Rotation Invariant Householder Parameterization for Bayesian PCA

Rajbir-Singh Nirwan

Dec 04, 2020

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

FIAS Frankfurt Institute for Advanced Studies

Outline

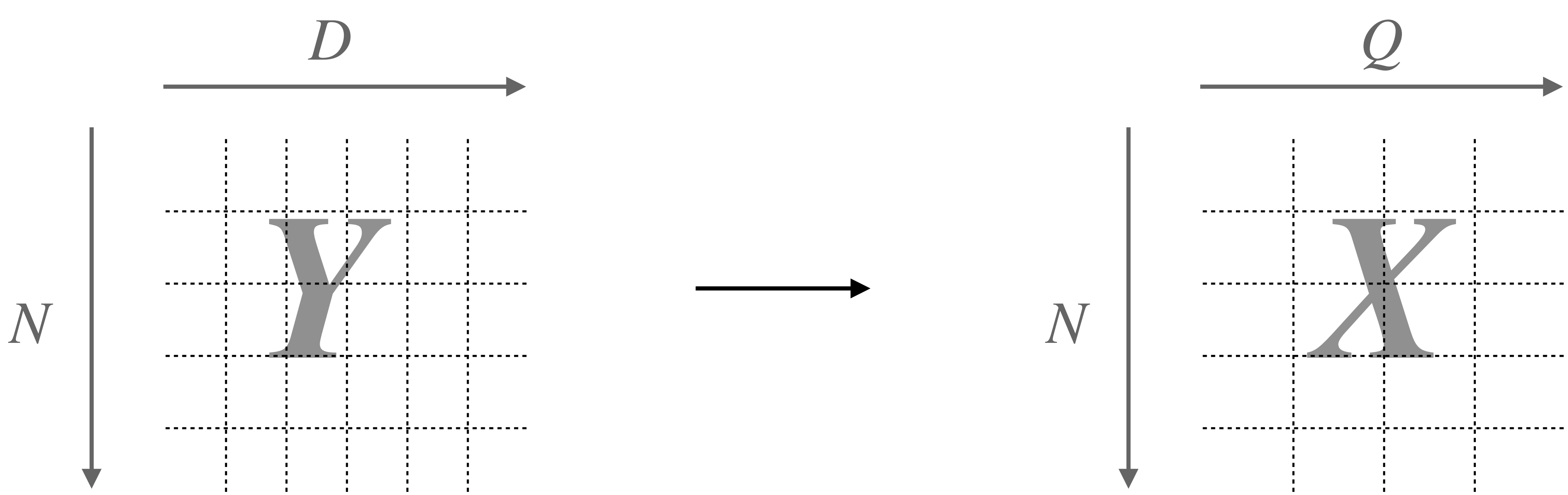
- Probabilistic PCA (PPCA)
- Non-identifiability issue of PPCA
- Conceptual solution to the problem
- Implementation
- Results

Probabilistic PCA

Classical PCA

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

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



Probabilistic PCA

Probabilistic PCA (PPCA)

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

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

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

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

Probabilistic PCA

Probabilistic PCA (PPCA)

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

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

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

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

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

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

Probabilistic PCA

Probabilistic PCA (PPCA)

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

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

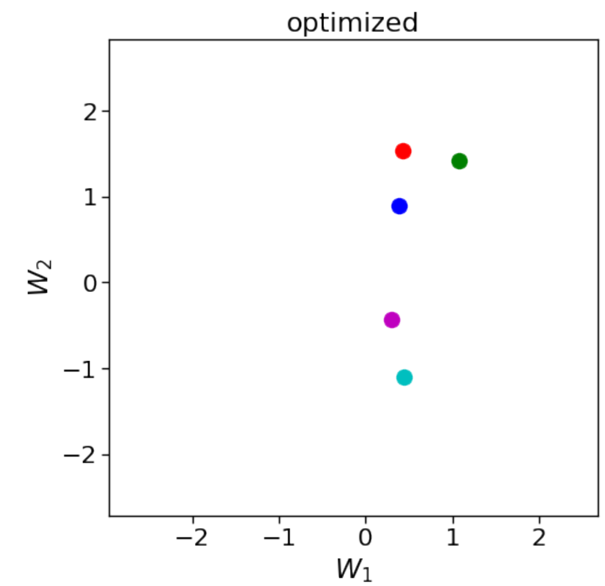
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$$

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

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

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

- Optimization for D=5, Q=2



$$\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \\ w_{41} & w_{42} \\ w_{51} & w_{52} \end{pmatrix}$$

Probabilistic PCA

Probabilistic PCA (PPCA)

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

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

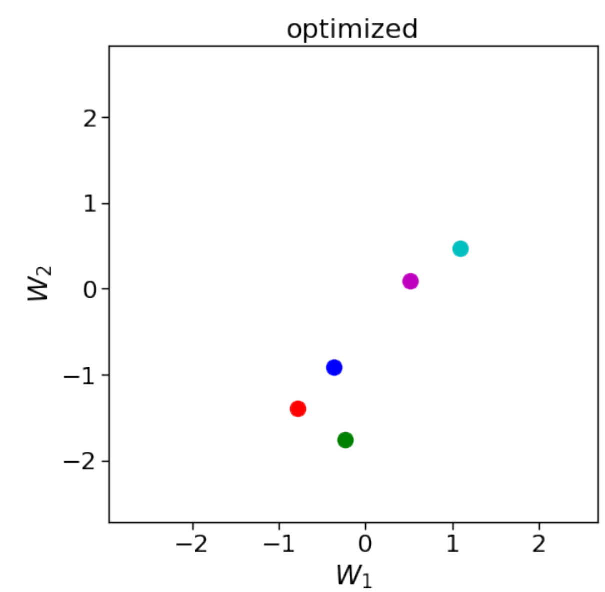
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$$

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

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

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

- Optimization for D=5, Q=2



$$\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \\ w_{41} & w_{42} \\ w_{51} & w_{52} \end{pmatrix}$$

Probabilistic PCA

Probabilistic PCA (PPCA)

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

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

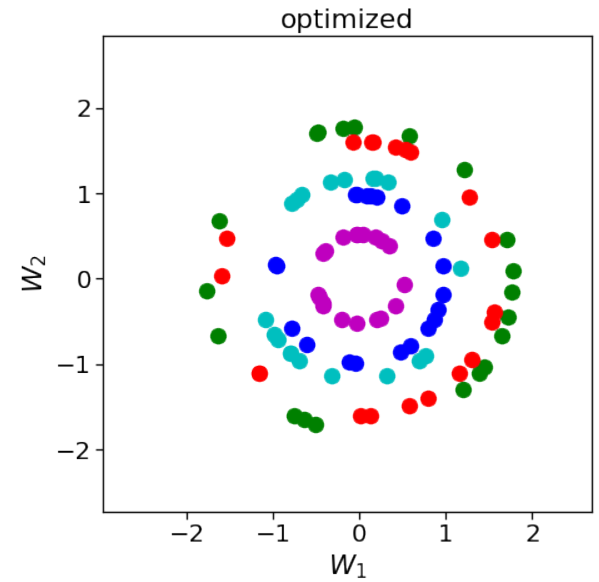
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$$

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

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

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

- Optimization for D=5, Q=2



$$\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \\ w_{41} & w_{42} \\ w_{51} & w_{52} \end{pmatrix}$$

Probabilistic PCA

Probabilistic PCA (PPCA)

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

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

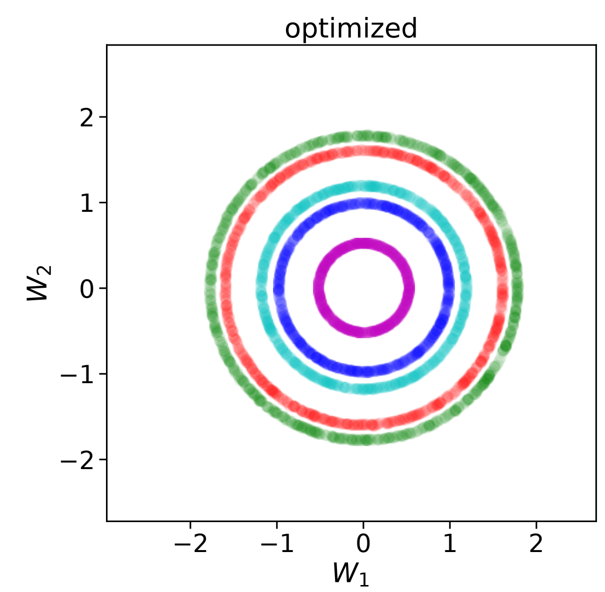
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$$

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

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

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

- Rotation invariant likelihood



$$\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \\ w_{41} & w_{42} \\ w_{51} & w_{52} \end{pmatrix}$$

Probabilistic PCA

Probabilistic PCA (PPCA)

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

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

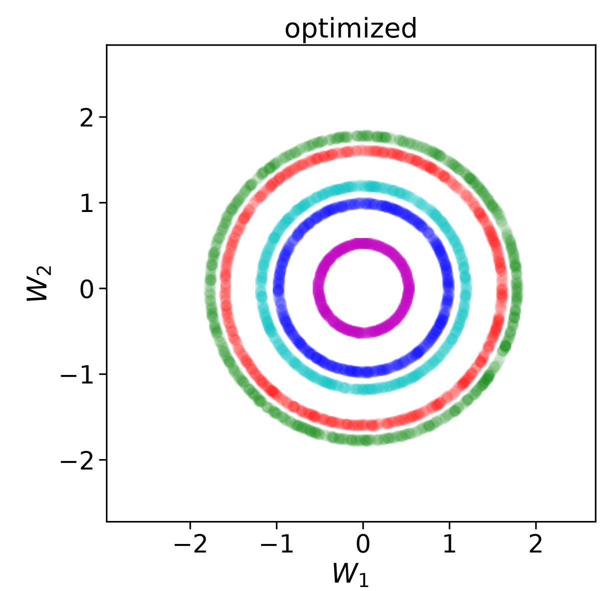
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$$

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

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

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

- Rotation invariant likelihood



$$\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \\ w_{41} & w_{42} \\ w_{51} & w_{52} \end{pmatrix}$$

- Solution

$$\boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\left(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\right)^T = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^T$$

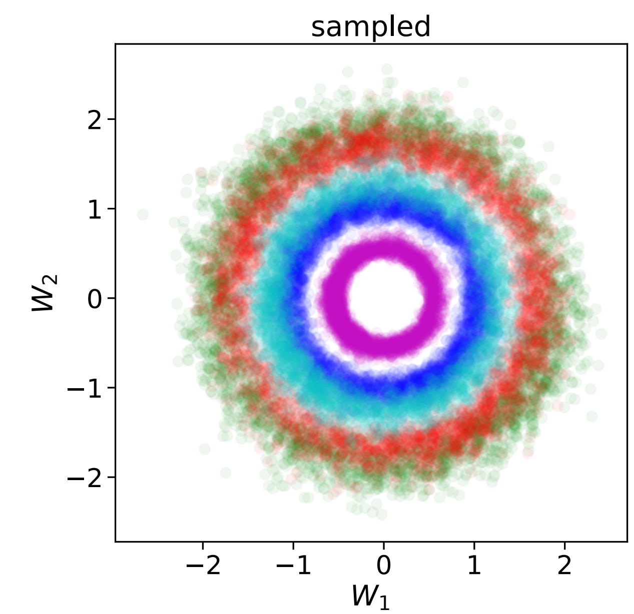
$$p(\boldsymbol{Y}\,|\,\boldsymbol{U}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{Y}_{n,:}\,|\,\boldsymbol{0}, \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^T + \sigma^2\boldsymbol{I}\right)$$

Bayesian approach to PPCA

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

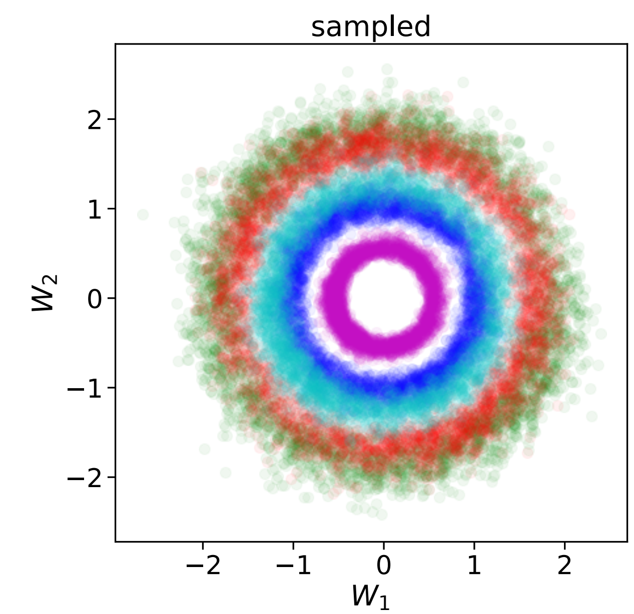
Bayesian approach to PPCA

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



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

Solution

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

Outline of procedure

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

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

$\boldsymbol{U} \sim ?$
$\boldsymbol{\Sigma} \sim ?$ $\rightarrow$ $\boldsymbol{U\Sigma\Sigma}^T\boldsymbol{U}^T$ **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** $\mathscr{V}_{Q,D} = \left\{\boldsymbol{U} \in \mathbb{R}^{D \times Q} \mid \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}\right\}$

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

Theory

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

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

- $\boldsymbol{U}$, $\boldsymbol{\Sigma}^2$ 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))

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

Theory

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

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

- $\boldsymbol{U}, \boldsymbol{\Sigma}^2$ 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))

→ $\boldsymbol{U}$ **is uniformly distributed on the Stiefel manifold**

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

Theory

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

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

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

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

  → $\boldsymbol{U}$ **is uniformly distributed on the Stiefel manifold**

- Distribution of the ordered eigenvalue matrix $\boldsymbol{\Sigma}^2$ of $\boldsymbol{WW}^T$ 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

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

Theorem 2 *Let $\boldsymbol{v}_D, \boldsymbol{v}_{D-1}, \ldots, \boldsymbol{v}_1$ be uniformly distributed on the unit spheres $\mathbb{S}^{D-1}, \ldots, \mathbb{S}^0$ 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*

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

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

Mezzadri (2007)

for $n = D : 1$

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

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

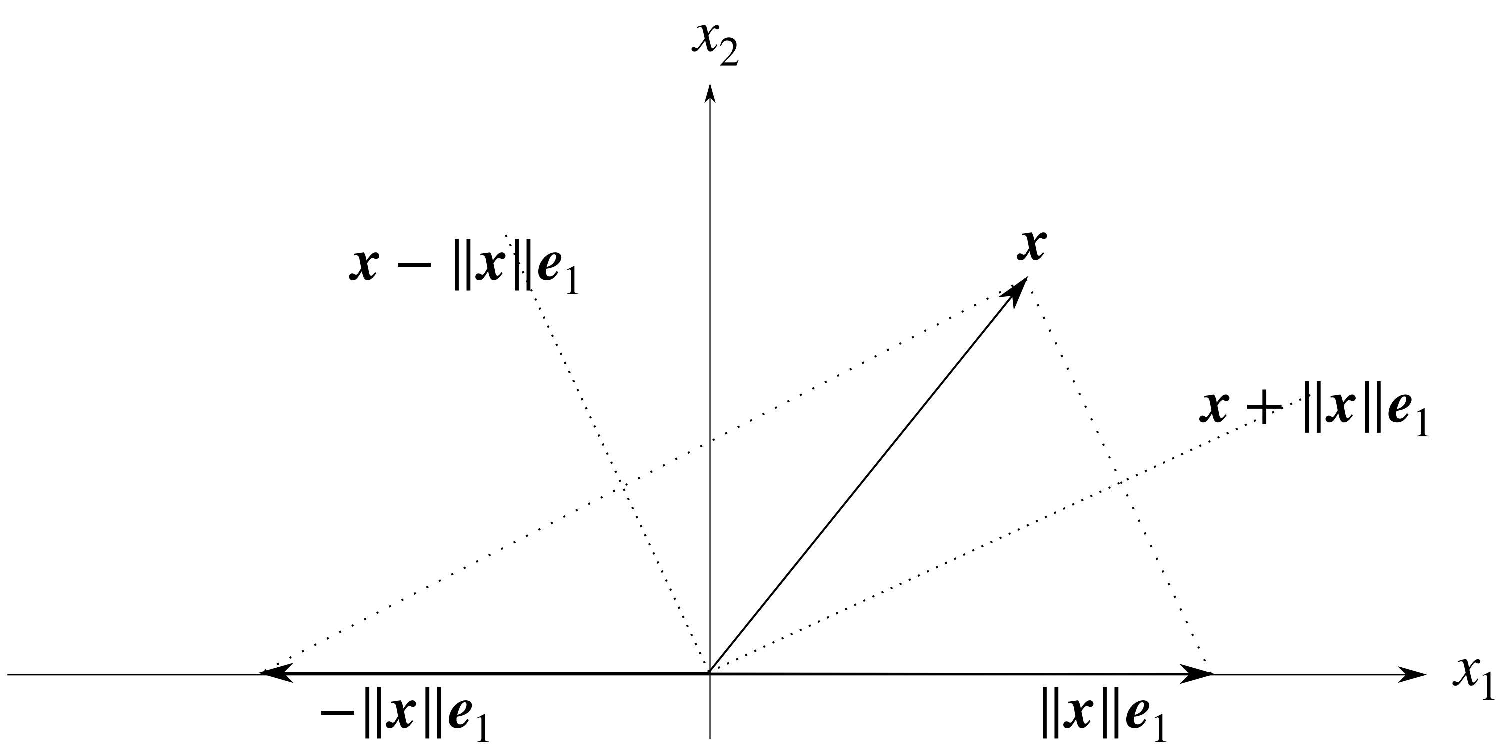
$$\tilde{\boldsymbol{H}}_{\boldsymbol{n}}\left(\boldsymbol{v}_n\right) = -\operatorname{sgn}\left(\boldsymbol{v}_{n1}\right)\left(\boldsymbol{I} - 2\boldsymbol{u}_n\boldsymbol{u}_n^T\right)$$

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

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

Householder Transformations

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



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

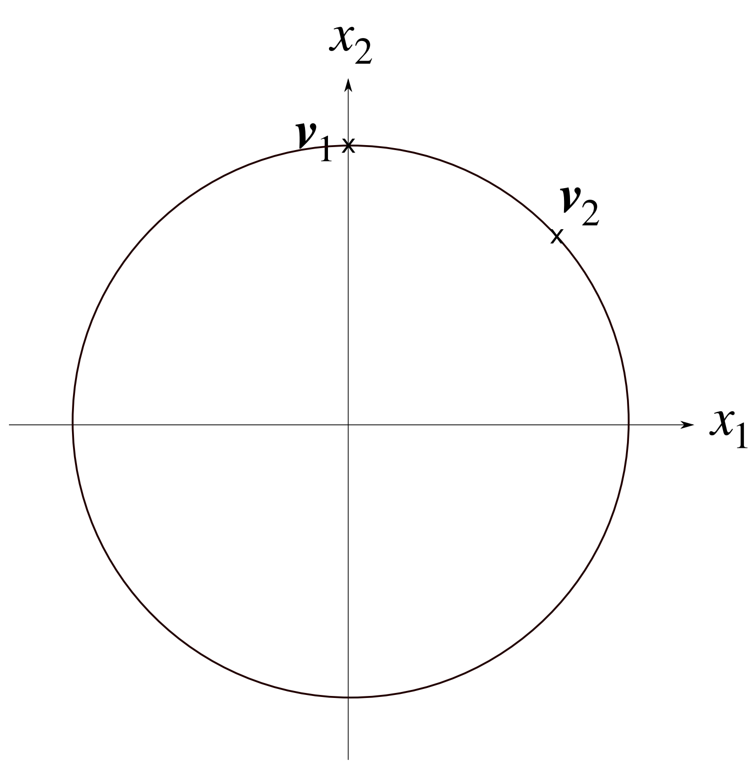
$$\boldsymbol{H} = \boldsymbol{1} - 2\hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^T$$

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

Householder Transformations

Example for $D = 2$

$$\boldsymbol{U} \in \mathscr{V}_{2,2} \qquad \boldsymbol{v}_1 = \begin{pmatrix} 0 \\ v_{11} \end{pmatrix} \quad \boldsymbol{v}_2 = \begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix}$$



for $n = D : 1$

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

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

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

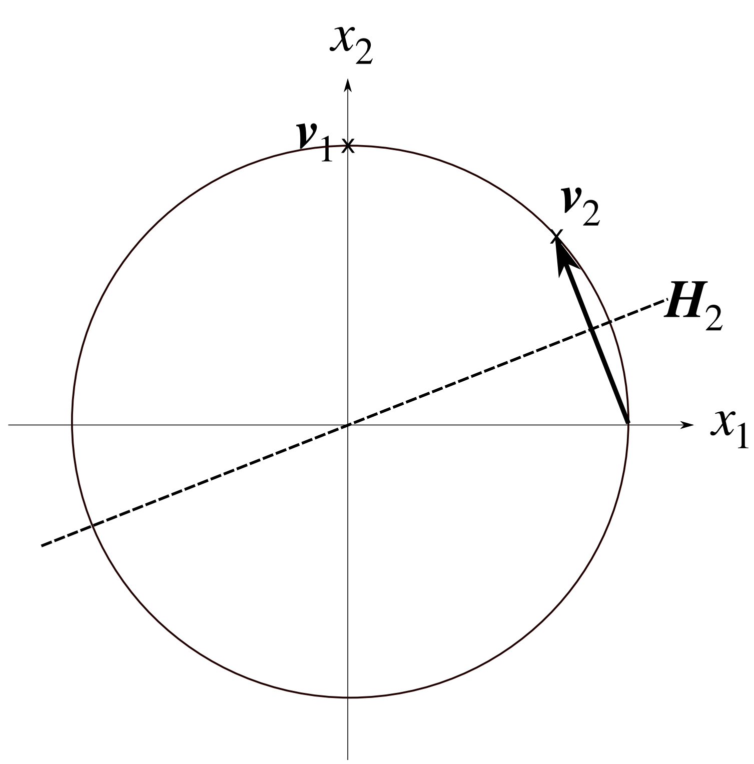
$$\boldsymbol{H}_n = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \tilde{\boldsymbol{H}}_{\boldsymbol{n}} \end{pmatrix}$$

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

Householder Transformations

Example for $D = 2$

$$\boldsymbol{U} \in \mathscr{V}_{2,2} \qquad \boldsymbol{v}_1 = \begin{pmatrix} 0 \\ v_{11} \end{pmatrix} \quad \boldsymbol{v}_2 = \begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix}$$



for $n = D : 1$

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

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

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

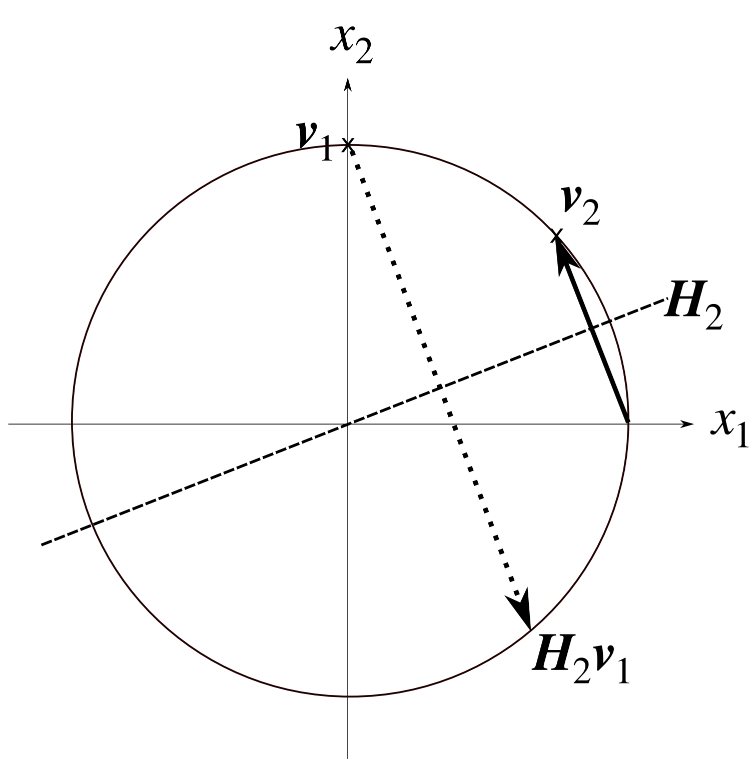
$$\boldsymbol{H}_n = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \tilde{\boldsymbol{H}}_{\boldsymbol{n}} \end{pmatrix}$$

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

Householder Transformations

Example for $D = 2$

$$\boldsymbol{U} \in \mathscr{V}_{2,2} \qquad \boldsymbol{v}_1 = \begin{pmatrix} 0 \\ v_{11} \end{pmatrix} \quad \boldsymbol{v}_2 = \begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix}$$



for $n = D : 1$

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

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

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

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

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

Householder Transformations

Example for $D = 2$

$$\boldsymbol{U} \in \mathscr{V}_{2,2} \qquad \boldsymbol{v}_1 = \begin{pmatrix} 0 \\ v_{11} \end{pmatrix} \quad \boldsymbol{v}_2 = \begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix}$$

for $n = D : 1$

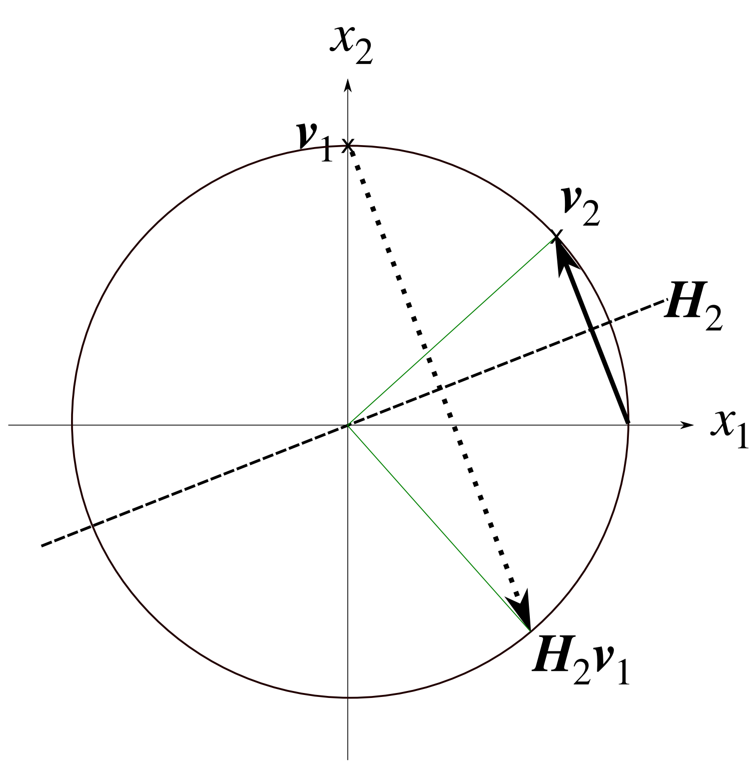
$\boldsymbol{v}_n \sim$ **uniform on** $\mathbb{S}^{n-1}$

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

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

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

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



Householder Transformations

Example for $D = 2$

$$\boldsymbol{U} \in \mathscr{V}_{2,2} \qquad \boldsymbol{v}_1 = \begin{pmatrix} 0 \\ v_{11} \end{pmatrix} \quad \boldsymbol{v}_2 = \begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix}$$

for $n = D : 1$

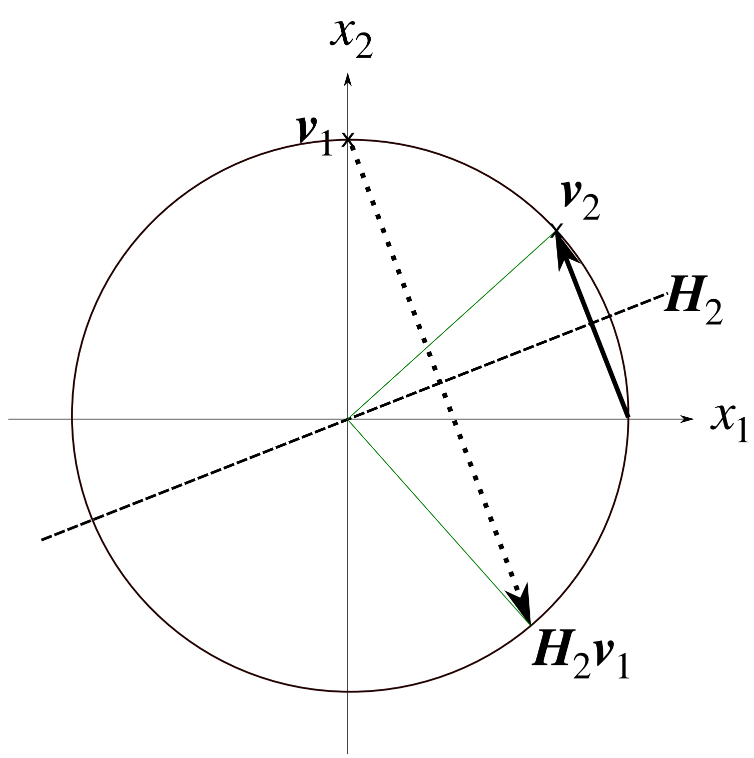
$\boldsymbol{v}_n \sim$ **uniform on** $\mathbb{S}^{n-1}$

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

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

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

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



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

Householder Transformations

Example for $D = 2$

$$\boldsymbol{U} \in \mathscr{V}_{2,2} \qquad \boldsymbol{v}_1 = \begin{pmatrix} 0 \\ v_{11} \end{pmatrix} \quad \boldsymbol{v}_2 = \begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix}$$

for $n = D : 1$

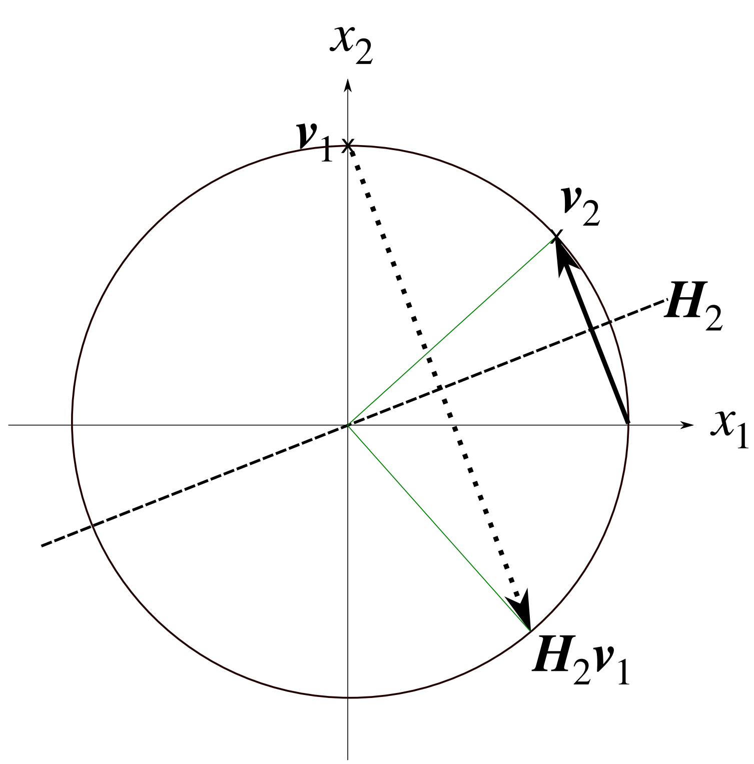
$\boldsymbol{v}_n \sim$ **uniform on** $\mathbb{S}^{n-1}$

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

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

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

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



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

$$\boldsymbol{U} = \boldsymbol{H}_3\boldsymbol{H}_2\boldsymbol{H}_1 = \begin{pmatrix} \boldsymbol{v}_3 & \boldsymbol{H}_3\boldsymbol{v}_2 & \boldsymbol{H}_3\boldsymbol{H}_2\boldsymbol{v}_1 \end{pmatrix}$$

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

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

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

Implementation

The full generative model for Bayesian PPCA:

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

Results

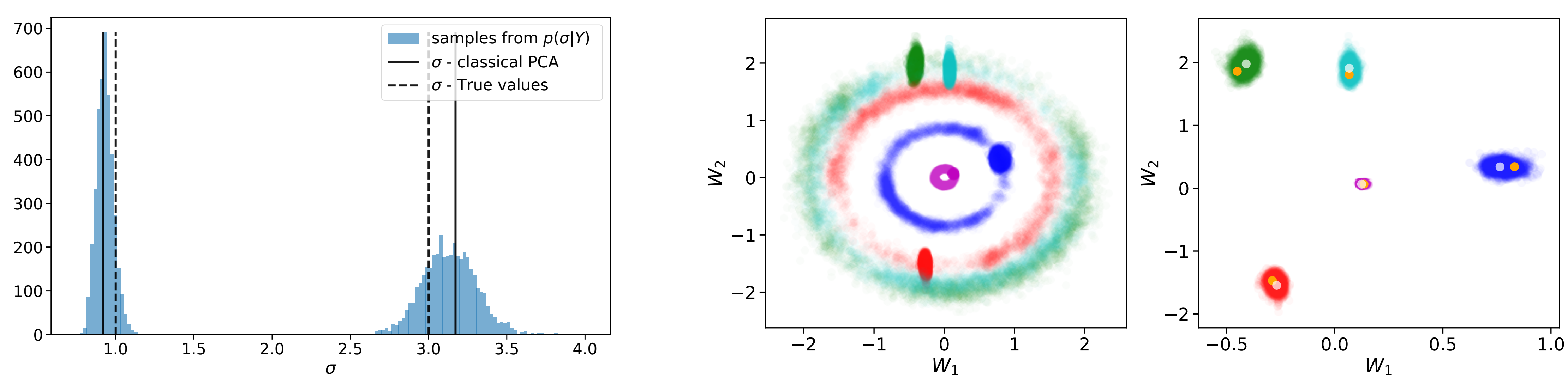
Synthetic Dataset

- Construction

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

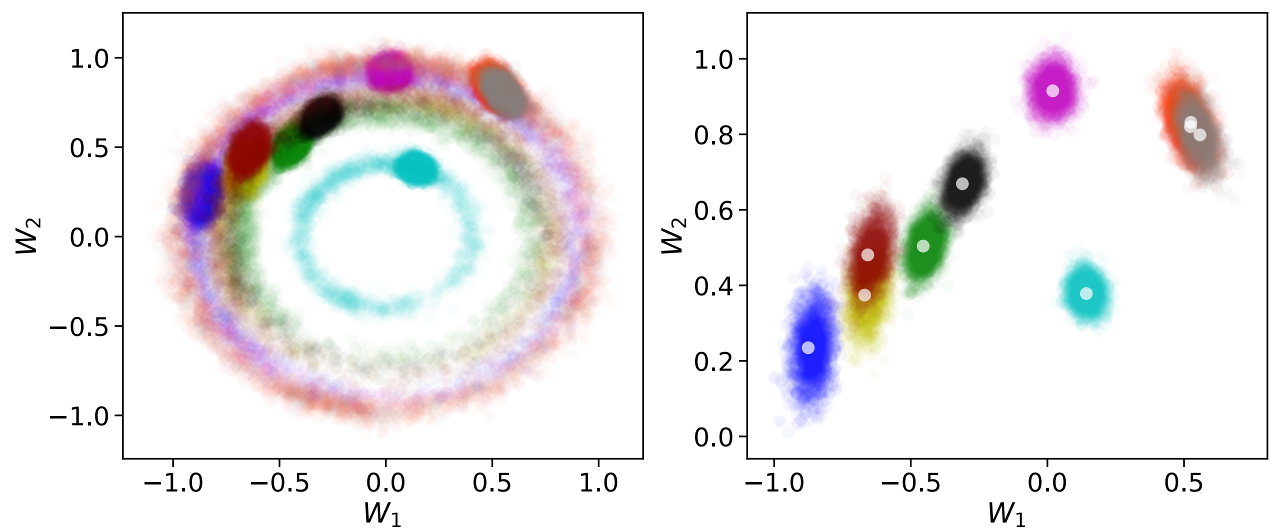
$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \in \mathbb{R}^{N \times Q}$$
$$\boldsymbol{U} \sim \textbf{uniform on Stiefel } \mathcal{V}_{Q,D}$$
$$\boldsymbol{\epsilon} \sim \mathcal{N}(0,\ 0.01) \in \mathbb{R}^{N \times D}$$
$$\boldsymbol{\Sigma} = \mathrm{diag}\left(\sigma_1, \sigma_2\right) = \mathrm{diag}\,(3.0,\ 1.0)$$
$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma} \in \mathbb{R}^{D \times Q}$$
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$$

- Inference



Results

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



Advantages

1. Breaks the rotation symmetry without changing the probabilistic model
2. Enrichment of the classical PCA solution with uncertainty estimates
3. Decomposition of $\boldsymbol{W}$ into rotation $\boldsymbol{U}$ and principle variances $\boldsymbol{\Sigma}$
    - 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

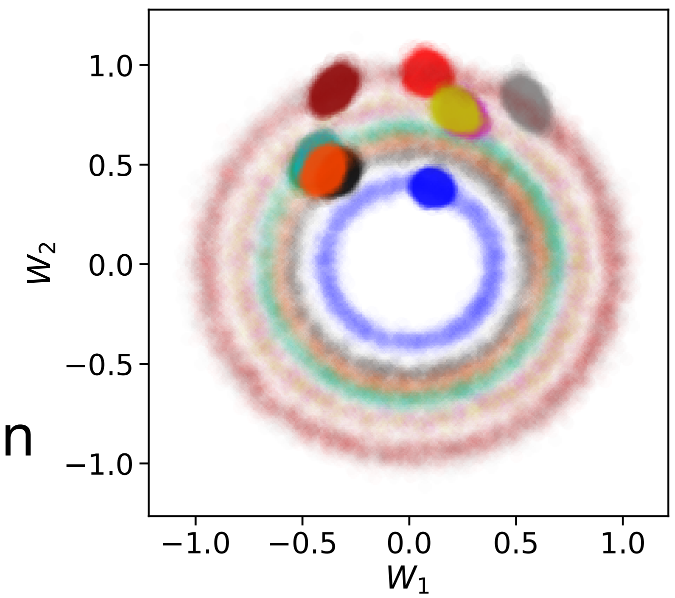
Results

Breast Cancer

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

Time House: 9.5 min

Time Standard: 25.6 min

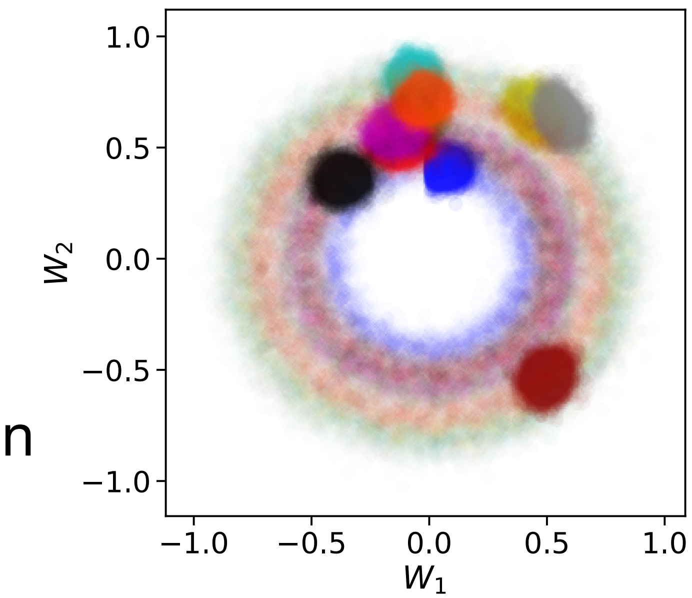


Diabetes

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

Time House: 8.9 min

Time Standard: 2.4 min

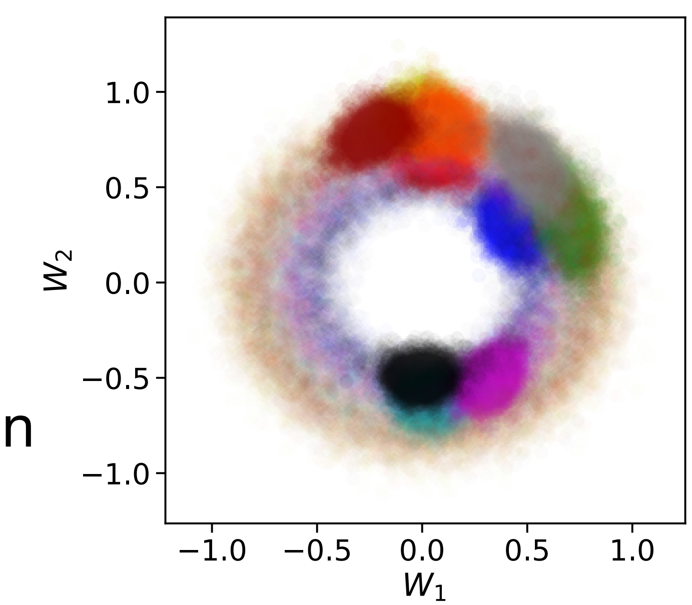


Wine

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

Time House: 0.4 min

Time Standard: 1.2 min

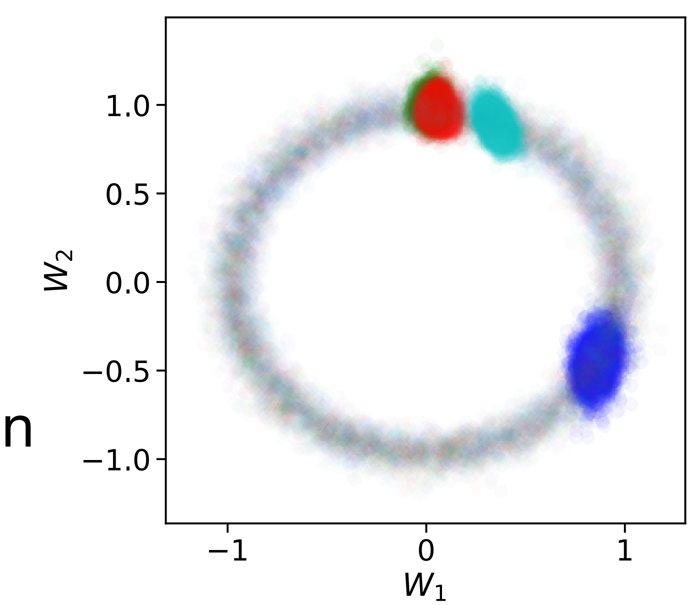


Iris

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

Time House: 0.2 min

Time Standard: 0.8 min



Results

for $n = D : 1$

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

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

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

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

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

Results

for $n = D : 1$

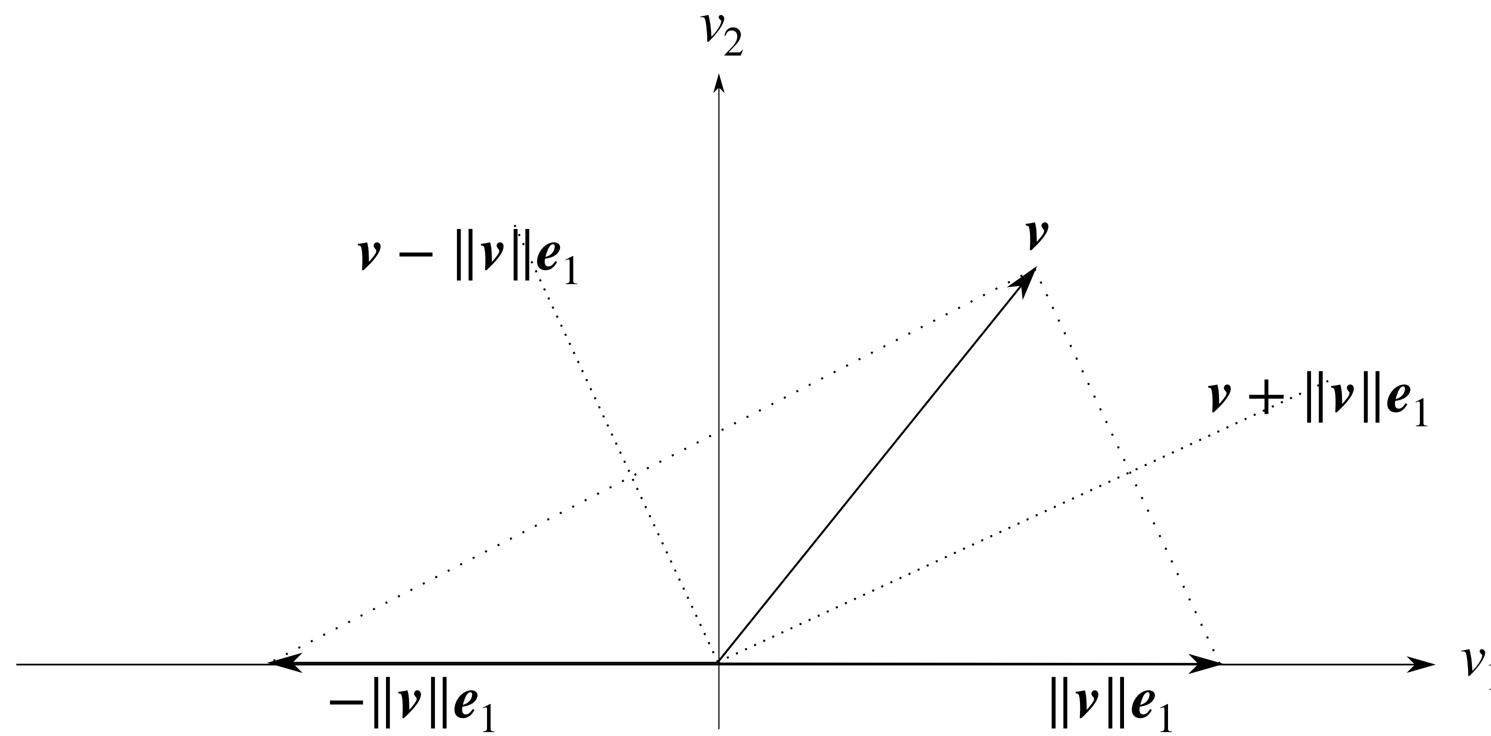
$\boldsymbol{v}_n \sim$ **uniform on** $\mathbb{S}^{n-1}$

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

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

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

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



Results

for $n = D : 1$

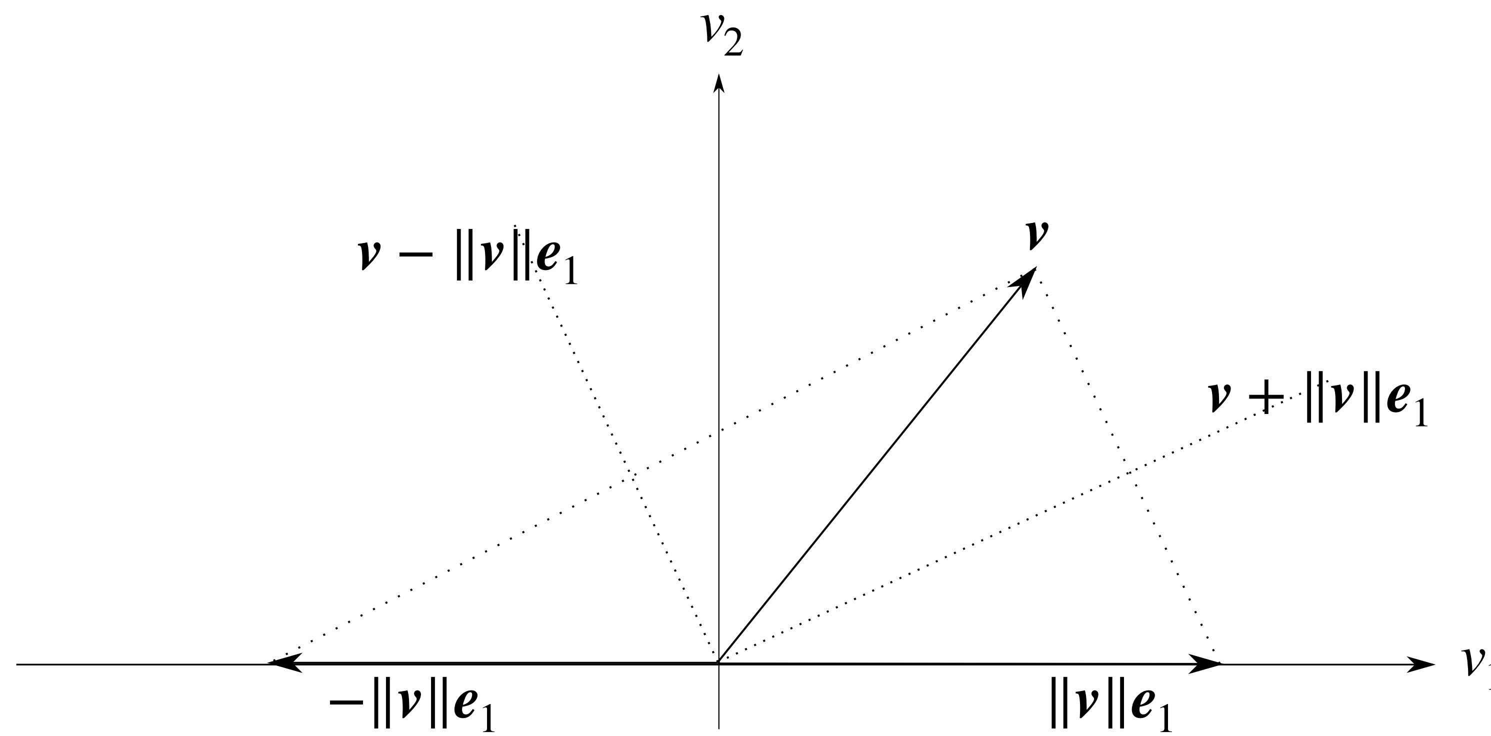
$v_n \sim$ **uniform on** $\mathbb{S}^{n-1}$

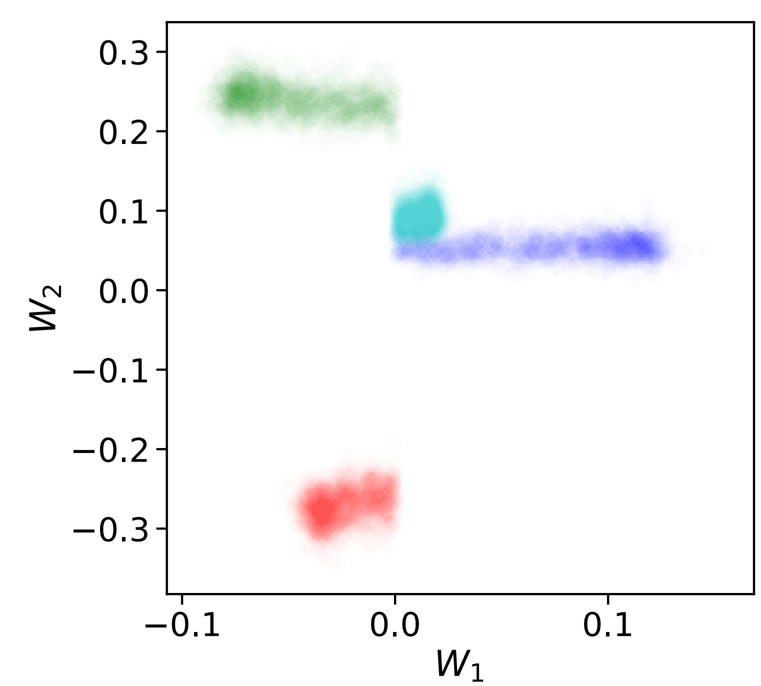
$$\boldsymbol{u}_n = \frac{\boldsymbol{v}_n + \operatorname{sgn}(\boldsymbol{v}_{n1}) \parallel \boldsymbol{v}_n \parallel \boldsymbol{e}_1}{\left\| \boldsymbol{v}_n + \operatorname{sgn}(\boldsymbol{v}_{n1}) \parallel \boldsymbol{v}_n \parallel \boldsymbol{e}_1 \right\|}$$

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

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

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





Extension to non-linear models

- GPLVM with the same rotation invariant problem

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

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

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

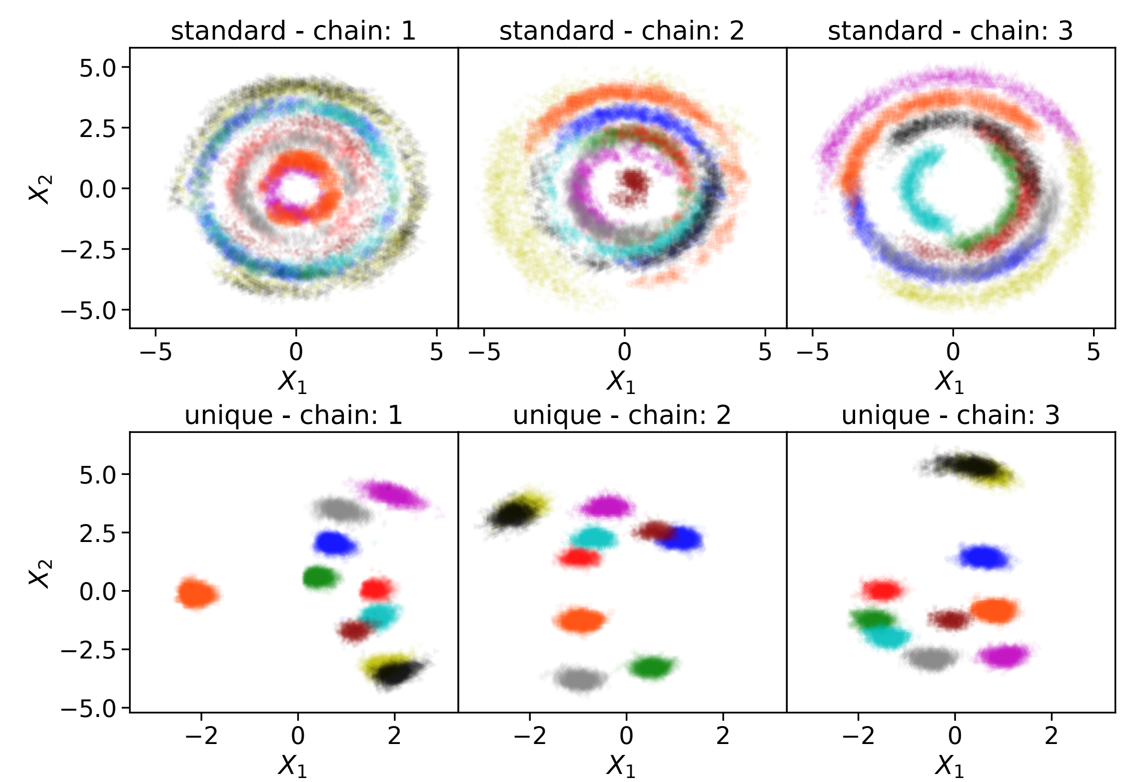
Extension to non-linear models

- GPLVM with the same rotation invariant problem

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

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

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



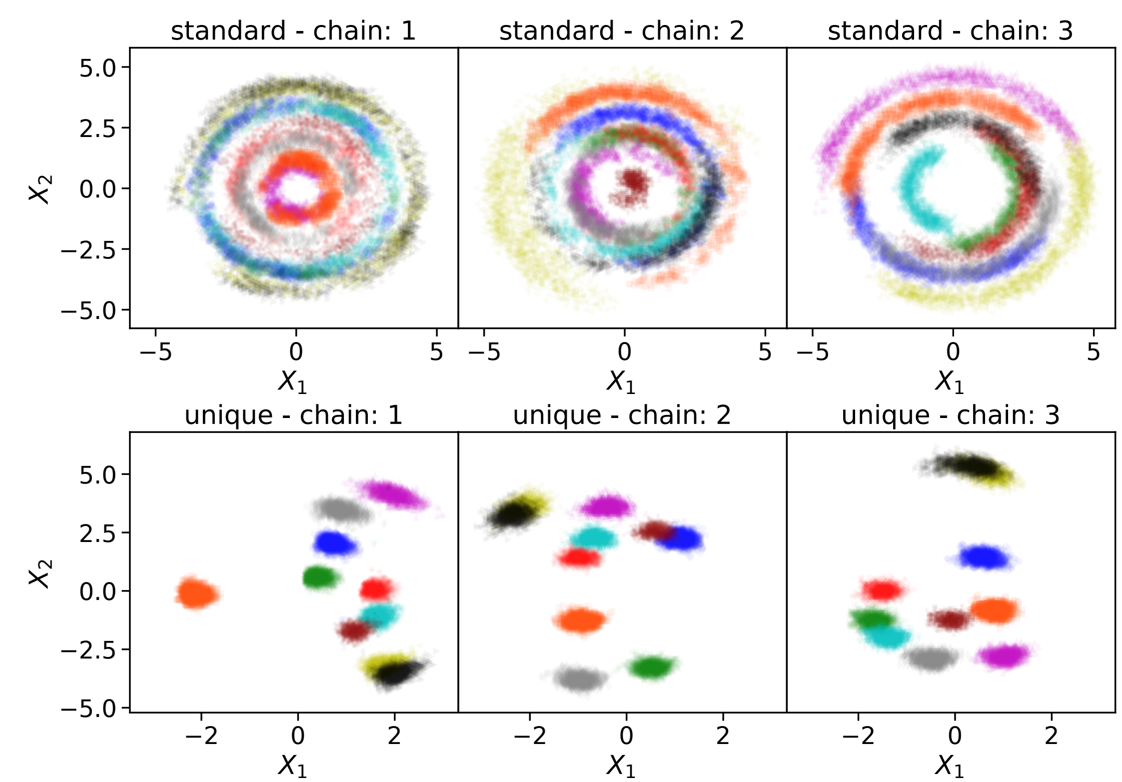
Extension to non-linear models

- GPLVM with the same rotation invariant problem

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

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

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



- 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 $\boldsymbol{W}$ 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

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

Thanks for your attention!

Supervisor: Prof. Dr. Nils Bertschinger