (\boldsymbol{\mu}_{\boldsymbol{\beta}_c})^T) \phi^T[\gamma_i(t_j)]) \right]}$,

where μ_{icj} and y_{ij} are the j-th element of $\boldsymbol{\phi}(\gamma_i(\mathbf{t})) \boldsymbol{\mu}_{\boldsymbol{\beta}_c}$ and $\boldsymbol{y}_i$, respectively.

Update $\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}} \left\| \boldsymbol{y}_i - a_i \boldsymbol{\phi}[\gamma_i(\mathbf{t})] \boldsymbol{\beta}_c - \mathbf{s}_i \right\|^2$ with μ_{a_i} and $\sigma_{a_i}^2$ updated,

$\mu_{s_i} \leftarrow \tilde{\mu}_{s_i} + \frac{\phi(\frac{-\phi - \tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}}) - \phi(\frac{\phi - \tilde{\mu}_{s_i}}{\sigma_{s_i}})}{\Phi(\frac{-\phi - \tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}}) - \Phi(\frac{\phi - \mu_{s_i}}{\tilde{\sigma}_{s_i}})} \tilde{\sigma}_{s_i}$,

$\sigma_{s_i}^2 \leftarrow \tilde{\sigma}_{s_i}^2 \left[1 + \frac{\frac{-\phi - \tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}} \phi(\frac{-\phi - \tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}}) - \frac{\phi - \tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}} \phi(\frac{\phi - \tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}})}{\Phi(\frac{\phi - \tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}}) - \Phi(\frac{-\phi - \tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}})} - \left(\frac{\phi(\frac{-\phi - \tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}}) - \phi(\frac{\phi - \tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}})}{\Phi(\frac{\phi - \tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}}) - \Phi(\frac{-\phi - \tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}})} \right)^2 \right]$, where $\tilde{\mu}_{s_i}$ and $\tilde{\sigma}_{s_i}^2$ are given in (5.14) and (5.15), respectively.

Update $\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}} \left\| \boldsymbol{y}_i - a_i \boldsymbol{\phi}[\gamma_i(\mathbf{t})] \boldsymbol{\beta}_c - s_i \right\|^2$ with μ_{s_i} and $\sigma_{s_i}^2$ updated.

For $c = 1, \ldots, C$,

$\boldsymbol{\mu}_{\beta_c} \leftarrow$

$\left(\mu_\tau \sum_i r_{ic} ((\sigma_{a_i}^2) + \mu_{a_i}^2) \boldsymbol{\phi}^T[\gamma_i(\mathbf{t})] \boldsymbol{\phi}[\gamma_i(\mathbf{t})] + \mathbf{I} \right)^{-1} \mu_\tau \sum_{i=1}^{N} r_{ic} \mu_{a_i} \boldsymbol{\phi}^T[\gamma_i(\mathbf{t})] (\boldsymbol{y}_i - \mathbf{1}_K \otimes \mu_{s_i})$,

$\Sigma_{\boldsymbol{\beta}_c} \leftarrow \left(\mu_\tau \sum_i r_{ic} (\sigma_{a_i}^2 + \mu_{a_i}^2) \boldsymbol{\phi}^T[\gamma_i(\mathbf{t})] \boldsymbol{\phi}[\gamma_i(\mathbf{t})] + \mathbf{I} \right)^{-1}$,

$\mu_{p_c} = \frac{\sum_{i=1}^{N} r_{ic} + \eta}{C\eta + \sum_{c=1}^{C} \sum_{i=1}^{N} r_{ic}}$, $\quad \mu_{\ln p_c} \leftarrow \psi(\sum_{i=1}^{N} r_{ic} + \eta) - \psi(C\eta + \sum_{c=1}^{C} \sum_{i=1}^{N} r_{ic})$,

where $\psi(\cdot)$ is the digamma function.

$\mu_\tau \leftarrow \frac{\frac{KN}{2} + \kappa}{\frac{1}{2} \sum_{i=1}^{N} \sum_{c=1}^{C} r_{ic} \left(\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}} \| \boldsymbol{y}_i - a_i \boldsymbol{\phi}[\gamma_i(\mathbf{t})] \beta_c - \mathbf{s}_i \|^2 \right) + \theta}$,

$\mu_{\ln \tau} \leftarrow \psi(\frac{KN}{2} + \kappa) - \ln \left(\frac{1}{2} \sum_{i=1}^{N} \sum_{c=1}^{C} r_{ic} \left(\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{s}} \left\| \boldsymbol{y}_i - a_i \boldsymbol{\phi}[\gamma_i(\mathbf{t})] \boldsymbol{\beta}_c - \mathbf{s}_i \right\|^2 \right) + \theta \right)$.

5.2 Choosing Initial Values

Similarly to its MCMC counterpart, our AVB method produces parameter estimates that are somewhat sensitive to the choice of initial values. In Chapter 2, we recommended switching $p\%$ of observations in each cluster every k iterations, which greatly improves the mixing and facilitates finding sensible initial values for our MCMC sampling. We propose a method of choosing initial values that mirrors to the approach for the full Bayesian algorithm.

For a given iteration, the expression of the (i, c)-th element of the responsibility matrix is given in (5.6), which essentially quantifies the likelihood of the i-th observation belonging to the c-th cluster. We classify the i-th observation into the cluster with the highest value among C responsibilities $r_{ic}, c = 1, \ldots, C$. Note that each row of the responsibility matrix has one element close to 1 and all other close to 0 after just several iterations. We adjust the responsibility matrix every m (typically 3 or 4) iterations as follows: We randomly select $p\%$ of observations in a cluster and randomly classify them into another cluster by setting the corresponding cluster responsibility value to 1. The same pattern is observed for the AVB method as we observed in Chapter 3: Should the switch result in a poorer clustering, we note based on experimentation that the algorithm can adjust itself and is likely to recover individual classifications of the previous partitions that were correct. Furthermore, one advantage of variational approximation is that the lower bound is monitored through iterations. A higher value of the lower bound indicates a better approximation to the posterior in general, so we may consider picking the initial values to be the parameter estimates corresponding to the highest lower bound. For computational efficiency, we disable the optimization part of updating γ_{ij} during the process of choosing initial values.

Sensible initial values are useful even for the procedure of choosing the initial values. We recommend using some existing clustering method, such as a model-

based or K-means method, to obtain reasonable starting values of the responsibility matrix and the B-spline coefficients for each cluster.

Finally, we run the full AVB method summarized in Algorithm 1 with the initial values described above.

5.3 Choosing the Number of Clusters

Using the variational method to choose the number of clusters when clustering multivariate normal data has proven to be successful in the literature. In Chapter 10, Bishop [2006] suggested starting with an excessive number of clusters and waiting for the number of nonempty clusters to drop throughout the iterations. Corduneanu and Bishop [2001] proposed an expectation-maximization type algorithm. Their algorithm also starts with an excessive number of clusters. We take a similar approach. In the E-step, the variational solution of the approximated posterior distributions is obtained via Algorithm 1 for all parameters except cluster probability vector $\boldsymbol{p}$ with only 1 iteration; in the M-step, the lower bound is maximized with respect to the cluster probability vector $(p_1, p_2, \ldots, p_{C^*})$, where C^* is the total number of non-empty clusters in the current iteration. By taking the gradient of the lower bound (D.1) with respect to $\boldsymbol{p}$ and setting it to 0, the solution of p_c is found to be

$$p_c = \frac{1}{N} \sum_{i=1}^{N} r_{ic},$$

where the expression of r_{ic} is given by (5.6). Once a cluster becomes empty, the next iteration excludes this cluster and the total number of clusters drops by 1. We adopt the idea of this algorithm and allow $p\%$ of observations to exchange memberships across different clusters every m iterations as described in the last section. We recommend running several extra steps without switching membership in the end to counteract any potential membership misspecification due to the membership switch.

For computational efficiency, we do not apply the optimization component of our proposed algorithm for choosing C.

5.4 Simulation Study

In this section, we apply our proposed adapted variational method on a set of simulated data to demonstrate its functionality for simultaneous clustering and registration. We specify 4 clusters, and thus generate 4 mean functions, which are

$$f_1(t) = 1.5\cos(5t) + 0.5\sin(5t), f_2(t) = \cos(10t), f_3(t) = \sin(5t), \text{ and } f_4(t) = t^2.$$

We assign 5, 6, 5, and 7 observations (23 total observations) to each cluster, respectively, and generate 23 warping functions with 20 steps distributed as $Dir(\boldsymbol{\alpha} = (1, \ldots, 1))$. We assume that 30 equally spaced measurements on $\mathcal{T}$ are taken from each curve. The simulated warping functions are applied to the clock time, and the underlying process times are obtained for each observation. For the i-th observation, we evaluate the B-spline function at its corresponding process times. A set of stretching/shrinking factors of size 23 is generated as independent $N(1, 0.05^2)$ and multiplied to the mean values of the corresponding observations. Finally, we add white noise with $\sigma^2 = 0.04$ to each observation at each time point. A vertical shift generated from $Unif\,(-1, 1)$ is added to each observation. A plot of the simulated dataset is shown in the top left panel of Figure 5.2.

To analyze the simulated data, we model the underlying mean curves via a B-spline representation with 6 basis functions of order 4 with equally spaced knots. The means $\boldsymbol{\beta}_0$ of the B-spline are taken to be $\mathbf{0}$, and we assume those coefficients are independent with variance $\mathbf{1}$, i.e., $\boldsymbol{\beta}|\boldsymbol{\beta}_0, \Gamma \sim N(\mathbf{0}, \mathbf{I})$. For the hyperparameters, we choose $\kappa = 25, \theta = 1$ for the precision, $\phi = 1$ for the vertical shifts, and $\alpha = 1$ for the warping functions. Following the method of choosing the number of clusters in section 5.3, we start with $C = 8$ and the algorithm correctly specifies $C = 4$

within 100 iterations and remains at 4 clusters for the rest of the iterations. The final variational results are obtained via applying Algorithm 1 with 10 iteration with the initial values specified by the estimates from the procedure of choosing C. The lower bounds across iterations are shown in Figure 5.1. Note that the 0-th value of the lower bound is the last lower bound obtained in the procedure of choosing C.

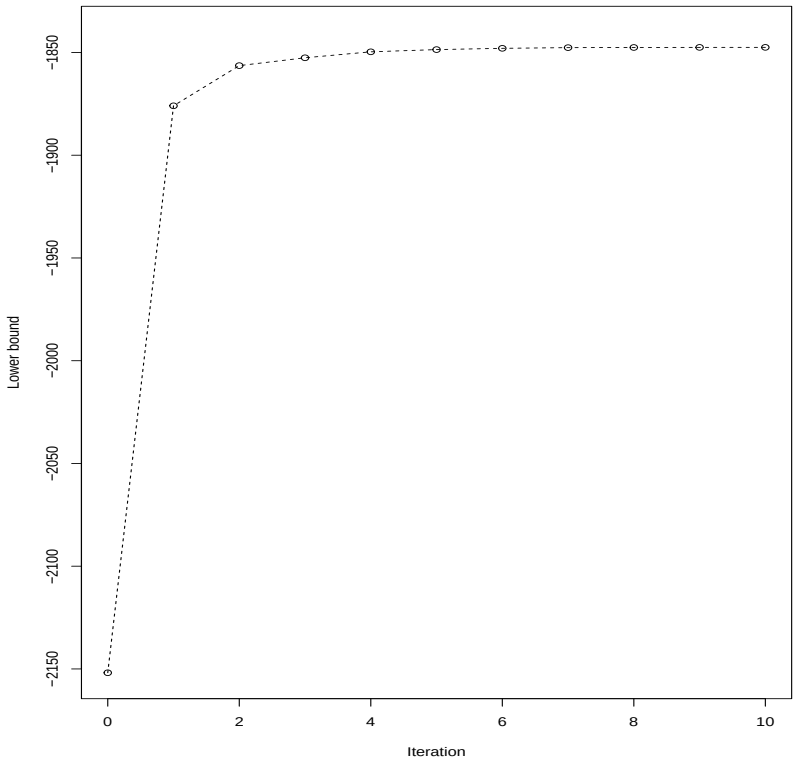

Figure 5.1 Lower bounds of the simulation study

For this set of simulated data, our proposed method specifies all the cluster memberships correctly. The curves after we remove the phase variations and vertical shifts are shown in the upper right panel of Figure 5.2. The lower left panel of Figure 5.2 displays the true signal curves (gray) and our posterior estimated signal curves (black). The estimated warping functions are shown in the lower right panel.

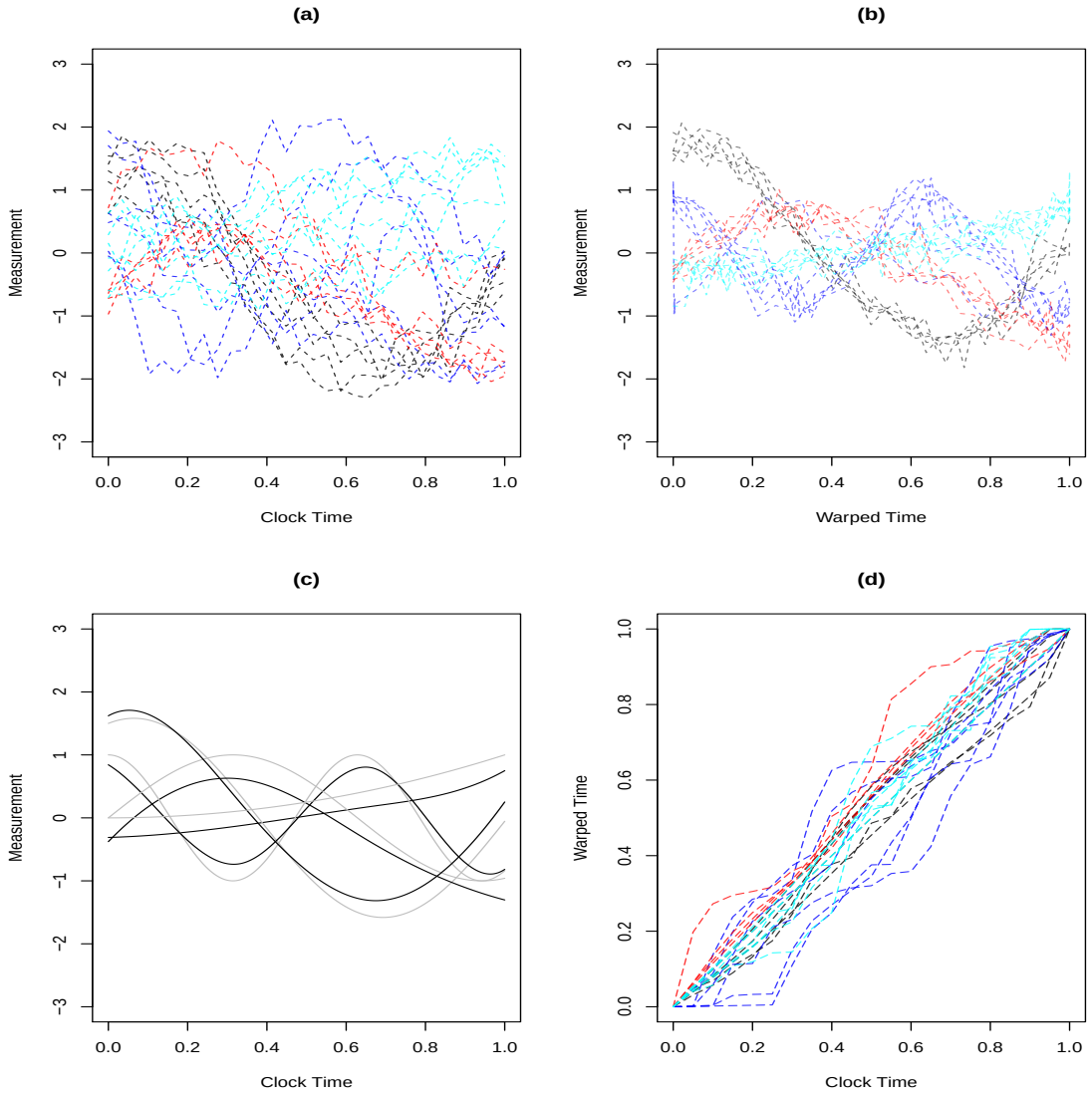

Figure 5.2 (a) A set of 23 simulated observations with 4 clusters. (b) Simulated data with phase variation removed, with superimposed posterior estimated mean curves (solid black). (c) True mean curves (gray) and estimated mean curves (black). (d) Estimated warping functions for all 4 clusters.

5.5 Real Data Analysis

Berkeley Growth Data

We apply the AVB method to the Berkeley growth acceleration data. To compare the results with what we obtained in Chapter 2, we choose the same parameters and

hyperparameters as in the Chapter 2. Namely, we model the signal functions with 8 B-spline basis functions of order 6 defined on a equally spaced knot sequence. We choose $\alpha = 4$, $\kappa = 50$, $\theta = 10$, $\sigma_a^2 = 0.05^2$, and $\phi = 1.2$. In order to compare the result with the underlying clustering structure in terms of gender, we assume $C = 2$ clusters throughout all the iterations.

To pick a set of initial values, we run the algorithm with 50 iterations. We switch 20% of the observations in each cluster every 3 iterations and let the algorithm run as a usual variational method for the last 10 iterations. After obtaining the initial values, we run our proposed AVB algorithm for 10 iterations. The lower bounds are shown in Figure 5.3. Note that iteration 0 corresponds to the lower bound of the last iteration in the procedure of obtaining the initial values.

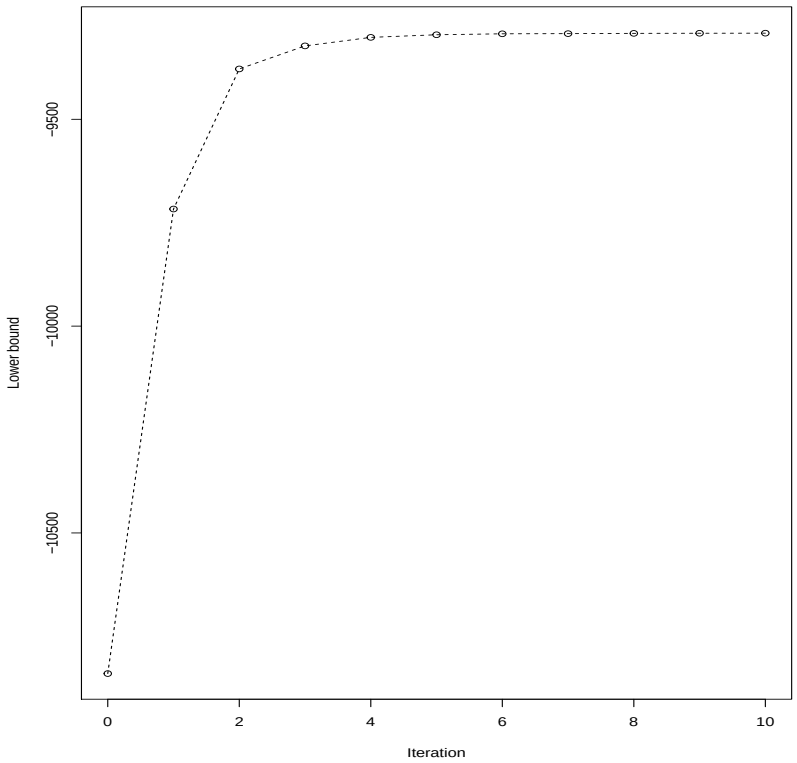

Figure 5.3 Lower bound of AVB for Berkeley acceleration data

The clustering results are shown in Table 5.1; nine females and two males are misclassified to the opposite gender, yielding overall cRate 88.17%. The clustering results are plotted in the second row of Figure 5.4; the solid blue curves represent those boys who are misclassified as girls and the pink solid curves represent the misclassified girls in panels (c) and (d), respectively. The estimated warping functions are shown in the last row of Figure 5.4.

Table 5.1 Clustering results for Berkeley acceleration curves by AVB.

	True cluster	
	Male	Female
Cluster I	37	2
Cluster II	9	45

Response of Human Fibroblasts to Serum

We apply the AVB method to the human fibroblasts to serum data described in Section 5.5. We model the signal functions with 6 B-spline basis functions of order 4 defined on a equally spaced knot sequence. We choose $\alpha = 5$, $\kappa = 1$, $\theta = 1$, $\sigma_a^2 = 0.1^2$, and $\phi = 3$. Following the method of choosing the number of clusters in section 5.3, we start with $C = 10$ and the algorithm ends up with $C = 5$ within 150 iterations. To pick a set of initial values, we run the algorithm with 50 iterations. We switch 20% of the observations in each cluster every 3 iterations and let the algorithm run as a usual variational method for the last 10 iterations. After obtaining the initial values, we run our proposed AVB algorithm for 10 iterations. The raw data with estimated cluster memberships are shown in the right panel of Figure 5.5.

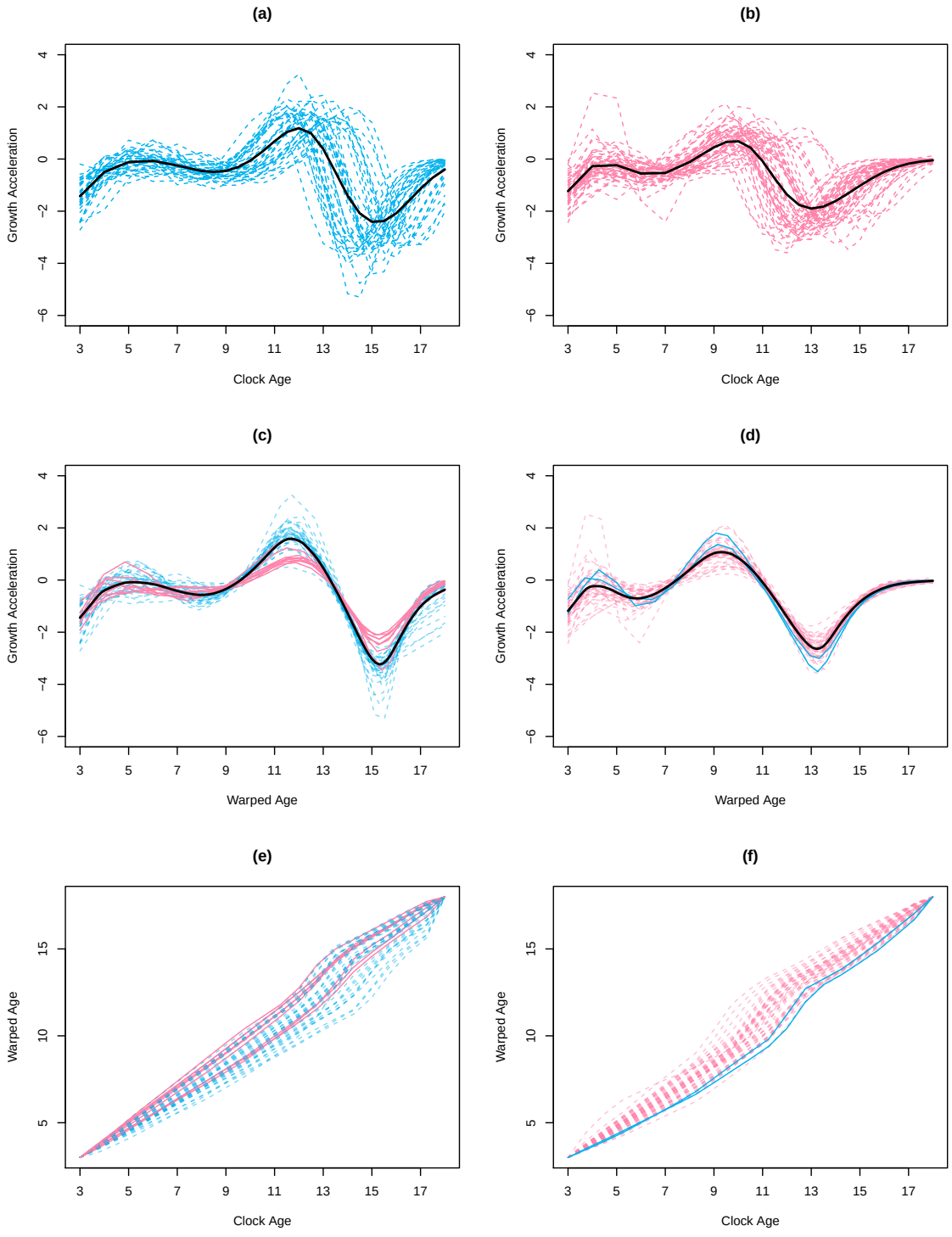

Figure 5.4 (a)-(b) Unregistered growth acceleration data for 39 boys (blue dashed) and 54 girls (pink dashed) with cross-sectional mean superimposed. (c) Registered cluster 1 with 36 boys (blue dashed) and 9 girls (pink solid). (d) Registered cluster 2 with 45 girls (pink dashed) and 3 boys (blue dashed). (e)-(f) Estimated warping functions for cluster 1 and cluster 2, respectively.

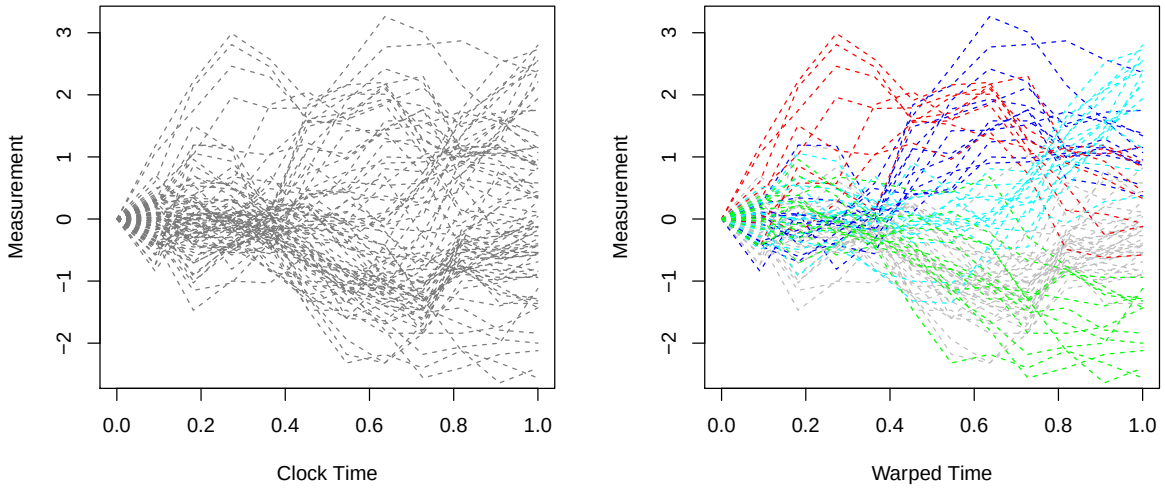

Figure 5.5 Left: Raw HFS data. Right: Raw HFS data with estimated membership.

The lower bounds are shown in Figure 5.6. Note that iteration 0 corresponds to the lower bound of the last iteration in the procedure of obtaining the initial values.

The clustering result is given in Figure 5.7. The registered curves shows a clearer pattern after the vertical shifts are removed, as shown in Figure 5.7 (a). All curves in the same cluster roughly follow the same pattern as shown in panels (b)-(f) of Figure 5.7. The estimated warping functions are shown in Figure 5.8.

5.6 Discussion

We have developed a variational approximation of the full Bayesian approach proposed in Chapter 3. Demonstrated by a simulation study and real data analysis, our proposed method produces promising results in terms of cluster accuracy and registrations in a relatively short amount of time. This adapted variational approach could serve as an independent clustering tool for functional observations with time warping or, at the very minimum, could help us choose the number of clusters and initial values for our full Bayesian inference.

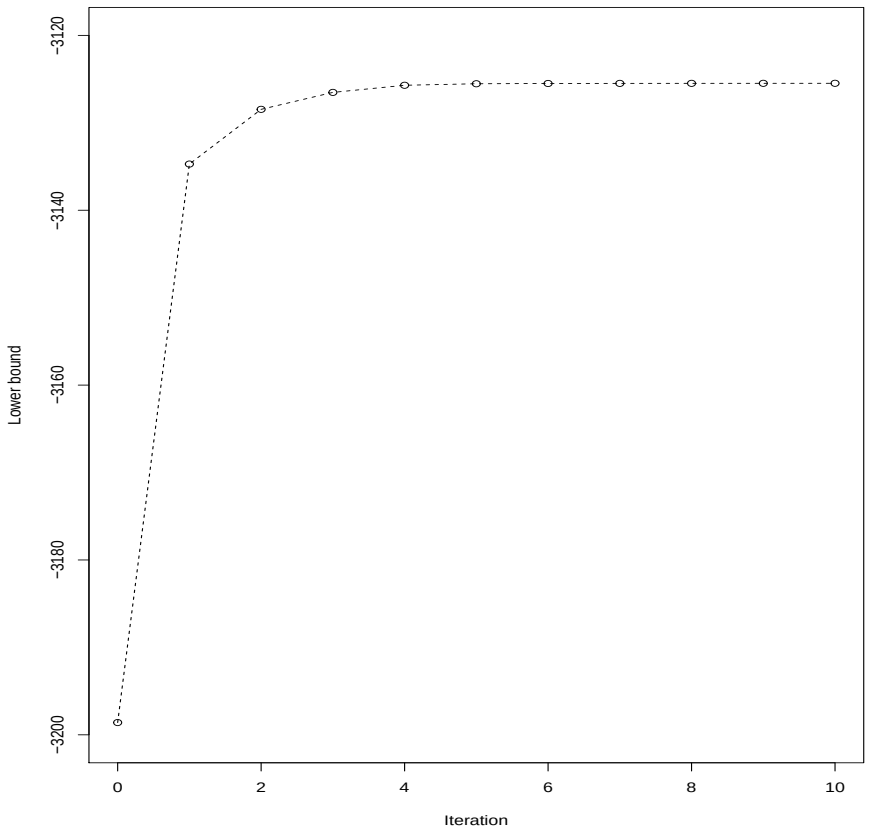

Figure 5.6 Lower bound of AVB for HFS data

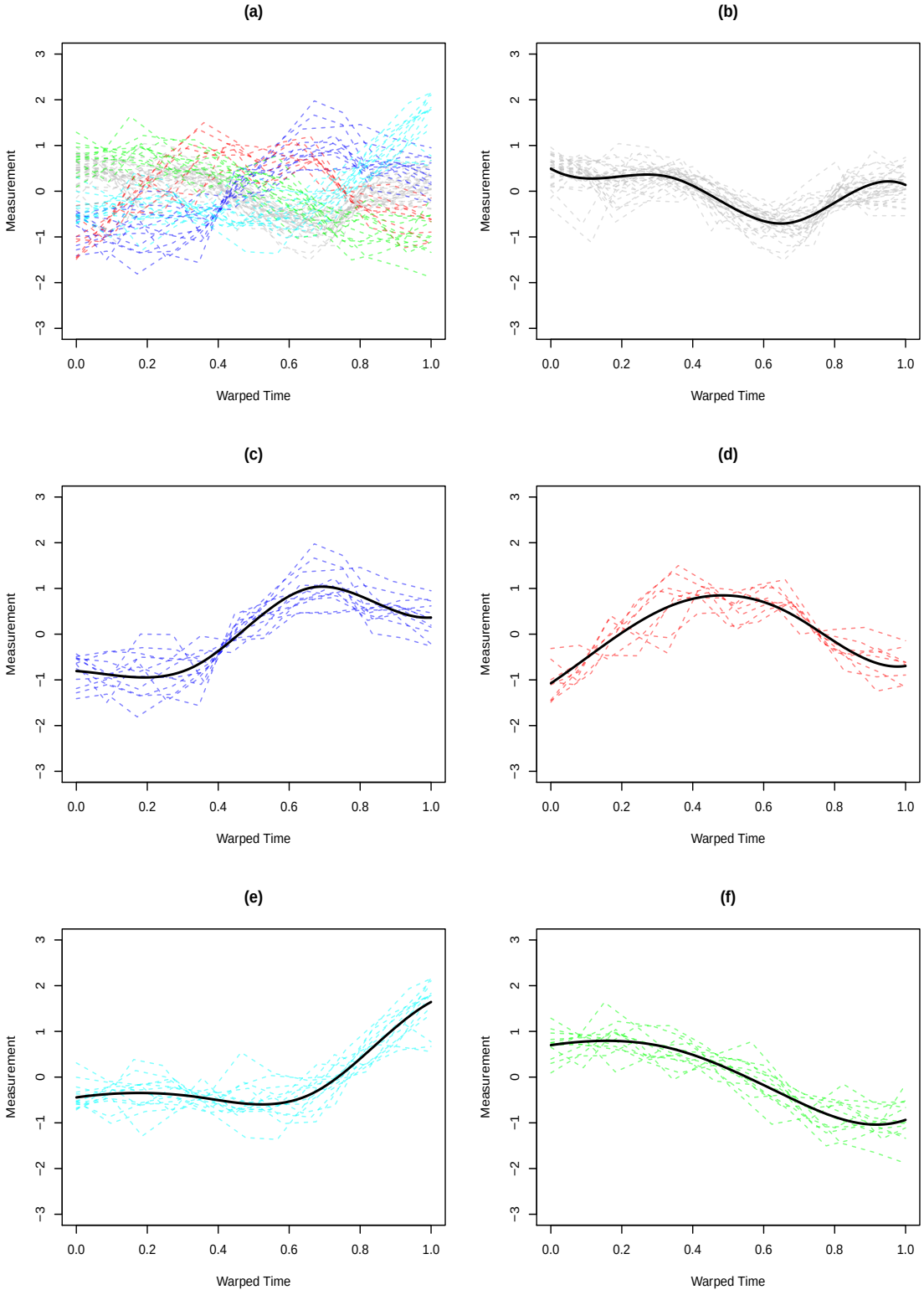

Figure 5.7 (a) Registered curves with vertical shifts removed. (b)-(f) Registered five clusters with their estimated mean curves superimposed.

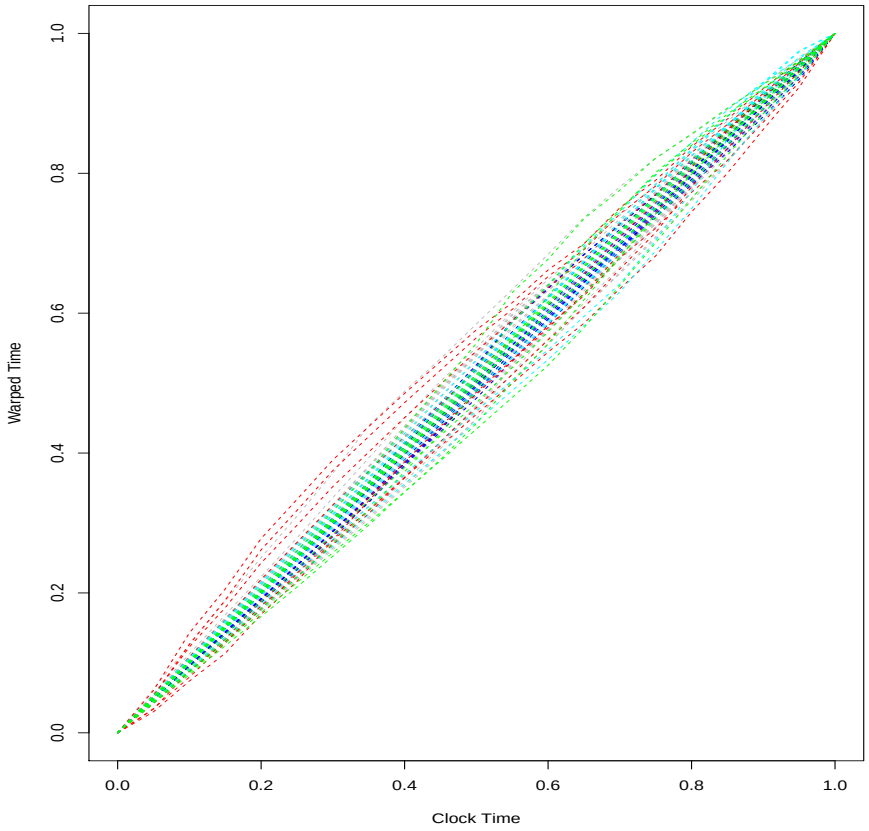

Figure 5.8 Estimated warping functions for HFS data

Bibliography

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

Orly Alter, Patrick O Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics: Theory and Methods*, 3(1):1–27, 1974.

Ming-Yen Cheng, Jianqing Fan, James S Marron, et al. On automatic boundary corrections. *The Annals of Statistics*, 25(4):1691–1708, 1997.

Wen Cheng, Ian L Dryden, and Xianzheng Huang. Bayesian registration of functions and curves. *In press: Bayesian Analysis doi: 10.1214/15-BA957*, 2015.

Adrian Corduneanu and Christopher M Bishop. Variational Bayesian model selection for mixture distributions. In *Artificial intelligence and Statistics*, volume 2001, pages 27–34. Morgan Kaufmann Waltham, MA, 2001.

Carl De Boor. *A Practical Guide to Splines*. New York: Springer-Verlag, 2001.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

David B Dunson and Amy H Herring. Semiparametric Bayesian latent trajectory models. *Proceedings ISDS Discussion Paper*, 16, 2006.

Cecilia Earls and Giles Hooker. Bayesian covariance estimation and inference in latent gaussian process models. *Statistical Methodology*, 18:79–100, 2014.

Cecilia Earls and Giles Hooker. Adapted variational bayes for functional data registration, smoothing, and prediction. *arXiv preprint arXiv:1502.00552*, 2015.

B Everitt, S Landau, M Leese, and D Stahl. *Cluster Analysis*. Chichester, UK: Wiley, 2011.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458): 611–631, 2002.

Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Philip Heidelberger and Peter D Welch. A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, 24(4):233–245, 1981.

Torsten Hothorn and Brian S Everitt. *A handbook of statistical analyses using R*. CRC press, 2014.

Vishwanath R Iyer, Michael B Eisen, Douglas T Ross, Greg Schuler, Troy Moore, Jeffrey CF Lee, Jeffrey M Trent, Louis M Staudt, James Hudson, Mark S Boguski, et al. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283(5398):83–87, 1999.

Gareth M James and Catherine A Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462):397–408, 2003.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, pages 79–86, 1951.

Xueli Liu and Mark CK Yang. Simultaneous curve registration and clustering for functional data. *Computational Statistics and Data Analysis*, 53(4):1361–1376, 2009.

Yihui Luan and Hongzhe Li. Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, 19(4):474–482, 2003.

John T Ormerod and MP Wand. Explaining variational approximations. *The American Statistician*, 64(2):140–153, 2010.

James O Ramsay and Xiaochun Li. Curve registration. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 351–363, 1998.

James O Ramsay and B W Silverman. *Functional Data Analysis*. Springer: New York, 2005.

JO Ramsay, Hadley Wickham, and Maintainer JO Ramsay. Package 'fda'. 2013.

Laura M Sangalli, Piercesare Secchi, Simone Vantini, and Valeria Vitelli. K-mean alignment for curve clustering. *Computational Statistics and Data Analysis*, 54 (5):1219–1233, 2010.

Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

Paul T Spellman, Gavin Sherlock, Michael Q Zhang, Vishwanath R Iyer, Kirk Anders, Michael B Eisen, Patrick O Brown, David Botstein, and Bruce Futcher. Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9 (12):3273–3297, 1998.

Rong Tang and Hans-Georg Müller. Time-synchronized clustering of gene expression trajectories. *Biostatistics*, 10(1):32–45, 2009.

Read D Tuddenham and Margaret M Snyder. Physical growth of California boys and girls from birth to eighteen years. *University of California Publications in Child Development*, 1(2):183–364, 1953.

Dimitris G Tzikas, Aristidis C Likas, and Nickolaos P Galatsanos. The variational approximation for bayesian inference. *Signal Processing Magazine, IEEE*, 25(6): 131–146, 2008.

Yafeng Zhang and Donatello Telesca. Joint clustering and registration of functional data. *arXiv:1403.7134*, 2014.

Appendix A

Choosing the Number of Clusters C

The following algorithm determines whether we accept the decrease of the number of clusters by 1. Let the number of non-empty clusters at iteration t be denoted by $C^{[t]}$.

Algorithm 2: Accept or reject the change of the number of clusters.

At iteration t, $C^{[t-1]} = C^*$ and $C^{[t]} = C^* - 1$;

if $avg\ \log \mathcal{L}_{C^*} > avg\ \log \mathcal{L}_{C^*+1}$ **then**
  accept $C^{[t]} = C^* - 1$;

else
  reset $C^{[t]} = C^* + 1$;
  reset the cluster membership to the first iteration in the most recent block of iterations where the number of clusters is $C^* + 1$;

The following algorithm determines whether we increase the number of clusters by 1 if the Markov chain stays at the same number of non-zero clusters for a long period.

Algorithm 3: Accept or reject the change the number of clusters when $M_{C^*} > M_0$.

At iteration t, $C^{[t]} = C^*$ and $M_{C^*} > M_0$;

if $avg\ \log \mathcal{L}_{C^*} > avg\ \log \mathcal{L}_{C^*+1}$ **then**
  keep $C^{[t]} = C^*$;

else
  reset $C^{[t]} = C^* + 1$;
  reset the cluster membership to the first iteration in the most recent block of iterations where the number of clusters is $C^* + 1$;

Appendix B

Bayesian Registration for One Cluster

B.1 Likelihood and Bayesian Analysis

The full details of our Bayesian registration and clustering algorithm is given in Section 3.2. Assuming one cluster ($C = 1$), we provide a summary of the algorithm that aligns curves in that cluster.

To estimate the warping function h_i for the i-th observation, a discrete approximation generated by a Dirichlet distribution is utilized [Cheng et al., 2015]. Without loss of generality, let us assume that the time domain $\mathcal{T} = [0, 1]$. Any general time domain $[T_1, T_2]$ may be converted into $[0, 1]$ by the transformation $g(t) = (t - T_1)/(T_2 - T_1)$. Let $\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iM} \sim Dir(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is a M-vector of positive parameters.

For the Dirichlet distribution, we have $\sum_j \gamma_{ij} = 1$, which suggests that the linear interpolation of the cumulative sum over γ_{ij} can serve as a discrete approximation of the continuous warping h_i. The parameter M controls the smoothness of the approximation. A large M results in a smoother approximation, but more computational burden.

The hyperparameter $\boldsymbol{\alpha}$ can be chosen to affect the "concentration" of the warping functions relative to the 45° reference line, which corresponds to no warping. Small values in $\boldsymbol{\alpha}$ allow more variability in each step of the approximation, and vice versa.

We assume that spline coefficient $\boldsymbol{\beta} \sim \text{MVN}(\boldsymbol{\beta}_0, \Gamma)$. It will be seen later that the full conditional distribution of $\boldsymbol{\beta}$ is still multivariate normal. We model the precision parameter $\tau = 1/\sigma^2$ with a (conjugate) gamma prior, i.e., $\tau \sim Gamma(\kappa, \theta)$.

Our prior model assumes the vertical shift S_i for the i-th observation is $Unif\ (-\phi, \phi)$ for some positive ϕ. On the stretching/shrinking factors a_i, we place independent $N(1, \sigma_a^2)$ priors, $i = 1, 2, \ldots, N$.

Under the preceding model assumptions in conjunction with the above prior distributions on the parameters, the joint distribution of the data and parameters is

$$
\begin{aligned}
& \mathcal{L}(\boldsymbol{\beta}, \ldots, \gamma_N, \tau, s_1, \ldots, s_N, a_1, \ldots, a_N, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) \\
= & \prod_{i=1}^{N} \mathcal{P}(\boldsymbol{y}_i|\boldsymbol{\beta}, \tau, s_1, \ldots, s_N, a_1, \ldots, a_N)\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{\beta}_0, \Gamma) \prod_{i=1}^{N} \mathcal{P}(\gamma_i|\boldsymbol{\alpha})\mathcal{P}(\tau|\kappa, \theta) \prod_{i=1}^{N} \mathcal{P}(s_i|\phi) \\
& \prod_{i=1}^{N} \mathcal{P}(a_i|\sigma_a^2) \\
\propto & \prod_{i=1}^{N} \tau^{K/2} \exp\left\{-\frac{1}{2}\tau \left\|\boldsymbol{y}_i - a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta} - \mathbf{s}_i\right\|^2\right\} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\Gamma^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\} \\
& \prod_{i=1}^{N} \prod_{m=1}^{M} \gamma_{im}^{\alpha_m - 1} \tau^{\kappa+1} \exp\{-\tau\theta\} \prod_{i=1}^{N} \mathbf{1}_{\{-\phi < s_i < \phi\}} \prod_{c=1}^{N} \exp\left\{-\frac{1}{2}(a_i - 1)^2\right\} \\
\propto & \tau^{Kn/2} \exp\left\{-\frac{1}{2}\tau \sum_{i=1}^{N} \left\|\boldsymbol{y}_i - a_i\boldsymbol{\phi}[\gamma_i(\mathbf{t})]\boldsymbol{\beta} - \mathbf{s}_i\right\|^2\right\} \prod_{c=1}^{C} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\Gamma^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\} \\
& \prod_{i=1}^{N} \prod_{m=1}^{M} \gamma_{im}^{\alpha_m - 1} \exp\{-\tau\theta\} \prod_{i=1}^{N} \mathbf{1}_{\{-\phi < s_i < \phi\}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}(a_i - 1)^2\right\}.
\end{aligned}
$$

This joint distribution will be used in Section 4 to obtain the relevant full conditional distributions for the MCMC algorithm.

B.2 Sampling Algorithm

Due to the complexity of the proposed model, an analytical posterior derivation is intractable, so our inference is based on MCMC sampling of the posterior distribution. At iteration t, the MCMC algorithm is as follows:

- **Metropolis-Hastings Algorithm for Sampling Warping γ_i**

  We update $\gamma_{i1}, \ldots, \gamma_{iM-1}$. The two endpoints satisfy the conditions $\gamma_{i0} = 0$, and $\gamma_{iM} = 1 - \sum_{j=1}^{M-1} \gamma_{ij}$, because of the constraints of the warping function, and hence are not involved in the updating procedure. After updating the

$\boldsymbol{z}_i$, we propose a value of γ_{ij}^* from a truncated normal with mean $\gamma_{ij}^{[t-1]}$ and variance σ_γ^2 on $[0, \gamma_{iM} + \gamma_{ij}]$ to guarantee a positive γ_{ij}^* and γ_{iM}^*. We accept the proposed value with probability

$$\lambda = \min\left\{1, \frac{\exp\left\{-\frac{1}{2}\tau^{[t-1]}\left\|\boldsymbol{y}_i - a_i^{[t-1]}\boldsymbol{\phi}[\gamma_i^{*(j)}(\mathbf{t})]\prod_{c=1}^{C}\boldsymbol{\beta}_c^{[t-1]z_{ic}^{[t]}} - \mathbf{s}_i^{[t-1]}\right\|^2\right\}}{\exp\left\{-\frac{1}{2}\tau^{[t-1]}\left\|\boldsymbol{y}_i - a_i^{[t-1]}\boldsymbol{\phi}[\gamma_i^{(j-1)}(\mathbf{t})]\prod_{c=1}^{C}\boldsymbol{\beta}_c^{[t-1]z_{ic}^{[t]}} - \mathbf{s}_i^{[t-1]}\right\|^2\right\}} \times \frac{(\gamma_{ij}^*)^{\alpha_j-1}(\gamma_{iM}^*)^{\alpha_M-1}\left[\Phi\left(\frac{r_{ij}^{[t]}-\gamma_{ij}^*}{\sigma_\gamma}\right) - \Phi\left(\frac{-\gamma_{ij}^*}{\sigma_\gamma}\right)\right]}{(\gamma_{ij}^{[t-1]})^{\alpha_j-1}(\gamma_{iM}^{[t-1]})^{\alpha_M-1}\left[\Phi\left(\frac{r_{ij}^{[t]}-\gamma_{ij}^{[t-1]}}{\sigma_\gamma}\right) - \Phi\left(\frac{-\gamma_{ij}^{[t-1]}}{\sigma_\gamma}\right)\right]}\right\}$$

where $\gamma_i^{(j)}$ is the warping function with the jump updated through the j-th element, and Φ is the standard normal CDF.

- **Gibbs Sampling for Spline Coefficients $\boldsymbol{\beta}$**

  After updating the γ_i's , let n denote the number of observations in the sample, the full conditional of $\boldsymbol{\beta}$ is given by

$$\begin{aligned}
&\mathcal{P}(\boldsymbol{\beta}|\text{rest}) \\
&\propto \exp\left\{-\frac{1}{2}\tau^{[t-1]}\sum_{l=1}^{n}\left\|\boldsymbol{y}_l - a_i^{[t-1]}\boldsymbol{\phi}[\gamma_l^{[t]}(\mathbf{t})]\boldsymbol{\beta} - \mathbf{s}_l^{[t-1]}\right\|^2\right\}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\Gamma^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\boldsymbol{\beta}'\underbrace{\left(\tau^{[t-1]}\sum_{l=1}^{n}\left[\left(a_l^{[t-1]}\right)^2\boldsymbol{\phi}'[\gamma_l^{[t]}(\mathbf{t})]\boldsymbol{\phi}[\gamma_l^{[t]}(\mathbf{t})] + \Gamma^{-1}\right]\right)}_{\text{call it } \mathbf{A}}\boldsymbol{\beta} - \right. \\
&\qquad\qquad \left. \boldsymbol{\beta}'\underbrace{\left(\tau^{[t-1]}\sum_{l=1}^{n}a_l^{[t-1]}\boldsymbol{\phi}'[\gamma_l^{[t]}(\mathbf{t})](\boldsymbol{y}_l - \mathbf{s}_l^{[t-1]})' + \Gamma^{-1}\boldsymbol{\beta}_0\right)}_{\text{call it } \mathbf{C}}\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{A}^{-1}\mathbf{C})'\mathbf{A}(\boldsymbol{\beta}-\mathbf{A}^{-1}\mathbf{C})\right\}.
\end{aligned}$$

  Therefore,

$$\boldsymbol{\beta}|\text{rest} \sim \text{MVN}(\mathbf{A}^{-1}\mathbf{C}, \mathbf{A}^{-1}).$$

- **Gibbs Sampling for Precision τ**

  After updating the γ_i's and $\boldsymbol{\beta}$'s, the full conditional distribution of τ is given

by

$$\begin{aligned}
&\mathcal{P}(\tau|\text{rest}) \\
\propto\ & \tau^{Kn/2}\exp\left\{-\frac{1}{2}\tau\sum_{i=1}^{n}\left\|\boldsymbol{y}_i - a_i^{[t-1]}\boldsymbol{\phi}(\gamma_i^{[t]}(\mathbf{t}))\boldsymbol{\beta}^{[t]} - \mathbf{s}_i^{[t-1]}\right\|^2\right\}\tau^{\kappa-1}\exp\{-\tau\theta\} \\
\propto\ & \tau^{Kn/2+\kappa-1}\exp\left\{-\tau\left(\frac{1}{2}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i - a_i^{[t-1]}\boldsymbol{\phi}(\gamma_i^{[t]}(\mathbf{t}))\boldsymbol{\beta}^{[t]} - \mathbf{s}_i^{[t-1]}\right\|^2 + \theta\right)\right\}.
\end{aligned}$$

It follows that

$$\tau|\text{rest} \sim Gamma\left(Kn/2+\kappa, \frac{1}{2}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i - a_i^{[t-1]}\boldsymbol{\phi}(\gamma_i^{[t]}(\mathbf{t}))\boldsymbol{\beta}^{[t]} - \mathbf{s}_i^{[t-1]}\right\|^2 + \theta\right).$$

- **Gibbs Sampling for Vertical Shift** S_i

  After updating the γ_i's, $\boldsymbol{\beta}$, and τ, the full conditional distribution of S_i is given by

$$\begin{aligned}
&\mathcal{P}(s_i|\text{rest}) \\
\propto\ & \exp\left\{-\frac{1}{2}\tau^{[t]}\left\|\boldsymbol{y}_i - a_i^{[t-1]}\boldsymbol{\phi}(\gamma_i^{[t]}(\mathbf{t}))\boldsymbol{\beta}^{[t]} - \mathbf{s}_i\right\|^2\right\}\mathbf{1}_{\{-\phi<s_i<\phi\}}.
\end{aligned}$$

To simplify the notation, let us define d_l as the l-th element of the vector $\boldsymbol{y}_i - a_i^{[t-1]}\boldsymbol{\phi}(\gamma_i^{[t]}(\mathbf{t}))\boldsymbol{\beta}_c^{[t]}$. The posterior then is

$$\begin{aligned}
\mathcal{P}(s_i|\text{rest}) \ &\propto\ \exp\left\{-\frac{1}{2}\tau^{[t]}\sum_{l=1}^{K}(s_i - d_l)^2\right\}\mathbf{1}_{\{-\phi<s_i<\phi\}} \\
&\propto\ \exp\left\{-\frac{1}{2}\tau^{[t]}\sum_{l=1}^{K}(s_i^2 - d_l s_i)^2\right\}\mathbf{1}_{\{-\phi<s_i<\phi\}} \\
&\propto\ \exp\left\{-\frac{1}{2}\tau^{[t]}K(s_i - \sum_{l=1}^{K}d_l/K)^2\right\}\mathbf{1}_{\{-\phi<s_i<\phi\}}
\end{aligned}$$

The normal kernel indicates that the posterior distribution of the vertical shift S_i is a truncated normal with mean $\sum_{l=1}^{K}d_l/K$, and variance $1/(\tau^{[t]}K)$, i.e.,

$$S_i|\text{rest} \sim N\left(\frac{\sum_{l=1}^{K}d_l}{K}, \frac{1}{\tau^{[t]}K}\right)\mathbf{1}_{\{-\phi<s_i<\phi\}}.$$

- **Gibbs Sampling for Stretching/Shrinking Factor** a_i

  After updating the γ_i's, $\boldsymbol{\beta}$, τ, and s_i's, the full conditional distribution of a_i is given by

$$\begin{aligned} & \mathcal{P}(a_i|\text{ rest}) \\ \propto\ & \exp\left\{-\frac{1}{2}\tau^{[t]}\left\|\boldsymbol{y}_i - a_i\boldsymbol{\phi}(\gamma_i^{[t]}(\mathbf{t}))\boldsymbol{\beta}^{[t]} - \mathbf{s}_i^{[t]}\right\|^2\right\}\exp\left\{-\frac{1}{2\sigma_a^2}(a_i-1)^2\right\}. \end{aligned}$$

  For economy of notation, let us denote the l-th element of $\boldsymbol{\phi}(\gamma_i^{[t]}(\mathbf{t}))\boldsymbol{\beta}^{[t]}$ and $\boldsymbol{y}_i$ by $\mu_{il}^{[t]}$ and y_{il}, respectively. The posterior becomes

$$\begin{aligned} & \mathcal{P}(a_i|\text{ rest}) \\ \propto\ & \exp\left\{-\frac{1}{2}\tau^{[t]}\left[\sum_{l=1}^{K} a_i^2(\mu_{il}^{[t]})^2 - \sum_{l=1}^{K} 2a_i\mu_{il}^{[t]}(y_{il}-s_i^{[t]})\right]\right\}\exp\left\{-\frac{1}{2\sigma_a^2}a_i^2 + \frac{1}{\sigma_a^2}a_i\right\} \\ \propto\ & \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma_a^2} + \tau^{[t]}\sum_{l=1}^{K}(\mu_{il}^{[t]})^2\right]a_i^2 + \left[\tau^{[t]}\sum_{l=1}^{K}\mu_{il}^{[t]}(y_{il}-s_i^{[t]}) + \frac{1}{\sigma_a^2}\right]a_i\right\}. \end{aligned}$$

  By completing the square, we have

$$a_i|\text{ rest} \sim N\left(\frac{\tau^{[t]}\sum_{l=1}^{K}\mu_{il}^{[t]}(y_{il}-s_i^{[t]}) + 1/\sigma_a^2}{1/\sigma_a^2 + \tau^{[t]}\sum_{l=1}^{K}(\mu_{il}^{[t]})^2}, \frac{1}{1/\sigma_a^2 + \tau^{[t]}\sum_{l=1}^{K}(\mu_{il}^{[t]})^2}\right).$$

Appendix C

Derivation of the Gradient for Optimizing $\boldsymbol{\gamma}_i$

From the numerical experiments of the full Bayesian algorithm, the choice of $M = 20$ usually provides a good approximation of warping functions for most cases. Although the Nelder-Mead method offers numerical differentiation in the `constrOptim` function in `R`, it works poorly if the dimension of the problem is high. In this section, we derive the derivative of (5.7) with respect to γ_{ik} for the i-th observation.

First notice the line between the $(l-1)$-st jump and the l-th jump is given by

$$\begin{aligned} \gamma_i(t) &= \sum_{j=1}^{l-1} \gamma_{ij} + \gamma_{il} \frac{t - \frac{l-1}{M}}{\frac{1}{M}} \\ &= \sum_{j=1}^{l-1} \gamma_{ij} + \gamma_{il}(tM - l + 1), \end{aligned} \tag{C.1}$$

for $t \in [\frac{l}{M}, \frac{l+1}{M}]$, and $1 \leq l \leq M-2$. If t falls into the last segment, then

$$\begin{aligned} \gamma_i(t) &= \sum_{j=1}^{M-1} \gamma_{ij} + (1 - \sum_{j=1}^{M-1} \gamma_{ij})(Mt - M + 1) \\ &= (M - Mt) \sum_{j=1}^{M-1} \gamma_{ij} + (Mt - M + 1) \end{aligned} \tag{C.2}$$

Based on (C.1) and (C.2), there are four possible ways a point t_n could be invovled in $\partial\gamma_i(t)/\partial\gamma_{ik}$ depending on the location of t_n. A pictorial illustration of these four cases is given in Figure C.1. Let us examine these four cases one by one.

1. Suppose $\frac{m-1}{M} \leq t_n < \frac{m}{M}$, and $m < k$. Then t_n is not involved in the derivative. That is,

$$\frac{\partial}{\partial r_{ik}} \gamma_i(t_n) = 0.$$

2. Suppose $\frac{k-1}{M} \leq t_n < \frac{k}{M}$, then

$$\frac{\partial}{\partial r_{ik}}\gamma_i(t_n) = Mt_n - k + 1.$$

3. Suppose $\frac{m-1}{M} \leq t_n \leq \frac{m}{M}$, and $k < m \leq M-1$, then

$$\frac{\partial}{\partial r_{ik}}\gamma_i(t_n) = 1.$$

4. Suppose $\frac{M-1}{M} < t_n \leq 1$, i.e., t_n is in the last segment of the approximation, then

$$\frac{\partial}{\partial r_{ik}}\gamma_i(t_n) = M - Mt_n.$$

Now, let us derive the derivative of $\phi[\gamma_i(t_n)](\Sigma_{\beta_c} + \boldsymbol{\mu}^*_{\beta_c}(\boldsymbol{\mu}^*_{\beta_c})^T)\phi^T[\gamma_i(t_n)]$ with respect to γ_{ik}. Denote $\mathbf{A}^c = \Sigma_{\beta_c} + \boldsymbol{\mu}^*_{\beta_c}(\boldsymbol{\mu}^*_{\beta_c})^T$. Denote $\phi_j(\cdot)$ and $\phi'_j(\cdot)$ as the j-th basis function and its derivative, respectively. Then

$$\begin{aligned}
& \frac{\partial}{\partial r_{ik}}\left\{\phi[\gamma_i(t_n)]\mathbf{A}^c\phi^T[\gamma_i(t_n)]\right\} \\
= & \frac{\partial}{\partial r_{ik}}\left\{\sum_{j=1}^{q}\sum_{l=1}^{q}\mathbf{A}^c_{(jl)}\phi_j[\gamma_i(t_n)]\phi_l[\gamma_i(t_n)]\right\} \\
= & \sum_{j=1}^{q}\sum_{l=1}^{q}\mathbf{A}^c_{(jl)}\left\{\phi'_j[\gamma_i(t_n)]\phi_l[\gamma_i(t_n)] + \phi'_l[\gamma_i(t_n)]\phi_j[\gamma_i(t_n)]\right\}\frac{\partial}{\partial r_{ik}}\gamma_i(t_n) \\
= & 2\left[\frac{\partial}{\partial r_{ik}}\gamma_i(t_n)\right]\phi[\gamma_i(t_n)]\mathbf{A}^c\left\{\phi'[\gamma_i(t_n)]\right\}^T, \qquad \text{(C.3)}
\end{aligned}$$

where $\phi'(\cdot)$ is a vector of the derivatives of the basis function. By a similar argument, we have

$$\frac{\partial}{\partial r_{ik}}\left\{\phi[\gamma_i(t_n)]\boldsymbol{\mu}^*_{\beta_c}\right\} = \left[\frac{\partial}{\partial r_{ik}}\gamma_i(t_n)\right]\phi'[\gamma_i(t_n)]\boldsymbol{\mu}^*_{\beta_c}. \qquad \text{(C.4)}$$

Putting (C.3) and (C.4) together, the derivative is given by

$$
\begin{aligned}
&\frac{\partial}{\partial r_{ik}} \ln q^*(\gamma_i) \\
= \; & -\frac{1}{2}\sum_{c=1}^{C}\Bigg\{ r_{ic}\mu_\tau^* \Bigg[\sum_{j:\frac{k-1}{M}\le t_j<\frac{k}{M}} \Big[(Mt_j-k+1)[(\sigma_{a_i}^2)^* + (\mu_{a_i}^*)^2]\phi[\gamma_i(t_n)]\mathbf{A}^c\left\{\phi'[\gamma_i(t_n)]\right\}^T \\
& -2\mu_{a_i}^* y_{ij}(Mt_j-k+1)\phi'[\gamma_i(t_n)]\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^* + 2\mu_{a_i}^*\mu_{s_i}^*(Mt_j-k+1)\phi'[\gamma_i(t_n)]\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^* \Big] \\
& + \sum_{j:\frac{K+1}{M}\le t_j<\frac{M-1}{M}} \Big[[(\sigma_{a_i}^2)^* + (\mu_{a_i}^*)^2]\phi[\gamma_i(t_n)]\mathbf{A}^c\left\{\phi'[\gamma_i(t_n)]\right\}^T \\
& -2\mu_{a_i}^* y_{ij}\phi'[\gamma_i(t_n)]\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^* + 2\mu_{a_i}^*\mu_{s_i}^*\phi'[\gamma_i(t_n)]\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^* \Big] \\
& + \sum_{j:\frac{M-1}{M}\le t_j\le 1} \Big[(M-Mt_j)[(\sigma_{a_i}^2)^* + (\mu_{a_i}^*)^2]\phi[\gamma_i(t_n)]\mathbf{A}^c\left\{\phi'[\gamma_i(t_n)]\right\}^T \\
& -2\mu_{a_i}^* y_{ij}(M-Mt_j)\phi'[\gamma_i(t_n)]\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^* + 2\mu_{a_i}^*\mu_{s_i}^*(M-Mt_j)\phi'[\gamma_i(t_n)]\boldsymbol{\mu}_{\boldsymbol{\beta}_c}^* \Big] \Bigg] \Bigg\} \\
& +(\alpha_0-1)\frac{1}{\gamma_{ik}} - (\alpha_0-1)\frac{1}{1-\sum_{j=1}^{M-1}\gamma_{ij}}. \qquad \text{(C.5)}
\end{aligned}
$$

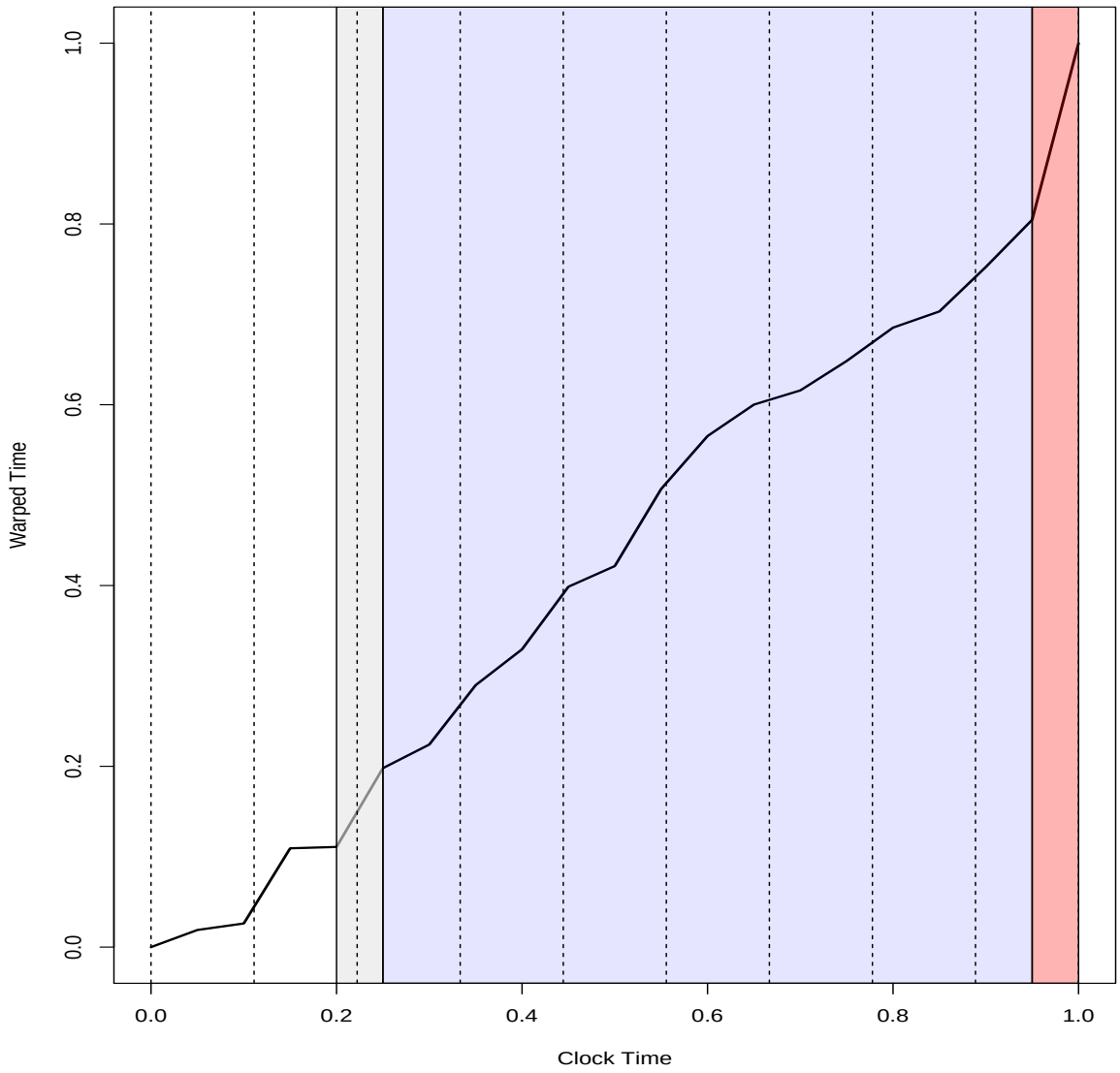

Figure C.1 An example how to calculate $\frac{\partial}{\partial \gamma_{i5}} \gamma_i(t_j)$: the four cases are illustrated in the white, gray, purple, and red segments, respectively. The dotted vertical lines represent time $\mathbf{t}$. we approximate the warping function with $M = 20$.

Appendix D

Convergence Criterion

Let us denote the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{Z}, \boldsymbol{p}, \boldsymbol{a}, \boldsymbol{s}, \tau)$. The logarithm of the lower bound of the approximation is given by

$$
\begin{aligned}
& \mathbb{E}_{q(\boldsymbol{\theta}_{-\boldsymbol{\gamma}})}[\ln \mathcal{P}(\mathbf{Y}, \boldsymbol{\gamma}, \boldsymbol{\theta}_{-\boldsymbol{\gamma}}) - \ln q(\boldsymbol{\theta}_{-\boldsymbol{\gamma}})] \\
= \; & \mathbb{E}[\ln \mathcal{P}(\mathbf{Y}|\boldsymbol{\theta})] + \mathbb{E}[\ln \mathcal{P}(\boldsymbol{\beta})] + \mathbb{E}[\ln \mathcal{P}(\mathbf{Z})] + \mathbb{E}[\ln \mathcal{P}(\boldsymbol{p})] + \mathbb{E}[\ln \mathcal{P}(\boldsymbol{a})] + \mathbb{E}[\ln \mathcal{P}(\boldsymbol{s})] \\
& + \mathbb{E}[\ln \mathcal{P}(\tau)] - \mathbb{E}[\ln q(\boldsymbol{\beta})] - \mathbb{E}[\ln q(\mathbf{Z})] - \mathbb{E}[\ln q^*(\boldsymbol{p})] - \mathbb{E}[\ln q^*(\boldsymbol{a})] - \mathbb{E}[\ln q^*(\boldsymbol{s})] \\
& + \mathbb{E}[\ln q^*(\tau)]. \qquad \text{(D.1)}
\end{aligned}
$$

The value of (D.1) should be increasing across iterations, and we monitor it until a certain convergence criterion is satisfied. Monitoring the lower bound not only provides us a stopping criterion for the iterations but also helps us check the mathematical derivations [Bishop, 2006, Ormerod and Wand, 2010]. Note that we have omitted the subscript of the expectation in (D.1) to unclutter the notation, the expectations are taken with respect to $q^*(\boldsymbol{\theta}_{-\boldsymbol{\gamma}})$ of the current iteration. Let us derive all expectations in (D.1) term by term.

$$
\begin{aligned}
& \mathbb{E}[\ln \mathcal{P}(\mathbf{Y}|\boldsymbol{\theta})] \\
= \; & \mathbb{E}\left(-\frac{KN}{2} \ln(2\pi) + \frac{KN}{2} \ln(\tau) - \frac{1}{2}\tau \sum_{i=1}^{N} \sum_{c=1}^{C} z_{ic} \left\| \boldsymbol{y}_i - a_i \boldsymbol{\phi}[\gamma_i(\mathbf{t})] \boldsymbol{\beta}_c - \mathbf{s}_i \right\|^2 \right) \\
= \; & -\frac{KN}{2} \ln(2\pi) + \frac{KN}{2} [\psi(\kappa^*) - \ln(\theta^*)] - \frac{1}{2} \mu_\tau^* \sum_{i=1}^{N} \sum_{c=1}^{C} r_{ic} \mathbb{E}_{\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{s}} \left\| \boldsymbol{y}_i - a_i \boldsymbol{\phi}[\gamma_i(\mathbf{t})] \boldsymbol{\beta}_c - s_i \right\|^2]
\end{aligned}
$$

where κ^*, θ^*, and the expectation of the norm are given by (5.10), (5.11), and (5.5), respectively.

For the terms involving $\boldsymbol{\beta}_c$, assuming $\Gamma_c = \mathbf{I}$ and $\boldsymbol{\beta}_c^0 = \mathbf{0}$, we have

$$\begin{aligned}
\mathbb{E}[\ln \mathcal{P}(\boldsymbol{\beta}_c)] &= \mathbb{E}\left(-\frac{q}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{I}| - \frac{1}{2}\boldsymbol{\beta}_c^T\boldsymbol{\beta}_c\right) \\
&= -\frac{q}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}[tr(\boldsymbol{\beta}_c\boldsymbol{\beta}_c^T)] \\
&= -\frac{q}{2}\ln(2\pi) - \frac{1}{2}tr(\mathbf{A}_c^{-1} + \mathbf{A}_c^{-1}\mathbf{c}_c\mathbf{c}_c^T\mathbf{A}_c^{-1}),
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\ln q^*(\boldsymbol{\beta}_c) \\
=\ &\mathbb{E}\left(-\frac{q}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}_c^{-1}| - \frac{1}{2}(\boldsymbol{\beta}_c - \mathbf{A}_c^{-1}\mathbf{c}_c)^T\mathbf{A}_c(\boldsymbol{\beta}_c - \mathbf{A}_c^{-1}\mathbf{c}_c)\right) \\
=\ &-\frac{q}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}_c^{-1}| - \frac{1}{2}\mathbb{E}(\boldsymbol{\beta}_c^T\mathbf{A}_k\boldsymbol{\beta}_c - 2\boldsymbol{\beta}_c^T\mathbf{c}_c + \mathbf{c}_c^T\mathbf{A}_c^{-1}\mathbf{c}_c) \\
=\ &-\frac{q}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}_c^{-1}| - \frac{1}{2}\left\{\mathbb{E}\left[(\mathbf{A}^{\frac{1}{2}}\boldsymbol{\beta}_c)^T(\mathbf{A}^{\frac{1}{2}}\boldsymbol{\beta}_c)\right] - 2\mathbf{c}_c^T\mathbf{A}_c^{-1}\mathbf{c}_c + \mathbf{c}^T\mathbf{A}_c^{-1}\mathbf{c}_c\right\} \\
=\ &-\frac{q}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}_c^{-1}| - \frac{1}{2}tr\left[\mathbb{E}\left(\mathbf{A}_c^{\frac{1}{2}}\boldsymbol{\beta}_c\boldsymbol{\beta}_c^T\mathbf{A}_c^{\frac{1}{2}}\right)\right] + \frac{1}{2}\mathbf{c}_c^T\mathbf{A}_c^{-1}\mathbf{c}_c \\
=\ &-\frac{q}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}_c^{-1}| - \frac{1}{2}tr\left(\mathbf{I} + \mathbf{A}_c^{-\frac{1}{2}}\mathbf{c}_c\mathbf{c}_c^T\mathbf{A}_c^{-\frac{1}{2}}\right) + \frac{1}{2}\mathbf{c}_c^T\mathbf{A}_c^{-1}\mathbf{c}_c \\
=\ &-\frac{q}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}_c^{-1}| - \frac{q}{2}.
\end{aligned}$$

For the membership vector $\mathbf{Z}$,

$$\begin{aligned}
\mathbb{E}\ln\mathcal{P}(\boldsymbol{z}_i) &= \mathbb{E}\left(\sum_{c=1}^{C} z_{ic}\ln p_c\right) \\
&= \sum_{c=1}^{C}\mathbb{E}(z_{ic})\mathbb{E}(\ln p_c) \\
&= \sum_{c=1}^{C}\left\{r_{ic}\left[\psi\left(\alpha_c\right) - \psi\left(\hat{\alpha}\right)\right]\right\},
\end{aligned}$$

where α_c and $\hat{\alpha}$ are given by (5.8) and (5.9), respectively.

$$\begin{aligned}
\mathbb{E}\ln q^*(\boldsymbol{z}_i) &= \mathbb{E}\left(\sum_{c=1}^{C} z_{ic}\ln r_{ic}\right) \\
&= \sum_{c=1}^{C}\mathbb{E}(z_{ic})\ln r_{ic} \\
&= \sum_{c=1}^{C}\left[r_{ic}\ln r_{ic}\right].
\end{aligned}$$

For the cluster probability $\boldsymbol{p}$. If we define $C(\boldsymbol{\eta}) = \Gamma(c\eta)/\prod_{c=1}^{C}\Gamma(\eta)$, we have

$$\begin{aligned}\mathbb{E}\ln\mathcal{P}(\boldsymbol{p}) &= \mathbb{E}\left(\ln C(\boldsymbol{\eta}) + (\eta-1)\sum_{c=1}^{C}\ln p_c\right)\\ &= \ln C(\boldsymbol{\eta}) + (\eta-1)\sum_{c=1}^{C}\left[\psi\left(\alpha_c\right) - \psi\left(\hat{\alpha}\right)\right],\end{aligned}$$

where $\boldsymbol{\eta} = (\eta,\ldots,\eta)$ is a vector of hyperparameters for $\boldsymbol{p}$. Let us define

$$C(\boldsymbol{\alpha}) = \Gamma(C\eta + \sum_{c=1}^{C}\sum_{i=1}^{N} r_{ic})/\prod_{c=1}^{C}\Gamma(\sum_{i=1}^{N} r_{ic} + \eta),$$

then

$$\begin{aligned}\mathbb{E}\ln q^*(\boldsymbol{p}) &= \mathbb{E}\left(\ln C(\boldsymbol{\alpha}) + \sum_{c=1}^{C}(\alpha_c - 1)\ln p_c\right)\\ &= \ln C(\boldsymbol{\alpha}) + \sum_{c=1}^{C}(\alpha_c-1)\left[\psi\left(\alpha_c\right) - \psi\left(\hat{\alpha}\right)\right].\end{aligned}$$

For the stretching/shrinking factor a_i, we have

$$\begin{aligned}&\mathbb{E}\ln\mathcal{P}(a_i)\\ =\ &\ln\frac{1}{\sqrt{2\pi}} - \ln\sigma_a - \frac{1}{2\sigma_a^2}\mathbb{E}(a_i^2 - 2a_i + 1)\\ =\ &\ln\frac{1}{\sqrt{2\pi}} - \ln\sigma_a - \frac{1}{2\sigma_a^2}\left[(\sigma_{a_i}^2)^* + (\mu_{a_i}^*)^2 - 2\mu_{a_i}^* + 1\right],\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}\ln q^*(a_i)\\ =\ &\ln\frac{1}{\sqrt{2\pi}} - \ln\sigma_{a_i}^* - \frac{1}{2(\sigma_{a_i}^*)^2}\mathbb{E}\left[a_i^2 - 2a_i\mu_{a_i}^* + (\mu_{a_i}^*)^2\right]\\ =\ &\ln\frac{1}{\sqrt{2\pi}} - \ln\sigma_{a_i}^* - \frac{1}{2(\sigma_{a_i}^2)^*}\left[(\sigma_{a_i}^2)^* + (\mu_{a_i}^*)^2 - 2(\mu_{a_i}^*)^2 + (\mu_{a_i}^*)^2\right]\\ =\ &\ln\frac{1}{\sqrt{2\pi}} - \ln\sigma_{a_i}^* - \frac{1}{2}.\end{aligned}$$

The expressions for $\mu_{a_i}^*$ and $(\sigma_{a_i}^2)^*$ are given in (5.12) and (5.13), respectively.

For the shift s_i,

$$\begin{aligned}\mathbb{E}\ln\mathcal{P}(s_i) &= \mathbb{E}\left(\ln\mathbf{1}_{\{-\phi<s_i<\phi\}}\frac{1}{2\phi}\right)\\ &= -\ln(2\phi).\end{aligned}$$

For the truncated normal $q^*(s_i)$, the density is given by

$$q^*(s_i) = \frac{\frac{1}{\sqrt{2\pi}\tilde{\sigma}_{s_i}} \exp\left\{-\frac{1}{2}\frac{(s_i-\tilde{\mu}_{s_i})^2}{\tilde{\sigma}_{s_i}^2}\right\}}{\Phi(\frac{\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}}) - \Phi(\frac{-\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}})}\mathbf{1}_{\{-\phi<s_i<\phi\}}.$$

Let us denote $Z_{s_i} = \Phi(\frac{\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}}) - \Phi(\frac{-\phi-\tilde{\mu}_{s_i}}{\tilde{\sigma}_{s_i}})$. Then,

$$\begin{aligned}
\mathbb{E}\ln q^*(s_i) &= \mathbb{E}\left(-\ln\tilde{\sigma}_{s_i} - \frac{1}{2}\ln(2\pi) - \ln Z_{s_i} - \frac{1}{2\tilde{\sigma}_{s_i}^2}(s_i^2 - 2s_i\tilde{\mu}_{s_i} + \tilde{\mu}_{s_i}^2)\right) \\
&= -\ln\tilde{\sigma}_{s_i} - \frac{1}{2}\ln(2\pi) - \ln Z_{s_i} - \frac{1}{2\tilde{\sigma}_{s_i}^2}\left((\sigma_{s_i}^2)^* + (\mu_{s_i}^*)^2 - 2\tilde{\mu}_{s_i}\mu_{s_i}^* + \tilde{\mu}_{s_i}^2\right).
\end{aligned}$$

The expressions for $\tilde{\mu}_{s_i}$, $\tilde{\sigma}_{s_i}^2$, $\mu_{a_i}^*$, and $(\sigma_{a_i}^*)^2$ are given by (5.14), (5.15), (5.16), and (5.17), respectively.

Finally, for the expectations related to τ, we have

$$\begin{aligned}
\mathbb{E}[\ln\mathcal{P}(\tau)] &= \mathbb{E}(\kappa\ln\theta - \ln\Gamma(\kappa) + (\kappa-1)\ln\tau - \theta\tau) \\
&= \kappa\ln\theta - \ln\Gamma(\kappa) + (\kappa-1)[\psi(\kappa^*) - \ln(\theta^*)] - \theta\frac{\kappa^*}{\theta^*},
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\ln q^*(\tau)] &= \mathbb{E}(\kappa^*\ln\theta^* - \ln\Gamma(\kappa^*) + (\kappa^*-1)\ln\tau - \theta^*\tau) \\
&= \kappa^*\ln\theta^* - \ln\Gamma(\kappa^*) + (\kappa^*-1)[\psi(\kappa^*) - \ln(\theta^*)] - \kappa^*.
\end{aligned}$$