Manifold-valued Dirichlet Processes

Hyunwoo J. Kim[†] HWKIM@CS.WISC.EDU
Jia Xu[†] JIAXU@CS.WISC.EDU
Baba C. Vemuri[§] VEMURI@CISE.UFL.EDU
Vikas Singh[†] VSINGH@BIOSTAT.WISC.EDU
[†]University of Wisconsin-Madison, Madison, WI 53706, USA
[§]University of Florida, Gainesville, FL 32611, USA
http://pages.cs.wisc.edu/~hwkim/projects/dp-mglm/

Abstract

Statistical models for manifold-valued data permit capturing the intrinsic nature of the curved spaces in which the data lie and have been a topic of research for several decades. Typically, these formulations use geodesic curves and distances defined *locally* for most cases — this makes it hard to design parametric models *globally* on smooth manifolds. Thus, most (manifold specific) parametric models available today assume that the data lie in a small neighborhood on the manifold. To address this 'locality' problem, we propose a novel nonparametric model which unifies multivariate general linear models (MGLMs) using multiple tangent spaces. Our framework generalizes existing work on (both Euclidean and non-Euclidean) general linear models providing a recipe to globally extend the locally-defined parametric models (using a mixture of local models). By grouping observations into sub-populations at multiple tangent spaces, our method provides insights into the hidden structure (geodesic relationships) in the data. This yields a framework to group observations and discover geodesic relationships between covariates X and manifold-valued responses Y, which we call Dirichlet process mixtures of multivariate general linear models (DP-MGLM) on Riemannian manifolds. Finally, we present proof of concept experiments to validate our model.

1. Introduction

The regression problem is amongst the most fundamental statistical tools in data analysis. If $x \in X$ is a set of covariates and $y \in Y$ is a measured response variable, the inference task is to identify the function $\ell(\cdot)$ such that $y = \ell(x) + \epsilon$, where ϵ is the noise term. Linear regression corresponds to the setting where $\ell(\cdot)$ is a linear function. Depending on other forms of $\ell(\cdot)$ and/or distributional assumptions on the response variables, we obtain progressively richer formulations such as logistic and Poisson regression. Often, one wants to determine non-linear relationships between the response variable and the covariates. While occasionally, applying a non-linearity to the output of a linear function is sufficient, such models fall short of characterizing arbitrarily shaped response functions — for instance, a mixture of simple (e.g., linear) models which pertain to clusters of the covariates which have similar relationships to the response variable, y. One solution is to impose a non-parametric Bayesian prior on a set of linear models. A constructive example of this idea is the Dirichlet Process Mixtures of Generalized Linear Models (DP-GLM) (Hannah et al., 2011).

While the family of linear regression models is very well studied, they are not directly applicable when the response variables y do not live in a vector space. Various scientific disciplines routinely acquire measurements where y is manifold-valued. For instance, the response variable may be a probability distribution function, a parametric family such as a multinomial, a covariance matrix or samples drawn from a high dimensional unit sphere. Such data arise routinely in machine learning (Lebanon, 2005; Ho et al., 2013b; Cherian & Sra, 2011; Sra & Hosseini, 2013), medical imaging (Cetingul & Vidal, 2011; Lenglet et al., 2006) and computer vision (Srivastava et al., 2007; Porikli et al., 2006; Cherian & Sra, 2014). Even when performing a basic statistical analysis on such datasets, we cannot apply vector-space operations (such as addition and multiplication) because the manifold is not a vector space. Forcibly assuming a Euclidean structure on such response variables may yield poor goodness of fit and/or weak statistical power for a fixed sample size. Driven by these motiva-

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

tions, there is a rapidly developing body of theoretical and applied work which generalizes classical tools from multivariate statistics to the Riemannian manifold setting.

Various statistical constructs have been successfully extended to Riemannian manifolds: these include regression (Zhu et al., 2009), classification (Xie et al., 2010), margin-based and boosting classifiers (Lebanon, 2005), interpolation, convolution, filtering (Goh et al., 2009), dictionary learning (Ho et al., 2013b; Cherian & Sra, 2011), and sparse coding (Cherian & Sra, 2014). Further, projective dimensionality reduction has also been studied in depth. For instance, the generalization of Principal Components analysis (PCA) via the so-called Principal Geodesic Analysis (PGA) (Fletcher et al., 2004), Geodesic PCA (Huckemann et al., 2010), Exact PGA (Sommer et al., 2013), Horizontal Dimension Reduction (Sommer, 2013), CCA on manifolds (Kim et al., 2014a), and an extension of PGA to tensor fields, a Riemannian manifold with product space structure (Xie et al., 2010). While these set of results significantly expand the operating range of multivariate statistics to the Riemannian manifold setting, methods that can reliably identify *non-linear* relationships between covariates and manifold valued response variables have not been as well studied. Many of these constructions fit a *single* model to the data, which is problematic if all of the data are not within the injectivity radius (Do Carmo, 1992). By allowing our formulation to characterize the samples as a mixture of simpler (e.g., linear) models, we resolve this limitation for complete, simply connected non-positively curved Riemannian manifolds. Our nonparametric extension is however still valid (within the injectivity radius) for other Riemannian manifolds, see (Afsari, 2011) for bounds on injectivity radius.

Specifically, we propose a new Bayesian model to extend the mixture of GLMs on the manifold of symmetric positive-definite (SPD) matrices using a Dirichlet prior. The clustering effect of the DP mixture leads to an infinite mixture of GLMs which effectively identifies proper local regions (tangent spaces) in which covariates exhibit geodesic relationship with manifold-valued responses. The goal here is to provide a comprehensive statistical framework for Dirichlet Process Mixtures Models where x lives in Euclidean space but y is manifold-valued. Specifically, to make our presentation concrete, we will study the setting for the SPD(n) manifolds (Bhatia, 2009) while noting that our techniques, carry through to other related Riemannian manifolds which share similar geometric properties (i.e., complete, simply connected and non-positively curved).

Related Work. There are several research results in literature that are related to and/or motivate this work. Separate from algorithms for multivariate statistics on manifolds (Lenglet et al., 2006), a distinct body of literature corresponds to statistical machine learning papers on nonparametric Bayesian techniques. Particularly, DP mixture models for prediction are closely related to some of our results and include the generalized linear models (Mukhopadhyay & Gelfand, 1997; Hannah et al., 2011), infinite SVM (iSVM) via DP priors (Zhu et al., 2011) and DP multinomial logit model (dpMNL) (Shahbaba & Neal, 2009). Our development is also loosely related to some old and new work in statistics on matrix-variate distributions. We use various basic concepts from the seminal book on this topic (Gupta & Nagar, 1999), as well as more recent work including distributions specifically related to medical imaging (Schwartzman, 2006), matrix-stick breaking process (MSBP) (Dunson et al., 2008), Dirichlet process mixture models (DPMM) on the unit sphere (Straub et al., 2015), SPD using Wishart distribution (Cherian et al., 2011) and matrix-variate Dirichlet priors (Zhang et al., 2014). Finally, our work is inspired by the DP mixtures estimation schemes in (Neal, 2000), which are related to Hybrid Monte Carlo or Hamiltonian Monte Carlo (HMC) algorithms (Duane et al., 1987; Neal, 2011). The reader will shortly recognize that the heart of our algorithm is a new HMC method for manifold-valued parameters, which may be of independent interest. Note that the Riemann manifold Langevin and Hamiltonian Monte Carlo (RMHMC) and Stochastic gradient Riemannian Langevin dynamics (SGRLD) methods have been proposed via a Riemannian metric on the probability space (Girolami & Calderhead, 2011), but they are not directly applicable for *parameters* on Riemannian manifolds – the setting considered here.

The main contributions of this work are: **a)** First, we present a new class of non-parameteric Bayesian mixture models which seamlessly combine both manifold-valued data and Euclidean representations. **b)** We investigate distributions on the SPD manifold and propose a specialized HMC algorithm which efficiently estimates manifold-valued parameters. **c)** We propose a new distribution to obtain a pair of parameters for models on the SPD manifold and its tangent space.

2. Preliminaries

General Linear Model. We start with the well-known multivariate general linear model (MGLM). Given pairs of covariates $\boldsymbol{x}_i \in \mathbf{R}^d$ and response variables $\boldsymbol{y}_i \in \mathbf{R}^{d'}$, we solve, $\boldsymbol{y}_i = \boldsymbol{\beta}^0 + \boldsymbol{\beta}^1 x_i^1 + \ldots + \boldsymbol{\beta}^d x_i^d + \epsilon$, where $\{\boldsymbol{\beta}^j\}_{j=0}^d \subset \mathbf{R}^{d'}$ are regression coefficients. It is known that the MGLM model assumes that $\boldsymbol{x}_i$ the covariates relate to $\boldsymbol{y}_i$ the responses via a linear function. If desired, one may apply non-linearity to the output but this cannot be a direct function of the covariates. To address this limitation and allow the response to be non-linearly related to the covariates, we may write a modified version as,

$$\boldsymbol{y}_i = \boldsymbol{\beta}_i^0 + \boldsymbol{\beta}_i^1 x_i^1 + \boldsymbol{\beta}_i^2 x_i^2 + \ldots + \boldsymbol{\beta}_i^d x_i^d + \epsilon \qquad (1)$$

where $\{\boldsymbol{\beta}_i^j\}_{j=0}^d \subset \mathbf{R}^{d'}$ are the unknown regression coefficients for each i. In this formulation, we allow each instance to have its own regression parameters, which offers advantages but creates an overfitting problem. The main flexibility offered by (1) is that the nonlinearity can be achieved by a mixture of an infinite number of linear models. On the other hand, fitting this model is ill-posed unless the regression parameters are constrained. Fortunately, the latter issue can be addressed by imposing a Dirichlet process (DP) prior as in (Hannah et al., 2011; Zhang et al., 2014). The DP mixture model is given by

$$(\boldsymbol{x}_i, \boldsymbol{y}_i)|\theta_i \sim F(\theta_i),\ \theta_i|G \sim G,\ G \sim DP(G_0, \nu). \quad (2)$$

where G_0 is a base distribution and ν is a concentration parameter. Using (2), a DP mixture of multivariate general linear models (DP-MGLM) is simply obtained by plugging in a d'-dimensional response Y into a DP mixture of generalized linear models (DP-GLM) studied in (Hannah et al., 2011; Mukhopadhyay & Gelfand, 1997). Specifically, we assume that the covariates X are modeled by a mixture of normal distributions, and that the responses Y are modeled by MGLMs conditioned on the covariates. The models are connected by associating a set of MGLM coefficients θ_y with each mixture component θ_x. Let $\theta = (\theta_x, \theta_y)$ be the set of parameters over X and $Y|X$, and let G_0 be a base distribution on θ. Then the DP-MGLM model, a special case of (Hannah et al., 2011), is given by,

$$\begin{aligned} \boldsymbol{y}_i|\boldsymbol{x}_i, \theta_{y_i} &\sim \mathcal{N}(\hat{\boldsymbol{y}}_i, \boldsymbol{\sigma}_y^2), \text{ where } \hat{\boldsymbol{y}}_i = MGLM(\boldsymbol{x}_i, \theta_{y_i}) \\ \boldsymbol{x}_i|\theta_{x_i} &\sim \mathcal{N}(\boldsymbol{\mu}_{x_i}, \boldsymbol{\sigma}_{x_i}^2), \text{ where } \theta_{x_i} = (\boldsymbol{\mu}_{x_i}, \boldsymbol{\sigma}_{x_i}^2) \\ \theta_i|G &\sim G,\ G \sim DP(G_0, \nu), \text{ where } \theta_i = (\theta_{x_i}, \theta_{y_i}). \end{aligned} \quad (3)$$

What if Y is manifold-valued? Observe that the MGLM in (3) assumes that the response variable Y is in a *vector space*. The main goal of this paper is to study the statistical inference task when Y are samples from a curved Riemannian manifold. Here, even the most basic fitting (and error distribution) assumptions are violated. For instance, symmetric positive-definite (SPD) matrix-valued response variables do not live in a vector space, and a linear combination in general may not yield an SPD matrix. The second issue is that the likelihood function of MGLM, critical in designing a sampling strategy for (3) is defined by a distance metric in the ambient (Euclidean) space. It ignores the underlying intrinsic geometry of the manifold-valued data. We will provide a solution to this problem shortly.

Basic Differential Geometry Notations. First, we introduce some concepts and notations of differential geometry (Do Carmo, 1992). On Riemannian manifolds, the geodesic curve (shortest path) from y_i to y_j can be parameterized by a tangent vector in the tangent space at y_i with an exponential map $\text{Exp}(y_i, \cdot) : T_{y_i}\mathcal{M} \to \mathcal{M}$ (mapping from the tangent space to the manifold). The inverse of the exponential map is the logarithm map, $\text{Log}(y_i, \cdot) : \mathcal{M} \to T_{y_i}\mathcal{M}$ (i.e., manifold to tangent space). Note that we will assume that the key conditions needed for these maps to exist (Pennec, 2006) are satisfied. The geodesic distance is measured by the length of tangent vector. Also, in this paper we use $\exp(\cdot)$ and $\log(\cdot)$ to denote matrix exponential and logarithm respectively.

Let $B \in \mathcal{M}$ be an anchor (base) point and let $\{V^j\}_{j=1}^d \subset T_B\mathcal{M}$ denote tangent vectors. They correspond to $\boldsymbol{\beta}^0$ and $\{\boldsymbol{\beta}^j\}_{j=1}^d$ resp. in (3). A model for MGLM on Riemannian manifolds (Kim et al., 2014b) is,

$$\boldsymbol{y} = \text{Exp}(\text{Exp}(B, \sum_{j=1}^d V^j x^j), \epsilon), \quad (4)$$

As briefly described above, DP-MGLM on Riemannian manifolds will allow each example i to have its own regression parameters. That is, each example $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathbf{R}^d \times \mathcal{M}$ has parameters $(B_i, \boldsymbol{V}_i)$. To reduce notational clutter, we will use the short-hand $\boldsymbol{V}\boldsymbol{x} := \sum_{j=1}^d \boldsymbol{V}^j x^j$, where $\boldsymbol{x} \in \mathbf{R}^d$.

Geometry of SPD(n). We will present our ideas in the context of DP-MGLM on the space of $n \times n$ symmetric positive definite matrices SPD(n). To do so, we briefly describe some basic geometric concepts related to this manifold, see (Moakher, 2005; Bhatia, 2009) for more details. The tangent space of SPD(n) is the space of $n \times n$ symmetric matrices, Sym(n). When the manifold is equipped with a GL-invariant metric, the geodesic distance between two SPD matrices B and Y is $d(B, Y)^2 = \text{tr}(\log^2(B^{-1/2}YB^{-1/2}))$. The exponential map and logarithm map are given by $\text{Exp}(B, V) = B^{1/2}\exp(B^{-1/2}VB^{-1/2})B^{1/2}$ and $\text{Log}(B, Y) = B^{1/2}\log(B^{-1/2}YB^{-1/2})B^{1/2}$ (Moakher, 2005; Cheng & Vemuri, 2013).

3. DP-MGLM on Riemannian manifolds

In this section, we specify an end to end model for DP-MGLM on the SPD manifold. To do this, we need a few key technical ingredients:

Step (a). First, we need to model the cluster of covariates, X which follows from an adaptation of existing work on DP-GLM (Hannah et al., 2011).

Step (b). Next, we need to characterize the conditional distribution $\mathbb{P}(y|x)$ specifically for the case where $y \in$ SPD(n). This requires two key steps. **i)** We need to specify the parameters for DP-MGLM for the SPD manifold setting. In particular, we should identify which space (i.e., the manifold) each parameter corresponds to when $y \in$ SPD(n). **ii)** Then, we must make appropriate distributional assumptions for the respective spaces so that the follow-up inference scheme is both statistically sound and computationally feasible.

We first discuss Step (a). To model the relationship between x and y, we non-parametrically model the joint distribution $\mathbb{P}(x, y|\theta) = \mathbb{P}(y|x, \theta)\mathbb{P}(x|\theta)$, using a Dirichlet process mixture (θ is a cluster model parameter). Within each cluster, the relationship between y and x is expressed using an MGLM. Note that the covariates X live in a Euclidean space $\mathbf{R}^d$. The parameters for X are $\theta_x = (\boldsymbol{\mu}_x, \boldsymbol{\sigma}_x^2)$, same as in (3). So, we can model a cluster of covariates X by a Gaussian distribution with parameters $(\boldsymbol{\mu}_x, \boldsymbol{\sigma}_x^2)$. The prior for these parameters is given by a DP-prior.

We now describe Step (b). For the Riemannian setting, we first give the corresponding expression for (3) for parameters of the MGLM, i.e., $\theta_y = (B, \boldsymbol{V})$, where $B \in \text{SPD}(n)$ and $\boldsymbol{V} \in \text{Sym}(n)^d$. Here, $\text{Sym}(n)$ denotes the space of symmetric matrices of size $n \times n$ and we have d separate V's in $\boldsymbol{V}$. Recall that in a GLM, noise is modeled as a Normal distribution so that the Maximum Likelihood estimate (MLE) minimizes the least squares error. In the current setting, ideally, the MLE must minimize the geodesic distance-based error. So, we need an analogous form (for the Normal distribution) for manifold-valued y's. The solution to this is to use the "generalized Normal" distribution on the manifold (Cheng & Vemuri, 2013). Then, the maximum likelihood estimator of the MGLM turns out to be equivalent to the minimization of a least squares geodesic-distance error, given the covariance parameter σ_y^2. In the next section, we will discuss explicit forms of the density function of the generalized Normal distribution and the equivalence between the log likelihood function and squared geodesic error. So, the joint distribution in one cluster, i.e., $F(\theta_i)$ in (2), is given by,

$$\begin{aligned} Y_i|\boldsymbol{x}_i, \theta_{y_i} &\sim \mathcal{N}_{\text{SPD}}(\hat{Y}_i, \sigma_y^2), \text{ where } \hat{Y}_i = \text{Exp}(B_i, \boldsymbol{V}_i \boldsymbol{x}_i) \\ \boldsymbol{x}_i|\theta_{x_i} &\sim \mathcal{N}(\boldsymbol{\mu}_{x_i}, \boldsymbol{\sigma}_{x_i}^2), \text{ where } \theta_{\boldsymbol{x}_i} = (\boldsymbol{\mu}_{\boldsymbol{x}_i}, \boldsymbol{\sigma}_{x_i}^2) \end{aligned} \tag{5}$$

where, $\mathcal{N}$ is a Normal distribution for $\boldsymbol{x} \in \mathbf{R}^d$, and $\mathcal{N}_{\text{SPD}}$ is the "generalized Normal" distribution for $Y \in \text{SPD}(n)$. The next step is to define the base distribution G_0 over $\theta = (\boldsymbol{\mu}_x, \boldsymbol{\sigma}_x^2, B, \boldsymbol{V})$ where σ_y is assumed to be given (or empirically estimated). To make it analytically feasible, we use a Normal (or log normal) distribution.

$$\begin{aligned} \boldsymbol{\mu}_x|\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0^2),\ \log(\boldsymbol{\sigma}_x^2)|M_\sigma, \Sigma_\sigma \sim \mathcal{N}(M_\sigma, \Sigma_\sigma^2) \\ B \sim \mathcal{N}_{\text{SPD}}(\mu_B, \sigma_B^2),\ \boldsymbol{V} \sim \mathcal{N}_{\text{Sym}}(\mu_V, \boldsymbol{\sigma}_V^2)^d, \end{aligned} \tag{6}$$

where $\mathcal{N}_{\text{Sym}}$ is a symmetric matrix-variate Normal distribution over $V \in \text{Sym}(n)$ defined later in (8).

Remark. For a SPD matrix-valued variable B, other distributions such as log normal, Wishart or inverse Wishart distribution can also be used within G_0. However, these distributions do not necessarily yield a sample B around mean or mode of the distribution with respect to a GL-invariant metric. So, if one has knowledge of a highly probable B (e.g., the Fréchet mean) and its neighbors w.r.t. the geodesic distance, then a log Normal or the generalized Normal distribution in (9) is more suitable. Using a log Normal distribution is useful because it is easier to sample (compared to generalized Normal). However, the Jacobian of the matrix exponential varies as a function of the sample location, which makes it harder to deal with the derivative of its log likelihood. We provide candidate distributions for the base distribution over $\text{Sym}(n)$ and $\text{SPD}(n)$ and the corresponding density functions and their log likelihood in the extended version of this paper, which are useful in deriving the final HMC algorithm.

4. Posterior Sampling

In this section, we describe our proposed method for posterior inference. To place our contribution in context, we first summarize the conventional approach and then the key modifications needed.

If the base measure G_0 is *conjugate*, then it yields an efficient sampling procedure called the "collapsed Gibbs sampling" (Neal, 2000). Unfortunately, the distributions in (6) are not known to be conjugate. To address the above problem, we instead use Gibbs sampling with auxiliary parameters by adapting Algorithm 8 in (Neal, 2000). This requires sampling cluster parameters for each cluster such that the distribution remains invariant — in our setting, this is simpler for $\theta_x = (\boldsymbol{\mu}_x, \boldsymbol{\sigma}_x^2)$ but more involved for $\theta_y = (B, \boldsymbol{V})$. For θ_x, we use a simple slice sampling for updating the parameters (Neal, 2000). Updating the regression parameters, θ_y is more challenging. This is because while slice sampling can be performed for each dimension independently, this is not true for the manifold-valued B. So, for a more effective sampling, we generalize the HMC method, used in Dirichlet process mixtures of multinomial logit model (dpMNL), a special case of DP-GLM (Hannah et al., 2011; Shahbaba & Neal, 2009).

The HMC algorithm needs to be generalized for the MGLM on Riemannian manifolds. Note that our formulation here is distinct from the Riemann manifold Langevin and Hamiltonian Monte Carlo (RMHMC) technique in (Girolami & Calderhead, 2011), which is Riemannian in the sense that it treats the joint probability space of the data as a Riemannian manifold. This is done by defining a Riemmanian metric (e.g., the Fisher-Rao metric) and the negative Hessian of the log-prior. However, the *data* itself are *not* assumed to lie on a manifold.

When the parameters lie in Euclidean space. Recall that conventional rejection sampling (such as Metropolis-Hastings) suffers from a low acceptance rate. However, HMC provides an ergodic Markov chain capable of achieving both large transitions and high acceptance rate. The underlying theory of HMC relies on Hamiltonian dynamics. Hamiltonian dynamics operates on a d-dimensional

position vector q and a d-dimensional momentum vector p, so that the full state space has $2d$-dimensions. For HMC, we usually use Hamiltonian functions written as $H(q,p) = U(q)+K(p)$. Here, $U(q)$ is the *potential energy* and $K(p)$ is the *kinetic energy*. Generally, the posterior distribution for the model parameters is the usual object of interest and hence these parameters take the role of the position, q. The potential energy is $U(q) = \log[\pi(q)L(q|D)]$, where $\pi(q)$ is the prior density, and $L(q|D)$ is the likelihood function, given the data D. The kinetic energy is defined by $K(p) = p^T M^{-1} p/2$, where, p is the auxiliary variable which can be interpreted as momentum and M is the "mass matrix". HMC proposes transitions $\theta \to \theta^*$, which are then accepted with probability based on Hamiltonian functions $\min\{1, \exp(H(q,p) - H(q^*,p^*)\}$, where q^* and p^* are proposed parameters and their momentum respectively (Neal, 2011).

Manifold setting. Defining the *potential energy* function for the HMC algorithm is simple – we can use the negative log of the joint probability. To define the *kinetic energy*, we must account for manifold-valued parameters; $B \in \mathcal{M}$ for the *intercept* and a set of tangent vectors $\boldsymbol{V}$ for the *slope*. To this end, the following description provides solutions to the main questions, **(a)** How to define the change of parameters B and $\boldsymbol{V}$? **(b)** How to update the parameters? **(c)** How to transport objects (such as momentum) to the appropriate tangent space? **(d)** How to sample the initial momentum?

First, we define the potential energy. To do so, we introduce the explicit form of probability density functions. The density function of the Normal distribution as a prior over $\text{Sym}(n)$ (definition 3.1.3 in (Schwartzman, 2006)) is

$$f_{\text{Sym}}(V; \mu_V, B) = \frac{1}{Z} \exp\left(-\frac{1}{2}\text{tr}[((V-\mu_V)B^{-1})^2]\right) \quad (7)$$

where $Z = (2\pi)^{q/2}|B|^{(n+1)/2}$, $|B|$ is the determinant of B and $q = n(n+1)/2$. Also, the simpler version (definition 3.1.4) in (Schwartzman, 2006)) is

$$f_{\text{Sym}}(V; \mu_V, \sigma^2) = \frac{1}{(2\pi)^{q/2}\sigma^q} \exp\left(-\frac{1}{2\sigma^2}\text{tr}[(V-\mu_V)^2]\right). \quad (8)$$

Next, to define the likelihood of $y \in \text{SPD}$, we introduce an explicit form of the generalized Normal distribution.

$$f_{\text{SPD}}(y; \mu_y, \sigma_y^2) = \frac{1}{Z(\mu_y, \sigma_y)} \exp\left(-\frac{d(y,\mu_y)^2}{2\sigma^2}\right) \quad (9)$$

where $Z(\mu_y, \sigma_y) = \int_{\mathcal{M}} \exp\left(-\frac{d(y,\mu_y)^2}{2\sigma_y^2}\right) dy$. Here, it turns out that $Z(\mu_y, \sigma_y)$ is constant w.r.t. μ when $\mathcal{M}$ is a symmetric space (Fletcher, 2013). So, the negative log-likelihood of each cluster c takes the form,

$$-\log \mathcal{L}(\theta_c^*|D_c) = n_c \log Z(\sigma_y) + \frac{1}{2\sigma_y^2}\sum_{i\in c} d(y_i, \hat{y}_i)^2 \quad (10)$$

where $\hat{y}_i = \text{Exp}(B, \boldsymbol{V}x_i)$, c is a cluster, n_c is the number of its elements. Interestingly, because the normalization factor is constant, maximizing the log likelihood reduces to minimizing the least squares error. We can now define our potential function as

$$U(B, \boldsymbol{V}) := \frac{1}{\sigma^2} E(B, \boldsymbol{V}) - \log f_{\text{SPD}}(B) - \log f_{\text{Sym}}(\boldsymbol{V}) \quad (11)$$

where $E(B, \boldsymbol{V}) := \frac{1}{2}\sum_i d(y_i, \hat{y}_i)^2$. We must now account for the change of parameters. Notice that the change of manifold valued $B \in \mathcal{M}$ is represented by a tangent vector $\dot{B} \in T_B\mathcal{M}$. However, the change of tangent vectors, $\dot{V}$, live in $T_V(T_B\mathcal{M})$ (a tangent space of a tangent space). Fortunately, the natural isomorphism $T_V(T_B\mathcal{M}) \cong T_B\mathcal{M}$ allows us to let $\dot{V}$ be in $T_B\mathcal{M}$ (Fletcher, 2013). By construction, the priors for B and $\boldsymbol{V}$ are Gaussian and so the log of the prior density functions are quadratic forms whose derivatives can be obtained analytically. As described in the extended version, these are given by,

$$\nabla_B U \approx -\frac{1}{\sigma_y^2}\sum_{i=1}^N \Gamma_{\hat{y}_i \to B}\text{Log}(\hat{y}_i, y_i) - \nabla_B \log f_{\text{SPD}}(B)$$

$$\nabla_{V^j} U \approx -\frac{1}{\sigma_y^2}\sum_{i=1}^N x_i^j \Gamma_{\hat{y}_i \to B}\text{Log}(\hat{y}_i, y_i) - \nabla_{V^j} \log f_{\text{Sym}}(V^j)$$

where Γ is the parallel transport operation.

Remarks. The least squares loss function is defined on a SPD manifold. If one uses the prior distribution over B which is defined in a Euclidean space instead of the generalized Normal distribution we use, then the gradient with respect to B needs to be separated into the derivative, $\nabla_B E$, along the curved surface (called covariant derivative) and the derivative along the ambient space $\nabla_B \log f_{\text{SPD}}$. Technically, these are not in the same space, which can be verified by comparing their respective update schemes. For instance, the next iterate B via $\nabla_B E$ is $\text{Exp}(B, \epsilon\nabla_B E)$ whereas the next iterate B suggested by $\nabla_B \log f_{\text{SPD}}$ is $B = B + \nabla_B \log f_{\text{SPD}}$. Fortunately, for V, the update schemes are identical. Both use the simple addition operation since $\nabla_{V^j} E$ and $\nabla_{V^j} \log f_{V^j}$ lie in vector spaces. A minor issue here is that their metrics might be different since $\nabla_{V^j} E$ lies in $T_B\mathcal{M}$ with a locally defined inner product $\langle U, V\rangle_B = \text{tr}(UB^{-1}VB^{-1})$ whereas $\nabla_{V^j} \log f_{\text{Sym}}(V^j) \in \text{Sym}(n)$ with the natural inner product $\langle U, V\rangle = \text{tr}(UV)$ in Euclidean space (and is independent of B) where a symmetric matrix-variate normal distribution is defined as (8). So, their scales might be different. In addition, there is no reason to expect that the samples drawn from this distribution in (8) are normally distributed in an arbitrary tangent space at B with respect to the GL-invariant metric. We provide a cleaner solution next.

Algorithm 1 HMC algorithm for DP-MGLM on Riemannian manifolds

Input: $(B_{cur}, \boldsymbol{V}_{cur}) \in \mathcal{M} \times T_B\mathcal{M}^n$, Leapfrog parameters $\epsilon \in \mathbf{R}_{++}, L \in \mathbf{Z}_{++}$
Output: $(B_{next}, \boldsymbol{V}_{next}) \in \mathcal{M} \times T_B\mathcal{M}^n$
Sample $(\dot{B}_{cur}, \dot{\boldsymbol{V}}_{cur}) \in T_B\mathcal{M} \times T_B\mathcal{M}^n$ from independent normal distribution w.r.t. Riemannian metric.
Initialize $(B, \boldsymbol{V}, \dot{B}, \dot{\boldsymbol{V}}) \leftarrow (B_{cur}, \boldsymbol{V}_{cur}, \dot{B}_{cur}, \dot{\boldsymbol{V}}_{cur})$
$\dot{B} \leftarrow \dot{B} - \frac{\epsilon}{2}\nabla_B U(B, \boldsymbol{V})$ and $\dot{\boldsymbol{V}} \leftarrow \dot{\boldsymbol{V}} - \frac{\epsilon}{2}\nabla_{\boldsymbol{V}} U(B, \boldsymbol{V})$
for $i \in \{1, \cdots, L\}$ **do**
  $B' \leftarrow B, B \leftarrow \text{Exp}(B, \epsilon\dot{B}), V \leftarrow V + \epsilon\dot{V}$
  $(\boldsymbol{V}, \dot{B}, \dot{\boldsymbol{V}}) \leftarrow (\Gamma_{B'\to B}\boldsymbol{V}, \Gamma_{B'\to B}\dot{B}, \Gamma_{B'\to B}\dot{\boldsymbol{V}})$
  /* Parallel transport [1] */
  if i !=L **then**
    $\dot{B} \leftarrow \dot{B} - \epsilon\nabla_B U(B, \boldsymbol{V})$ and $\dot{\boldsymbol{V}} \leftarrow \dot{\boldsymbol{V}} - \epsilon\nabla_{\boldsymbol{V}} U(B, \boldsymbol{V})$
  end if
end for
$\dot{B} \leftarrow \dot{B} - \frac{\epsilon}{2}\nabla_B U(B, \boldsymbol{V})$ and $\dot{\boldsymbol{V}} \leftarrow \dot{\boldsymbol{V}} - \frac{\epsilon}{2}\nabla_{\boldsymbol{V}} U(B, \boldsymbol{V})$
Accept $(B, \boldsymbol{V})$ with probability $\min[1, \exp(H(\dot{B}_{cur}, \dot{\boldsymbol{V}}_{cur}, B_{cur}, \boldsymbol{V}_{cur}) - H(\dot{B}, \dot{\boldsymbol{V}}, B, \boldsymbol{V}))]$

4.1. Defining an alternative distribution for both the base point B and a set of tangent vectors $\boldsymbol{V}$

As a solution, we propose a new distribution for $(B, V) \in \mathcal{M} \times T_B\mathcal{M}$ by conditionally combining two distributions.

$$B|\mu_B, \sigma_B^2 \sim \mathcal{N}_{\text{SPD}}(B|\mu_B, \sigma_B^2), V|\mu_V, B \sim \mathcal{N}_{\text{Sym}}(V|\mu_V, B) \tag{12}$$

This is more of a "Normal like" distribution for both B and V w.r.t a GL-invariant metric. Lemma 4.1 (proof in the extended version) shows,

Lemma 4.1. *Let $(B, V) \in SPD(n) \times Sym(n)$ be a sample drawn using (12) , then V is Normally distributed w.r.t. a GL-invariant metric at the tangent space $T_B\mathcal{M}$ at B. For each B, the probability density function of V is proportional to $\exp(-\frac{1}{2}\|V\|_B^2)$) at $T_B\mathcal{M}$, when $\mu_V = 0$.*

Note that it is not exactly a Normal distribution because of the dependence on $|B|$. More details of these distributions are provided in the extended version. With these components, our final **HMC algorithm** is given by Algorithm 1.

Some additional details. We use the exponential map for parameter updates for $B \in \text{SPD}(n)$. For all parameters in the vector space ($T_B\mathcal{M}$), the vector addition operation suffices. However, once the base point B_{old} changes to a new B, then the objects $\dot{B}, \dot{V}, V$ do not belong to the tangent space of B anymore. So, they need to be parallel transported from the old anchor point B_{old} to the new anchor point B. Then, the kinetic energy at each time point can be properly measured by the sum of squared norms of the tangent vectors in the new tangent space at B. Finally, we point out that the initial momentum is set by finding a random direction in the tangent space at B; its magnitude is given by the length w.r.t. the Riemannian inner product. Let D denote the measurements (or data). For the prediction of response Y, the conditional distribution of $Y|X = x, D$ is $f(Y|X = x, D) \approx \frac{1}{S}\sum_{s=1}^{S} f(Y|X = x, \theta^{(s)})$. Thus, the prediction $\mathbb{E}[Y|X = x, D] = \mathbb{E}[\mathbb{E}[Y|X = x, \theta]|D]$ is approximated by the posterior samples $\{\theta^{(s)}\}_{s=1}^{S}$. Since Y is on $\mathcal{M}$, the expectation is the Fréchet mean. This can be updated in an online manner for the SPD manifold (Ho et al., 2013a).

[1]Parallel transport: Let $\mathcal{M}$ be a differentiable manifold with an affine connection ∇ and I be an open interval. Let $c : I \to \mathcal{M}$ be a differentiable curve in $\mathcal{M}$ and let V_0 be a tangent vector in $T_{c(t_0)}\mathcal{M}$, where $t_0 \in I$. Then, there exists a unique parallel vector field V along c, such that $V(t_0) = V_0$. Here, $V(t)$ is called the *parallel transport* of $V(t_0)$ along c. We denote the parallel transport from y to y' of V as $\Gamma_{y\to y'}V$. Intuitively, parallel transport of V_0 along curve c can be interpreted as the parallel translation of V_0 on manifolds preserving the angle between $V(t)$ and c.

5. Experiments

To evaluate the proposed model, we conduct a set of experiments on synthetic and real-world data.

5.1. Experiments on synthetic data

DP mixtures of MGLM on SPD. We first evaluate whether our algorithm can simultaneously find a set of geodesic relationships between the covariates and the manifold-valued response variables. We follow the experimental protocol from (Hannah et al., 2011) which is broadly used in the literature, but with the distinction that now we have $Y \in \text{SPD}(n)$. To do this, we simulate data from multiple geodesic curves which are parameterized by the covariates — this gives heteroscedasticity properties where DP-GLM approaches are known to be effective. The number of "local" models in this synthetic data varies between 2 to 5. Our sample size is 300. We perform a few hundred realizations where the number of MCMC samples in each realization is 1000. We set the burn-in period to 100 epochs. When the data is sampled from a single local model, one should expect both a manifold-valued MGLM and our model to perform well and estimate the parameters correctly. However, when the samples are drawn from a mixture of multiple local models, the flexibility offered by our framework must yield improvements. Since visualizing the model fit on the SPD manifold is not possible, we perform a Principal Geodesic analysis (PGA) to pick a prominent direction of variance and project the original data onto this axis for evaluation. As shown in Fig. 1 (multiple datasets), in nearly all cases, the model provides a good fit and is able to identify a very good estimate of the real local relationships in the data, exactly as desired.

For quantitative evaluations, we compute the mean squared error (MSE) as well as the R-squared statistic, which are standard measures to evaluate goodness of fit. As shown

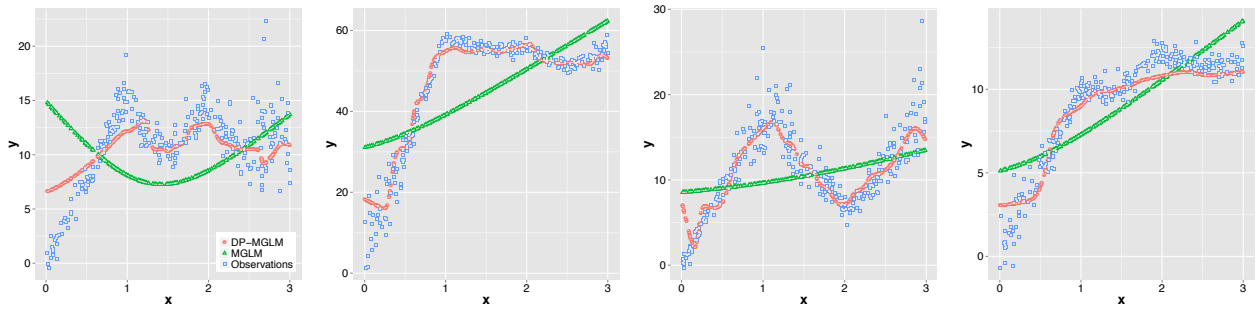

Figure 1. The figure shows the models fitted in the PGA axes space versus the covariates. The prediction of DP-MGLM is shown using a single sample from the posterior, $\theta^{(i)}$. To visualize the response variable $Y \in \text{SPD}(3)$, we project the variables onto the axis obtained by PGA (y-axis). The x-axis is the covariate $\boldsymbol{x} \in \mathbf{R}$. Red and blue correspond to our predictions and the measurements respectively.

in Table 1, at least in part due to the locality problem we described to motivate the paper, our DP-MGLM achieves much smaller MSE while consistently obtaining better $\mathbf{R}^2$ statistic, compared with a manifold-valued MGLM (and a slightly improved variant which centers the covariate). Our framework does not require centering the covariates.

Estimating Models for Spatially-based Covariates. A number of applications motivating the need for statistics on manifold-valued responses come from image analysis. To evaluate our model in this setup, we synthesized an experiment where the responses form a distribution on SPD whereas the corresponding covariates are grid points on an image lattice. The ability to estimate such models faithfully offers numerous advantages including clustering and the ability to draw samples from the estimated model, e.g., for performing downstream hypothesis tests. We test these scenarios next in the context of estimating $\mathbb{E}(y|x)$.

Our generating function is a mixture of models with spatially localized support. Each voxel is a manifold-valued measurement $Y \in \text{SPD}(3)$ (such as in diffusion tensor imaging) whose grid locations are the covariates. For ease of visual assessment, each perceptual region in Fig. 2 (left column) is generated by a single function. The top-left patch has two regions (the circle at the center and the background). Within the background region the measurements change gradually depending on horizontal coordinate. The vertical coordinate is a nuisance variable. The center left patch also has two functions and simulates diverging flow streams. That is, the orientations across the two local models are the same at the bottom and as the vertical coordinate increases, the orientations of the ellipses change (conditioned on which of the two models they came from). For both these examples, our estimated model accurately uncovers the local geodesic relationships and we obtain a clustering as a result. The assignment (from a single sample of the posterior distribution) is shown in the right column (first two rows). Note that since the two models move apart slowly (bottom to top), a simpler clustering scheme based on the product space of covariate $\boldsymbol{x}$ and the responses, e.g., k-means, and DPMM does not capture the structure without significant parameter adjustment though we acknowledge more specialized clustering methods can be used (Medvedovic & Sivaganesan, 2002). Finally, we ran a qualitative experiment where the measurements together with the covariates corresponds to a visual concept,

Table 1. Mean squared errors and R-squared (R^2) statistic w.r.t the intrinsic metric on SPD(3) for eight synthetic datasets. MGLMc denotes MGLM with centered covariate $\boldsymbol{x}$.

	Mean Squared Error		R^2	
Model	Train	Test	Train	Test
DP-MGLM	**1.18 ± 0.99**	**1.19 ± 1.04**	**0.80 ±0.06**	**0.79 ± 0.08**
MGLMc	3.40 ± 2.43	3.28 ± 2.14	0.39 ± 0.16	0.38 ± 0.16
MGLM	4.94 ± 3.40	4.80 ± 3.09	0.10 ± 0.04	0.10 ± 0.04

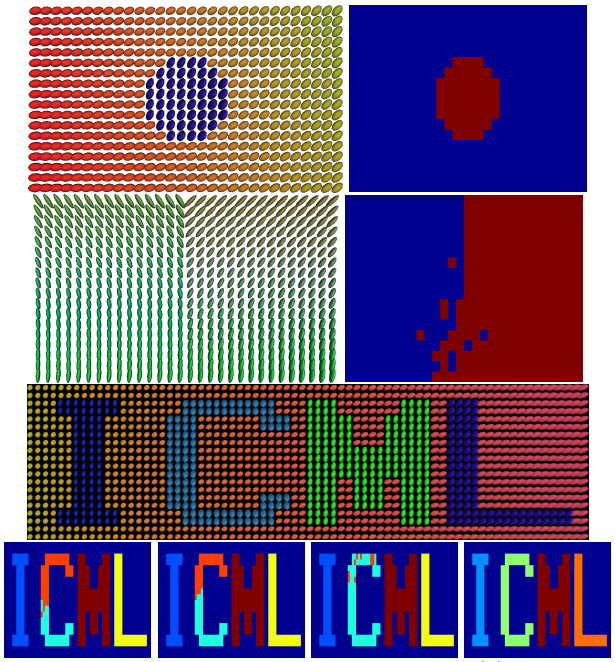

Figure 2. (Rows 1–2, col 1) Each voxel is a SPD(3) matrix; the covariates are the grid positions (horizonal, vertical coordinates). (Rows 1–2, col 2) shows a clustering result. (Row 3) is a glyph figure where the global mixtures of local models is "ICML". (Row 4) A clustering based on the posterior samples $\theta^{(i)}$.

Figure 3. The top two rows show 6 sample faces with ages ranging from $20 \sim 80$. The bottom row (left image) shows 40 landmarks (indexed by numbers) on an example image . The second image of the bottom row shows correlation magnitude of the landmark's variation with age as a heat map. **Best viewed in color.**

see Fig. 2 (third row). The goal here was to assess whether additional samples from the posterior distribution visually correspond to the same concept. We noticed that while the samples are smoothed, the clustering indices on these samples shown in Fig. 2 (fourth row) suggest that our estimated model generalizes well.

5.2. Experiments on real-world data

Next, we conduct an experiment on facial datasets which are derived from the important biometric task of face recognition and age estimation. In particular, we attempt to assess: how do facial landmark appearances evolve with age? Which age ranges/periods are most correlated with which face regions? This problem is important for facial age estimation (Guo et al., 2013). Since we expect that changes in different face regions will likely correspond to different age periods, it exhibits nice heteroscedasticity properties. We used the Lifespan database (Minear & Park, 2004), which contains 580 subjects with ages ranging from 18–93. To avoid the influence of facial expressions, we focus only on the "Neutral" subset which contains images without facial expressions and human labeled landmark points are provided (Guo et al., 2013). These include 40 points overall, see Fig. 3. We used the covariance descriptors common in image processing, computed from the feature vector $[r, c, R_{rc}, G_{rc}, B_{rc}, I_r, I_c]$, where r (and c) is row (and column) index, R, G, B are colors and I_r, I_c are intensity derivatives. The covariance matrix for an image patch (size 20×20) centered at each landmark is a 7×7 SPD, the response variable, $Y \in \mathcal{M}$. The age of the person associated with each image is the covariate, x.

We run Algorithm 1 on each landmark. The algorithm provides a set of local models for each landmark; here, these local models correspond to age ranges. In the manifold setting, each 'local' cluster (or model) can be interpreted as a geodesic explaining the relationship between the covariates (age range) and evolution in the covariance descriptor in that period. For each landmark, there are multiple clusters — we simply measure the length of the corresponding tangent vectors and pick the median as the representative. After normalization to $[0, 1]$, we show it as a color coded heat map in Fig. 3 shown in the bottom right of the figure. We see that our algorithm found that regions around the center of the eye (numbered as $2 \sim 5$, $7 \sim 10$) and nose ($27 \sim 29$) exhibit *no* meaningful relationship with age (shown in blue). On the other hand, regions around the brow ($12 \sim 18$), cheeks ($34 \sim 40$) and forehead ($21 \sim 23$) exhibit a much *stronger* relationship (e.g., wrinkles) shown in red. This is consistent with prior findings (Montillo & Ling, 2009), which identified similar landmarks as the most distinguishing identifiers for age.

6. Conclusion

We have presented a novel algorithm for Dirichlet process mixtures of multivariate general linear models on Riemannian manifolds. The formulation globally extends the locally-defined parametric models on Riemannian manifolds using a mixture of local models, thereby solving the "locality" problem pervasive in various parametric formulations for a class of Riemannian manifolds. We derive specific sampling schemes for the SPD manifold but the ideas should apply to other manifolds with similar geometries (e.g., non-positively curved). We also studied and proposed a new distribution to get a pair of parameters for models on the SPD manifold and its tangent space. On the algorithm side, we derived a specialized HMC algorithm which efficiently estimates manifold-valued parameters, which may be of independent interest. While our development here is primarily on the theoretical side, we believe that the proposal will lead to practical sampling and inference schemes for various problems in medical imaging, machine learning and computer vision that involve statistical tasks on the SPD manifold.

Acknowledgment

This work was supported in part by NIH grants AG040396 (VS), NS066340 (BCV), NSF CAREER award 1252725 (VS). Partial support was also provided by the Center for Predictive Computational Phenotyping (CPCP) at UW-Madison (AI117924). We are grateful to Michael A. Newton, Vamsi K. Ithapu and WonHwa Kim for various discussions related to the content presented in this paper.

References

Afsari, Bijan. Riemannian L^p center of mass: Existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011.

Bhatia, Rajendra. *Positive definite matrices*. Princeton University Press, 2009.

Cetingul, H. E. and Vidal, R. Sparse Riemannian manifold clustering for HARDI segmentation. In *ISBI*, pp. 1750–1753, 2011.

Cheng, Guang and Vemuri, Baba C. A novel dynamic system in the space of SPD matrices with applications to appearance tracking. *SIAM journal on imaging sciences*, 6(1):592–615, 2013.

Cherian, A., Morellas, V., Papanikolopoulos, N., and Bedros, S. Dirichlet process mixture models on SPD matrices for appearance clustering in video surveillance applications. In *CVPR*, pp. 3417–3424, 2011.

Cherian, Anoop and Sra, Suvrit. Generalized dictionary learning for SPD matrices with application to nearest neighbor retrieval. In *ECML*, pp. 318–332, 2011.

Cherian, Anoop and Sra, Suvrit. Riemannian sparse coding for positive definite matrices. In *ECCV*, pp. 299–314, 2014.

Do Carmo, Manfredo P. *Riemannian geometry*. Springer, 1992.

Duane, Simon, Kennedy, Anthony D, Pendleton, Brian J, and Roweth, Duncan. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

Dunson, David B, Xue, Ya, and Carin, Lawrence. The matrix stick-breaking process. *JASA*, 103(481):317–327, 2008.

Fletcher, P Thomas. Geodesic regression and the theory of least squares on Riemannian manifolds. *IJCV*, 105(2): 171–185, 2013.

Fletcher, P Thomas, Lu, Conglin, et al. Principal geodesic analysis for the study of nonlinear statistics of shape. *TMI*, 23(8):995–1005, 2004.

Girolami, Mark and Calderhead, Ben. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Goh, Alvina, Lenglet, Christophe, Thompson, Paul M, and Vidal, René. A nonparametric Riemannian framework for processing HARDI. In *CVPR*, pp. 2496–2503, 2009.

Guo, Guodong, Guo, Rui, and Li, Xin. Facial expression recognition influenced by human aging. *IEEE Trans. Affective Computing*, 4(3):291–298, 2013.

Gupta, Arjun K and Nagar, Daya K. *Matrix variate distributions*, volume 104. CRC Press, 1999.

Hannah, Lauren A, Blei, David M, and Powell, Warren B. Dirichlet process mixtures of generalized linear models. *JMLR*, 12:1923–1953, 2011.

Ho, J., Cheng, G., Salehian, H., and Vemuri, B. Recursive Karcher expectation estimators and geometric law of large numbers. In *AISTATS*, pp. 325–332, 2013a.

Ho, J., Xie, Y., and Vemuri, B. C. On a nonlinear generalization of sparse coding and dictionary learning. In *ICML*, pp. 1480–1488, 2013b.

Huckemann, Stephan, Hotz, Thomas, et al. Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica*, pp. 1–100, 2010.

Kim, Hyunwoo J, Adluru, Nagesh, Bendlin, Barbara B, Johnson, Sterling C, Vemuri, Baba C, and Singh, Vikas. Canonical correlation analysis on riemannian manifolds and its applications. In *ECCV*, pp. 251–267. Springer, 2014a.

Kim, Hyunwoo J., Adluru, Nagesh, Collins, Maxwell D., Chung, Moo K., Bendlin, Barbara B., Johnson, Sterling C., Davidson, Richard J., and Singh, Vikas. Multivariate general linear models (MGLM) on Riemannian manifolds with applications to statistical analysis of diffusion weighted images. In *CVPR*, Columbus, Ohio, June 2014b.

Lebanon, Guy. *Riemannian geometry and statistical machine learning*. PhD thesis, 2005.

Lenglet, C., Rousson, M., Deriche, R., and Faugeras, O. Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor MRI processing. *JMIV*, 25(3):423–444, 2006.

Medvedovic, Mario and Sivaganesan, Siva. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206, 2002.

Minear, Meredith and Park, Denise C. A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, 36(4):630–633, 2004.

Moakher, Maher. A differential geometric approach to the geometric mean of SPD matrices. *SIAM Journal on Matrix Analysis and Applications*, 26(3):735–747, 2005.

Montillo, Albert and Ling, Haibin. Age regression from faces using random forests. In *ICIP*, pp. 2465–2468, 2009.

Mukhopadhyay, Saurabh and Gelfand, Alan E. Dirichlet process mixed generalized linear models. *JASA*, 92 (438):633–639, 1997.

Neal, R. MCMC using Hamiltonian dynamics. *Handbook of MCMC*, pp. 113–162, 2011.

Neal, Radford M. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

Pennec, Xavier. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006.

Porikli, F., Tuzel, O., and Meer, P. Covariance tracking using model update based on Lie algebra. In *CVPR*, pp. 728–735, 2006.

Schwartzman, Armin. *Random ellipsoids and false discovery rates: Statistics for diffusion tensor imaging data*. PhD thesis, Stanford University, 2006.

Shahbaba, B. and Neal, R. Nonlinear models using Dirichlet process mixtures. *JMLR*, 10:1829–1850, 2009.

Sommer, Stefan. Horizontal dimensionality reduction and iterated frame bundle development. In *Geometric Science of Information*, pp. 76–83. Springer, 2013.

Sommer, Stefan, Lauze, François, and Nielsen, Mads. Optimization over geodesics for exact principal geodesic analysis. *Advances in Computational Mathematics*, pp. 283–313, 2013.

Sra, S. and Hosseini, R. Geometric optimisation on positive definite matrices with application to elliptically contoured distributions. In *NIPS*, pp. 2562–2570, 2013.

Srivastava, Anuj, Jermyn, Ian, and Joshi, Shantanu. Riemannian analysis of probability density functions with applications in vision. In *CVPR*, pp. 1–8, 2007.

Straub, Julian, Chang, Jason, Freifeld, Oren, and Fisher III, John W. A Dirichlet process mixture model for spherical data. In *AISTAT*, pp. 930–938, 2015.

Xie, Yuchen, Vemuri, Baba C, et al. Statistical analysis of tensor fields. In *MICCA*, pp. 682–689. 2010.

Zhang, Zhihua, Wang, Dakan, Dai, Guang, and Jordan, Michael I. Matrix-variate Dirichlet process priors with applications. *Bayesian Analysis*, 9:259–289, 2014.

Zhu, Hongtu, Chen, Yasheng, Ibrahim, Joseph G, et al. Intrinsic regression models for positive-definite matrices with applications to DTI. *JASA*, 104(487), 2009.

Zhu, Jun, Chen, Ning, and Xing, Eric P. Infinite svm: a Dirichlet process mixture of large-margin kernel machines. In *ICML*, pp. 617–624, 2011.

Manifold-valued Dirichlet Process
(*Supplementary Material*)

1. Introduction

We provide the proof of Lemma 4.1. in the main paper. Additional discussion with specific forms of candidate distributions for priors and their derivatives are given. We also present more details of the implementation.

2. Proof of Lemma 4.1

Lemma 4.1. *Let $(B, V) \in SPD(n) \times Sym(n)$ be a sample drawn using the expression in Eq. (12) , then V is Normally distributed with respect to a GL-invariant metric at the tangent space $T_B\mathcal{M}$ at B. In that, for each B, the probability density function of V is proportional to $\exp(-\frac{1}{2}\|V\|_B^2))$ at $T_B\mathcal{M}$, when $\mu_V = 0$.*

Proof. We will derive an expression for the density. By inspection, we have

$$\int\int f(B;\mu_B,\sigma_B^2)f(V;\mu_V,B)dVdB = \int f(B;\mu_B,\sigma_B^2)\left[\int f(V;\mu_V,B)dV\right]dB = 1 \tag{1}$$

Let $q = n(n+1)/2$. Given the density functions Eq. (7) and (8) in the main paper, the density of the proposed distribution $f_{\text{SPD,Sym}}((B,V)|\mu_B,\sigma_B^2,\mu_V)$ is the product of density functions given by

$$\begin{aligned} f((B,V)|\mu_B,\sigma_B^2,\mu_V) &= \frac{1}{Z(\mu_B,\sigma_B^2)}\exp\left(-\frac{1}{2\sigma_B^2}\mathrm{d}(B,\mu_B)^2\right)\frac{1}{(2\pi)^{q/2}|B|^{(n+1)/2}}\exp\left(-\frac{1}{2}\mathrm{tr}[((V-\mu_V)B^{-1})^2]\right) \\ &= \frac{1}{Z(\mu_B,\sigma_B^2)}\exp\left(-\frac{1}{2\sigma_B^2}\mathrm{d}(B,\mu_B)^2\right)\frac{1}{(2\pi)^{q/2}|B|^{(n+1)/2}}\exp\left(-\frac{1}{2}\|V-\mu_V\|_B^2\right) \\ &= f(B;\mu_B,\sigma_B^2)\frac{1}{(2\pi)^{q/2}|B|^{(n+1)/2}}\exp\left(-\frac{1}{2}\|V\|_B^2\right), \text{ when } \mu_V = 0 \in \mathrm{Sym}(n) \end{aligned} \tag{2}$$

where the inner product of $U, V \in T_B\mathcal{M}$ is $\langle U, V\rangle_B = \mathrm{tr}(B^{-1/2}UB^{-1}VB^{-1/2})$. □

3. Distributions and their derivatives for DP-MGLM on SPD manifolds

3.1. Prior distributions for SPD matrix

Wishart distribution over $n \times n$ SPD X with V a (fixed) positive definite matrix and df degrees of freedom.

$$\begin{aligned} f(X|V,df) &= \frac{1}{2^{\frac{n\times df}{2}}|V|^{\frac{df}{2}}\Gamma_n(\frac{df}{2})}|X|^{\frac{df-n-1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}(V^{-1}X)\right) \\ \log f(X|V,df) &= -\log Z(V,df) + \frac{df-n-1}{2}\log\det(X) - \frac{1}{2}\mathrm{tr}(V^{-1}X) \\ \frac{\partial}{\partial X}\log f(X|V,df) &= \frac{df-n-1}{2}X^{-1} - \frac{1}{2}V^{-1} \end{aligned} \tag{3}$$

Since X is a symmetric positive definite matrix, we have $\frac{\partial}{\partial X}\log\det(X) = X^{-1}$, see A.4.1 in (Boyd & Vandenberghe, 2004). Also we know that $\frac{\partial}{\partial X}\mathrm{tr}(AX^T) = A$ (Petersen & Pedersen, 2012). So we can ensure that the derivative is a symmetric matrix.

Generalized normal distribution over $n \times n$ SPD X with (fixed) mean positive definite matrix M and $\sigma \in \mathbf{R}$.

$$\begin{aligned} f(X|M,\sigma) &= \frac{1}{Z(M,\sigma)} \exp\left(-\frac{1}{2\sigma^2}\mathrm{d}(X,M)^2\right) \\ \log f(X|M,\sigma) &= -\log Z(\sigma) - \frac{1}{2\sigma^2}\mathrm{d}(X,M)^2 \\ \nabla_X \log f(X|V,n) &= \frac{\mathrm{Log}(X,M)}{\sigma^2} \end{aligned} \tag{4}$$

The second equality holds since Z is constant w.r.t M on SPD manifolds. Note that this derivative is in $T_p\mathcal{M}$.

3.2. Prior distributions for symmetric matrix

Normal distribution 1. (definition 3.1.2 in (Schwartzman, 2006)) over $X \in Sym(n)$ with mean matrix 0 and covariance matrix I with respect to Lebesque measure on $\mathbf{R}^q$ is given by

$$f(X) = \frac{1}{(2\pi)^{q/2}} \exp(-\frac{1}{2}\mathrm{tr}(X^2)) \tag{5}$$

where $q = n(n+1)/2$. This is equivalent to multivariate normal distribution with the appropriate reshaping function. For example, for $p = 3$, Z is constructed as

$$Z = \begin{pmatrix} N(0,1) & N(0,1/2) & N(0,1/2) \\ * & N(0,1) & N(0,1/2) \\ * & * & N(0,1) \end{pmatrix} \tag{6}$$

Normal distribution 2. (definition 3.1.3 in (Schwartzman, 2006)) over $X \in Sym(p)$ with mean matrix M and covariance matrix Σ

$$\begin{aligned} f(X;M,\Sigma) &= \frac{1}{(2\pi)^{q/2}|\Sigma|^{(p+1)/2}} \exp\left(-\frac{1}{2}\mathrm{tr}((X-M)\Sigma^{-1})^2\right) \\ \log f(X|M,\Sigma) &= -\log Z(\Sigma) - \frac{1}{2}\mathrm{tr}[((X-M)\Sigma^{-1})^2] \\ \frac{\partial}{\partial X}\log f(X|M,\Sigma) &= -\frac{1}{2}\frac{\partial}{\partial X}\mathrm{tr}\left[(X-M)\Sigma^{-1}(X-M)\Sigma^{-1}\right] \\ &= -\frac{1}{2}\frac{\partial}{\partial X}\left[\mathrm{tr}(X\Sigma^{-1}X\Sigma^{-1}) - 2\mathrm{tr}(\Sigma^{-1}M\Sigma^{-1}X) + \mathrm{tr}(M\Sigma^{-1}M\Sigma^{-1})\right] \\ &= \Sigma^{-1}(M-X)\Sigma^{-1} \end{aligned} \tag{7}$$

The last equality is obtained by $\frac{\partial}{\partial X}\mathrm{tr}(AX^T) = A$ and $\frac{\partial}{\partial X}\mathrm{tr}(AXBX) = A^TX^TB^T + B^TX^TA^T$.

We showed above few candidates for prior distributions over B and $\boldsymbol{V}$. In our implementation, we used the (8) and (9) in our main paper. As we mentioned in the main paper, (12) can be used for Algorithm 1 in the main paper.

The derivative of (11) in the main paper is obtained by the derivative of MGLM (Kim et al., 2014) and the derivatives of distributions f_{SPD} and f_{Sym}.

4. Intrinsic mean and prediction

For prediction, we average the Monte Carlo samples of the expectation conditioned on θ. In general, there is no notion of addition on manifolds. So instead of arithmetic mean, by computing the intrinsic mean of MCMC realizations on manifolds, the prediction can be obtained as we discussed in section 4. in our main paper. In the main paper, we call the intrinsic mean "Fréchet mean". Also, it is the same as the Karcher mean for a geodesically complete manifold SPD(n). For more discussion about intrinsic means, we refer (Afsari, 2011). The Karcher mean is given by

$$\bar{y} = \arg\min_{y\in\mathcal{M}} \sum_{i=1}^{N} w_i d^2(y, y_i),$$

where the w_i is weight of data $y_i \in \mathcal{M}$.

Karcher mean is obtained by Algorithm 1, where α denotes the step size ($\alpha = 1$ was used).

Algorithm 1 Karcher mean

Input: $y_1, \ldots, y_N \in \mathcal{M}$, α
Output: $\bar{y} \in \mathcal{M}$
$\bar{y}_0 = y_1$
while $\| \sum_{i=1}^{N} \mathrm{Log}(\bar{y}_k, y_i) \| > \epsilon$ **do**
  $\Delta\bar{y} = \frac{\alpha}{N} \sum_{i=1}^{N} \mathrm{Log}(\bar{y}_k, y_i)$
  $\bar{y}_{k+1} = \mathrm{Exp}(\bar{y}_k, \Delta\bar{y})$
end while

5. Comparison with RMHMC

In our main paper, we briefly discussed RMHMC. We compare our algorithm with RMHMC. First, we revisit some detail of HMC algorithm. HMC algorithm can be implemented by Leapfrog (or Strömer-Verlet) integrator (Duane et al., 1987).

$$\begin{aligned} p(t+\epsilon/2) &= p(t) - (\epsilon/2)\nabla_q U(q(t)) \\ q(t+\epsilon) &= q(t) + \epsilon M^{-1} p(t+\epsilon/2) \\ p(t+\epsilon) &= p(t+\epsilon/2) - (\epsilon/2)\nabla_q U(q(t+\epsilon)) \end{aligned} \tag{8}$$

HMC algorithm requires a prefixed number of step (L) and predetermined step size ϵ for leapfrog integrator. In the special case where only one deterministic step is used, it is called the Langevin algorithm, which is a discrete time approximation to the Langevin diffusion process (Ishwaran, 1999). The performance of HMC is dependent on L and ϵ. It is known that when L is reasonably large, the benefits of hybrid Monte Carlo can be fully exploited (Neal, 1995). RMHMC elegantly addresses this by providing automated adaptation mechanisms.

RMHMC uses the sum of expected Fisher information matrix and the negative Hessian of the log-prior as the metric tensor $G(\theta)$ instead of M in equation 7 in our main paper. In other words, RMHMC uses $\|\dot{\theta}\|_\theta^2 = \dot{\theta}^T G(\theta)\dot{\theta} = p^T G(\theta)^{-1} p$, where $p = M\dot{\theta}$. So the kinetic energy is naturally defined by the half of the square norm of each $\dot{\theta}$ (Girolami & Calderhead, 2011). It adopts to the local geometry of joint probability. It allows position-specific distance metric which may yield more effective transitions within MCMC scheme. However that algorithm is not specifically designed for the manifold-valued parameters. It updates parameters by the vector-space operations which is not directly applicable for manifold values. For more details, we refer the reader to (Girolami & Calderhead, 2011).

6. Hamiltonian function for DP-MGLM

For our DP-MGLM, the Hamiltonian function is written as

$$H(B, \boldsymbol{V}, \dot{B}, \dot{\boldsymbol{V}}) = U(B, \boldsymbol{V}) + K(\dot{B}, \dot{\boldsymbol{V}}) \tag{9}$$

The potential function is given as

$$U(B, \boldsymbol{V}) := \frac{1}{\sigma^2} E(B, \boldsymbol{V}) - \log f_{\mathrm{SPD}}(B) - \log f_{\mathrm{Sym}}(\boldsymbol{V}) \tag{10}$$

where $E(B, \boldsymbol{V}) := \frac{1}{2} \sum_i d(y_i, \hat{y}_i)^2$. The Kinetic energy of DP-MGLM is defined by

$$K(\dot{B}, \dot{\boldsymbol{V}}) := \frac{1}{2}\|\dot{B}\|_B + \frac{1}{2} \sum_{j=1}^{d} \|\dot{\boldsymbol{V}}^j\|_B \tag{11}$$

where the covariate is in $\mathbf{R}^d$.

References

Afsari, Bijan. Riemannian L^p center of mass: Existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011.

Boyd, Stephen P and Vandenberghe, Lieven. *Convex optimization*. Cambridge university press, 2004.

Duane, Simon, Kennedy, Anthony D, Pendleton, Brian J, and Roweth, Duncan. Hybrid monte carlo. *Physics letters B*, 195 (2):216–222, 1987.

Girolami, Mark and Calderhead, Ben. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Ishwaran, Hemant. Applications of hybrid monte carlo to bayesian generalized linear models: Quasicomplete separation and neural networks. *Journal of Computational and Graphical Statistics*, 8(4):779–799, 1999.

Kim, Hyunwoo J., Adluru, Nagesh, Collins, Maxwell D., Chung, Moo K., Bendlin, Barbara B., Johnson, Sterling C., Davidson, Richard J., and Singh, Vikas. Multivariate general linear models (MGLM) on Riemannian manifolds with applications to statistical analysis of diffusion weighted images. In *CVPR*, Columbus, Ohio, June 2014.

Neal, Radford M. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, 1995.

Petersen, K. B. and Pedersen, M. S. The matrix cookbook, Nov 2012.

Schwartzman, Armin. *Random ellipsoids and false discovery rates: Statistics for diffusion tensor imaging data*. PhD thesis, Stanford University, 2006.