Givens Transform Approach for Efficient Probabilistic Principle Component Analysis for Bayesian Dimensionality Reduction (GT-PPCA)

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

We develop scalable and flexible Probabilistic Principal Component Analysis (PPCA) methods for determining posterior distributions of spanning frames based on a Givens Representation of the PCA which we term (GT-PPCA). This addresses significant challenges that arise with latent variable in a traditional formulation of PPCA. For sampling posterior distributions we develop Hamiltonian Monte-Carlo Methods (HMC) for sampling on the Stiefel Manifold the PCA orthogonal frame sets. We demonstrate our approach on several challenging example problems including tests problems XYZ and problems arising in our recent work on understanding medical patient data associated with coagulopathy (factors influencing blood clotting). We show our methods provides ways to identify when data sets contain a mixture of low dimensional structures that would not be resolved with traditional PCA approaches. We further show how our approach can be used to develop heirarchical models in terms of low dimensional structures learned from the data sets or to develop prior distributions useful in generalizing low dimensional structures to new settings. To facilitate use of our GT-PPCA method we provide a package with the widely-used Stan statistics package.

1 Introduction

Probabilistic PCA (PPCA) [18] posits a probabilistic generative model where high-dimensional data is determined by a linear function of some low-dimensional latent state. Conducting inference on PPCA can be interpreted geometrically as finding the closest low-dimensional hyper-plane to a cloud of data points. This probabilistic approach is attractive because it enables a straightforward methodology, via Bayesian inference, to quantify the uncertainty in our estimates (to prevent overfitting) and conduct hypothesis testing. For example, given high dimensional medical data for patients with two different type of injuries, it would be desirable to find posterior distributions of low-dimensional subspaces that describe the data, then find the probability given the data that these subspaces are different for the two groups of patients. While uncertainty quantification of the latent factors in PPCA has been explored in the literature [9, 2, 3], there are currently no out-of-the-box solutions available to researchers.

In addition to enabling uncertainty quantification and hypothesis testing, probabilistic models are amenable to expansion and can serve as modules within larger probabilistic graphical models. This is important in real-world settings where we seek to utilize any known prior information in our inference or when true generative models do not necessarily follow the simple generative process set forth by PPCA. If we believe our latent factors to be sparse, we can add Laplace or Cauchy priors to our PPCA model yielding a probabilistic sparse PCA[19]. Following the medical example, we may believe subspaces for different groups of patients come from some common prior distribution of subspaces, in which case we can build a hierarchical model to do transfer learning. Similarly, we can expand

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

the PPCA graphical model to conduct non-linear dimensionality reduction via Mixtures of Factor Analyzers [8]. To handle binary or discrete data we can expand PPCA using a link function as in Bayesian Exponential Family PCA (BXPCA) [15].

While expanded PPCA models have shown promise on a variety of problems, they have not been fully explored because their implementation remains elusive, and most inference schemes such as Expectation Maximization (EM) only provide point estimates. The availability of PPCA in a simple framework for building probabilistic graphical models like Stan [4] would allow rapid building and prototyping of such models in a fully Bayesian way that provides uncertainty around any point estimates.

Bayesian inference of orthonormal matrices Many of the difficulties in conducting full Bayesian inference on PPCA and related models stem from having to infer one or more unknown orthonormal matrix parameters. This is difficult because $n \times p$ orthonormal matrices form a rather particular subset (also submanifold) of all possible $n \times p$ matrices; this is analogous to three-dimensional unit vectors which lie on the sphere (a submanifold of $\mathbb{R}^3$). Thus we require a probability distribution on a sub-manifold within the full space of orthonormal matrices. More specifically, the prior and posterior distributions of an orthonormal matrix W must have support over the set of $n \times p$ orthonormal matrices i.e. they should assign zero probability to sets of non-orthonormal matrices. This set of orthonormal matrices is known as the Stiefel manifold and denoted $V_{n,p}$ [16]. If we naively conduct inference over the elements of W without paying mind to orthonormality constraints, we have no way of obtaining valid posteriors with support over the Stiefel Manifold.

Transformed random variables Posterior distributions for constrained parameters in probabilistic graphical models are routinely inferred by transforming such parameters to an unconstrained space and seeking posterior distributions over the transformed parameter [4, 11]. This requires a smooth one-to-one transformation $f : \operatorname{supp}(z_{\text{constr}}) \rightarrow \mathbb{R}^D$, where $\operatorname{supp}(z_{\text{constr}})$ is the support of the constrained random variable z_{constr}. To our knowledge, no such transformation has been proposed to map orthonormal matrices to a comparable unconstrained space.

GT-PPCA In this paper we draw on techniques from Differential Geometry and Numerical Analysis to introduce a novel and geometrically elegant way to represent orthonormal matrices. In our approach we express orthonormal matrices in terms of a sequence of fundamental rotations through given angles. This gives insight into the geometry of the Stiefel manifold, and results in a transform we call the Givens Transform, that maps orthonormal matrices to an unconstrained space. We apply the Givens Transform to inference of PPCA-based models, and collectively refer to this as GT-PPCA.

GT-PPCA is straightforward to implement in stand-alone inference schemes, but is particularly useful in the context of probabilistic programming framework like Stan [4], where we can use it for uncertainty quantification and hypothesis testing of PPCA models, as well as extending PPCA to more complex probabilistic graphical models, two previously intractable tasks. We provide Stan code for our example models allowing for use and expansion by researchers and scientists out-of-the-box. We demonstrate how inference of these models in Stan yields good empirical performance on large probabilistic graphical models that were previously intractable to implement especially if fully-Bayesian posterior analysis is desired. Specifically we present a hierarchical subspace model for grouped, multi-view medical data and a PPCA Hidden Markov Model (HMM) for disease-network data.

In addition to opening the door for straightforward implementation of large PPCA models, GT-PPCA yields insight in to novel and useful ways to work with and interpret our models. For instance, the elegant geometric representation lets us see how by limiting the range of the parameters in GT-PPCA, we can naturally avoid issues of unidentifiability and multi-modal posteriors that arise in other methods. GT-PPCA also allows us new and creative ways to generate and use prior distributions on orthonormal matrices. In the setting of using the matrix directly this task has previously been rather complicated and rather intractable for even small problem sizes. This is linked to the difficulty of evaluating densities of orthonormal matrix distributions in other representations. As we shall discuss in more detail, our GT representation provides a rather natural way to specify prior distributions comparable to the Matrix Langevin prior [16].

Related work While previous authors have developed methods for posterior sampling of distributions orthonormal matrices, these methods can at times suffer from numerical issues, they are difficult to implement on large probabilistic graphical models, and they can not be used in any general inference scheme such as VI or Maximum A-Posteriori (MAP) estimation like GT-PPCA can. Brubaker et al. [2] and Byrne and Girolami [3] used separate approaches to modify the Leap-Frog integrator typically used in Hamiltonian Monte Carlo (HMC), so that Hamiltonian exploration, and thus MCMC samples of posteriors, satisfied any necessary constraints at all times. Specifically, Brubaker et al. [2] uses the SHAKE integrator [13] to simulate Hamiltonian dynamics and generate proposals. The integrator works by repeatedly taking a step forward that may be off the manifold using ordinary leap frog, then projecting back down to the nearest point on the manifold. This projection is done via Newton iterations, which may converge to the wrong local minimum in practice or perhaps not converge at all, possibly jeopardizing the ergodicity of a Markov Chain, and the integrity of samples [1]. Byrne and Girolami [3] took a different approach, exploiting the fact that closed form solutions are known for the geodesic equations over the Stiefel manifold in the embedded coordinates, W. While this method is completely explicit, requiring no Newton iterations, in practice we found that for larger step sizes, the integrator steps off the Stiefel manifold, due to the numerical imprecision of the matrix exponential function. Because these methods use modified integrators for constrained parameters, in practice they require keeping track of the support of each variable and which type of integrator to use on each variable. This adds an extra layer of implementation complexity, especially for large complex probabilistic graphical models, that makes it difficult to implement these methods within a probabilistic programming language such as Stan. This precludes the rapid prototyping and building of models as well as the flexibility to use different inference schemes that Stan provides. Lastly, we remark that for inference on orthonormal matrices, these methods can lead to multi-modal posteriors, that can be avoided in a straight-forward way using the Givens transform.

Paper outline We give a brief overview of probabilistic dimensionality reduction in Section 2. We discuss the geometry of the Stiefel Manifold in Section 3, before finally introducing the Givens Transform (GT) in Section 4. Finally, we present various empirical studies where we used GT-PPCA in Stan for practical uncertainty quantification and hypothesis testing, as well as for building complex probabilistic graphical models.

2 Probabilistic Principle Component Analysis (PPCA)

In probabilistic principle component analysis (PPCA) one starts by considering a collection of data points in a typically high-dimensional vector space and seeks to find a posterior distribution over a reduced representations of the data in the form of a lower dimensional subspace. The central postulate is that for a data vector $\mathbf{x} \in \mathbb{R}^n$ there exists an unknown low-dimensional latent representation $\mathbf{z} \in \mathbb{R}^p$ where $p < n$, (ideally with $p \ll n$). The two representations are related to each other by a single unknown linear transformation $\mathbf{x} \to \mathbf{z}$. Mathematically, we consider a finite collection of sampled data vectors $\mathbf{x}_i \in \mathbb{R}^n$, $i = 1, \cdots, N$ and try to estimate this subspace. Formally, PPCA consists of the following generative process

$$\begin{aligned} p(\mathbf{z}_i) &\sim \mathcal{N}_p(0, I) \\ p(\mathbf{x}_i|\mathbf{z}_i, W, \Lambda, \sigma^2) &\sim \mathcal{N}_n(W\Lambda\mathbf{z}_i, \sigma^2 I). \end{aligned} \tag{1}$$

The W is an $n \times p$ orthonormal matrix and Λ is a $p \times p$ diagonal matrix with positive elements. For simplicity in our presentation of PPCA, we have assumed here that the data has only zero mean but the more general case can also readily be considered [17, chapt. 12.1].

Quantifying uncertainty Inference for PPCA is typically conducted by obtaining a point estimate for W via a closed-form estimator, or for expanded PPCA models via (EM) [17, chapt. 12.2], neither of which provide a notion of uncertainty for our point estimates. Without information regarding the uncertainty of our estimates these point estimates could be far from the true value of W and thus mislead our conclusions, especially for larger models and/or when there is relatively little data available. Furthermore, point estimates do now allow for hypothesis testing e.g. statistically testing whether two different groups of observations lie in the same subspace. We show with examples how GT-PPCA in Stan makes it easy achieve these tasks as we show in section 5.

Expanding models As alluded to previously, PPCA generative model can be flexibly expanded in several ways as modelers see fit. To build a probabilistic sparse PCA, one can place a Laplace or Cauchy prior over the elements of W. If we have meaningfully grouped data, such as data from hospital patients with different types of injury, it might be desirable to designate a separate W parameter (subspace) for each group, then place prior over these subspaces to garner the benefits of hierarchical modeling [7, chapt. 5]. Mohamed et al. [15] showed that we can model non-Gaussian data, $\mathbf{x}_i$, by replacing equation 1 with an exponential family member whose natural parameters are given by $\mathrm{Expon}(W\Lambda\mathbf{z}_i)$ where $\mathrm{Expon}(\cdot)$ is an appropriate link function. Again, in the context of a probabilistic programming language such as Stan, these extensions to the base PPCA model become trivial to implement as we illustrate with examples in section 5.

Importance of the Orthonormality Condition The orthonormal constraint on the matrix W plays an important role in obtaining robust methods for making inferences in probabilistic PCA because it alleviates identifiability and numerical issues. If one were to relax the orthonormality constraint the likelihood function would assign identical probability to a whole equivalence class of matrices $W \sim V$ where the span is the same linear subspace $\text{span}\{W\} = \text{span}\{V\}$ Murphy [17, chapt. 12.1.3]. Besides resulting in an unidentifiable model, in practice this presents a number of major challenges. This first is that the matrices in a given equivalence class are not all equally well-conditioned numerically and round-off errors and truncation errors become problematic in practical calculations. Secondly, these issues with the representation further manifest in the log-likelihood objective function where regions arise of particularly large curvature as pointed out by [10]. This causes significant numerical issues for variational inference (VI) in nonlinear optimization methods and in Monte-Carlo (MC) approaches with samplers having slow mixing times [10]. We note that, while most identifiability issues and numerical issues are alleviated by constraining inference to orthonormal matrices, the PPCA likelihood is equivalent for an orthonormal matrix W and any permutation of the columns of W being negative as pointed out by both Murphy [17, chapt. 12.1.3] and Holbrook et al. [10]. As such, even the methods of Brubaker et al. [2] and Byrne and Girolami [3] will lead to multi-modal posteriors, that can be avoided in a straight-forward way by appealing to insights revealed by the Givens Transform, as we explain in Section 4.

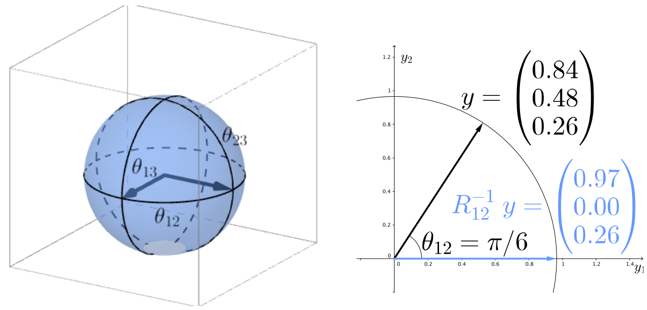
3 Geometry of the Stiefel Manifold

The set of $n \times p$ orthonormal matrices $V_{n,p}$, form a sub-manifold in the space of general $n \times p$ matrices known as the Stiefel Manifold [16] and formally defined as

$$V_{n,p} := \{Y \in \mathbb{R}^{n \times p} : YY^T = I\}. \tag{2}$$

Intuitively, the elements of $V_{n,p}$ can be thought of not as orthonormal matrices, but as p-frames which are comprised of p orthonormal vectors that lie in n-dimensional space. To move about the Stiefel manfiold, one can rigidly rotate the vectors in the p-frame about any combination of axes an arbitrary number of times. In the case where $n = 3$ and $p = 2$, this is almost identical to sphere (Figure 1a), but with an extra angle, θ_{23} that controls how much the second basis vector is rotated about the first. For a three-dimensional set of points forming a flat, pancake-like cloud, PPCA can be thought of as finding the best 2-frame that aligns with this cloud.

While $n \times p$ orthonormal matrices are represented by np elements, the Stiefel Manifold $V_{n,p}$, has an intrinsic dimension of $np - p(p+1)/2$. This arises from the constraints on the columns of the matrix that impose orthonormality. This dimensionality can be seen by observing, that the first column of $Y \in V_{n,p}$ must have norm one and hence has one constraint placed on it. The second column must also have norm one and also must be orthogonal to the first column hence with two constraints placed on it. Continuing to the third column through the n^{th} one arrives at the conclusion that each point of the Stiefel Manifold has only $np - (1 + 2 + \cdots + p) = np - p(p+1)/2$ degrees of freedom. The Givens transform can be thought of as an $np - p(p+1)/2$-dimensional set of coordinates Θ, that represent elements of the Stiefel manifold.



(a) (b) (c)

Figure 1: Visualizing the Givens Transform. (a) How the Givens Reduction “zeros out” a column vector. (b) A geometric view of the Stiefel manifold, two-frame in three dimensions. (c) Sampling without a proper measure adjustment.

4 Givens Transform (GT) approach to PPCA (GT-PPCA)

For several types of constrained parameters, posterior distributions are in practice rather routinely inferred using both MCMC and VI by transforming the constrained variables to an unconstrained space using a one-to-one mapping $T : \mathrm{supp}(z_{\mathrm{constr}}) \to \mathbb{R}^D$, where $\mathrm{supp}(z_{\mathrm{constr}})$ is the support of the constrained random variable z_{constr}. One can obtain a posterior over the unconstrained parameter that corresponds to the original constrained parameter of interest, then map inferences back to the original constrained space. This procedure requires computing the Jacobian, $J_{T^{-1}}$ of the transformation, to obtain $f_Y(y) = f_{z_{\mathrm{constr}}}(T^{-1}(y))J_{T^{-1}}(y)$ where Y is an unconstrained random variable with probability density function (PDF) f_Y and $f_{z_{\mathrm{constr}}}$ is the probability density of z_{constr}, which for PPCA comes from equation 1. The extra Jacobian term accounts for how the a unit volume under the transformation changes [11]. Without this extra Jacobian factor, inference between the two spaces is incomparable. For example, uniformly sampling in spherical coordinates (unconstrained space) does not correspond to uniformly sampling on the sphere (constrained space), unless we include an appropriate term accounting for how volumes are warped under the transformation (see Figure 1c). Intuitively, ares that are near the poles get shrunk far more than areas near the equator, so when mapped back on to the sphere, points will congregate closer to the poles of the sphere. Conducting inference in a transformed space is most notably used in ADVI and Stan’s HMC routines [4, 11]. In sub-section 4.1 we briefly discuss Givens Reductions, motivating the Givens Transform. In sub-section 4.2 we discuss geometric aspects of the Givens transform such as avoiding multi-modality and including a term that measures how volume is changed under the transform that is analogous to the Jacobian described above.

4.1 Givens Reductions and the Givens Transform

The Givens Reduction is a numerical analysis technique for reducing a square matrix A to upper-triangular form [14]. The technique works by applying a series of rotation of matrices to A such that elements below the diagonal are “zeroed out" starting with the second element of the first column, and moving down the first column before zeroing out the appropriate elemnents of the subsequent columns. For example if A is a 3×3 matrix and its first column is $(0.84, 0.48, 0.26)^T$, the Givens Reduction would would apply a rotation in the $(1, 2)$-plane, R_{12}^{-1} so as to annhilate the second element of this column (Figure 1b). For an $n \times p$ orthonormal matrix Y, applying the Givens Reduction requires mulltiplication by $(n-1)+(n-2)+\cdots+(n-p) = np - p(p+1)/2$ rotations matrices each with their own respective angles and results in the matrix$I_{n,p}$, whose columns are the first p standard basis vectors.

$$(R_{pn}^{-1} \cdots R_{p,p+2}^{-1} R_{p,p+1}^{-1}) \cdots (R_{2n}^{-1} \cdots R_{24}^{-1} R_{23}^{-1})(R_{1n}^{-1} \cdots R_{13}^{-1} R_{12}^{-1})Y = I_{n,p}, \tag{3}$$

From the perspective of p-frames, the Givens Reductions maps all p-frames to the canonical p-frame $I_{n,p}$. Geometrically, this is because if Y is already orthonormal then applying rotation matrices will rigidly rotate all columns of Y at once preserving their orthogonality, and leaving their length unchanged. Because rotations are invertible we can rewrite 3 as

$$Y = (R_{12} \cdots R_{1n}) \cdots (R_{23} \cdots R_{2n})(R_{p+1,n} \cdots R_{pn}) I_{n,p}. \tag{4}$$

which we refer to as the Givens Representation of an orthonormal matrix Y. Since each of the $np - p(p+1)/2$ rotation matrices have an associated angle $(\theta_{12} \cdots \theta_{1n}) \cdots (\theta_{23} \cdots \theta_{2n})(\theta_{p+1,n} \cdots \theta_{pn})$, that we collectively refer to as Θ, we can use these angles to represent any $n \times p$ orthonormal matrix. In this way we have reparameterized all $n \times p$ orthonormal matrices [1], a constrained space, in terms of unconstrained angles [2], using a transform $\Theta : V_{n,p} \to \mathbb{R}^{np-p(p+1)/2}$. In a probabilistic programming framework like Stan we can treat Θ as an unknown parameter and $Y(\Theta)$ is a transformed variable we are free to use in a likelihood such as the likelihood from 1. We also mention that multiplication by rotation matrices are inexpensive to compute as they are highly sparse (especially in large dimensions) and when applied to a matrix, they only modify two rows of that matrix at a time. We refer to 4 as the Givens representation or Givens Transform.

4.2 Geometry of the Givens Transform

Topologically, $V_{n,p}$ is locally equivalent to Euclidean space, but not globally equivalent, meaning it is impossible to find a one-to-one map between the Stiefel manifold and Euclidean space. Technically speaking, the Givens transform can map angles to all of $V_{n,p}$ except for a subset $S \subset V_{n,p}$, that in the $n = 3, p = 2$ case corresponds to a sliver when $\theta_{12} \in (-\pi, \pi)$, $\theta_{13} \in (-\pi/2, \pi/2)$, and $\theta_{23} \in (-\pi/2, \pi/2)$. Luckily this set is of measure zero (under the proper measure for the Stiefel manifold, and thus, with probability one, the orthonormal matrix that describes the true subspace our data lie in will not be in that set.

In practice, we actually limit the angle θ_{12} to an interval of length π rather than an integral of length 2π, that traverses the entire Stiefel manifold. Examining the angles of the Givens transform reveals geometrically, the insight that in the latter case, two equivalent bases that are the negation of each other can be reached, resulting in a multi-modal posterior that makes sampling and VI more difficult and harder to interpret. To avoid this multi-modality using the modified integrator methods would require a mechanism to avoid boundaries, which are not as intuitively defined in the default embedded coordinates as in the Givens transform.

Lastly, we note that if the true bases lies near a pole, i.e. θ_{ij} is close to $-\pi/2$ or $\pi/2$, then posteriors will tend to be multi-modal as the region in parameter space close to the boundaries will be close to equally valid, while the region near zero, will not be valid and thus contain little probability mass. In these cases, one can simply change the chart so that $\theta_{ij} \in (0, \pi)$, creating a uni-modal posterior in the new coordinate system, and alleviating numerical issues. In Stan this is straight-forward, as one simply has to change the lower and upper bound of the angle parameter.

An analogous Jacobian term using differential forms As stated earlier, conducting inference on a transform space requires a Jacobian term accounting for how volumes are warped by the transform, but in the case of the Givens Transform this poses a problem because an $n \times p$ orthonormal matrix is np-dimensional, but the Givens transform, $\Theta(Y)$, maps this set to an $np - p(p+1)2$-dimensional set of angles Θ. In this more general scenario, one can not simply take the determinant of the Jacobian as the volume morphing factor, because the Jacobian is not even square and hence the determinant is undefined. To obtain the correct factor one must appeal to the calculus of differential forms.

Intuitively, differential forms measure how a transform warps an infinitesimal volume from one space to another, but they are more general in that they can be applied irrespective of the coordinates we use to describe either space. For example, spherical coordinates $(\theta, \varphi) \mapsto (x(\theta, \varphi), y(\theta, \varphi), z(\theta, \varphi))$ map points in the flat plane, $\mathbb{R}^2$, to points in $\mathbb{R}^3$ that lie on the sphere. $d\theta \wedge d\varphi$ represents a small area in the plane that can be rewritten as a small patch in $\mathbb{R}^3$ by finding $d\theta$ and $d\varphi$ in terms of dx, dy, and dz and applying the well defined rules of a wedge product.

[1]other than a set of measure zero, that is thus negligible

[2]the angles are themselves constrained to lie in certain intervals e.g. $[0, \pi)$ but these sorts of constraints are routine to deal with using a one-to-one diffeomorphism between intervals and the real line e.g. the sigmoid transform

One can then integrate this over a sub-area of the sphere to obtain the measure of that sub-area. The beauty of differential forms is that a different coordinate system such as polar coordinates, simply correspond to using a different basis for representing forms, which again are vectors. In fact writing a form in a different bases simply involves taking partial derivatives, e.g. $dx = (\partial x/\partial\theta)\, d\theta + (\partial x/\partial\varphi)\, d\varphi$. If we substitute the forms in the angle bases in to 5 and simplify using the well-defined anti-symmetric and distributive properties of wedge products and differential forms (see [16]), we obtain a proper area-form 5 in spherical coordinates, $S(x(\theta,\varphi), y(\theta,\varphi), z(\theta,\varphi))\, d\theta \wedge d\varphi$, that can then be integrated using a double integral in polar coordinates. The absolute value of this area-form evaluated at some point specified in angle coordinates, (θ_0, φ_0), intuitively measures how much the area of tiny square in angle space gets shrunk or stretched when the area is mapped to an area on the sphere.

Analogously, we can measure volumes on the Stiefel manifold. For $n \times p$ orthonormal matrices, there are only $np - p(p+1)/2$ free parameters and so the proper form to measure sets of orthonormal matrices is in fact a $np - p(p+1)/2$-form. For an orthonormal, $n \times p$ matrix, Y, we can find an orthonormal $n \times n$ matrix G such that $G^T Y = I_{n,p}$. In fact G just comes from the product of the appropriate rotation matrices that comes from the Givens Reduction 3. Muirhead [16] shows that the correct form comes from wedging the elements of the $n \times p$ matrix $G^T dY$ that lie below the diagonal i.e.

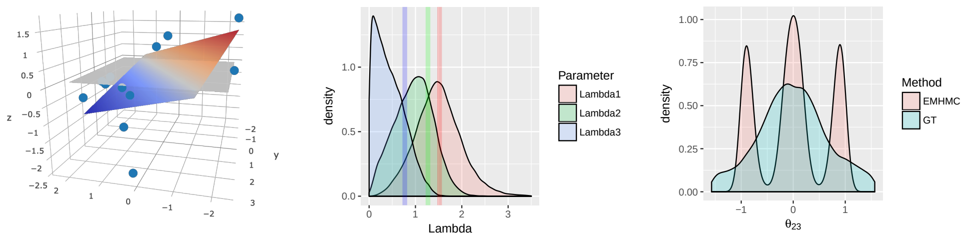
$$\bigwedge_{i=1}^{p} \bigwedge_{j=i+1}^{n} G_j^T\, dY_i \tag{5}$$

where G_j is the jth column of G and Y_i is the ith column of Y. To obtain the form in angle coordinates we simply obtain dY in angle coordinates. dY_i can be obtained in terms of the angle coordinates by the following relationship, $dY_i = J_{Y_i}(\Theta)\, d\Theta$, where J_{Y_i} is the Jacobian of Y_i with respect to the angle coordinates. Once we obtain the form 6 in terms of the angle coordinates, the result is a wedge product of $np - p(p+1)/2$ vectors that are $np - p(p+1)/2$ dimensional, which reduces to the determinant of these vectors aligned side by side as a $np - p(p+1)/2 \times np - p(p+1)/2$ matrix. This determinant is analogous to and serves the same purpose as Jacobian adjustment that comes from transforming random variables. We can insert it in to the log-probability of a model to avoid the sort of unintended sampling behavior depicted in Figure 1c. We incorporate the form 6 in to the log-probability of all of our Stan examples.

5 Empirical Studies

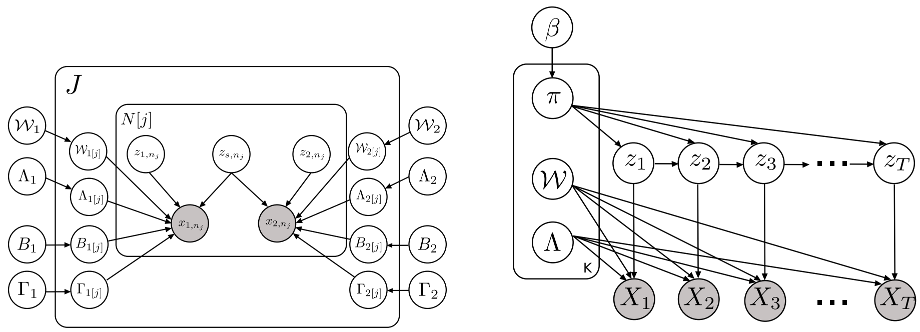
5.1 Synthetic Data

We generated a synthetic, three-dimensional dataset that lies on a two-dimensional plane with $N = 15$ observations according the generative process of PPCA 1. We chose $\mathrm{diag}(\Lambda) = \mathrm{diag}(1, 1)$, $\sigma^2 = 1$, and W to be $I_{3,2}$ which in the Givens representation corresponds to $\theta_{12} = \theta_{13} = \theta_{23} = 0$. This example illustrates how simply running GT-PPCA in Stan can alleviate overfitting issues in the common use-case where one seeks to carry out dimensionality reduction in a low-observation regime. A standard classical PCA analysis yields the singular values $\mathrm{diag}(\hat{\Lambda}) = (1.52,\ 1.27,\ 0.77)$, possibly suggesting that our data lie close to some two-dimensional plane, since the third singular value has a larger drop off from the first two than the second has from the first. Figure 2a illustrates geometrically a point estimate of the subspace found by PCA. This corresponds to the subspace spanned by the PCA point estimates of the latent factor loadings. Because of relatively low signal to noise ratio and modest sample size, the point estimate is drastically affected by only a few observations and is characteristically different from the flat plane, which we know to be the truth in this case. The PCA point estimate θ_{13}, which if we recall from Figure 1a is the Givens Transform angle that controls the upwards tilt of the plane, is $\hat{\theta}_{13} = -0.15$. Meanwhile, posterior HMC samples from GT-PPCA in Stan yields a median value of -0.24 and a 95% posterior interval of $(-1, 0.78)$. This lets us know that there is high uncertainty around our point estimate given the data, and suggests that any conclusions drawn from point estimates may be overfit to the data, thus protecting us from concluding false-positive results and suggesting to the experimenter that more data is needed to make a conclusive statement. Alternatively, we can incorporate any prior knowledge we have about the problem, such as knowledge about the structure of W or knowledge about the W of a closely related



(a) (b) (c)

Figure 2: Inferences for three-dimensional synthetic data. (a) Three-dimensional points, true subspace (grey), and classical PCA point estimate of subspace (colored). (b) Estimated densities from posterior draws of Λ parameters A.K.A the singular values, and point estimates from classical PCA show as colored bars. (c) Avoidance of multi-modal behavior in GT-PPCA versus EMHMC.



(a) (b)

Figure 3: Probabilistic graphical models for (a) Hierarchical CCA Model (b) Count Subspace HMM for Network Data.

group of samples, in the form of a prior distribution of our angles in our Stan model as we do in the following subsection.

The fully Bayesian approach provided by GT-PPCA in Stan also allows us to examine posterior draws of Λ to make probabilistic statements about the inherent dimensionality in our data. Figure 2b shows estimated densities from posterior draws of Λ. The posterior of Λ_3 for example places considerable mass close to zero (58% of samples were less that 0.5), providing strong evidence that our data is inherently two, not three, dimensional. This is as oppose to classical PCA where we heuristically assess dimensionality based solely on the magnitude of our point estimates.

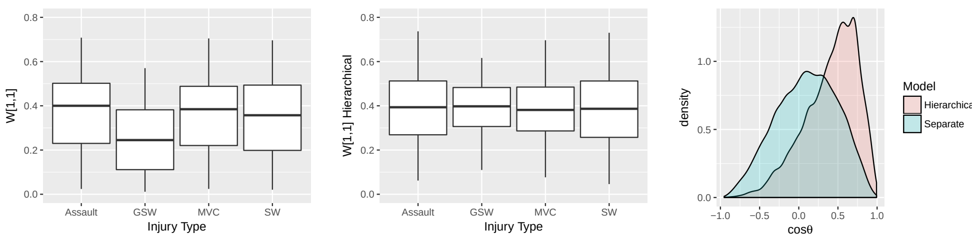
Lastly, Figure 2c compares posterior samples of our synthetic data from Embedded Manifold HMC (EMHMC) and GT-PPCA in Stan. As explained in section, EMHMC explores the entire Stiefel manifold which includes multiple equivalent modes, where as with GT-PPCA we can eliminate this multi-modal behavior by simply constraining the Givens Transform angle parameters. This is useful both for interpretation and in higher dimensional problems where the number of modes grows exponentially and HMC can not visit all of them.

5.2 Coagulopathy using hierarchical subspace models

Include figure for CCA probabilistic graphical model.

5.3 School Network

Show W for each of the three states then posterior probabilities of what state you're in. Show probabilistic graphical model. Point how this can be used for disease networks and also recognizing states in fMRI data.



(a) (b) (c)

Figure 4: Inferences for Hierarchical CCA model.

6 Discussion

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References

[1] Diagnosing biased inference with divergences. `http://mc-stan.org/documentation/case-studies/divergences_and_bias.html`. Accessed: 2017-05-11.

[2] Marcus Brubaker, Mathieu Salzmann, and Raquel Urtasun. A family of mcmc methods on implicitly defined manifolds. In *Artificial Intelligence and Statistics*, pages 161–172, 2012.

[3] Simon Byrne and Mark Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

[4] Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 20, 2016.

[5] Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. A generalization of principal component analysis to the exponential family. In *Nips*, volume 13, page 23, 2001.

[6] Alan Edelman. 18.325: Finite random matrix theory volumes and integration. 2005.

[7] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[8] Zoubin Ghahramani, Geoffrey E Hinton, et al. The em algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto, 1996.

[9] Peter D Hoff. Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2): 438–456, 2009.

[10] Andrew Holbrook, Alexander Vandenberg-Rodes, and Babak Shahbaba. Bayesian inference on matrix manifolds for linear dimensionality reduction. *arXiv preprint arXiv:1606.04478*, 2016.

[11] Alp Kucukelbir, Rajesh Ranganath, Andrew Gelman, and David Blei. Fully automatic variational inference of differentiable probability models. In *NIPS Workshop on Probabilistic Programming*, 2014.

[12] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[13] Benedict Leimkuhler and Sebastian Reich. *Simulating hamiltonian dynamics*, volume 14. Cambridge University Press, 2004.

[14] Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 2. Siam, 2000.

[15] Shakir Mohamed, Zoubin Ghahramani, and Katherine A Heller. Bayesian exponential family pca. In *Advances in neural information processing systems*, pages 1089–1096, 2009.

[16] Robb J Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009.

[17] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[18] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[19] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.