SAND2015-1440J

Bayesian Karhunen-Loève Expansions: A random subspace approach

Kenny Chowdhary[a], Habib N. Najm[a]

[a] *Sandia National Laboratories, Livermore, CA*

Abstract

One of the most widely-used statistical procedures for dimensionality reduction of high dimensional random fields is Principal Component Analysis (PCA), which is based on the Karhunen-Loève expansion (KLE) of a stochastic process with finite variance. The KLE is analogous to a Fourier series expansion for a random process, where the goal is to find an orthogonal transformation for the data such that the projection of the data onto this orthogonal subspace is optimal in the L^2 sense, i.e, which minimizes the mean square error. In practice, this orthogonal transformation is determined by performing an SVD (Singular Value Decomposition) on the sample covariance matrix or on the data matrix itself. Sampling error is typically ignored when quantifying the principal components, or, equivalently, basis functions of the KLE. Furthermore, it is exacerbated when the sample size is much smaller than the dimension of the random field. In this paper, we introduce a Bayesian KLE procedure, allowing one to obtain a probabilistic model on the principal components, which can account for inaccuracies due to limited sample size. The probabilistic model is built via Bayesian inference, from which the posterior becomes the matrix Bingham density over the space of orthonormal matrices. We use a modified Gibbs sampling procedure to sample on this space and then build a probabilistic Karhunen-Loève expansions over random subspaces to obtain a set of low-dimensional surrogates of the stochastic process. We illustrate this probabilistic procedure with a finite dimensional stochastic process inspired by Brownian motion.

Keywords: Karhunen-Loève expansion, Principal Component Analysis, uncertainty quantification, Bayesian inference, matrix Bingham density, Gibbs sampling, Markov Chain Monte Carlo.

1. Introduction

The Karhunen-Loève theorem briefly states that a square integrable stochastic process can be represented by a linear combination of orthogonal functions, typically taken to be eigenfunctions of the covariance function of the stochastic process, with random coefficients [1]. To find the eigenfunctions of the covariance function, one can employ solvers for the Fredholm integral equation of the second kind [2], or, in the case of a discrete finite process, utilize Principal Component Analysis (PCA), which is a ubiquitous statistical procedure for model reduction of high-dimensional random data [3]. In this paper, we will consider the latter by only looking at discrete stochastic processes (or discretized versions of continuous stochastic processes).

The eigenfunctions in the Karhunen-Loève expansion (KLE) are known as the principal components or directions, and in many cases, since the covariance function is not known exactly, they are approximated from the sample covariance matrix. This can be done by performing an eigenvalue/ eigenvector decomposition of the sample covariance matrix or, more efficiently, by performing a singular value decomposition on the data itself. A consequence of working with sample data is that there is an underlying statistical uncertainty in computing these principal components. For different sets of realizations of the stochastic process, the sample covariance will change and thus the principal components will vary. Often this statistical variance, which is worse when the sample size is much smaller than the dimensionality of the stochastic process, is ignored, but a single set of principal components is still used. In the context of uncertainty quantification, it is important to understand and utilize the full probabilistic structure of the underlying quantity of interest - in this case being a stochastic process, which includes the principal components. Ignoring the full probabilistic structure while only using first or second order the summary statistics, e.g. means and/or variances, can lead to drastic under or over-estimates of quantities of interest [4]. More simply, it is incorrect to model the principal components deterministically. Other methods do exist, which take a slightly different route and place a distribution on the covariance matrix itself, incorporating the use of the Wishart and inverse Wishart distributions [5]. Often these methods require the sample covariance matrix to be fully ranked, and so one must explore algorithms that condition the covariance matrix, e.g., shrinkage methods, etc [6]. The method described in this paper does not require any type of conditioning or inversion of the covariance matrix directly.

In this paper, we derive the matrix Bingham density for the principal components, which is a measure on the space of orthonormal matrices, i.e., the Stiefel manifold. This gives us a probabilistic characterization of the principal directions which best explain the data. In order to derive this density, we utilize the classical minimum reconstruction procedure for determining the principal components. Once we obtain the appropriate density on the Stiefel manifold, we introduce a modified Gibbs sampling procedure, similar to the algorithm introduced by Hoff [7], to obtain samples of the principal components which can account for statistical uncertainty due to limited sample size. From this we obtain a collection of random subspaces onto which we can project our data and obtain random Karhunen-Loève expansions, i.e., low dimensional representations of our data.

Similar work can be found in topics related to probabilistic PCA algorithms and Factor Analysis, which formalize the problem in a more classical Bayesian framework, but utilize Expectation-Maximization algorithms to arrive at a single set of principal components [8]. Our work differs in that we can obtain multiple samples of the principal components on the manifold which describes the density of these principal components.

This paper is structured as follows. In Section 2 we briefly describe the mathematical setup and derivation of classical PCA. In Section 3 we derive the posterior density on the principal components in a Bayesian setting. Section 4 details the Gibbs sampling procedure used to sample from the density derived in the previous section. Section 5 describes how one can use the Bayesian framework to arrive at random KLE's. Finally, Section 6 illustrates these sampling methods on low and high-dimensional random processes.

2. Setup and Derivation of the Principal Components

Consider the space of all $m \times R$ real, orthonormal matrices, referred to as the Stiefel manifold, denoted by $\mathcal{V}_{R,m}$. The Stiefel manifold consists of matrices whose columns live on $\mathcal{S}^{m-1}$, i.e., the surface of the m-dimensional unit sphere. Given a collection of n realizations of an m-dimensional zero-mean stochastic process, x, or random field, denoted by $\{x_1, \dots, x_n\}$, $x_i \in \mathbb{R}^m$, the principal components can be derived using the following minimum reconstruction argument. Note that if x is not a zero-mean process, we can simply consider $x - \bar{x}$, where $\bar{x}$ is the exact or sample mean. Let $\Phi \in \mathbb{R}^{m \times R}$ be some element in $\mathcal{V}_{R,m}$ so that its columns form an orthonormal basis for some R-dimensional subspace in an

m-dimensional space. Consider the reconstruction or projection error of the data onto this R-dimensional subspace Φ, i.e., the mean square error:

$$\mathrm{Err}(\Phi) \quad = \quad \sum_{i=1}^{n} \|x_i - \Phi\Phi^T x_i\|^2, \tag{1}$$

where $\Phi\Phi^T \in \mathbb{R}^{m\times m}$ is the projection matrix, and $\|\cdot\|^2$ is the usual Euclidean norm. PCA attempts to find an orthonormal matrix Φ which has the least projection error,

$$\Phi^* \quad = \quad \arg\min_{\Phi\in\mathcal{V}_{R,m}} \sum_{i=1}^{n} \|x_i - \Phi\Phi^T x_i\|^2, \tag{2}$$

where the columns of $\Phi^* \in \mathbb{R}^{m\times R}$ are referred to as the R principal components. Equivalently, in the derivation of the the Karhunen-Loève expansion, one tries to find the eigenfunctions which minimize the mean square error. In practice, in order to determine the principal components, one can solve (2) analytically. Expanding the sum for Err(Φ) gives

$$\mathrm{Err}(\Phi) \quad = \quad \sum_{i=1}^{n} x_i^T x_i - x_i^T \Phi\Phi^T x_i. \tag{3}$$

Since the minimization in (2) is only over $\Phi \in \mathcal{V}_{R,m}$, the first term on the right hand side can be ignored. This results in

$$\Phi^* \quad = \quad \arg\max_{\Phi\in\mathcal{V}_{R,m}} \sum_{i=1}^{n} x_i^T \Phi\Phi^T x_i. \tag{4}$$

Let $X \in \mathbb{R}^{m\times n}$ denote the data matrix where the i^{th} column is x_i and recall that the trace of a matrix is the sum of the diagonal elements. Then, (4) can be more compactly written as

$$\Phi^* \quad = \quad \arg\max_{\Phi\in\mathcal{V}_{R,m}} \mathrm{tr}(\Phi^T(nS)\Phi), \tag{5}$$

where tr denotes the matrix trace and $S \in \mathbb{R}^{m\times m}$ is the sample covariance matrix, i.e., $S = n^{-1}XX^T$, where we have assumed x is a zero-mean stochastic process. Now, using (5) it can be shown that Φ^* is exactly the set of eigenvectors of S which have the R largest eigenvalues [3]. It can also be shown that $\mathrm{tr}(\Phi^T S\Phi)$ represents the sum of the variance along each orthonormal column in Φ. Thus, (5) shows an equivalence between the minimum reconstruction derivation of PCA (2) and the maximum variance derivation.

One can equate the optimization problem (5) to an iterative optimization procedure where we solve for Φ^* one column at a time. We briefly detail this approach since it will be illustrative in understanding how to sample matrices on the Stiefel manifold while retaining orthogonality. In order to solve (5) one column at a time, we can first find the m-dimensional orthonormal vector, $\phi_1 \in \mathbb{R}^m$ with $\phi_1^T \phi_1 = 1$, that maximizes $\phi_1^T S \phi_1$. ϕ_1 represents the direction onto which the data exhibits maximum variance, which turns out to be the eigenvector of S with the largest eigenvalue. Next, in order to find the second principal component, we seek another orthonormal vector, $\phi_2 \in \mathbb{R}^m$ such that $\phi_2^T \phi_2 = 1$ and $\phi_1^T \phi_2 = 0$, which maximizes $\phi_2^T S \phi_2$. The method of Lagrange multipliers yields the eigenvector with the second largest eigenvalue. This process can be repeated in order to determine the remaining columns of Φ^*. In the next section, when we introduce the Bayesian procedure for obtaining samples on the principal components. We will utilize this iterative Bayesian procedure to obtain orthogonal samples.

3. Bayesian approach to PCA

Let us assume that the projection error can be modeled by i.i.d white noise. That is,

$$x - \Phi\Phi^T x \quad = \quad \eta, \tag{6}$$

where $\eta \in \mathbb{R}^m$ and $\eta \sim \mathcal{N}(0, \sigma^2 I)$. Let us define the conditional density for the data, x, given the principal components, Φ, to be proportional to the projection error, i.e., $p(x|\Phi) \propto p(x - \Phi\Phi^T x)$. Then,

$$p(x|\Phi, \sigma) \quad \propto \quad \exp\left(-\frac{1}{2\sigma^2}\eta^T \eta\right), \tag{7}$$

which follows from (6). We refer to $p(x|\Phi, \sigma)$ as the likelihood of the data. One can think of the likelihood, $p(x|\Phi, \sigma)$, for a fixed x and σ, as purely a function of Φ, denoted by $g(\Phi) \doteq p(x|\Phi, \sigma)$. Then, $g(\Phi)$ can be interpreted as a penalization or cost function for Φ, which we can use to minimize the mean square error.

Let $\pi(\Phi)$ be the uniform density on the Stiefel manifold for $\Phi \in \mathcal{V}_{R,m}$. That is,

$$\pi(\Phi) \quad \propto \quad 1_{\{\Phi \in \mathcal{V}_{R,m}\}}(\Phi), \tag{8}$$

where

$$1_{\{\Phi \in \mathcal{V}_{R,m}\}}(\Phi) \quad = \quad \begin{cases} 1, & \Phi \in \mathcal{V}_{R,m} \\ 0, & \Phi \notin \mathcal{V}_{R,m} \end{cases}$$

Bayes' rule then gives

$$p(\Phi|x,\sigma) \quad \propto \quad p(x|\Phi,\sigma)\pi(\Phi), \tag{9}$$

where the proportionality is up to a constant, which only depends on the data, and the noise term, $\sigma \in \mathbb{R}^+$, is fixed for now. If we let X be the data matrix where the columns are i.i.d. realizations of x, then (9) can be written more explicitly as

$$p(\Phi|X,\sigma) \quad \propto \quad \prod_{i=1}^{n} p(x_i|\Phi,\sigma)\pi(\Phi), \tag{10}$$

Plugging (7) into (10), gives

$$p(\Phi|X,\sigma) \quad \propto \quad \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\|x_i - \Phi\Phi^T x_i\|^2\right) 1_{\{\Phi\in\mathcal{V}_{R,m}\}}(\Phi). \tag{11}$$

Finally, after some algebra, one can show that

$$p(\Phi|X,\sigma) \quad \propto \quad \mathrm{etr}(\Phi^T(nS)\Phi/2\sigma^2)1_{\{\Phi\in\mathcal{V}_{R,m}\}}(\Phi), \tag{12}$$

where etr is the exponential trace of a matrix, $S \doteq n^{-1}XX^T$, and the prior probability distribution is over the Stiefel manifold, $\mathcal{V}_{R,m}$. The density in (12) is referred to as the matrix Bingham density, which is a density on $\mathcal{S}^{m-1}$ (see [9] for more details on the normalization constant).

Sampling from this density can be tricky because samples from the matrix Bingham density must be $(m\times R)$-dimensional orthonormal matrices. To sample from (12) we propose breaking up the density into conditional densities over the individual columns of Φ, utilizing the chain rule for probability densities. Thus, we can write the posterior as

$$p(\Phi|X,\sigma) \quad \propto \quad p(\phi_1|X,\sigma)p(\phi_2|\phi_1,X,\sigma)\cdots p(\phi_R|\phi_1,\ldots,\phi_{R-1},X,\sigma), \tag{13}$$

where $\Phi = [\phi_1 \cdots \phi_R] \in \mathbb{R}^{m\times R}$, subject to orthonormality between $\phi_1,\ldots,\phi_R$. To be clear, the conditional densities in (20) are as follows:

$$\begin{aligned} p(\phi_1|X,\sigma) &\propto \exp(\phi_1^T(nS)\phi_1/2\sigma^2), \quad \|\phi_1\|_2 = 1 \qquad (14)\\ p(\phi_2|\phi_1,X,\sigma) &\propto \exp(\phi_2^T(nS)\phi_2/2\sigma^2), \quad \text{s.t. } \phi_2 \perp \phi_1, \|\phi_2\|_2 = 1\\ &\vdots\\ p(\phi_R|\phi_1,\ldots,\phi_{R-1},X,\sigma) &\propto \exp(\phi_R^T(nS)\phi_R/2\sigma^2), \quad \text{s.t. } \phi_R \perp \phi_2,\ldots,\phi_{R-1}, \|\phi_R\|_2 = 1. \end{aligned}$$

For each conditional density we will use a Gibbs sampling procedure. Each conditional density is not independent of its predecessors so some maneuvering must be taken in order to sample over the correct orthogonal space, which is explained in the next section. In short, we will first show how to sample $p(\phi_1|X,\sigma)$, then use a simple linear transformation using the left null space of ϕ_1 to generate samples from $p(\phi_1|\phi_2,X,\sigma)$, and so on.

4. Sampling the Posterior Density

In this section, we will show how to sample from (20) where each conditional density is defined by (14) using Hoff's Gibbs sampling algorithm introduced in [7]. Note that the algorithm presented here is a slight modification of the aforementioned algorithm, which will give us a set of ordered vectors, analogous to retrieving principal components in order of decreasing variance. Without this modification, we loose the ordering of the principal components in order of decreasing variance. Nonetheless, both algorithms will produce principal components from the matrix Bingham density defined in (12).

First, in Section 1 we will describe how to sample the vector Bingham density, i.e. (12) for $R = 1$. Then, in Section 4.2 we will introduce a fairly simple trick to sample from the full matrix Bingham density using a left null space transformation.

4.1. Sampling the vector Bingham density

Consider the vector Bingham density on the m-dimensional sphere with respect to the uniform distribution over the unit sphere,

$$p(\phi|A) \quad \propto \quad \exp(\phi^T A\phi)1_{\{\phi\in\mathcal{V}_{1,m}\}}(\phi). \tag{15}$$

Without loss of generality, let us assume that $A \in \mathbb{R}^{m\times m}$ is symmetric. For the purposes of this paper, A is in fact proportional to the sample covariance matrix, i.e., $A = n\sigma^{-2}S/2$, which is always semi-positive definite. Since this matrix A is symmetric, it always admits an eigenvector/ eigenvalue decomposition

$$A \quad = \quad U\Lambda U^T, \tag{16}$$

where $U \in \mathbb{R}^{m\times m}$ is unitary and $\Lambda \in \mathbb{R}^{m\times m}$ is diagonal and has non-negative values. If we transform our density under the isometric mapping U, letting $y = U^T\phi$, then we can write (15) under the new variable y as

$$p(y|E,\Lambda) \quad \propto \quad \exp\left(\sum_{i=1}^{m} \lambda_i y_i^2\right) 1_{\{y\in\mathcal{V}_{1,m}\}}(y), \tag{17}$$

where λ_i's are the eigenvalues of S. Note that since the change of variables is given by a linear, unitary mapping U, the determinant of the Jacobian is 1. Furthermore, we can write the probability density function for $s(y) \doteq 1_{\{y\in\mathcal{V}_{1,m}\}}(y)$ explicitly as

$$s(y) \quad \propto \quad s(y) = \left(1 - \sum_{i=1}^{m-1} y_i^2\right)^{-1/2}, \quad \text{s.t.} y_m^2 = 1 - \sum_{i=1}^{m-1} y_i^2. \tag{18}$$

Note that the uniform density over the sphere $\mathcal{S}^{\Updownarrow-\infty}$ only has $m-1$ degrees of freedom due to the normality constraint. Thus, the density in (17) can be explicitly written as

$$p(y|E,\Lambda) \quad \propto \quad \exp\left(\sum_{i=1}^{m} \lambda_i y_i^2\right)\left(1 - \sum_{i=1}^{m-1} y_i^2\right)^{-1/2}, \quad \text{s.t.} \quad y_m^2 = 1 - \sum_{i=1}^{m-1} y_i. \tag{19}$$

A Gibbs sampling procedure can be performed to sample from (19), which means that we need to derive the one-dimensional conditional densities for (19). Hoff suggests performing a simple transformation before deriving the conditional densities for $p(y|E,\Lambda)$, in order to improve the mixing of the Markov Chain [7, 10]. We briefly go over the suggested transformation.

In a straightforward Gibbs sampling procedure for (19), we need to sample from $p(y_i|y_{-i},E,\Lambda)$ where

$$y_{-i} \quad = \quad (y_1,\dots y_{i-1}, y_{i+1},\dots y_m) \in \mathbb{R}^{m-1}.$$

This conditional density is hard to sample from in practice, so we perform the following transformation. Let $\theta \doteq y_i^2$ and define

$$q \quad \doteq \quad \frac{1}{1-\theta}(y_1^2,\dots y_{i-1}^2, y_{i+1}^2,\dots y_m^2),$$

so that $\{y_i^2, y_{-i}^2\} = \{\theta, (1-\theta)q_{-i}\}$. Then, after some calculation, Hoff shows that the conditional density, $p(y_i|y_{-i},E,\Lambda)$, in terms of $\theta \doteq y_i^2$

$$p(\theta|q_{-i},E,\Lambda) \quad = \quad \exp(\theta[\lambda_i - q_{-i}^T\lambda_{-i}])\theta^{1/2}(1-\theta)^{(m-3)/2}, \tag{20}$$

where λ_i is the i^{th} diagonal element of Λ. Since we are making a change of variables, we will also need the determinant of the Jacobian which is given by the following:

$$\left|\tfrac{d\theta}{dy_i}\right| \quad = \quad 2|y_i| \quad = \quad 2\theta^{1/2} \quad \left|\tfrac{dq_j}{dy_j}\right| \quad = \quad 2\tfrac{|y_j|}{1-y_{j2}} \quad = \quad 2q_j^{1/2}(1-\theta)^{1/2,} \quad ,j \neq i. \tag{21}$$

In order to sample (20) over $\theta \in (0,1)$, we can proceed in either of two directions. The first is the most straightforward, but not the most efficient. In the first approach, we can either build the inverse cumulative distribution function (CDF) via interpolation and then sample based on the inverse CDF method, or evaluate $p(\theta|q_{-i}, E, \Lambda)$ at a set of uniform grid points, weight them according to their PDF value, and then draw samples from this discrete density. A more efficient alternative is to use a rejection sampler. The target density (20) is of the form

$$p(\theta|q_{-i}, E, \Lambda) \quad \propto \quad \theta^{-1/2}(1-\theta)^{k-1}e^{\theta a}, \tag{22}$$

where $k = (m-3)/2$ and $a = \lambda_i - q_{-i}^T\lambda_{-i}$. In order to obtain a proper rejection sampler, we need a proposal density $f(\theta)$, also known as an envelope function, s.t $Mf(\theta) > p(\theta|q_{-i}, E, \Lambda)$ for some fixed constant $M > 0$. Since (22) is very similar to a beta density, Hoff proposes using a beta$(1/2, 1 + k \wedge [(k-a) \vee -1/2])$ envelope which works well for many choices of k and a [7]. Note that choosing the constant M is not trivial in practice. Please see the companion R implementation to [7] for a proper choice of M.

Under the Gibbs sampling approach the above procedure generates a Markov chain in $\{y_1^2, \ldots, y_m^2\}$ with a stationary distribution equal to $p(y_1^2, \ldots, y_m^2|E, \Lambda)$ [7]. The sign of y_i does not actually effect the density so it can be randomly assigned. The algorithm is summarized below.

```
**Input**: $A$, $\phi^{(0)} = e_1$, where $e_1$ is the canonical unit vector
**Output**: Gibbs sample from $p(\phi|A)$
Let $A = E^T \Lambda E$ and set $y = E^T \phi$;
**for** $i = 1, \dots, m$, *in random order* **do**
    Set $\{q_1, \dots, q_m\} = \{y_1/(1-y_1)^2, \dots, y_m/(1-y_m)^2\}$ ;
    Sample $\theta \in (0,1)$ from $p(\theta|q_{-i}, E, \Lambda) \propto \theta^{-1/2}(1-\theta)^k e^{\theta a}$ with $k = (m-3)/2$ and $a = \lambda_i - q_{-i}^T \lambda_{-i}$;
    Sample $s_i$ on $\{-1, +1\}$ using a binomial with $p = .5$;
    Transform $\theta$ back to $y$: $y_i = s_i \theta^{1/2}$ and $y_j = (1-\theta) q_j$ for $j \neq i$;
**end**
Transform $y$ back to $\phi$: $\phi = Ey$ ;
Add new sample to the Markov chain: $\phi^{(1)} = \phi$
```

Algorithm 1: Gibbs sampler for the vector Bingham density.

Algorithm 1 can be repeated to obtain a Markov chain of samples from the vector Bingham density. In practice, the mixing seems to work rather quickly, usually with a burn-in of only about five to ten samples. Figure 1 shows samples from a three-dimensional vector Bingham density using this Gibbs sampling procedure outlined above. Notice that the distribution is bi-modal because the vector Bingham density is antipodally symmetric, i.e., the density is invariant under a scalar multiplication by -1. In the next section, we explore how one obtains samples from the individual conditional densities in (14), i.e., a set of orthogonal vectors.

4.2. Sampling the matrix Bingham Density

In the previous section we discussed how to sample from the vector Bingham density. Now, we will explain how to sample from the conditional densities given in (14). In general, suppose we want to sample from the vector Bingham density (15) subject to $\phi \perp \Psi$ where $\Psi \in \mathbb{R}^{m \times k}$ is some set of k orthonormal columns, with $k < R$. This density can be written as

$$p(\phi|A, \phi \perp \Psi) \quad \propto \quad \exp(\phi^T A \phi) 1_{\{\phi \in \mathcal{V}_{1,m}, \phi \perp \Psi\}}(\phi) \tag{23}$$

In short, to sample from this density, we will write ϕ as a function of the left null space of Ψ

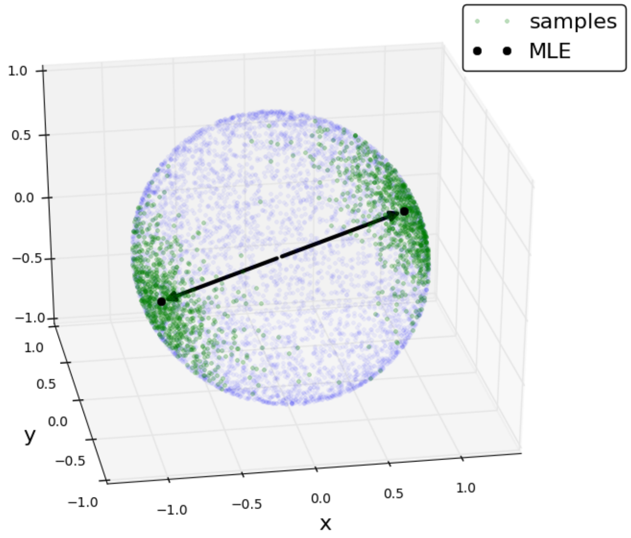

Figure 1: Samples of a three-dimensional vector Bingham density. Arrows indicate the direction of the principal eigenvector. Points in green indicate the samples from the vector Bingham density, while points in blue indicate samples from the Stiefel manifold, illustrating the surface of the sphere, $\mathcal{S}^2$.

and show that under this linear, isometric mapping, the density is again a vector Bingham density.

To show this, let $N \in \mathbb{R}^{m \times m-k}$ be an orthonormal basis for the left null space of $\Psi \in \mathbb{R}^{m \times k}$. That is, for $l \in \mathbb{R}^m$ in the span of the columns of N, we have $l^T\Psi = 0$. Now, if ϕ must be orthogonal to Ψ, then ϕ must be a linear combination of the columns of N, i.e., $\phi = Nz$ for some $z \in \mathcal{S}^{m-k}$, i.e., a vector on the $(m-k)$-dimensional unit sphere. Then we can perform the following change of variables for (23)

$$\begin{aligned} p(\phi = Nz|A) &\propto \exp((Nz)^T A (Nz)) 1_{\{z \in \mathcal{V}_{1,m-k}\}}(z) \\ &= \exp(z^T \tilde{A} z) 1_{\{z \in \mathcal{V}_{1,m-k}\}}(z), \qquad (24) \end{aligned}$$

where $\tilde{A} \doteq N^T A N$. Thus, (24) is again a vector Bingham density. This means we can use Algorithm 1 again. In particular, we can use Algorithm 1 in an iterative fashion to sample from the conditional densities defined in (14), with A defined as $n\sigma^{-2}S/2$.

In summary, consider matrix Bingham density

$$p(\Phi|A) \propto \exp(\Phi^T A \Phi) 1_{\{\Phi \in \mathcal{V}_{m,R}\}}(\Phi), \qquad (25)$$

subject to $\Phi \in \mathcal{V}_{R.m}$, where $\Phi = [\phi_1, \ldots, \phi_R]$. We can now sample from this density using the following algorithm.

```
**Input**: $A$, $\Phi^{(0)} = I$, where $I$ is the $m \times R$ identity matrix.
**Output**: Set of Gibbs samples from $p(\Phi|A)$
Using Algorithm 1 generate $M$ samples from $p(\phi|A)$, denoted by $\{\phi_1^{(1)}, \cdots, \phi_1^{(M)}\}$;
for $r = 2, \ldots, m$ do
    for $j = 1, \ldots, M$ do
        Let $\Psi_j \doteq [\phi_1^{(j)} \cdots \phi_r^{(j)}]$, s.t. $\Psi \in \mathcal{V}_{R,r-1}$;
        Compute the Null space of $\Psi_j$, $N_j \in \mathbb{R}^{m\times m-r+1}$;
        Set $\tilde{A}_j = N_j^T A N_j$;
        Use Algorithm (1) to get sample of $z \sim \exp(z^T \tilde{A} z)$;
        Transform $z$ to get sample of $\phi_r$: $\phi_r^{(j)} = N_j z$;
    end
end
```

Algorithm 2: Gibbs sampler for the matrix Bingham density using conditional densities.

Since the Gibbs sampling procedure for the vector Bingham density converges to the unique stationary distribution, giving us exact samples from each conditional density in (14), Algorithm (2) will converge to the target matrix Bingham density as well (see [11, 7] for more details about convergence of the Gibbs sampler).

In summary, Algorithm (2) obtains samples of the matrix Bingham density by breaking up the joint density into conditional densities, using the chain rule, for which each conditional density can be sampled via the vector Bingham algorithm. This allows flexibility in obtaining the orthonormal vectors Φ in two ways. First, by obtaining samples of the conditional densities, one can choose to increase R, i.e., the number of basis elements of Φ, adaptively if more basis terms are needed. Secondly, if one is probabilistically certain of a particular subset of the columns of Φ, i.e., one might know the first few columns exactly, then Algorithm (2) can be used to sample over the space orthogonal to the know subspace. This allows one to obtain a probability distribution on select columns of Φ only.

Hoff proposes a more classical Gibbs procedure, which is perfectly valid for a fixed choice of R [7]. Essentially, Hoff's Gibbs sampling algorithm runs over each individual column of Φ, while fixing all other columns simultaneously. He shows that this algorithm indeed generates a reversible, irreducible, aperiodic Markov chain for $R < m$. The difference between the algorithms introduced in this paper and Hoff's algorithm is that the principal components obtained in the latter formulation will not necessarily be ordered in decreasing projection error. In contrast, Algorithm (2) will indeed give us a set of ordered principal directions. This is more consistent with computational techniques involving SVD used to compute principal component vectors, which return the orthonormal vectors in order of decreasing variance. This ordering is useful in determining which basis terms to keep in the Karhunen-Loève expansion. Typically, the basis terms are chosen so that the cumulative energy, given by $\Sigma_{i=1}^{j}\lambda_i^2/\Sigma_{i=1}^{m}\lambda_i^2$ where $j \leq m$ is the number of basis terms retained and λ_i's are the variances along the respective ϕ_i directions, is above a prescribed threshold, i.e. 90%. In order to illustrate this point, Figure 2 shows samples of a three-dimensional matrix Bingham density with $m = 3$ and $R = 2$. Both algorithms provide samples from the same density, but Algorithm (2) provides samples in the *correct* ordering. Note that both algorithms provide subspaces which are more-or-less equivalent under rotation.

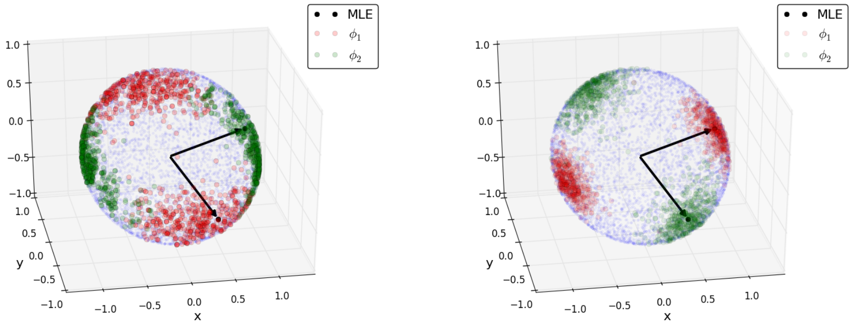

Figure 2: The samples on the left provide samples around the vectors determined by PCA, while Hoff's algorithm on the right provide equivalent samples, but under different rotations.

5. Random Karhunen-Loève Expansions

The Karhunen-Loève theorem states that one can represent any square integrable stochastic process as a linear combination of deterministic, orthogonal vectors, whose coefficients are uncorrelated, but not necessarily independent, random variables [12, 13]. To be precise, for a zero-mean, m-dimensional stochastic process $W \in \mathbb{R}^m$, and any set of m orthonormal basis functions $\Psi = \{\psi_1, \ldots, \psi_m\}$ where $\psi_i \in \mathbb{R}^m$, one can write

$$W \quad \overset{d}{=} \quad \sum_{i=1}^{m} \alpha_i \psi_i, \tag{26}$$

where $\{\alpha_i = \langle W, \psi_i\rangle\}_{i=1}^m$ constitute a set of uncorrelated random variables induced by the projection of W onto each ψ_i, and equality is given in distribution. Typically, one chooses the basis functions, ψ_i, to be the eigenvectors of the covariance matrix of W (or the sample covariance matrix if one is only given samples of the stochastic process), where λ_i are the corresponding eigenvalues. In this way, the expansion (26) can be optimally chosen in the L^2 sense if one chooses eigenvectors in decreasing order of eigenvalues, i.e. ψ_i are ordered such that $\lambda_1 \geq \cdots \geq \lambda_m$. In particular, Parseval's theorem gives

$$\mathbb{E}\left[\left(W - \sum_{i=1}^{R} \alpha_i \psi_i\right)^2\right] \quad = \quad \sum_{i=k+1}^{m} \lambda_i^2,$$

where $R \leq m$. Thus, for $R < m$, if the residual eigenvalues are small, the k-dimensional approximation, $\sum_{i=1}^{R} \alpha_i \psi_i$, to W may be a sufficient representation of the original stochastic process, at least in distribution.

In the present context, we no longer have a single set of principal components, but rather a set of M random subspaces $\{\Phi^{(1)}, \ldots, \Phi^{(M)}\}$, where $\Phi^{(i)} \in \mathcal{V}_{R,m}$ are sampled from (5) via the Gibbs sampling procedure outlined in Algorithms (1) and (2). Thus, each set of R orthonormal basis functions, $\Phi^{(i)}$, admits M Karhunen-Loeve approximates to the stochastic process W:

$$\left\{ \sum_{j=1}^{R} \alpha_j^{(i)} \phi_j^{(i)} \right\}_{i=1}^{M}, \tag{27}$$

where $\Phi^{(i)} = \{\phi_1^{(i)}, \ldots, \phi_R^{(i)}\}$ contains R orthonormal m-dimensional columns. For each fixed i, the distribution on coefficients, $\{\alpha_1^{(i)}, \ldots, \alpha_R^{(i)}\}$ can be determined by projecting samples of W onto each basis vector $\phi_j^{(i)}$. That is, samples of $\{\alpha_1^{(i)}, \ldots, \alpha_R^{(i)}\}$ are given by

$$\{\langle \phi_1^{(i)}, x_j\rangle, \ldots, \langle \phi_R^{(i)}, x_j\rangle\}_{j=1}^{M}. \tag{28}$$

It is important to note that even though the individual $\alpha_j^{(i)}$'s are uncorrelated random variables, they may not be independent. Thus, the full joint density must be determined in most cases, unless independence is known. For a Gaussian process W, uncorrelated, in fact, implies independence, and, moreover, one can show that the $\alpha_j^{(i)}$'s are independent, zero-mean normal random variables with variance $\mathbb{E}[\langle \phi_j^{(i)}, W\rangle^2]$. In all other cases, in order to obtain the full joint density on $\{\alpha_1^{(i)}, \ldots, \alpha_R^{(i)}\}$ from the projection samples (28), one could use kernel density estimation (KDE) along with the inverse Rosenblatt transformation, to produce a polynomial chaos expansion (PCE) for the $\alpha_j^{(i)}$'s [14, 15]. A thorough discussion of alternative methods for inferring the full joint density on the coefficients goes beyond the scope of this paper, so we refer the readers to the references.

6. Examples

In this section, we illustrate the approaches introduced in the previous sections on a discretized version of a continuous time, square integrable stochastic process, given by

$$W_t \quad = \quad \sum_{k=1}^{3} \frac{\xi_k}{\sqrt{\pi\left(k-\frac{1}{2}\right)}} \sqrt{2} \sin\left(\left(k-\frac{1}{2}\right)\pi t\right), \tag{29}$$

for $t \in [0,1]$, $\xi_k \sim \mathcal{N}(0,1)$, where $\left\{\sqrt{2}\sin\left(\left(k-\frac{1}{2}\right)\pi t\right)\right\}_{k=1}^{3}$ are the first three eigenfunctions of the covariance function for standard Brownian motion, and $\left\{1/\sqrt{\pi\left(k-\frac{1}{2}\right)}\right\}_{k=1}^{3}$ are the corresponding eigenvalues. We discretize in time to obtain an m-dimensional approximation, $W_m = (w_1, \ldots, w_m) \in \mathbb{R}^m$, to (29), where

$$w_j \quad = \quad \sum_{k=1}^{3} \frac{\xi_k}{\sqrt{\pi\left(k-\frac{1}{2}\right)}} \sqrt{2} \sin\left(\left(k-\frac{1}{2}\right)\pi \frac{j}{N}\right), \tag{30}$$

for $j = 1, \ldots, m$. Figure 3 shows the the first three PCA modes, in absolute value, of W_m with $m = 100$, alongside realizations of this finite-dimensional stochastic process.

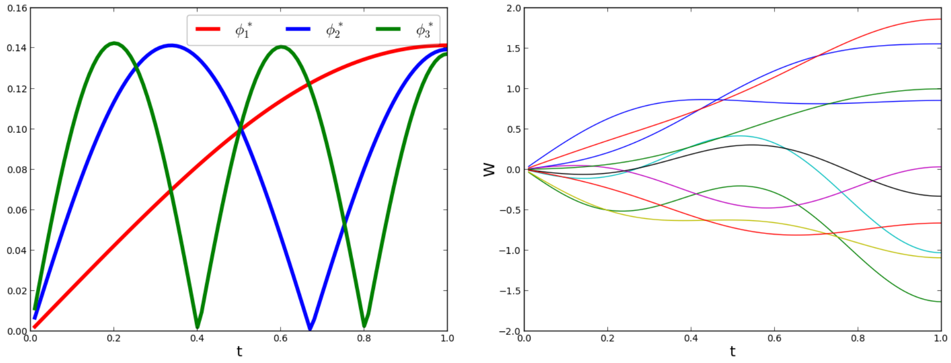

Figure 3: (left) First three PCA modes for W_m. Absolute value is shown since the principal vectors are invariant under scalar multiplication by -1. (right) Realizations of the stochastic process W_m.

6.1. Samples from the matrix Bingham density

Suppose we are given n realizations of our stochastic process, where $n \ll m$ (this is not a requirement but is indicative of a scenario in which we have very few samples of a discrete random process relative to the dimensionality). Assuming a fixed noise parameter value, σ, we use the Bayesian KLE approach outlined in Algorithm (2) to obtain samples of the principal components. Figure 4 shows the spread of samples from the matrix Bingham density (12), illustrated by the shaded regions, color coded for each principal mode. Figure 4 also displays a single realization from the matrix Bingham density, which lives on the Stiefel manifold $\mathcal{V}_{3,100}$.

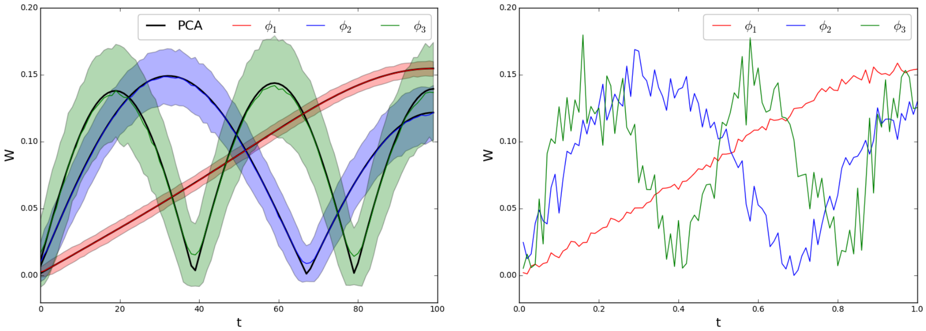

Figure 4: (left) Shaded regions represent ±2σ error bars for principal components sampled from the matrix Bingham density. Black lines represent the PCA modes, which are the eigenvectors of the sample covariance which are computed from performing an SVD. (right) One sample from the matrix Bingham density. In this example, $n = 25$, $m = 100$, and $\sigma = .1$. Again, absolute values are shown since the density is antipodally symmetric.

If the noise parameter, σ, is not known, the Gibbs sampling procedure makes it fairly easy to obtain a posterior on the noise, given an appropriate choice of a prior. In fact, by choosing the conjugate prior on $1/\sigma^2$ to be gamma(α, β), then

$$p(1/\sigma^2|\Phi, X) \quad \sim \quad \text{gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n} \|x_i - \Phi\Phi^T x_i\|^2\right), \tag{31}$$

where $p(\Phi|X, \sigma)$ is defined in (12). Figure 5 shows samples of the matrix Bingham density for $n = 25$ when we impart a gamma prior on $1/\sigma^2$. Note the similarity to the results shown in Figure 4. In general, if one does not choose a conjugate prior on σ, one can perform a Metropolis-Hastings accept/ reject scheme for σ, for every sample of Φ, and then iterate between the two. This approach is known as Metropolis-within-Gibbs sampling or block MCMC [16].

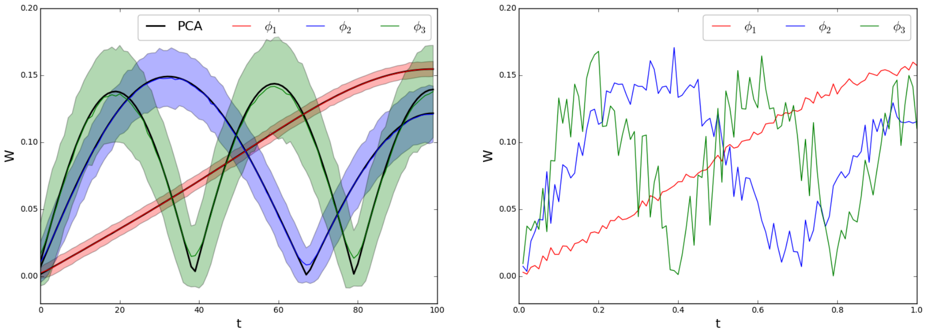

Figure 5: (left) Shaded regions represent ±2σ error bars at each point in time for samples from the matrix Bingham density. Black lines represent the PCA modes, which are the eigenvectors of the sample covariance. (right) Samples from the matrix Bingham density. In this example, $n = 25$, $m = 100$, and the prior on $1/\sigma^2$ is given by beta$(100, .1)$. Again, absolute values are shown since the density is antipodally symmetric.

An important observation from these figures is that the samples from the matrix Bingham density exhibit a large amount of fluctuation, compared with the vectors obtained via traditional PCA, which, in general, seems far smoother as a function of t. This is not a consequence of the Gibbs sampling algorithm, but rather a consequence of the assumption on the data likelihood in (6), which assumes that the projection error is i.i.d Gaussian white noise. This assumption is equivalent to having a matrix Bingham density for the principal directions (12), and, furthermore, indicates that the manifold defined by the matrix Bingham density does not impose any smoothness constraint, and, in fact, gives more weight to *noisy* realizations. In other words, samples from the matrix Bingham density are inherently *noisy* as illustrated by Figures 5 and 6. If one does require some degree of regularity on the principal component samples, one can either use a different conjugate prior on the noise enforcing smaller σ^2 values, or impose a prior on Φ in (12) which tends to smooth the samples. For example, if D is the m-dimensional, first-order, finite difference operator, then one might consider

$$p(\Phi|X,\sigma) \quad \propto \quad \mathrm{etr}(\Phi^T n(S - \delta D^T D)\Phi/2\sigma^2), \tag{32}$$

where $\delta > 0$ is a tunable parameter which penalizes the columns of Φ for having a high total variation or squared difference. This is by no means the only prior that imposes regularity, but, rather we want to emphasize that the Bayesian framework allows for regularization via an appropriately chosen prior.

Before we move on to the random Karhunen-Loeve expansions, we illustrate how the uncertainty about the principal components decreases as the sample size, n, increases. In particular, Figure 6 shows the uncertainty in the principal vectors when n is multiplied by a factor of ten. Note the reduction in the spread of the samples, shown in shaded regions of color, compared with Figure 5.

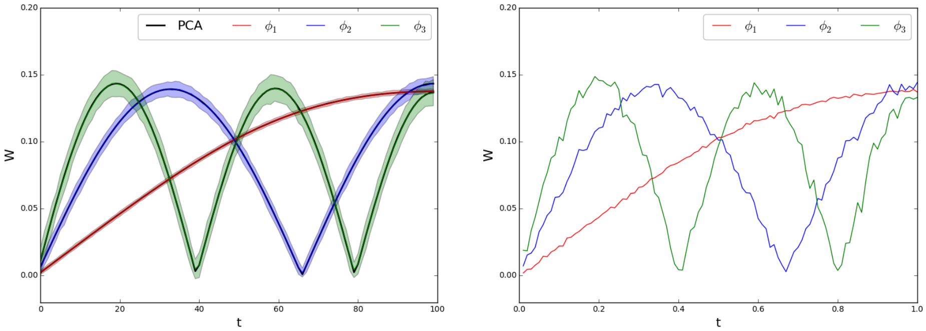

Figure 6: (left) Shaded regions represent ±2σ error bars at each point in time for samples from the matrix Bingham density. Black lines represent the PCA modes, which are the eigenvectors of the sample covariance. (right) Samples from the matrix Bingham density. In this example, $n = 250$, $m = 100$, and the prior on $1/\sigma^2$ is given by gamma(100, .1). Compare this with Figures 4 and 5, where $n = 25$.

6.2. Random KLE

Each sample from the matrix Bingham density, $\Phi^{(i)}$, for $i = 1, \dots, M$, admits a Karhunen-Loeve expansion, where the coefficients are independent, zero-mean Gaussians (this is only valid because W_m is a Gaussian process). To determine the variance for each projection coefficient, we use (28) to obtain samples and then compute the empirical variance. The set of random approximates (27) can now be used to generate new sample data, where each set lives on the subspace defined by $\Phi^{(i)}$. Figure 7 shows the original 25 samples versus a set of samples generated from (27).

Figure 8 shows samples from the Karhunen-Loeve approximates when we multiply the number of samples. by a factor of ten. Note that realizations exhibit slightly less variation due to higher certainty in the principal components (see Figure 6).

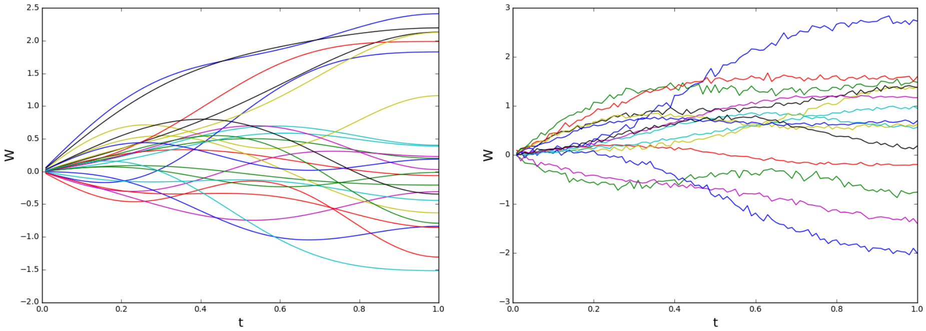

Figure 7: (left) $n = 25$ samples from the stochastic process, W. (right) Samples from the random Karhunen-Loeve approximates. We take $m = 100$ and impose a gamma prior on the noise (See Figure 5).

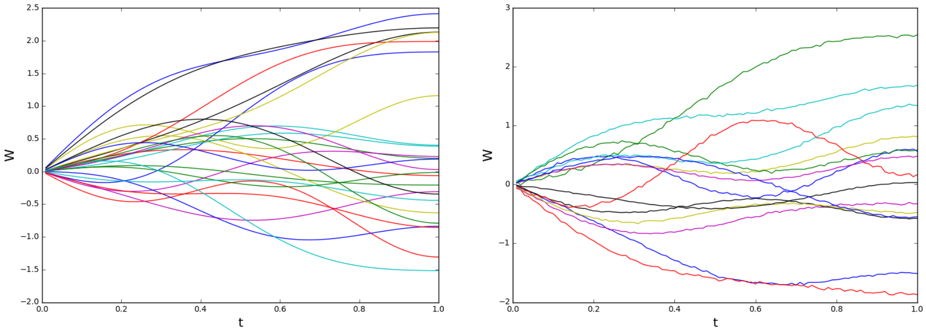

Figure 8: (left) $n = 250$ samples from the stochastic process, W. (right) Samples from the random Karhunen-Loeve approximates. We take $m = 100$ and, again, impose a gamma prior on the noise (See caption of Figure 5).

As per the discussion in the previous section, because the principal components exhibit variability (See Figures 4 and 5 (right)), the samples from the KLE, which are linear combinations of the principal components, exhibit similar variation (see Figure 7 (right)). Again, these fluctuations can be mitigated by decreasing the noise term σ^2 in (12) or by imposing a smoothing prior for Φ (32). However, as previously discussed, this is a natural consequence of the matrix Bingham density, which stems from assumption that our projection error is i.i.d white noise (6). Regardless, summary statistics such as means, variances, and correlations can still be computed and can reasonably approximate the statistics of the true, underlying stochastic process. In fact, for most types of summary statistics, for which *smoothness* is not a necessity, e.g., $P(W(t) \in [a, b] | t \in [t_1, t_2])$, one can interpret these *noisy* realizations in Figures 4 and 5 (right) as samples of the stochastic process from non-contiguous realizations. Therefore, a more appropriate plot of the realizations, which would help compute these types of summary statistics, might look like the following (see Figure 9).

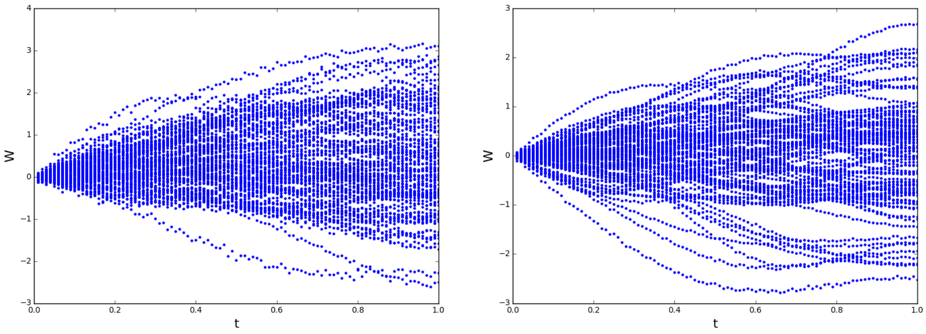

Figure 9: Plot of 100 realizations of the random KLE where each blue data point is plotted as non-contiguous realizations with $n = 25$ (left) and $n = 250$ (right) in (12) and a gamma(100, .1) prior for $1/\sigma^2$. All data points are shown in blue to emphasize the non-contiguous interpretation of the realizations.

7. Conclusion

In this paper we formulate a Bayesian procedure to obtain the basis functions for the Karhunen-Loève Expansion of a square-integrable stochastic process, which allows for uncertainty in the principal components as a function of the sample size. We derive the matrix Bingham density on the Stiefel manifold as the posterior density in this Bayesian approach by making an assumption that the projection error is i.i.d. Gaussian white noise, and introduce a modified Gibbs sampling procedure, based on work by Hoff [7], to sample from this density. The modification allows us to obtain an orthonormal basis in order of decreasing variance, similar to how the principal components are computed numerically via classical Singular Value Decomposition (SVD). Moreover, the Bayesian framework allows flexibility in the form of priors on the noise and the principal components themselves. After samples are obtained from the matrix Bingham density, we can compute random Karhunen-Loeve expansions to generate realizations of the original stochastic process. This probabilistic characterization of the principal components is important in the context of uncertainty quantification so that we can accurately predict the affects of sample size for any quantities of interest which depend on these stochastic processes.

8. Acknowledgement

Support for this work was provided through the Scientific Discovery through Advanced Computing (SciDAC) program funded by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94-AL85000.

References

[1] O. P. L. Maitre, O. M. Knio, Spectral Methods for Uncertainty Quantification with applications to computational fluid dynamics, Springer, New York, 2010.

[2] O. Moklyachuk, Simulation of random processes with known correlation function with the help of karhunen-loeve decomposition, Theory of Stochastic Processes 13 (29) (2008) 163–169.

[3] I. T. Jolliffe, Principal Component Analysis, Springer, New York, 2002.

[4] T. Needham, A visual explanation of Jensen's inequality, The American Mathematical Monthly 100 (8) (1993) 768–771.

[5] M. Bouriga, O. Fron, Estimation of covariance matrices based on hierarchical inverse-wishart priors, Journal of Statistical Planning and Inference 143 (4) (2013) 795 – 808.

[6] G. Cao, L. R. Bachega, C. A. Bouman, The sparse matrix transform for covariance estimation and analysis of high dimensional signals, Image Processing: IEEE Transactions 64 (1) (2011) 27–43.

[7] P. D. Hoff, Simulation of the matrix Bingham-von Mises-Fisher distribution with applications for multivariate and relational data, Journal of Computational and Graphical Statistics 18 (2) (2013) 438–456.

[8] M. E. Tipping, C. M. Bishop, Probabilistic principal component analysis, Journal of the Royal Statistical Society 61 (3) (1999) 611–622.

[9] Y. Chikuse, Statistics on special manifolds, Springer, New York, 2003.

[10] A. Kume, S. G. Walker, Sampling from compositional and directional distributions, Statistics and Computing 16 (3) (2006) 261–265.

[11] G. Casella, G. E. I., Explaining the gibbs sampler, The American Statistician 46 (3) (1992) 167–174.

[12] R. G. Jaimez, J. C. Ruiz, M. J. Valderrama, On the numerical expansion of a second order stochastic process, Applied stochastic models and data analysis 8 (2) (1992) 67–77.

[13] R. G. Jaimez, M. J. Bonnet, On the karhunen-loeve expansion for transformed processes, Trabajos de Estadistica 2 (2) (1987) 81–90.

[14] R. G. Jaimez, M. J. Bonnet, On the karhunen-loeve expansion for transformed processes, Trabajos de Estadistica 2 (2) (1987) 81–90.

[15] K. Sargsyan, B. Debusschere, H. Najm, O. L. Matre, Spectral representation and reduced order modeling of the dynamics of stochastic reaction networks via adaptive data partitioning, SIAM Journal on Scientific Computing 31 (6) (2010) 4395–4421.

[16] S. Chib, E. Greenberg, Understanding the metropolis-hastings algorithm, The american statistician 49 (4) (1995) 327–335.