Rotation Invariant Householder Parameterization for Bayesian PCA

Rajbir-Singh Nirwan

Sept 04, 2019

Outline

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

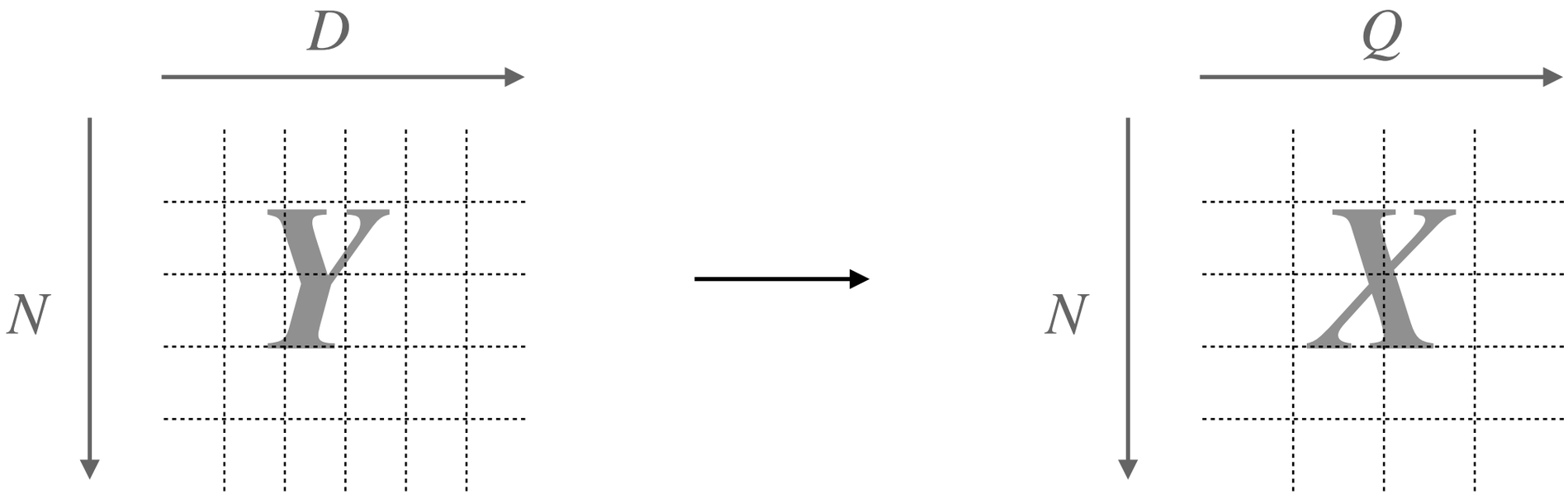
Probabilistic PCA

- Classical PCA

  Formulated as a projection from data space $\boldsymbol{Y}$ to a lower dimensional latent space $\boldsymbol{X}$

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

  Latent space: maximizes variance of projected data, minimizes MSE



Probabilistic PCA

- Classical PCA

  Formulated as a projection from data space $\boldsymbol{Y}$ to a lower dimensional latent space $\boldsymbol{X}$

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

  Latent space: maximizes variance of projected data, minimizes MSE

- Probabilistic PCA (PPCA)

  Viewed as a generative model, that maps the latent space $\boldsymbol{X}$ to the data space $\boldsymbol{Y}$

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

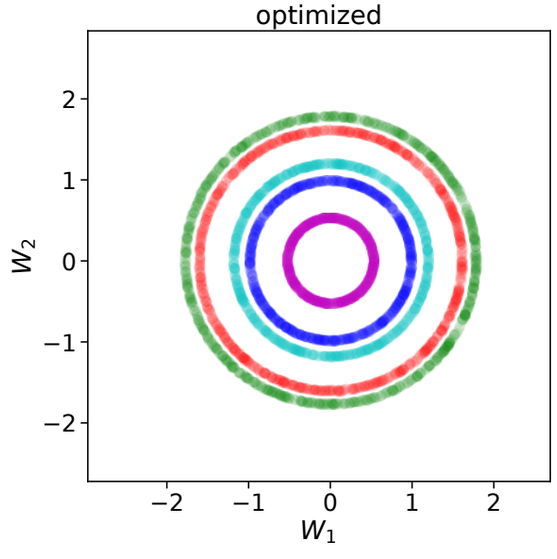
  $$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$$

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

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

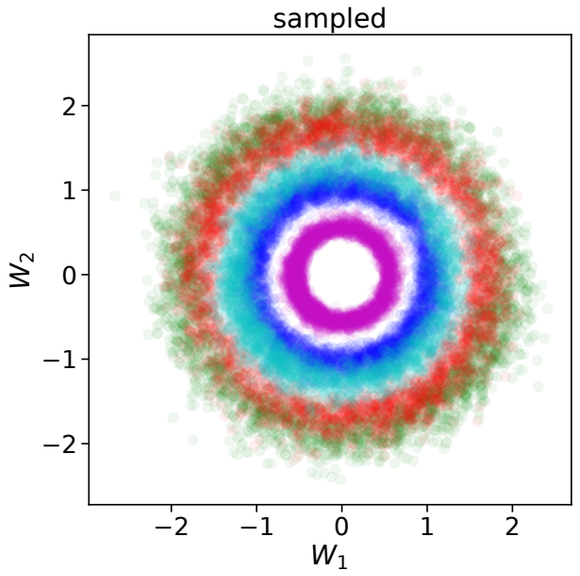
  $$\boldsymbol{W}\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{W}^T = \boldsymbol{W}\boldsymbol{W}^T \quad \forall \; \boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}$$

- Rotation invariant likelihood



Bayesian approach to PPCA

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



- 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 $\quad p(\boldsymbol{U}, \boldsymbol{\Sigma} \mid \boldsymbol{Y})$

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

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

$U \sim ?$
$\Sigma \sim ?$ $\rightarrow$ $U\Sigma\Sigma^T U^T$ **Wishart**

Theory

- Since $\boldsymbol{U}, \boldsymbol{\Sigma}$ is SVD of $\boldsymbol{W} \rightarrow$ $\boldsymbol{U}$ is a orthogonal matrix

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

- $\boldsymbol{U}, \boldsymbol{\Sigma}^2$ is eigenvalue decomposition of $\boldsymbol{W}\boldsymbol{W}^T$

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

→ U is uniformly distributed on the Stiefel manifold

- Distribution of the ordered eigenvalue matrix $\boldsymbol{\Sigma}^2$ of $\boldsymbol{W}\boldsymbol{W}^T$ is (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:

$U \sim$ **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*

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

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

Mezzadri (2007)

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

for $n = D : 1$

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

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

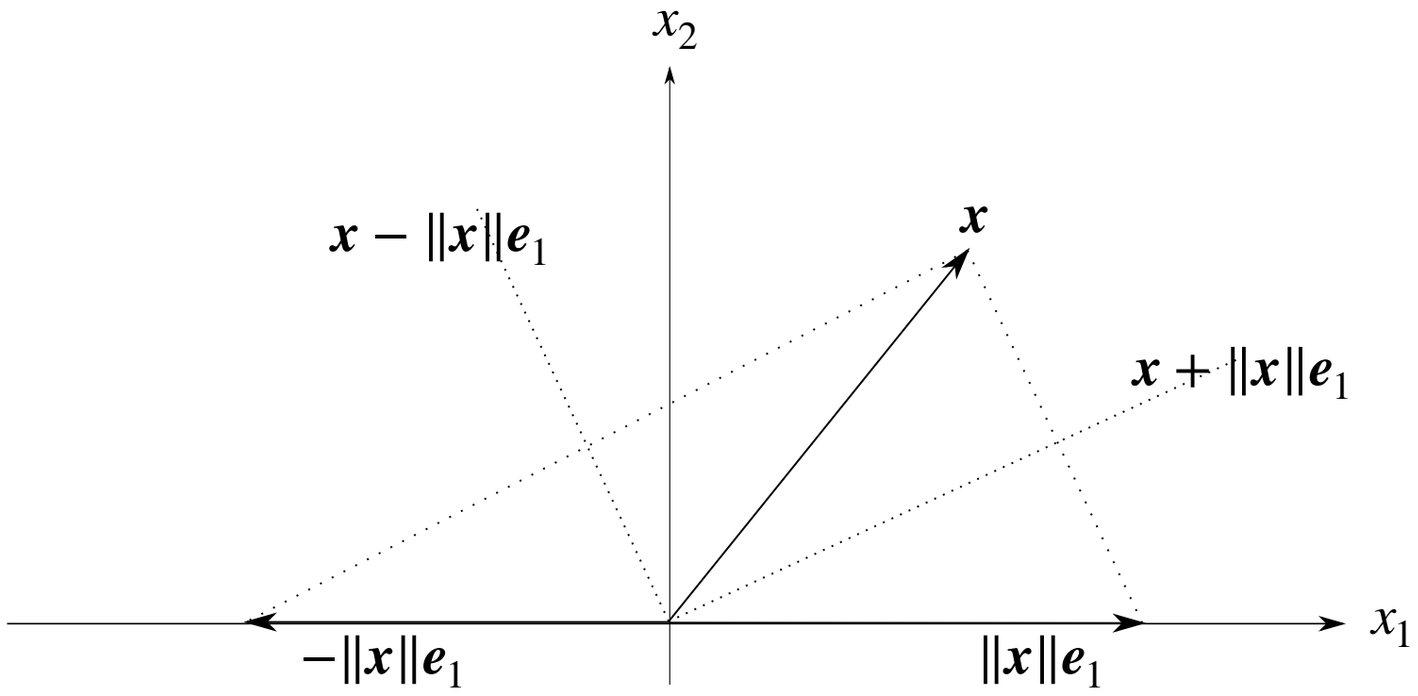
$$\tilde{\boldsymbol{H}}_{\boldsymbol{n}}(\boldsymbol{v}_n) = -\operatorname{sgn}(\boldsymbol{v}_{n1})\left(\boldsymbol{I} - 2\boldsymbol{u}_n\boldsymbol{u}_n^T\right)$$

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

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

Householder Transformations

- Used to reflect a vector in such a way that all coordinates but one disappear, e.g.: QR-decomposition



$$\boldsymbol{u} = \boldsymbol{x} \pm \|\boldsymbol{x}\|\boldsymbol{e}_1$$

$$\boldsymbol{H} = \mathbf{1} - 2\hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^T$$

$$\boldsymbol{H}\boldsymbol{x} = \|\boldsymbol{x}\|\boldsymbol{e}_1$$

Householder Transformations

Example for $D = 2$

$$\boldsymbol{v}_1 = \begin{pmatrix} 0 \\ v_{11} \end{pmatrix} \qquad \boldsymbol{v}_2 = \begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix}$$

Construction of $\boldsymbol{H}_1$

$$v_{11} \in \{-1,1\}$$

$$\tilde{\boldsymbol{H}}_1 = v_{11} \qquad \boldsymbol{H}_1 = \begin{pmatrix} 1 & 0 \\ 0 & v_{11} \end{pmatrix}$$

Construction of $\boldsymbol{H}_2$

$$\boldsymbol{v}_2 \in \mathbb{S}^1 \qquad \boldsymbol{v}_1 = \begin{pmatrix} 0 \\ v_{11} \end{pmatrix} \quad \boldsymbol{v}_2 = \begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix}$$

$$\boldsymbol{U} = \boldsymbol{H}_2\boldsymbol{H}_1 = \boldsymbol{H}_2 \begin{pmatrix} 1 & 0 \\ 0 & v_{11} \end{pmatrix} = \begin{pmatrix} \boldsymbol{v}_2 & \boldsymbol{H}_2\boldsymbol{v}_1 \end{pmatrix}$$

for $n = D : 1$

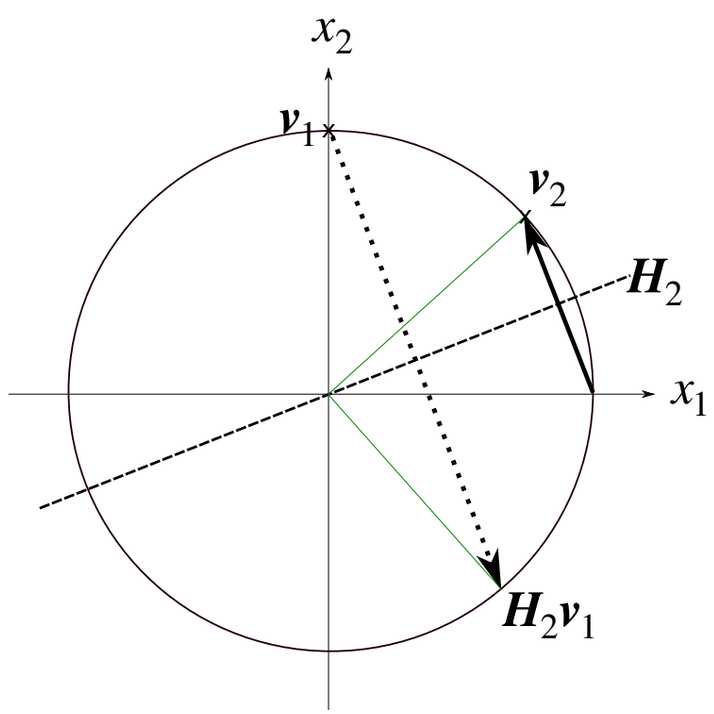
$$\boldsymbol{v}_n \sim \textbf{uniform on } \mathbb{S}^{n-1}$$

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

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

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

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



Implementation

The full generative model for Bayesian PPCA:

$$\boldsymbol{v}_D, \ldots, \boldsymbol{v}_{D-Q+1} \sim \mathcal{N}(0, \boldsymbol{I})$$

$$\boldsymbol{\sigma} \sim p(\boldsymbol{\sigma})$$

$$\boldsymbol{\mu} \sim p(\boldsymbol{\mu})$$

$$\boldsymbol{U} = \prod_{q=1}^{Q} \boldsymbol{H}_{D-q+1}\left(\boldsymbol{v}_{D-q+1}\right)$$

$$\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\sigma})$$

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}$$

$$\sigma_{\textbf{noise}} \sim p\left(\sigma_{\textbf{noise}}\right)$$

$$Y \sim \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{Y}_{n,:} | \boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2_{\textbf{noise}}\, \boldsymbol{I}\right)$$

Results

Synthetic Dataset

- Construction

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

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

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

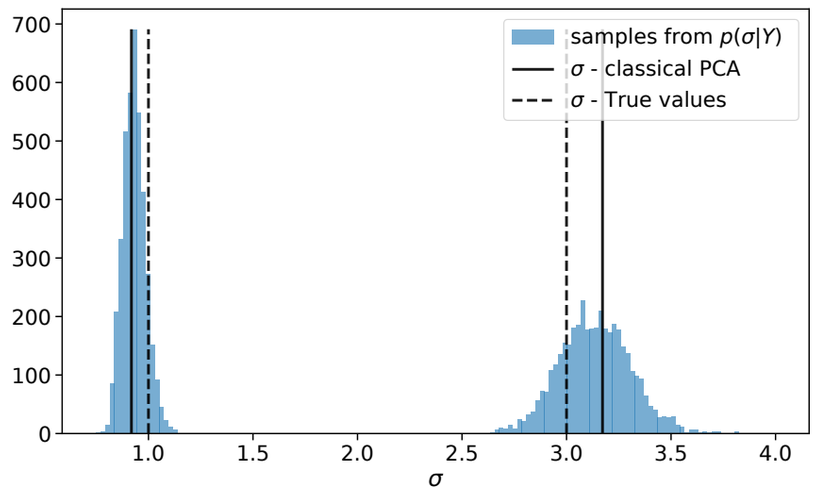
$$\boldsymbol{\epsilon} \sim \mathcal{N}(0,\ 0.01) \in \mathbb{R}^{N \times D}$$

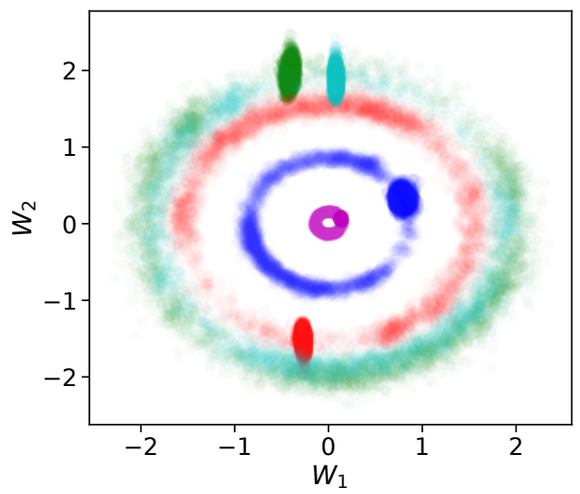
$$\boldsymbol{\Sigma} = \operatorname{diag}\left(\sigma_1, \sigma_2\right) = \operatorname{diag}\left(3.0,\ 1.0\right)$$

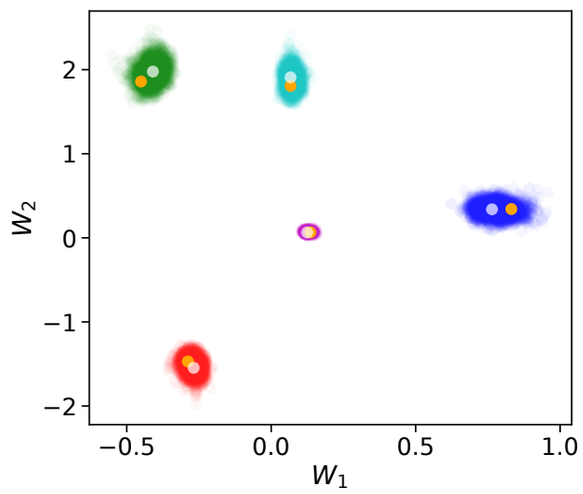
$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma} \in \mathbb{R}^{D \times Q}$$

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

- Inference



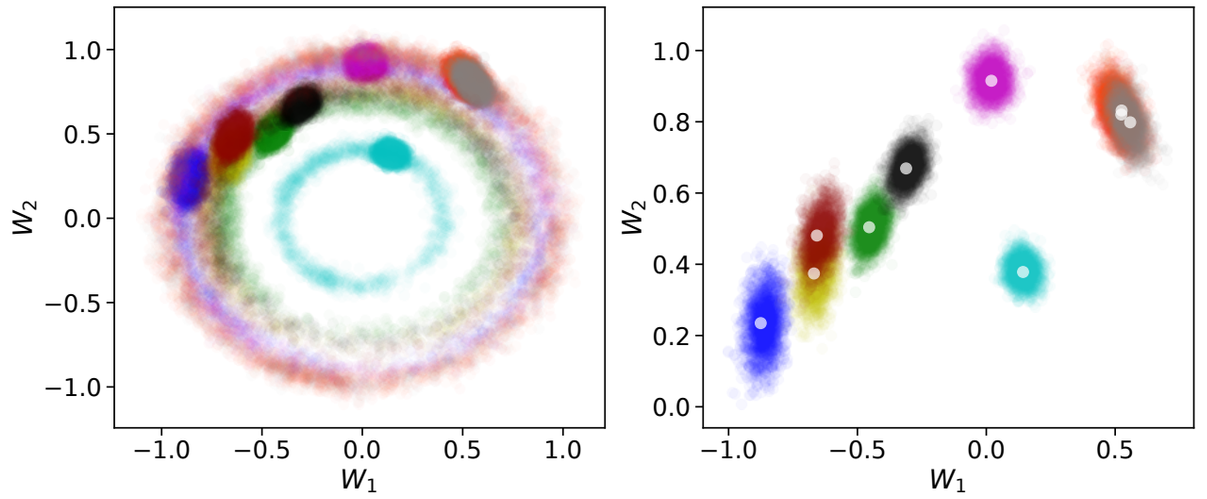




Results

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

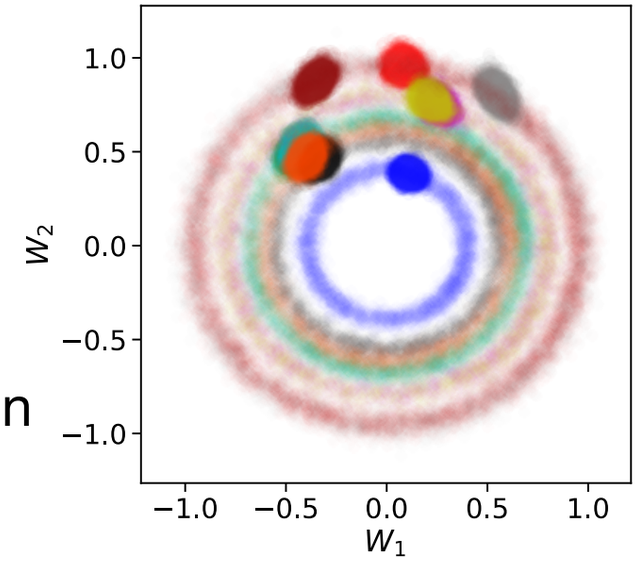
Results

Breast Cancer

$(N, D) = (569,30)$

Time House: 9.5 min

Time Standard: 25.6 min

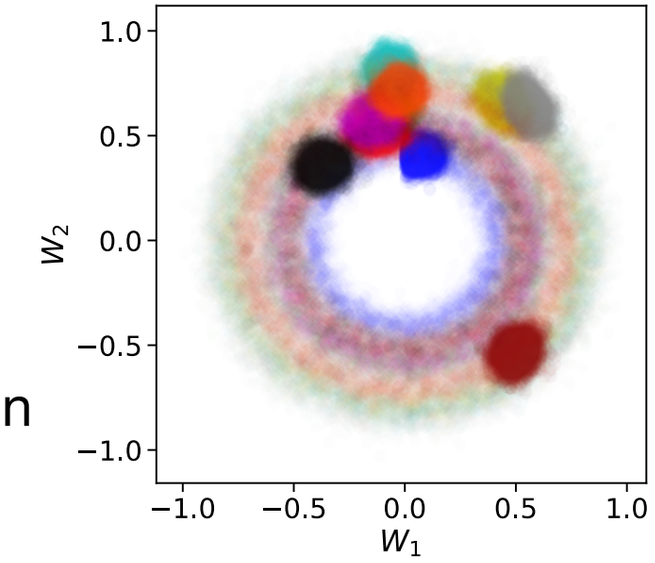


Diabetes

$(N, D) = (442,10)$

Time House: 8.9 min

Time Standard: 2.4 min

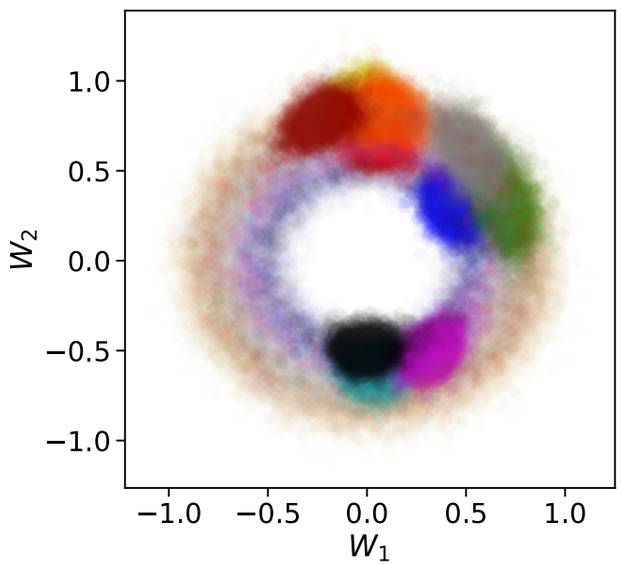


Wine

$(N, D) = (178,13)$

Time House: 0.4 min

Time Standard: 1.2 min

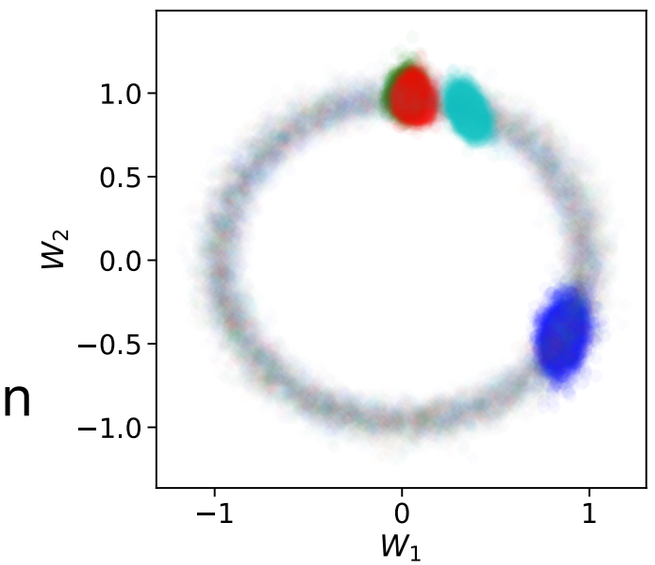


Iris

$(N, D) = (150,4)$

Time House: 0.2 min

Time Standard: 0.8 min



Results

for $n = D : 1$

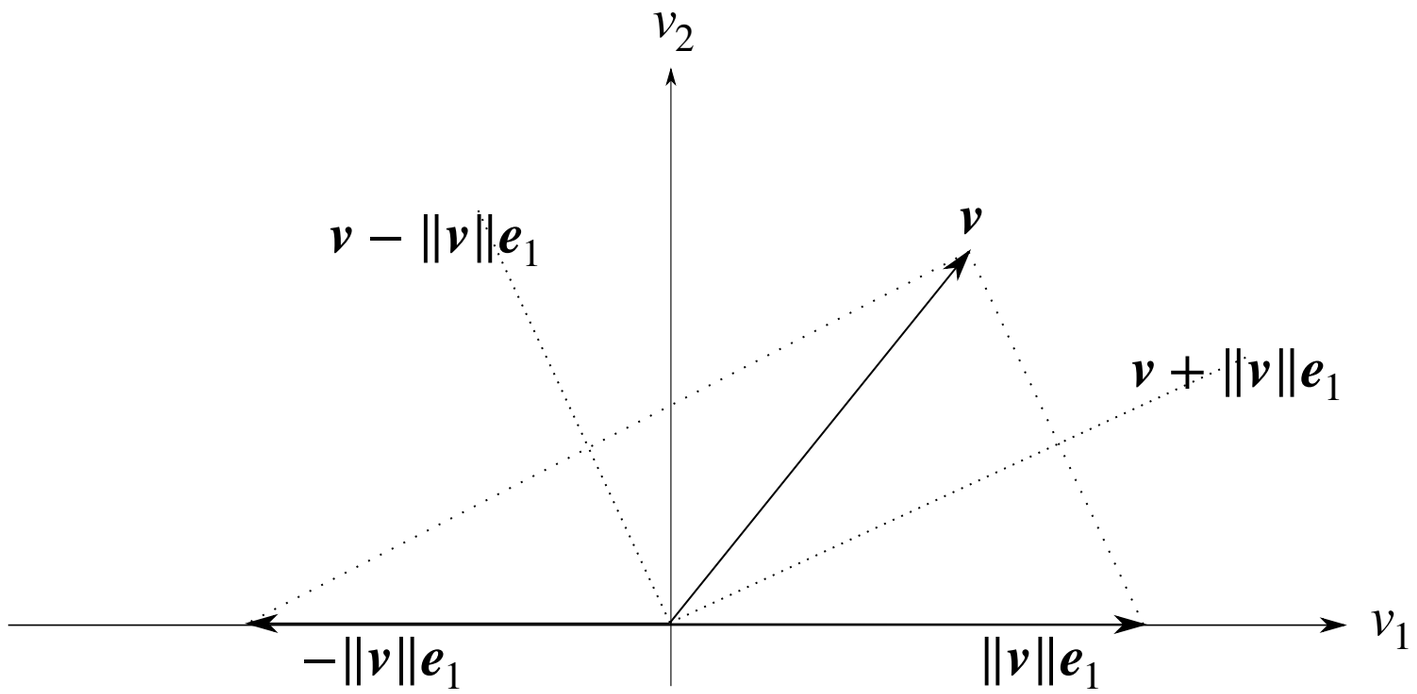
$$\boldsymbol{v}_n \sim \textbf{uniform on } \mathbb{S}^{n-1}$$

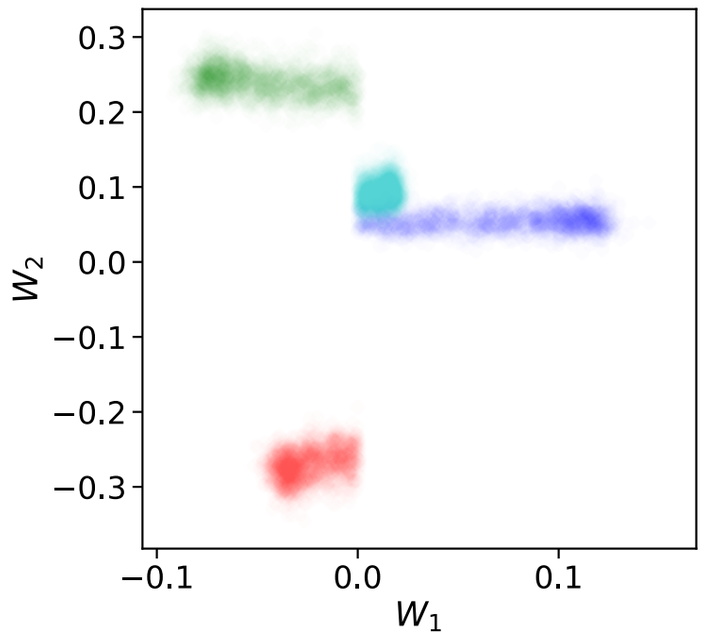
$$\boldsymbol{u}_n = \frac{\boldsymbol{v}_n + \operatorname{sgn}(\boldsymbol{v}_{n1}) \parallel \boldsymbol{v}_n \parallel \boldsymbol{e}_1}{\left\| \boldsymbol{v}_n + \operatorname{sgn}(\boldsymbol{v}_{n1}) \left\| \boldsymbol{v}_n \right\| \boldsymbol{e}_1 \right\|}$$

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

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

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





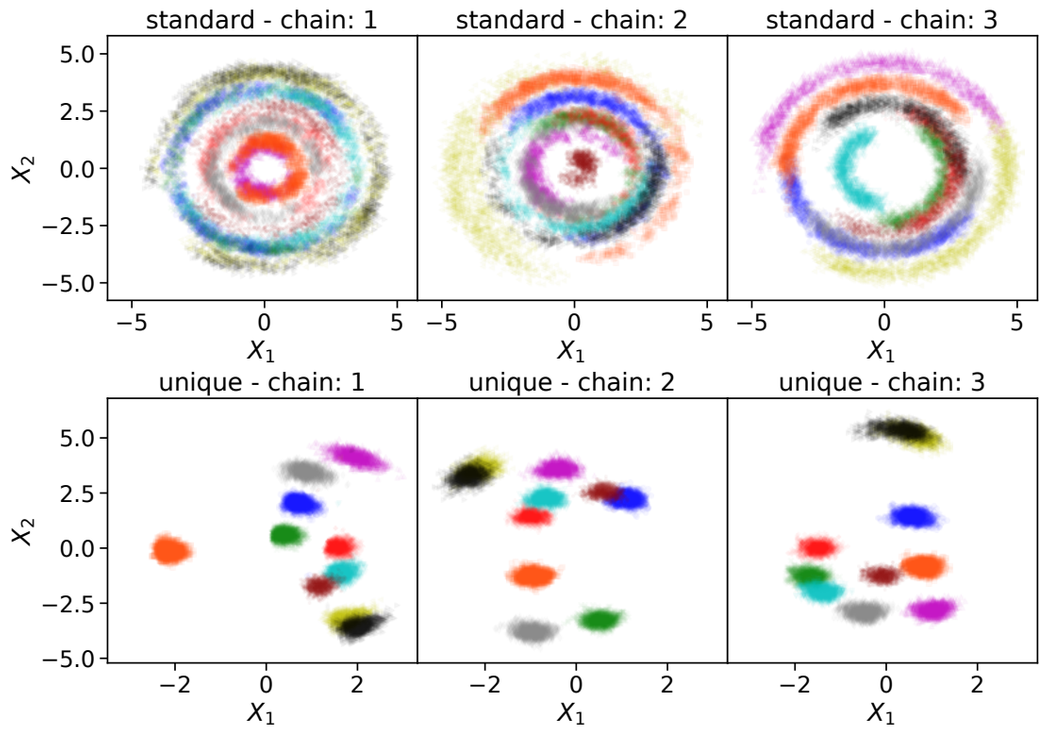
Extension to non-linear models

- GPLVM with the same rotation invariant problem

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

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

$$k_{\mathrm{SE}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_{\mathrm{SE}}^2 \exp\left(-0.5 \parallel \boldsymbol{x} - \boldsymbol{x}' \parallel_2^2 / l^2\right)$$



- 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 $\boldsymbol{W}$ in PPCA, which uniquely identifies principle components even though the likelihood is 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 $\boldsymbol{W}$

- 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

Github: https://github.com/RSNirwan/HouseholderBPCA

Thanks for your attention!

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