Rotation Invariant Householder Parameterization for Bayesian PCA

Rajbir S. Nirwan [1] **Nils Bertschinger** [1 2]

Abstract

We consider probabilistic PCA and related factor models from a Bayesian perspective. These models are in general not identifiable as the likelihood has a rotational symmetry. This gives rise to complicated posterior distributions with continuous subspaces of equal density and thus hinders efficiency of inference as well as interpretation of obtained parameters. In particular, posterior averages over factor loadings become meaningless and only model predictions are unambiguous. Here, we propose a parameterization based on Householder transformations, which remove the rotational symmetry of the posterior. Furthermore, by relying on results from random matrix theory, we establish the parameter distribution which leaves the model unchanged compared to the original rotationally symmetric formulation. In particular, we avoid the need to compute the Jacobian determinant of the parameter transformation. This allows us to efficiently implement probabilistic PCA in a rotation invariant fashion in any state of the art toolbox. Here, we implemented our model in the probabilistic programming language Stan and illustrate it on several examples.

1. Introduction

Modern algorithms and computational tools have vastly expanded the scope of Bayesian modeling over the last decades. In particular, Hamiltonian Monte-Carlo (HMC) sampling (see Betancourt (2017) for a thorough introduction) and variational inference (Bishop, 2006) allow to approximate high-dimensional posterior distributions. In addition, software tools such as probabilistic programming languages, e.g. Stan (Carpenter et al., 2017), or libraries for automatic differentiation, e.g. Tensorflow (Abadi et al., 2015), simplify the implementation. Thanks to these advances, researchers can focus on the statistical modeling as model implementation often requires just a few lines of code. Nevertheless, some models remain challenging. Well known examples include Probabilistic PCA (PPCA) and related factor models which are widely used in exploratory and confirmatory data analysis. These models are non-identifiable, which poses major problems when fitted parameters, e.g. factor loadings, are being interpreted. In this context, the non-identifiability arises from a rotational symmetry of the likelihood model. While identification can be restored by imposing constraints on the factor loadings (Jöreskog, 1969; Peeters, 2012), maximum likelihood estimation can be biased by the imposed constraints (Millsap, 2001).

Here, we consider PPCA from a Bayesian perspective. In this context, the rotational symmetry gives rise to a complicated posterior with continuous subspaces of equal density. Continuous symmetries can severely reduce sampling efficiency as many samples are required in order to explore the corresponding subspaces. Furthermore, interpretation of the sampled parameters becomes impossible as marginal posterior averages are meaningless when many different parameter combinations give rise to the same likelihood. This is akin to the discrete label switching symmetry observed in mixture models. In this case, the parameters of the individual mixture components cannot be averaged as the arbitrary component labels between different samples could be switched. Thus, a major advantage of Bayesian modeling, namely the ability to access estimation uncertainty, cannot be utilized as posterior means and variances not just reflect variation within, but also between components.

Therefore, it is desirable to remove the rotational symmetry of PPCA. Formally, the symmetry arises due to the choice of coordinate system for the principal components being arbitrary. In particular, a rotation of the coordinate axes leaves the model likelihood unchanged. Mathematically, the rotation symmetry is made explicit by formulating the model in terms of an orthogonal matrix (transforming into a fixed coordinate system) and a diagonal matrix containing the explained variances. To this end, we employ the singular value decomposition (SVD) and parameterize the Stiefel manifold of orthogonal matrices in terms of non-

[1]Department of Computer Science, Goethe University, Frankfurt, Germany [2]Frankfurt Institute for Advanced Studies, Frankfurt, Germany. Correspondence to: Rajbir S. Nirwan <nirwan@fias.uni-frankfurt.de>, Nils Bertschinger <bertschinger@fias.uni-frankfurt.de>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

redundant unconstrained parameters. The mapping between parameters and orthogonal matrices is provided in terms of a sequence of Householder transformations.

1.1. Related work

Special sampling methods for inference on statistical manifolds have been developed, in which the geometric structure of the manifold is respected and geodesic trajectories are traced out with respect to the metric of the manifold (Byrne & Girolami, 2013). In general, the geometry will be non-Euclidean, e.g. spherical for unit vectors. Also the geometry of the Stiefel manifold of orthogonal matrices can be handled in this fashion. Unfortunately, as the geometry needs to be build into the sampling algorithm, this approach requires substantive work. This also applies to approaches using analytic results for the matrix von Mises-Fisher distribution on the Stiefel manifold. Exploiting conditional conjugacy (Hoff, 2009) and (Šmídl & Quinn, 2007) have employed this distribution to implement Gibbs sampling and variational inference for Bayesian PPCA respectively. An alternative, which allows to utilize general purpose tool boxes, is to reparameterize the manifold in terms of unconstrained parameters[1]. Pourzanjani et al. (2017) achieved this via Givens rotations. They also demonstrate that the model can then be implemented in Stan without any changes to the underlying sampling routine and unrestricted by conjugacy requirements.

Ideally, reparameterizing a Bayesian model should not change the joint density defined by the model. In case of the transformation via Givens rotations, the parameter density corresponding to a uniform measure on the Stiefel manifold is not known. Thus, by change of measure the density needs to be corrected by the Jacobian determinant, which poses a major computational bottleneck in high-dimensional models. Here, we propose a different parameterization in terms of Householder transformations. In this case, results from random matrix theory allow us to obtain the induced parameter density corresponding to a Gaussian prior on the original parameters of PPCA, which include the rotational symmetry. In section 2 we recall the necessary mathematical results to do so. Then, we illustrate the resulting model on several examples (section 3 and 4) before we conclude in section 5.

2. Background

Here, we quickly review PPCA and explain why the model is rotationally symmetric, and how the symmetry can be removed via singular value decomposition (SVD). Finally, we discuss some results from random matrix theory in order to construct a suitable prior on the transformed parameters.

[1] In general, such a reparameterization cannot be achieved globally. Instead, the non-Euclidean geometry gives rise to singularities in the parameter transformation, e.g. at the north pole when mapping a sphere onto a plane, or a different mapping, i.e. chart, needs to be employed at different points.

2.1. Probabilistic Principal Component Analysis (PPCA)

PPCA (Tipping & Bishop, 1999; Bishop, 2006) relates a D-dimensional observation $\boldsymbol{y}$ to a Q-dimensional latent vector $\boldsymbol{x}$ via a linear mapping $\boldsymbol{W} \in \mathbb{R}^{D\times Q}$

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{\mu}$ allows the model to have a non-zero mean and $\boldsymbol{\epsilon}$ states the variance not explained by the first two terms as noise. PPCA assumes a standard Gaussian distribution on the latent space, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and a zero mean Gaussian noise with variance σ^2 for all dimensions, i.e. $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. The likelihood for $\boldsymbol{y}$ given $\boldsymbol{W}, \boldsymbol{\mu}$ and σ^2 takes the form

$$p(\boldsymbol{y}|\boldsymbol{W}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}). \tag{2.2}$$

For N observations, denoted as $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N)^T \in \mathbb{R}^{N\times D}$, equation (2.1) becomes

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}, \tag{2.3}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N)^T \in \mathbb{R}^{N\times Q}$ and $\boldsymbol{\epsilon} \in \mathbb{R}^{N\times D}$. The likelihood of $\boldsymbol{Y}$ given $\boldsymbol{W}$ then becomes

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{W}) &= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{Y}_{n,:}|\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}), \\ \ln p(\boldsymbol{Y}|\boldsymbol{W}) &= \frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{K}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{K}^{-1}\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T), \end{aligned} \tag{2.4}$$

where $\boldsymbol{K} = \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}$ and $\tilde{\boldsymbol{Y}}$ are the centered observations, i.e. $\tilde{\boldsymbol{Y}}_{n,:} = \boldsymbol{Y}_{n,:} - \boldsymbol{\mu}$. This expression can be maximized analytically and yields the solution (Tipping & Bishop, 1999)

$$\begin{aligned} \boldsymbol{\mu}_{\mathrm{ML}} &= \frac{1}{N}\sum_n \boldsymbol{Y}_{n,:}, \\ \boldsymbol{W}_{\mathrm{ML}} &= \boldsymbol{U}\left(\boldsymbol{\Lambda} - \sigma^2\boldsymbol{I}\right)^{1/2}\boldsymbol{R}, \end{aligned} \tag{2.5}$$

where $\boldsymbol{U} \in \mathbb{R}^{D\times Q}$ is an orthogonal matrix containing the principal eigenvectors of the empirical covariance matrix $\frac{1}{N}\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T$, $\boldsymbol{\Lambda} \in \mathbb{R}^{Q\times Q}$ is a diagonal matrix containing the corresponding eigenvalues $(\lambda_1, ..., \lambda_Q)$ of $\frac{1}{N}\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T$ on the diagonal and $\boldsymbol{R} \in \mathbb{R}^{Q\times Q}$ is an arbitrary rotation matrix. The maximum likelihood estimation for σ^2 is given by

$$\sigma^2_{\mathrm{ML}} = \frac{1}{D-Q}\sum_{i=Q+1}^{D}\lambda_i, \tag{2.6}$$

which is basically the variance of $\boldsymbol{Y}$ not picked up by the model. By projecting the data points $\boldsymbol{y}$ into the latent space, a representation of reduced dimensionality $Q < D$ can be obtained. The posterior mean of the latent vectors is given by (Bishop, 2006)

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}] = \left(\boldsymbol{W}_{\text{ML}}^T \boldsymbol{W}_{\text{ML}} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{W}_{\text{ML}}^T (\boldsymbol{y} - \boldsymbol{\mu}_{\text{ML}}) \,. \tag{2.7}$$

Note that the rotation symmetry in the likelihood caused by $\boldsymbol{R}$ in (2.5) makes the likelihood non-unique. Two different solutions $\boldsymbol{W}$ and $\tilde{\boldsymbol{W}} = \boldsymbol{W}\boldsymbol{R}$ will lead to the same likelihood, since $\tilde{\boldsymbol{W}}\tilde{\boldsymbol{W}}^T = \boldsymbol{W}\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{W}^T = \boldsymbol{W}\boldsymbol{W}^T$. For this reason, numerical optimization algorithms often converge to different results when started from different initial conditions. While model predictions are unaffected by this non-uniqueness, interpretation of parameters and the projected latent vectors becomes impossible. Especially the latter is problematic if PPCA is employed as a pre-processing step to reduce the dimensionality of the data and further analysis might be sensitive to the arbitrary choice of latent space rotation. Figure 2.1 (left side) shows the maximum likelihood result for 1000 different initial conditions for $\boldsymbol{W}$. One can clearly see the rotation symmetry.

PPCA is closely related to classical PCA. In particular, as shown by Tipping & Bishop (1999), the classical solution is recovered by letting $\sigma^2 \to 0$. In this case, the dimensionality reduction is achieved by the linear projection

$$\boldsymbol{x} = \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{U}^T (\boldsymbol{y} - \boldsymbol{\mu}_{\text{ML}}) \tag{2.8}$$

where $\boldsymbol{\mu}_{\text{ML}}, \boldsymbol{\Lambda}$ and $\boldsymbol{U}$ are defined as in (2.5). The eigenvalues λ_q are interpreted as the variance explained by the q-th principal component $\boldsymbol{U}_{:,q}$ in this context. From (2.7) the classical solution is obtained by assuming vanishing noise variance σ^2 and fixing the latent space rotation of the maximum likelihood solution at $\boldsymbol{R} = \boldsymbol{I}$.

Factor analysis is closely related to PPCA as well. In this case, PPCA is slightly generalized by assuming that the noise is distributed as $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi})$ with diagonal covariance matrix $\boldsymbol{\Psi}$. In particular, the likelihood has the same rotational symmetry as PPCA. Here, we just note that all our derivations in the following sections apply equally to this model.

2.2. Bayesian Approach to PPCA

The fully Bayesian way to solve (2.4) would be to impose prior distributions $p(\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2)$ and solve for the posterior

$$p(\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2) p(\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2)}{p(\boldsymbol{Y})}, \tag{2.9}$$

which is not tractable anymore. Therefore, we have to resort to other techniques, e.g. sampling the posterior. For

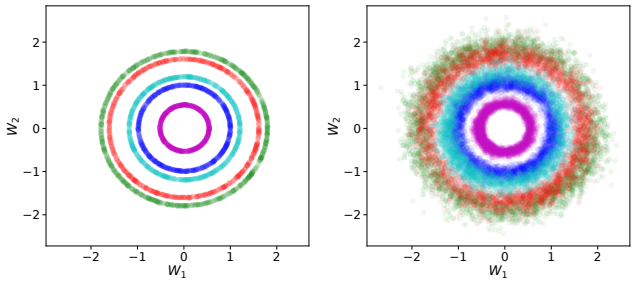

Figure 2.1. Data are created by sampling 50 5-dimensional points from the standard normal distribution and map them via a random linear mapping $\boldsymbol{W} \in \mathbb{R}^{5\times 2}$ to the data-space $\boldsymbol{Y}$. Elements of $\boldsymbol{W}$ are drawn from a standard normal distribution. Left side: Maximum likelihood solution for 1000 runs with different initial conditions of $\boldsymbol{W}$. Right side: 4000 samples from the posterior distribution. Different colors show different rows of $\boldsymbol{W}$.

a rotation invariant prior $p(\boldsymbol{W})$ (e.g. an isotropic Gaussian) the posterior will have the same rotation symmetry as well. Thus, the posterior has a complicated shape with a continuous subspace of equal density posing a considerable challenge for sampling algorithms.

Figure 2.1 (right side) shows 4000 posterior samples. The likelihood is given by (2.4) and the prior on $\boldsymbol{W}$ is a standard normal. We can clearly see the rotation symmetry again. Samples were drawn using the NUTS sampler of Stan (Carpenter et al., 2017), which is able to fully explore the challenging posterior. Yet, in higher dimensions $D > Q \gg 1$ the problem becomes more severe and a lot of samples are required to fully explore the symmetric solutions. Furthermore, statistics of the posterior samples, e.g. computing the mean or variance, do not have any significance in this case.

The question arises: Can we somehow take out the rotation symmetry without changing the model (i.e. likelihood or the mass of the posterior distribution)? The answer is yes: We can achieve this by reparametrizing $\boldsymbol{W}$. That way the results are identified uniquely. The aim is not to change the model by the reparameterization. Therefore, we need specific corresponding distributions for the parameters on the reparameterized model. In the next sections we explain how to do that.

2.3. Singular Value Decomposition

It is well known, that every positive-definite matrix $\boldsymbol{S}$ can be diagonalized by an orthogonal transformation $\boldsymbol{U}$ as $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, where $\boldsymbol{\Lambda}$ contains the eigenvalues of $\boldsymbol{S}$ on the diagonal and the columns of $\boldsymbol{U}$ are the corresponding normalized eigenvectors.

This is not true anymore for other matrices, but there exists a more general decomposition called the singular value

decomposition (SVD). A matrix $\boldsymbol{A} \in \mathbb{R}^{D\times Q}$ can be decomposed as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \tag{2.10}$$

where $\boldsymbol{U} = (\boldsymbol{u}_1, ..., \boldsymbol{u}_D)^T, \boldsymbol{V} = (\boldsymbol{v}_1, ..., \boldsymbol{v}_Q)$ are orthogonal matrices and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, ..., \sigma_Q) \in \mathbb{R}^{D\times Q}$ is a diagonal matrix containing the singular values. When $\boldsymbol{A}$ has rank $P < Q$, only P of them will be non-zero. It is easy to show that the square of the diagonal elements $(\sigma_1^2, ..., \sigma_P^2)$ are the non-zero eigenvalues of $\boldsymbol{A}\boldsymbol{A}^T$ and $\boldsymbol{A}^T\boldsymbol{A}$, which are similar matrices, and all $\boldsymbol{u}_j$'s and $\boldsymbol{v}_j$'s obey the relation

$$\begin{aligned} &\boldsymbol{A}\boldsymbol{v}_j = \sigma_j\boldsymbol{u}_j, \quad \boldsymbol{A}^T\boldsymbol{u}_j = \sigma_j\boldsymbol{v}_j, \quad \forall\, j = 1, ..., P \\ &\boldsymbol{A}\boldsymbol{v}_j = \boldsymbol{0}, \qquad \boldsymbol{A}^T\boldsymbol{u}_j = \boldsymbol{0}, \qquad \forall\, j > P \end{aligned} \tag{2.11}$$

We will decompose our mapping $\boldsymbol{W} \in \mathbb{R}^{D\times Q}$ via SVD as $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$. The likelihood only depends on the outer product of $\boldsymbol{W}$. Since the outer product does not depend on $\boldsymbol{V}$,

$$\boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^T, \tag{2.12}$$

we can neglect $\boldsymbol{V}$ [2], i.e., we assume it to be the identity matrix. That fixes the coordinate frame of the latent space and takes out the rotation symmetry.

By reparameterizing $\boldsymbol{W}$ by $\boldsymbol{U}$ and $\boldsymbol{\Sigma}$, we obtain a unique representation[3]. However, in the fully Bayesian case, we have a prior on $\boldsymbol{W}$. After reparameterization by $\boldsymbol{U}$ and $\boldsymbol{\Sigma}$, we have to make a Jacobian adjustment to the probability density, since otherwise the prior density mass on $\boldsymbol{W}$ will change as well. In the next sections, we consider a zero mean Gaussian prior on $\boldsymbol{W}$, and how to adjust the prior on $\boldsymbol{U}$ and $\boldsymbol{\Sigma}$, such that we do not change the distribution on $\boldsymbol{W}$.

Similarly, SVD provides a numerically stable way for classical PCA. In this case, the data matrix $\boldsymbol{Y}$ is decomposed as $\boldsymbol{Y} = \boldsymbol{U}_{PCA}\boldsymbol{\Sigma}_{PCA}\boldsymbol{V}_{PCA}^T$ and the desired dimensionality reduction is achieved by the linear map $\boldsymbol{\Sigma}_{PCA}\boldsymbol{U}_{PCA}^T$. Again, we see that the matrix of right singular vectors $\boldsymbol{V}_{PCA}$, corresponding to an arbitrary choice of latent coordinate rotation, is dropped.

2.4. Random Matrix Theory, Haar Measure

SVD decomposes our $\boldsymbol{W}$ into an orthogonal matrix $\boldsymbol{U}$, a diagonal matrix $\boldsymbol{\Sigma}$ and another orthogonal matrix $\boldsymbol{V}$, that we are neglecting in our analysis, since the outer product $\boldsymbol{W}\boldsymbol{W}^T$ does not depend on $\boldsymbol{V}$.

[2]Note, $\boldsymbol{V}$ is precisely the rotation symmetry that we want to get rid of.

[3]In general, the SVD is not unique. While the singular values can be ordered to ensure uniqueness of $\boldsymbol{\Sigma}$, the left singular vectors $\boldsymbol{U}$ are only determined up to the sign. As explained later, we enforce a sign convention on them.

For a standard Gaussian prior on $\boldsymbol{W}$, i.e. all elements of $\boldsymbol{W}$ are zero mean Gaussian with unit variance, $\boldsymbol{W}\boldsymbol{W}^T$ has a Wishart distribution and the following theorem holds (James & Lee, 2014):

Theorem 1 *Let the entries of $\boldsymbol{W} \in \mathbb{R}^{D\times Q}$ be i.i.d. Gaussian with zero mean and unit variance. The joint probability density of the ordered strictly positive eigenvalues of the Wishart matrix $\boldsymbol{W}^T\boldsymbol{W}$, $\lambda_1 \geq ... \geq \lambda_Q$, equals*

$$p(\boldsymbol{\lambda}) = ce^{-\frac{1}{2}\sum_{q=1}^{Q}\lambda_q}\prod_{q=1}^{Q}\left(\lambda_q^{\frac{D-Q-1}{2}}\prod_{q'=q+1}^{Q}|\lambda_q - \lambda_{q'}|\right), \tag{2.13}$$

where $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_Q)$ and c is a constant that depends on D and Q.

Note that the non-zero eigenvalues of $\boldsymbol{W}^T\boldsymbol{W}$ are the same as for $\boldsymbol{W}\boldsymbol{W}^T$, and therefore they have the same probability density function. The singular values σ_i of $\boldsymbol{W}$ are the square roots of $\boldsymbol{\lambda}$'s, $\sigma_i = \sqrt{\lambda_i}$. Thus, by change of measure we get[4]

$$p(\sigma_1, ..., \sigma_Q) = ce^{-\frac{1}{2}\sum_{q=1}^{Q}\sigma_q^2}\prod_{q=1}^{Q}\left(\sigma_q^{D-Q-1}\prod_{q'=q+1}^{Q}|\sigma_q^2 - \sigma_{q'}^2|\right)\prod_{q=1}^{Q}2\sigma_q, \tag{2.14}$$

where the last product in the above term is the Jacobian correction. Provided this density function as the prior distribution on the singular value matrix $\boldsymbol{\Sigma}$, we can easily sample from its posterior.

To calculate the prior distribution for $\boldsymbol{U}$, we need to dig deeper into random matrix theory. As mentioned earlier, a standard Gaussian prior on $\boldsymbol{W}$ gives rise to a Wishart distribution on $\boldsymbol{W}\boldsymbol{W}^T$. Equation (2.12) shows, that a Wishart matrix can be decomposed into a product of an orthogonal matrix $\boldsymbol{U}$ and a diagonal matrix $\boldsymbol{\Sigma}$. Furthermore, it is known that the eigenvectors $\boldsymbol{U}$ are distributed uniformly in the space of orthogonal matrices[5]. The set of orthogonal matrices is known as the Stiefel manifold $\mathcal{V}_{Q,D}$

$$\mathcal{V}_{Q,D} = \left\{\boldsymbol{U} \in \mathbb{R}^{D\times Q} | \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}\right\}. \tag{2.15}$$

The dimension of this manifold is $DQ - \frac{1}{2}Q(Q+1)$, accounting for the fact that the orthogonality constraint reduces the number of independent degrees of freedom. The

[4]Given a probability density $p_x(x)$ on x and an invertible map g, so that $y = g(x)$, the density $p_y(y)$ on y is given by $p_y(y) = p_x(g^{-1}(y))\left|\det\left(\frac{dg^{-1}(y)}{dy}\right)\right|$. Here the absolute determinant of the Jacobian accounts for the change of volume under g.

[5]Bai et al. (2007) and Uhlig (1994) mention that the eigenvector matrix of a Wishart distribution is Haar-distributed, which is also true for singular Wishart matrices, but we could not find a proof anywhere.

Stiefel manifold can be equipped with a uniform measure. This measure is an example of a *Haar* measure, as it is invariant under the action of the orthogonal group $O(D) \simeq \mathcal{V}_{D,D}$, i.e.

$$p(\boldsymbol{U}) = p(\boldsymbol{R}\boldsymbol{U}) \quad \forall \boldsymbol{R} \in O(D)\,. \tag{2.16}$$

To proceed, we need to find an unconstrained parameterization for orthogonal matrices along with a density on the parameters, such that the resulting matrix is Haar distributed. Shepard et al. (2014); Pourzanjani et al. (2017) and Mezzadri (2007) suggest ways to do that. The procedure is explained in detail in the next section.

2.5. QR-decomposition and Householder Transformations

Given a matrix $\boldsymbol{Z} \in \mathbb{R}^{D\times Q}$ the (thin) QR-decomposition decomposes $\boldsymbol{Z}$ into an orthogonal matrix $\boldsymbol{Q} \in \mathcal{V}_{Q,D}$ and an upper triangular matrix $\boldsymbol{R} \in \mathbb{R}^{Q\times Q}$. If the elements of $\boldsymbol{Z}$ are i.i.d Gaussian with mean zero and unit variance, $\boldsymbol{Q}$ is Haar distributed (Mezzadri, 2007). Note that this is only the case if a unique QR-decomposition is used. In practice, this is usually achieved by enforcing the convention that the diagonal elements of $\boldsymbol{R}$ are positive.

To compute the QR-decomposition of $\boldsymbol{Z}$, the so-called Householder Transformations $\boldsymbol{H}$ can be used. These transformations are reflections on the plane spanned by a vector $\boldsymbol{v}_n \in \mathbb{R}^n$. To decompose a D-by-Q matrix, Q of such transformations are needed and the resulting orthogonal matrix $\boldsymbol{Q}$ can be written as

$$\boldsymbol{Q} = \boldsymbol{H}_D(\boldsymbol{v}_D)\boldsymbol{H}_{D-1}(\boldsymbol{v}_{D-1})...\boldsymbol{H}_{D-Q+1}(\boldsymbol{v}_{D-Q+1}). \tag{2.17}$$

To construct $\boldsymbol{H}_n$, we define $\tilde{\boldsymbol{H}_n}(\boldsymbol{v}_n)$ as

$$\tilde{\boldsymbol{H}_n}(\boldsymbol{v}_n) = -\text{sgn}(\boldsymbol{v}_{n1})\left(\boldsymbol{I} - 2\boldsymbol{u}_n\boldsymbol{u}_n^T\right) \in \mathbb{R}^{n\times n}, \tag{2.18}$$

where

$$\boldsymbol{u}_n = \frac{\boldsymbol{v}_n + \text{sgn}(\boldsymbol{v}_{n1})\|\boldsymbol{v}_n\|\boldsymbol{e}_1}{\|\boldsymbol{v}_n + \text{sgn}(\boldsymbol{v}_{n1})\|\boldsymbol{v}_n\|\boldsymbol{e}_1\|} \tag{2.19}$$

and construct $\boldsymbol{H}_n$ by

$$\boldsymbol{H}_n = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \tilde{\boldsymbol{H}_n} \end{pmatrix}. \tag{2.20}$$

The algorithm for the QR-decomposition constructs a suitable sequence of vectors $\boldsymbol{v}_D, \ldots, \boldsymbol{v}_{D-Q+1}$ based on the entries of the successively rotated columns of $\boldsymbol{Z}$. The choice of the sign fulfills the requirements mentioned earlier, i.e. the diagonal elements of $\boldsymbol{R}$ are positive.

Here, we are only interested in the resulting orthogonal matrix $\boldsymbol{Q}$. In particular, by randomly drawing vectors $\boldsymbol{v}$, we obtain a random orthogonal matrix via (2.17). The following theorem tells us which distribution the $\boldsymbol{v}$s need to have in order to get an orthogonal matrix distributed according to the Haar measure (Mezzadri, 2007).

Theorem 2 *Let $\boldsymbol{v}_D, \boldsymbol{v}_{D-1}, ..., \boldsymbol{v}_1$ be uniformly distributed on the unit spheres $\mathbb{S}^{D-1}, ..., \mathbb{S}^0$ respectively, where $\mathbb{S}^{n-1}$ is the unit sphere in $\mathbb{R}^n$. Furthermore, let $\boldsymbol{H}_n(\boldsymbol{v}_n)$ be the n-th Householder transformation as defined in equation (2.20) The product*

$$\boldsymbol{Q} = \boldsymbol{H}_D(\boldsymbol{v}_D)\boldsymbol{H}_{D-1}(\boldsymbol{v}_{D-1})...\boldsymbol{H}_1(\boldsymbol{v}_1) \tag{2.21}$$

is a random orthogonal matrix with distribution given by the Haar measure on $O(D)$.

A corresponding draw from the Stiefel manifold $\mathcal{V}_{Q,D}$ can be obtained by just taking the first Q columns of $\boldsymbol{Q}$. Alternatively, this is achieved by drawing vectors $\boldsymbol{v}_D, \ldots, \boldsymbol{v}_{D-Q+1}$ uniformly from the respective unit sphere and construct the unitary matrix from the corresponding Householder transformations. By the theorem above, this gives the Haar measure on $\mathcal{V}_{Q,D}$ as the Householder transformations $\boldsymbol{H}_{D-Q}, \ldots, \boldsymbol{H}_1$ only effect the columns $D-Q, \ldots, 1$ by (2.20).

3. Unique PPCA

Here, we connect the results from the previous section to obtain a model for PPCA without the latent space symmetry that plaques the original formulation. We start by sampling uniformly from the unit spheres $\mathbb{S}^{n-1}$. This is most easily achieved by drawing an i.i.d. standard Gaussian vector of dimension n and normalizing its length. Note that this spends an additional parameter compared to the dimensionality $n-1$ of the unit sphere $\mathbb{S}^{n-1}$. Yet, the n-dimensional standard Gaussian distribution is easy to sample from and the vector length is sufficiently constrained by the Gaussian prior such that the sampler can move around the unit sphere effectively. The same parameterization is employed, for example, by Stan in order to support a `unit_vector` data type (Stan Development Team, 2018). Furthermore, the Householder transformation does not require the vectors $\boldsymbol{v}$ to be of unit length as they are normalized anyways via (2.19). Thus, we construct the orthogonal matrix $\boldsymbol{U}$ distributed with the Haar measure by successively transforming standard Gaussian random vectors.

Next, we sample the ordered singular values from the joint distribution (2.14). Again, this can be accomplished by transforming the ordered vector $(\sigma_1, \ldots, \sigma_Q)$ to an unconstrained space and correcting the joint density by the Jacobian determinant of the transformation. We refer to Stan Development Team (2018) for details of this transformation. Finally, from the orthogonal matrix $\boldsymbol{U}$ and the diagonal matrix $\boldsymbol{\Sigma}$, we construct $\boldsymbol{W}$ by $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}$, which we then use in the likelihood of the PPCA model. Overall, we obtain the

following generative model:

$$
\begin{aligned}
\boldsymbol{v}_D, \ldots, \boldsymbol{v}_{D-Q+1} &\sim \mathcal{N}(0, \boldsymbol{I}) \\
\boldsymbol{\sigma} &\sim p(\boldsymbol{\sigma}) \propto \text{ eq. (2.14)} \\
\boldsymbol{\mu} &\sim p(\boldsymbol{\mu}), \text{e.g. a broad Gaussian} \\
\boldsymbol{U} &= \prod_{q=1}^{Q} \boldsymbol{H}_{D-q+1}(\boldsymbol{v}_{D-q+1}) \\
\boldsymbol{\Sigma} &= \text{diag}(\boldsymbol{\sigma}) \\
\boldsymbol{W} &= \boldsymbol{U}\boldsymbol{\Sigma} \\
\sigma_{\text{noise}} &\sim p(\sigma_{\text{noise}}) \\
\boldsymbol{Y} &\sim \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{Y}_{n,:} | \boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2_{\text{noise}}\boldsymbol{I}\right).
\end{aligned}
$$

Note that this model defines the same distribution as the corresponding model with a standard Gaussian prior $\boldsymbol{W} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. In both cases, the sampling distribution is governed by the Wishart distributed matrix $\boldsymbol{W}\boldsymbol{W}^T$, even though the distribution on $\boldsymbol{W}$ is actually different. We implemented both models in the probabilistic programming language Stan. The code for the simulations is available on Github: https://github.com/RSNirwan/HouseholderBPCA.

Compared to previous approaches based on parameterizing the Stiefel manifold in terms of Givens rotations (Pourzanjani et al., 2017), our model has the following advantages: First, our Householder parameters $\boldsymbol{v}$ are unconstrained, in contrast to the angular parameters of Givens rotations where the sampler might hit the boundary of the space. Secondly, we avoid the computationally demanding computation of the Jacobian determinant (Shepard et al., 2014). Compared to other approaches employing the matrix von Mises-Fisher distribution (Hoff, 2009; Šmídl & Quinn, 2007), our model has the following advantages: First, we do not require conditional conjugacy allowing non-linear extensions of Bayesian PPCA as illustrated in section 4. Secondly, we do not need to resort to rejection sampling or variational approximations. Similar to these approaches, our parameterization introduces a combinatorial symmetry by the sign ambiguity of the SVD. This is akin to the label switching problem in Gaussian mixture models which usually poses few problems with the sampler simply getting stuck in a specific mode. Here, in order to compare results across different modes, we postprocess each sample such that the first entry of each column of $\boldsymbol{U}$ is made positive. In the following, we illustrate our model on some data sets and discuss possible extensions to non-linear models where similar symmetries arise.

3.1. Model Comparison

3.1.1. Synthetic Dataset

Here, we build our own synthetic dataset with known parameters and the goal is to reconstruct the parameter values.

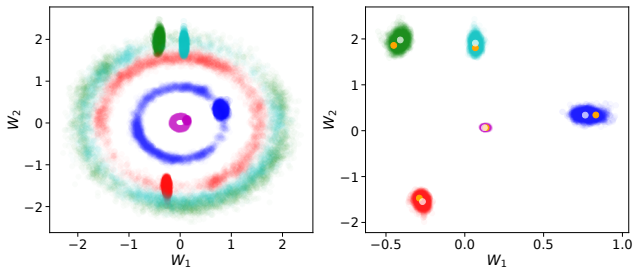

Figure 3.1. Results for synthetic dataset. Left: In the background, we see samples from the posterior of the standard parameterization. Samples from the $\boldsymbol{U}$ and $\boldsymbol{\Sigma}$ mapped to $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}$ are shown in a darker color. Right: Comparison of the suggested model to classical PCA. White dots are the classical PCA solution and orange dots are the true values. 4000 samples for each row of $\boldsymbol{W}$ are shown in different colors.

For $(N, D, Q) = (150, 5, 2)$ we sample $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ from a standard normal distribution and construct $\boldsymbol{W}$ by $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma} \in \mathbb{R}^{D \times Q}$, where $\boldsymbol{U}$ is sampled from the Stiefel manifold with Haar measure and we specify $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \sigma_2)$, where $(\sigma_1, \sigma_2) = (3.0, 1.0)$. Then, we get the observation $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ denotes the noise sampled from a zero mean Gaussian with a standard deviation of 0.01.

The left plot in Figure 3.1 compares the Bayesian inference for the standard model, where we directly assume a standard Gaussian prior on $\boldsymbol{W}$ with our suggested model, where we parameterize $\boldsymbol{W}$ by $\boldsymbol{U}$ and $\boldsymbol{\Sigma}$ and infer the posterior distribution of both. As expected, the standard model has the rotational symmetry in $\boldsymbol{W}$, whereas our model takes out the rotation[6]. On the right side (Figure 3.1), we compare the posterior distribution of our model with the classical PCA solution and as expected with the low observation noise, the solutions are quite similar. In contrast to classical PCA, the Bayesian approach provides a distribution for our parameters, accessing estimation uncertainty of the solution as well.

Figure 3.2 shows the posterior distribution of $\boldsymbol{\Sigma}$, that we fixed to $\text{diag}(3.0, 1.0)$. As we can see, the true values are recovered quite well.

3.1.2. Breast Cancer Wisconsin Dataset

We tested the model on the Breast Cancer Wisconsin dataset as well. The dataset was downloaded from the Python toolbox scikit-learn (Pedregosa et al., 2011) and contains 569 labeled datapoints with 30 features. We neglect the labels and take the standardized 569×30 matrix as the input to our model. For visualization purposes, we again map the data to $Q = 2$. Figure 3.3 shows the performance of different models. In the left plot, we see the posterior samples for the

[6]The additional sign ambiguity of the columns of $\boldsymbol{U}$ is removed afterwards by post processing each sample.

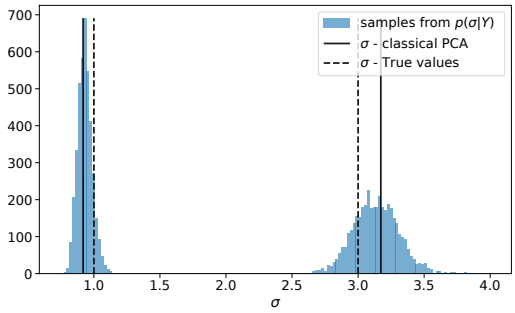

Figure 3.2. Histogram of the samples from the posterior distribution of $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2)$. The vertical solid black lines are the classic PCA solution $(3.17, 0.92)$ and dashed lines are the true values $(3.0, 1.0)$, which generated the data.

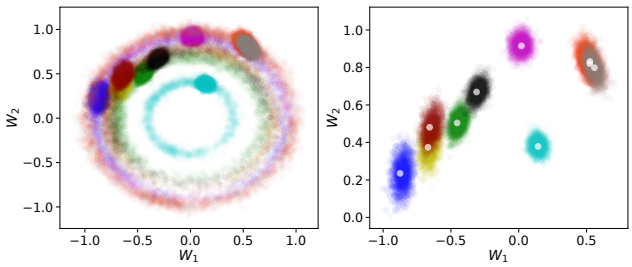

Figure 3.3. Results for the Breast Cancer Wisconsin dataset. Left: Posterior samples of the standard parameterization of $\boldsymbol{W}$ (in the background) and posterior samples of $\boldsymbol{W}$ parameterized by $\boldsymbol{U}$ and $\mathbf{\Sigma}$ (foreground). Right: Posterior samples of our parameterization and white dots are the classical PCA solution. 4000 samples for each row of $\boldsymbol{W}$ are shown in different colors. For clarity, we only show 10 randomly chosen rows instead of all 30.

direct parameterization of $\boldsymbol{W}$ and the parameterization of $\boldsymbol{W}$ by $\boldsymbol{U}$ and $\mathbf{\Sigma}$ and again, our model is able to uniquely identify $\boldsymbol{W}$. The right plot contains the uniquely identified $\boldsymbol{W}$ and the classical PCA solution (white dots). Again, our model enriches the classical PCA solution with uncertainty estimates. Samples were drawn from 4 independent chains which all converged to the same solution.

3.2. Computational Complexity

Compared to probabilistic PCA we compute the DxQ loading matrix $\boldsymbol{W}$ from parameters $\boldsymbol{v}$ for the principal rotation and $\boldsymbol{\sigma}^2$ for the principal variances. This is achieved via a sequence of Q Householder transformations which, as detailed in Shepard et al. (2014), can be achieved in $\mathcal{O}(DQ^2)$ steps. The same complexity applies when computing gradients with respect to the model parameters. Overall, this increases the total computation time of probabilistic PCA by a factor of Q, i.e. the number of latent dimensions, compared to the standard parameterization. The scaling with respect to the number of data points is unchanged.

When sampling from the model, the adaptive NUTS sampler, which is implemented in Stan, needs fewer leap-frog steps per sample because the rotational symmetry is removed from the posterior and does not need to be explored. Numerically, this leads to slightly lower wall clock times overall. From this perspective, our implementation is as efficient as other sampling methods proposed for probabilistic PCA or Bayesian factor analysis. If scalability becomes an issue, our model can easily be used together with more scalable approximation methods such as variational Bayes.

Pourzanjani et al. (2017) proposed an alternative parameterization which removes the rotational symmetry by means of Givens rotations. This approach requires a Jacobian adjustment to account for the change of measure in order to ensure the uniform Haar measure. This requires the computation of the absolute determinant of an $D \cdot Q$ dimensional matrix. Thus, compared to their approach we achieve a substantial speedup as, according to theorem 2, no Jacobian adjustment is required in our model.

3.3. Other than Gaussian Priors

The prior in the paper has been chosen to aid comparison with standard (classical) PCA. Using the SVD, we decompose the prior into a rotation, i.e. the principal axis, and a diagonal matrix, containing the principal components/variances. Thus, the parameters of our model are easily and well interpretable. For the Gaussian prior, which is often chosen in Bayesian factor analysis to support Gibbs sampling, we show that the rotation and variances are independent. Furthermore, the rotation is uniformly distributed according to the Haar measure.

Other priors of this structure can be constructed without issues. In particular, the interpretation of the variance parameters $\boldsymbol{\sigma}^2$ as the principal components helps to this end. For instance, automatic relevance determination is readily accomplished by putting a shrinkage/sparsity prior on the principle components. Interestingly, such a prior would be very different from priors arising from imposing sparsity on the factor loadings as suggested in sparse Bayesian factor analysis (Bhattacharya & Dunson, 2011). In this case, the induced distribution on the principal components is still sparse - at least numerically, as we are not aware of any theoretical results. More importantly though, the induced rotation is not distributed uniformly anymore. Thus, such a prior gives rise to an implicit a-priori preference for certain principal axes. Note that this does not break the rotational symmetry of the model which arises from the rotation of the latent space. Instead it imposes a preference for certain directions in data space which, as we believe, is actually undesirable in many cases. We view it as a strength of our model that interpretable priors can be choosen independently for the rotation and variances.

4. Extension to Non-linear Models

4.1. Gaussian Process Latent Variable Model

Since our method works well for the linear PPCA model, the question arises, whether we could as well improve the posterior sampling of non-linear dimensionality reduction techniques. Lawrence (2005) introduced an extension of PPCA called the Gaussian Process Latent Variable Model (GP-LVM). The model starts with the dual formulation of PPCA. In this case, instead of marginalizing over $\boldsymbol{X}$ in equation (2.3), we marginalize over $\boldsymbol{W} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The resulting likelihood then takes the form

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{d=1}^{D} \mathcal{N}(\boldsymbol{Y}_{:,d}|\boldsymbol{\mu}, \boldsymbol{K} + \sigma^2 I), \tag{4.1}$$

where $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^T$ and $\boldsymbol{K}_{ij} = \boldsymbol{X}_{i,:}^T \boldsymbol{X}_{j,:}$. Identifying this with the covariance kernel of Gaussian process (Rasmussen, 2006), the model is generalized by replacing the linear kernel $\boldsymbol{K}_{ij}$ with a non-linear one. In this way, a non-linear dimensionality reduction method is obtained, the GP-LVM. One of the most used covariance kernels is the exponentiated quadratic kernel (Rasmussen, 2006)

$$k_{\text{SE}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_{\text{SE}}^2 \exp(-0.5||\boldsymbol{x} - \boldsymbol{x}'||_2^2 / l^2), \tag{4.2}$$

where σ_{SE}^2 is the kernel variance and l the lengthscale. We will use this covariance function in our analysis as well.

As suggested by Lawrence (2005), we can optimize equation (4.1) $\log p(\boldsymbol{Y}|\boldsymbol{X})$ with respect to the latent positions and the hyperparameters. Optimization can easily lead to overfitting and a fully Bayesian treatment of the model would be preferable. Therefore, we need a prior distribution $p(\boldsymbol{X})$ on $\boldsymbol{X}$ and using Bayes rule, the posterior is given as $p(\boldsymbol{X}|\boldsymbol{Y}) = p(\boldsymbol{Y}|\boldsymbol{X})p(\boldsymbol{X})/p(\boldsymbol{Y})$. As in the case of PPCA the posterior is not analytically tractable and, as before, we resort to sampling in order to approximate it. Note that the kernel k_{SE} only depends on the distance between points $\boldsymbol{x}$ and $\boldsymbol{x'}$ and thus the model likelihood is again invariant under rotations of the latent space $\boldsymbol{X}$.

4.2. Model Comparison

To test our model in the non-linear case, we again take the Breast Cancer Wisconsin dataset. For the GP-LVM the input is the transposed matrix $\boldsymbol{Y} \in \mathbb{R}^{N \times D}$, where $N = 30$ and $D = 569$. We standardized the data and set σ_{SE} and l to one and only sample the latent space. First, we fit the data with the standard parameterization, where we directly have the $\boldsymbol{X}$ as parameters and then we reparameterize $\boldsymbol{X}$ by $\boldsymbol{U}$ and $\boldsymbol{\Sigma}$ and sample $\boldsymbol{v}$ and $\boldsymbol{\sigma}$ as described in sections 2.4 and 2.5.

Figure 4.1 shows the results for the exponentiated quadratic kernel for $Q = 2$. As before, we used Stan to sample from the posterior on three independent chains. The top row shows the samples from the different chains for the standard parameterization. One can clearly see the rotational symmetry. The bottom row shows the samples with our suggested parameterization. As expected, there is no rotational symmetry anymore. However, due to the models complexity/flexibility many local minima arise and the chains with both (the standard and our suggested) parameterizations do not converge to the same posterior.

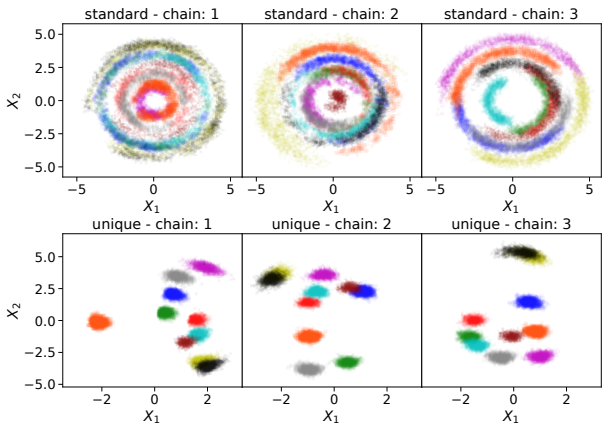

Figure 4.1. Posterior samples from for the GP-LVM. The top row shows the results for the standard parameterization for $\boldsymbol{X}$ and the bottom row shows the results of our suggested parameterization. 2000 samples per chain for each row of $\boldsymbol{X}$ are shown in different colors. Again, for clarity, we only show 10 randomly chosen rows instead of all 30.

5. Conclusion

We suggested a new parameterization for the mapping $\boldsymbol{W}$ in PPCA taking latent to observed points, which uniquely identifies the principal components even though the likelihood and the posterior (for the Bayesian case) are rotationally symmetric. We have shown how to parameterize the model via the singular vectors and values of $\boldsymbol{W}$ and how to set the prior on the new parameters such that the model is not changed compared to a standard Gaussian prior on $\boldsymbol{W}$ directly. Furthermore, we provided an efficient implementation via Householder transformations.

Thereafter, we tested the model on a synthetic and the Breast Cancer Wisconsin dataset. Our model was able to uniquely reconstruct the true parameters that created the synthetic data. On both datasets our model provided similar results as classical PCA, additionally containing the uncertainty of the estimates as well. In the end, we showed that our method can also be used to remove the symmetry in the latent space of a non-linear GP-LVM model. Overall, we believe that our approach, thanks to its known prior distribution and computational efficiency, can be quite useful for latent space models with rotation symmetric likelihoods.

Acknowledgements

The authors thank Dr. h.c. Maucher for funding their positions.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

Bai, Z. D., Miao, B. Q., and Pan, G. M. On asymptotics of eigenvectors of large sample covariance matrix. *Ann. Probab.*, 35(4):1532–1572, 07 2007.

Betancourt, M. A Conceptual Introduction to Hamiltonian Monte Carlo. *ArXiv e-prints*, January 2017.

Bhattacharya, A. and Dunson, D. B. Sparse bayesian infinite factor models. *Biometrika*, 98 2:291–306, 2011.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, New York, 1st edition, 2006.

Byrne, S. and Girolami, M. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40 (4):825–845, 2013.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32, 2017. ISSN 1548-7660.

Hoff, P. D. 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.

James, O. and Lee, H.-N. Concise Probability Distributions of Eigenvalues of Real-Valued Wishart Matrices. *ArXiv e-prints*, February 2014.

Jöreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2): 183–202, Jun 1969. ISSN 1860-0980.

Lawrence, N. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *J. Mach. Learn. Res.*, 6:1783–1816, December 2005. ISSN 1532-4435.

Mezzadri, F. How to generate random matrices from the classical compact groups. *Notices of the American Mathematical Society*, 54(5):592 – 604, 5 2007. ISSN 0002-9920.

Millsap, R. E. When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 8(1):1–17, 2001.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Peeters, C. F. W. Rotational uniqueness conditions under oblique factor correlation metric. *Psychometrika*, 77(2): 288–292, Apr 2012. ISSN 1860-0980.

Pourzanjani, A. A., Jiang, R. M., Mitchell, B., Atzberger, P. J., and Petzold, L. R. General Bayesian Inference over the Stiefel Manifold via the Givens Transform. *ArXiv e-prints*, October 2017.

Rasmussen, C. E. *Gaussian processes for machine learning*. MIT Press, 2006.

Shepard, R., Gidofalvi, G., and Brozell, S. R. The multifacet graphically contracted function method. II. a general procedure for the parameterization of orthogonal matrices and its application to arc factors. *The Journal of Chemical Physics*, 141(6):064106, 2014.

Stan Development Team. Stan modeling language users guide and reference manual, version 2.18.0, 2018.

Tipping, M. E. and Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61:611–622, 1999.

Uhlig, H. On singular wishart and singular multivariate beta distributions. *Ann. Statist.*, 22(1):395–405, 03 1994.

Šmídl, V. and Quinn, A. On bayesian principal component analysis. *Computational Statistics & Data Analysis*, 51 (9):4101–4123, 2007.