Rotation Invariant Householder Parameterization for Bayesian PCA

Rajbir-Singh Nirwan, Nils Bertschinger

June 11, 2019

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

FIAS Frankfurt Institute for Advanced Studies

Outline

- Probabilistic PCA (PPCA)
- Non-identifiability issue of PPCA
- Conceptual solution to the problem
- Implementation
- Results

Probabilistic PCA

- Classical PCA

  Formulated as a projection from data space ***Y*** to a lower dimensional latent space ***X***

  $$\boldsymbol{Y} \in \mathbb{R}^{N \times D} \quad \rightarrow \quad \boldsymbol{X} \in \mathbb{R}^{N \times Q}$$

  Latent space: maximizes variance of projected data, minimizes MSE

Probabilistic PCA

- Classical PCA

  Formulated as a projection from data space ***Y*** to a lower dimensional latent space ***X***

$$\boldsymbol{Y} \in \mathbb{R}^{N \times D} \quad \rightarrow \quad \boldsymbol{X} \in \mathbb{R}^{N \times Q}$$

  Latent space: maximizes variance of projected data, minimizes MSE

- Probabilistic PCA (PPCA)

  Viewed as a generative model, that maps the latent space ***X*** to the data space ***Y***

$$\boldsymbol{X} \in \mathbb{R}^{N \times Q} \quad \rightarrow \quad \boldsymbol{Y} \in \mathbb{R}^{N \times D}$$

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

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad \epsilon \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right)$$

$$p(\boldsymbol{Y} \mid \boldsymbol{W}) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{Y}_{n,:} \mid \boldsymbol{0}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}\right)$$

$$\boldsymbol{W}\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{W}^T = \boldsymbol{W}\boldsymbol{W}^T \quad \forall\ \boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}$$

Probabilistic PCA

- Classical PCA

  Formulated as a projection from data space ***Y*** to a lower dimensional latent space ***X***

$$\boldsymbol{Y} \in \mathbb{R}^{N \times D} \quad \rightarrow \quad \boldsymbol{X} \in \mathbb{R}^{N \times Q}$$

  Latent space: maximizes variance of projected data, minimizes MSE

- Probabilistic PCA (PPCA)

  Viewed as a generative model, that maps the latent space ***X*** to the data space ***Y***

$$\boldsymbol{X} \in \mathbb{R}^{N \times Q} \quad \rightarrow \quad \boldsymbol{Y} \in \mathbb{R}^{N \times D}$$

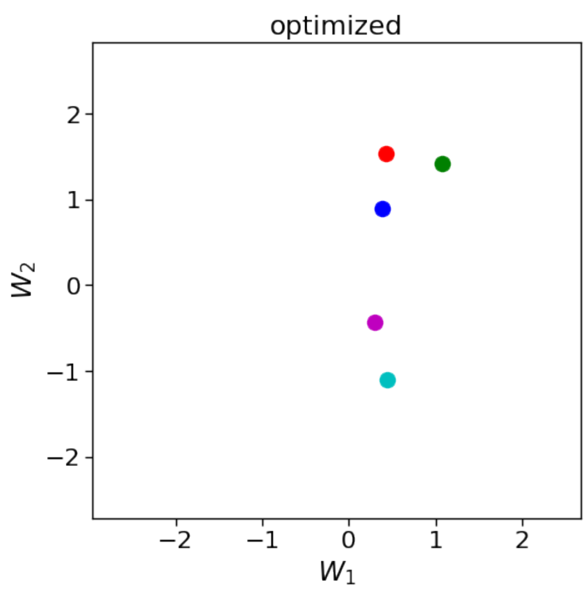
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$$

$$\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \quad \epsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right)$$

$$p(\boldsymbol{Y} \mid \boldsymbol{W}) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{Y}_{n,:} \mid \mathbf{0}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}\right)$$

$$\boldsymbol{W}\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{W}^T = \boldsymbol{W}\boldsymbol{W}^T \quad \forall\ \boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}$$

- Optimization for D=5, Q=2



Probabilistic PCA

- Classical PCA

  Formulated as a projection from data space ***Y*** to a lower dimensional latent space ***X***

$$\boldsymbol{Y} \in \mathbb{R}^{N \times D} \quad \rightarrow \quad \boldsymbol{X} \in \mathbb{R}^{N \times Q}$$

  Latent space: maximizes variance of projected data, minimizes MSE

- Probabilistic PCA (PPCA)

  Viewed as a generative model, that maps the latent space ***X*** to the data space ***Y***

$$\boldsymbol{X} \in \mathbb{R}^{N \times Q} \quad \rightarrow \quad \boldsymbol{Y} \in \mathbb{R}^{N \times D}$$

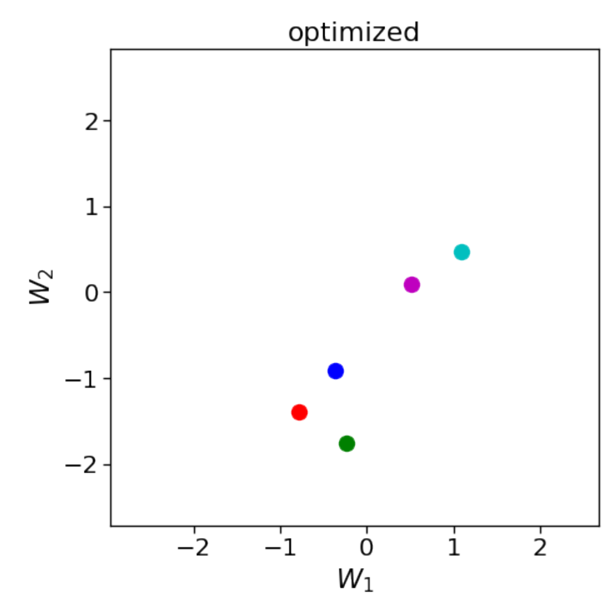
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$$

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad \epsilon \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{I}\right)$$

$$p(\boldsymbol{Y} \mid \boldsymbol{W}) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{Y}_{n,:} \mid \boldsymbol{0}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}\right)$$

$$\boldsymbol{W}\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{W}^T = \boldsymbol{W}\boldsymbol{W}^T \quad \forall\ \boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}$$

- Optimization for D=5, Q=2



Probabilistic PCA

- Classical PCA

  Formulated as a projection from data space ***Y*** to a lower dimensional latent space ***X***

$$\boldsymbol{Y} \in \mathbb{R}^{N \times D} \quad \rightarrow \quad \boldsymbol{X} \in \mathbb{R}^{N \times Q}$$

  Latent space: maximizes variance of projected data, minimizes MSE

- Probabilistic PCA (PPCA)

  Viewed as a generative model, that maps the latent space ***X*** to the data space ***Y***

$$\boldsymbol{X} \in \mathbb{R}^{N \times Q} \quad \rightarrow \quad \boldsymbol{Y} \in \mathbb{R}^{N \times D}$$

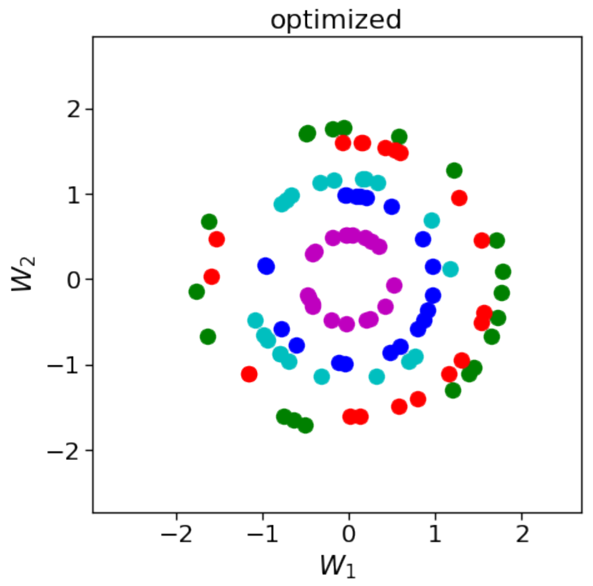
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$$

$$\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \quad \epsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right)$$

$$p(\boldsymbol{Y} \mid \boldsymbol{W}) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{Y}_{n,:} \mid \mathbf{0}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}\right)$$

$$\boldsymbol{W}\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{W}^T = \boldsymbol{W}\boldsymbol{W}^T \quad \forall\ \boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}$$

- Optimization for D=5, Q=2



Probabilistic PCA

- Classical PCA

Formulated as a projection from data space ***Y*** to a lower dimensional latent space ***X***

$$\boldsymbol{Y} \in \mathbb{R}^{N \times D} \quad \rightarrow \quad \boldsymbol{X} \in \mathbb{R}^{N \times Q}$$

Latent space: maximizes variance of projected data, minimizes MSE

- Probabilistic PCA (PPCA)

Viewed as a generative model, that maps the latent space ***X*** to the data space ***Y***

$$\boldsymbol{X} \in \mathbb{R}^{N \times Q} \quad \rightarrow \quad \boldsymbol{Y} \in \mathbb{R}^{N \times D}$$

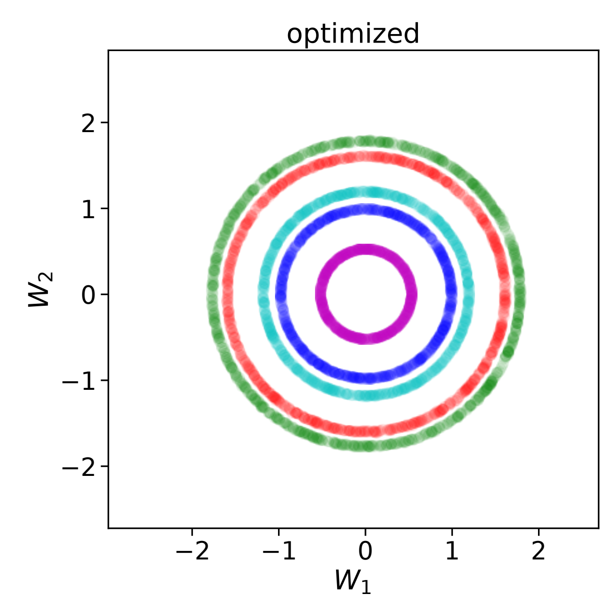
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$$

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad \epsilon \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right)$$

$$p(\boldsymbol{Y} \mid \boldsymbol{W}) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{Y}_{n,:} \mid \boldsymbol{0}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}\right)$$

$$\boldsymbol{W}\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{W}^T = \boldsymbol{W}\boldsymbol{W}^T \quad \forall\ \boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}$$

- Rotation invariant likelihood



Bayesian approach to PPCA

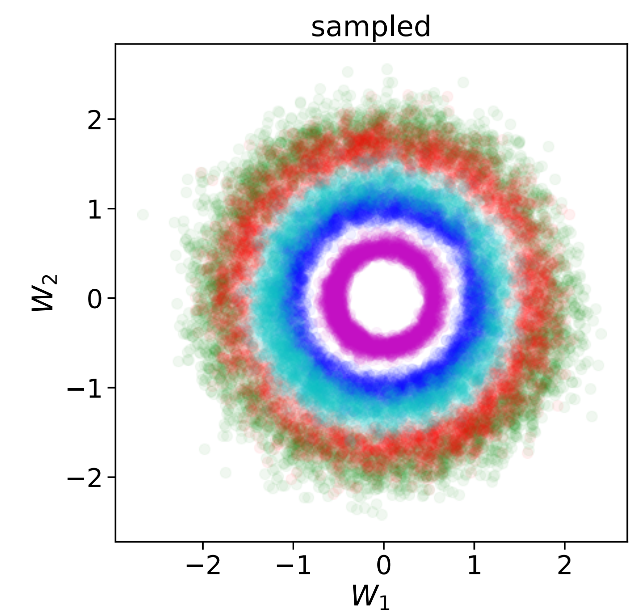
$$p\left(\boldsymbol{W} \mid \boldsymbol{Y}\right) = \frac{p\left(\boldsymbol{Y} \mid \boldsymbol{W}\right) p\left(\boldsymbol{W}\right)}{p(\boldsymbol{Y})}$$

- If prior does not break the symmetry, posterior will be rotation invariant as well

- Sampling will be challenging, posterior averages are meaningless and the interpretation of the latent space is almost impossible

Bayesian approach to PPCA

$$p\left(\boldsymbol{W} \mid \boldsymbol{Y}\right) = \frac{p\left(\boldsymbol{Y} \mid \boldsymbol{W}\right) p\left(\boldsymbol{W}\right)}{p(\boldsymbol{Y})}$$



- If prior does not break the symmetry, posterior will be rotation invariant as well
- Sampling will be challenging, posterior averages are meaningless and the interpretation of the latent space is almost impossible

Solution

- Find different parameterization of the model, such that the probabilistic model is not changed

Outline of procedure

- SVD of W $\quad \boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\right)^T = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^T$
- Fix coordinate system $\quad \boldsymbol{V} = \boldsymbol{I}$
- Specify correct prior $\quad p(\boldsymbol{U}, \boldsymbol{\Sigma})$
- Sample from $\quad p(\boldsymbol{U}, \boldsymbol{\Sigma} \mid \boldsymbol{Y})$

Solution

- Find different parameterization of the model, such that the probabilistic model is not changed

Outline of procedure

- SVD of W $\qquad \boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \left(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\right)^T = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^T$
- Fix coordinate system $\qquad \boldsymbol{V} = \boldsymbol{I}$
- Specify correct prior $\qquad p(\boldsymbol{U}, \boldsymbol{\Sigma})$
- Sample from $\qquad p(\boldsymbol{U}, \boldsymbol{\Sigma} \,|\, \boldsymbol{Y})$

$\boldsymbol{W} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \quad \rightarrow \quad \boldsymbol{W}\boldsymbol{W}^T$ **is Wishart distributed**

$\boldsymbol{U} \sim ?$
$\boldsymbol{\Sigma} \sim ? \quad \rightarrow \quad \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T$ **is Wishart distributed**

$$\begin{matrix} \boldsymbol{U} \sim ? \\ \boldsymbol{\Sigma} \sim ? \end{matrix} \rightarrow \boldsymbol{U\Sigma\Sigma}^T\boldsymbol{U}^T$$ **Wishart**

Theory

- Since ***U***, ***Σ*** is SVD of ***W*** and ***U***, ***Σ***^{2} is eigenvalue decomposition of ***WW***$^\mathrm{T}$ → ***U*** is eigenvector matrix

$\boldsymbol{U} \in \mathscr{V}_{Q,D}$ **Stiefel manifold** $\mathscr{V}_{Q,D} = \left\{\boldsymbol{U} \in \mathbb{R}^{D\times Q} \,|\, \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}\right\}$

Eigenvectors of Wishart matrix are distributed uniformly in space of orthogonal matrices (Blai (2007), Uhlig (1994))

→ ***U* is uniformly distributed on the Stiefel manifold**

$$\begin{matrix} \boldsymbol{U} \sim ? \\ \boldsymbol{\Sigma} \sim ? \end{matrix} \;\rightarrow\; \boldsymbol{U\Sigma\Sigma}^T\boldsymbol{U}^T$$ **wishart**

Theory

- Since ***U***, ***Σ*** is SVD of ***W*** and ***U***, ***Σ***^{2} is eigenvalue decomposition of ***WW***$^\mathrm{T}$ → ***U*** is eigenvector matrix

$\boldsymbol{U} \in \mathscr{V}_{Q,D}$ **Stiefel manifold** $\mathscr{V}_{Q,D} = \left\{ \boldsymbol{U} \in \mathbb{R}^{D\times Q} \,|\, \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I} \right\}$

Eigenvectors of Wishart matrix are distributed uniformly in space of orthogonal matrices (Blai (2007), Uhlig (1994))

→ *U* is uniformly distributed on the Stiefel manifold

- Square of ordered eigenvalue matrix ***Σ*** is distributed as (James & Lee (2014))

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

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

Implementation

- Need: $\boldsymbol{U} \sim$ **uniform on Stiefel** $\mathscr{V}_{Q,D}$
  $\boldsymbol{\Sigma} \sim p(\boldsymbol{\Sigma}) \leftarrow$ **easy, since we know the analytic exp for density**

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 $O(D)$.

Mezzadri (2007)

How to uniformly sample $\boldsymbol{U}$ on $\mathscr{V}_{Q,D}$

for $n = D : 1$

$\boldsymbol{v}_n \sim$ **uniform on** $\mathbb{S}^{n-1}$

$$\boldsymbol{u}_n = \frac{\boldsymbol{v}_n + \operatorname{sgn}(\boldsymbol{v}_{n1}) \, \| \boldsymbol{v}_n \| \, \boldsymbol{e}_1}{\| \boldsymbol{v}_n + \operatorname{sgn}(\boldsymbol{v}_{n1}) \, \| \boldsymbol{v}_n \| \, \boldsymbol{e}_1 \|}$$

$$\tilde{\boldsymbol{H}}_{\boldsymbol{n}}(\boldsymbol{v}_n) = -\operatorname{sgn}(\boldsymbol{v}_{n1}) \left(\boldsymbol{I} - 2\boldsymbol{u}_n\boldsymbol{u}_n^T\right)$$

$$\boldsymbol{H}_n = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \tilde{\boldsymbol{H}}_{\boldsymbol{n}} \end{pmatrix}$$

$$\boldsymbol{U} = \boldsymbol{H}_D(\boldsymbol{v}_D)\,\boldsymbol{H}_{D-1}(\boldsymbol{v}_{D-1})\ldots\boldsymbol{H}_1(\boldsymbol{v}_1)$$

Implementation

The full generative model for Bayesian PPCA:

$$\boldsymbol{v}_D, \ldots, \boldsymbol{v}_{D-Q+1} \sim \mathcal{N}(0, \boldsymbol{I})$$
$$\boldsymbol{\sigma} \sim p(\boldsymbol{\sigma})$$
$$\boldsymbol{\mu} \sim p(\boldsymbol{\mu})$$
$$\boldsymbol{U} = \prod_{q=1}^{Q} \boldsymbol{H}_{D-q+1}\left(\boldsymbol{v}_{D-q+1}\right)$$
$$\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma})$$
$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}$$
$$\sigma_{\textbf{noise}} \sim p\left(\sigma_{\textbf{noise}}\right)$$
$$Y \sim \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{Y}_{n,:} \,|\, \boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2_{\textbf{noise}}\boldsymbol{I}\right)$$

Results

Synthetic Dataset

- Construction

  $(N, D, Q) = (150{,}5{,}2)$

$$X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \in \mathbb{R}^{N \times Q}$$

$$\boldsymbol{U} \sim \textbf{uniform on Stiefel } \mathscr{V}_{Q,D}$$

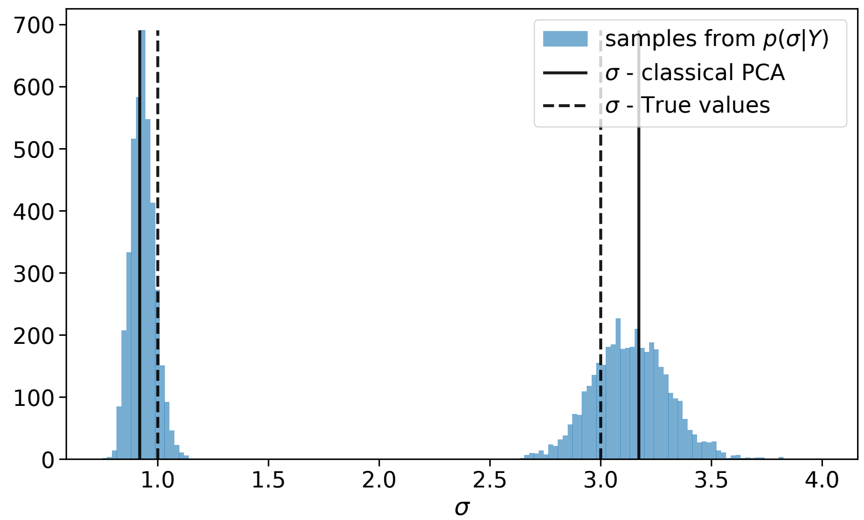
$$\boldsymbol{\epsilon} \sim \mathcal{N}(0,\ 0.01) \in \mathbb{R}^{N \times D}$$

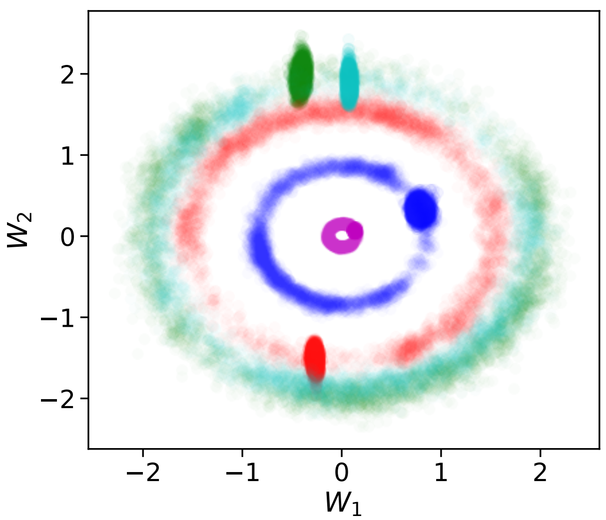
$$\boldsymbol{\Sigma} = \mathrm{diag}\left(\sigma_1, \sigma_2\right) = \mathrm{diag}\,(3.0,\ 1.0)$$

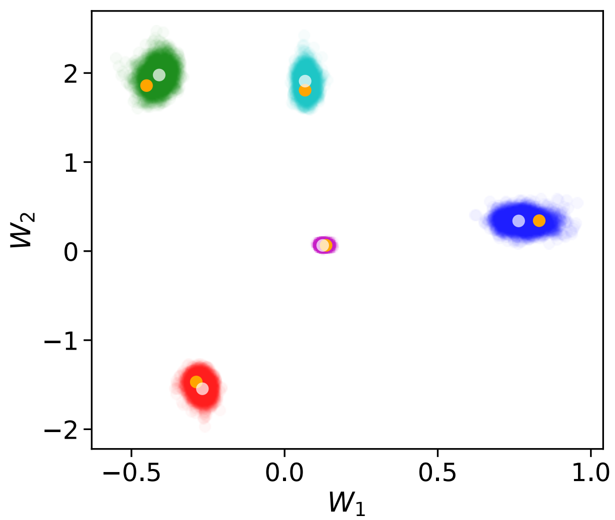
$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma} \in \mathbb{R}^{D \times Q}$$

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

- Inference



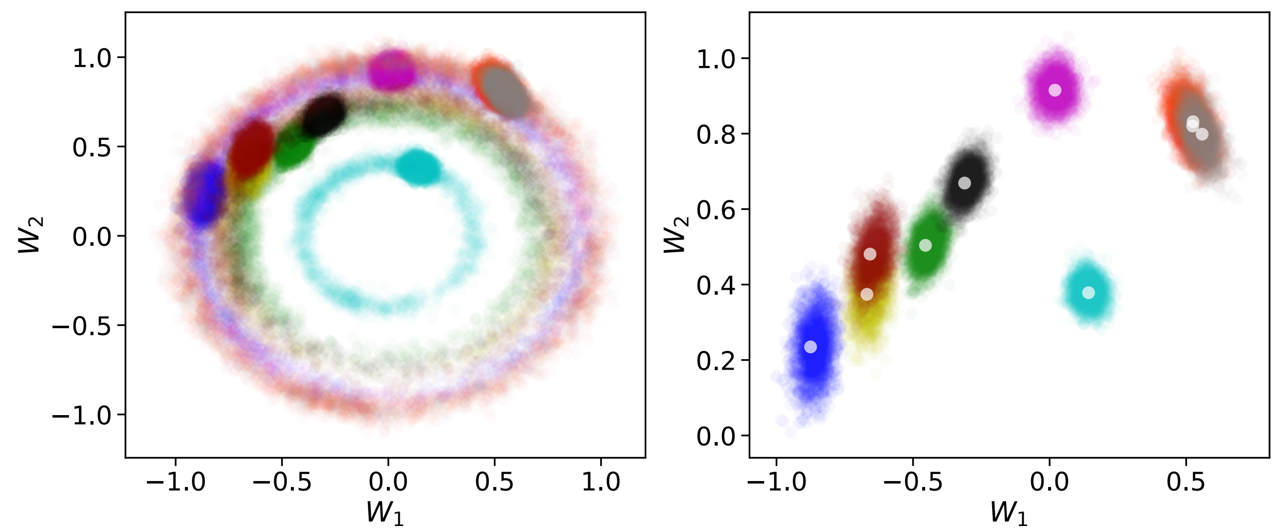




Results

Breast Cancer Wisconsin Dataset $(N, D) = (569, 30)$

- Bayesian PCA



- Advantages
  - Breaks the rotation symmetry without changing the probabilistic model
  - Enrichment of the classical PCA solution with uncertainty estimates
  - Decomposition of prior into rotation and principle variances
    - Allows to construct other priors without issues
    - Sparsity prior on principle variances without a-priori rotation preference
    - If desired a-priori rotation preference without affecting the variances

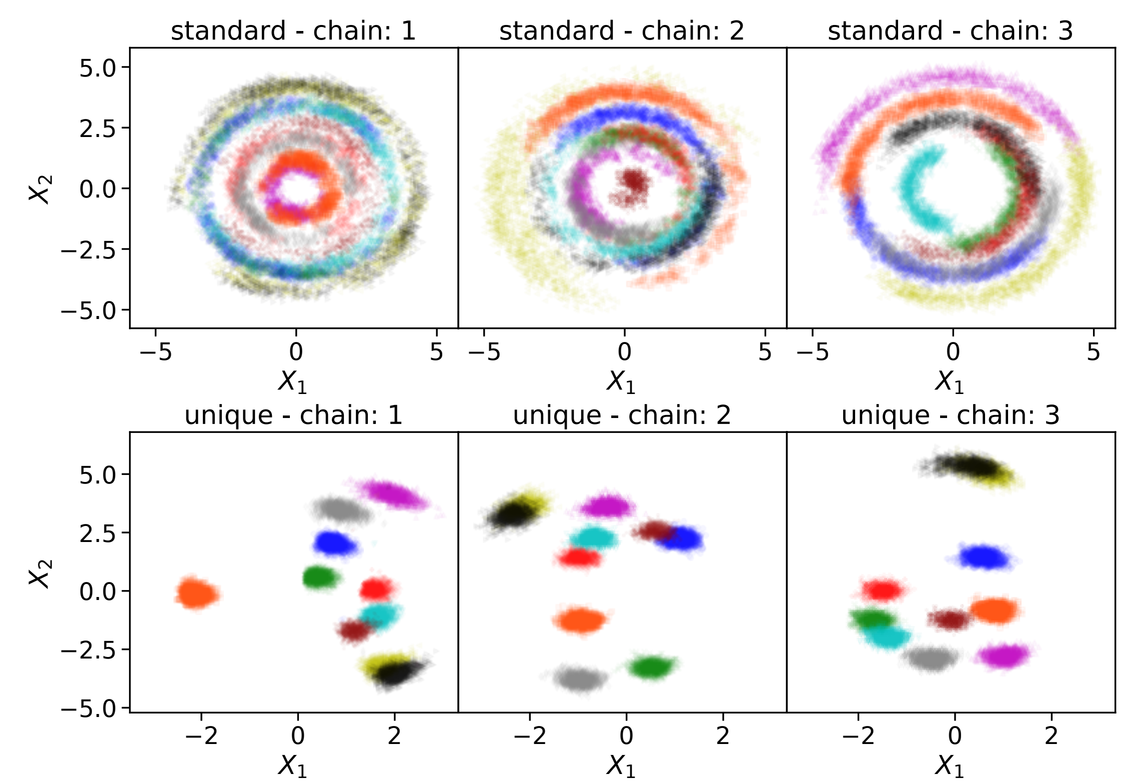
Extension to non-linear models

- GPLVM with the same rotation invariant problem

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

$$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^T, \quad \boldsymbol{K}_{ij} = \boldsymbol{X}_{i,:}^T \boldsymbol{X}_{j,:} = k\left(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:}\right)$$

$$k_{\mathrm{SE}}\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \sigma_{\mathrm{SE}}^2 \exp\left(-0.5 \,\|\, \boldsymbol{x} - \boldsymbol{x}' \,\|_2^2 / l^2\right)$$



- No rotation symmetry in the posterior for the suggested parameterization
- Different chains converge to different solutions due to increased model complexity

Conclusion

- Suggested new parameterization for ***W*** in PPCA, which uniquely identifies principle components even though the likelihood and the posterior are rotationally symmetric
- Showed how to set the prior on the new parameters such that the model is not changed compared to a standard Gaussian prior on ***W***
- Provided an efficient implementation via Householder transformations (no Jacobian correction needed)
- New parameterization allows for other interpretable priors on rotation and principle variances
- Extended to non-linear models and successfully solved the rotation problem there as well

Poster session: #235
Github: https://github.com/RSNirwan/HouseholderBPCA

Thanks for your attention!

Supervisor: Prof. Dr. Nils Bertschinger
Funder: Dr. h. c. Helmut O. Maucher