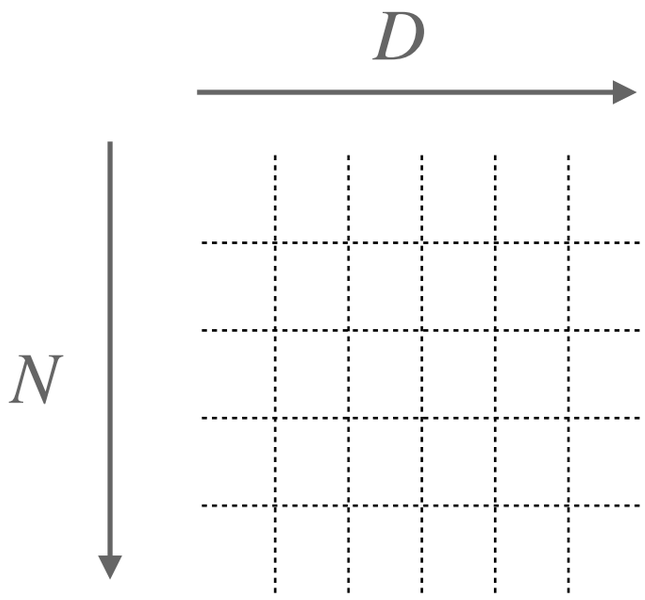
Latent Variable Models

Rajbir-Singh Nirwan

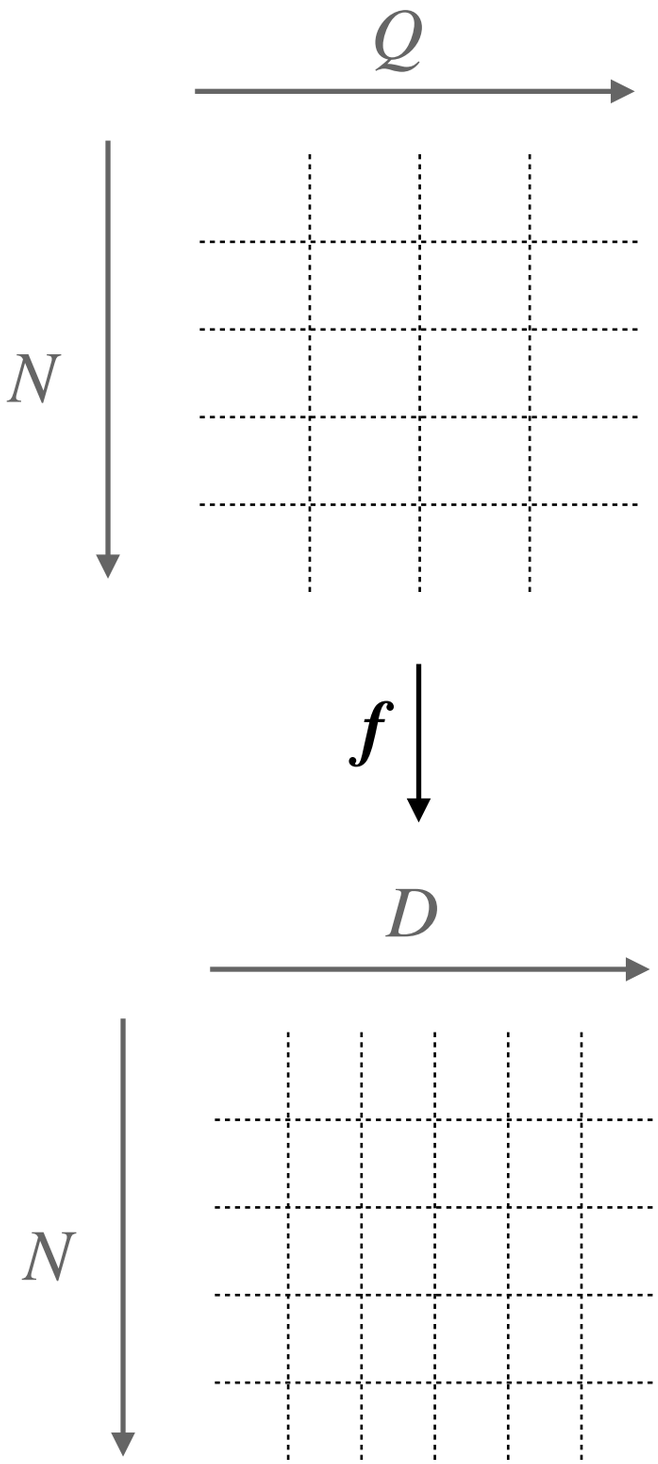
Aug 27, 2019

Outline

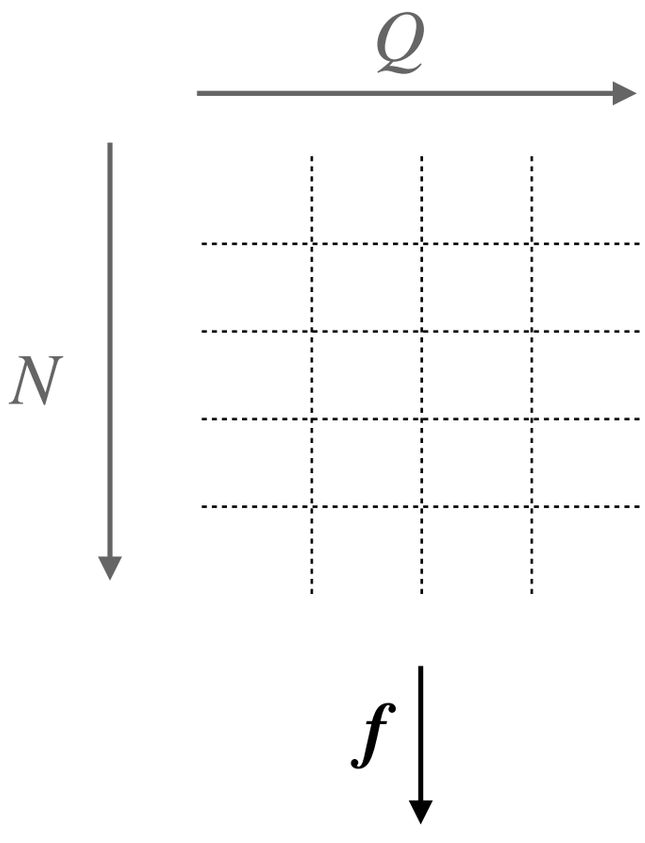
- Latent Variable Models
- Examples: PPCA, VAE
- Non-identifiability issue of PPCA
- Solution
- Summary


D
N

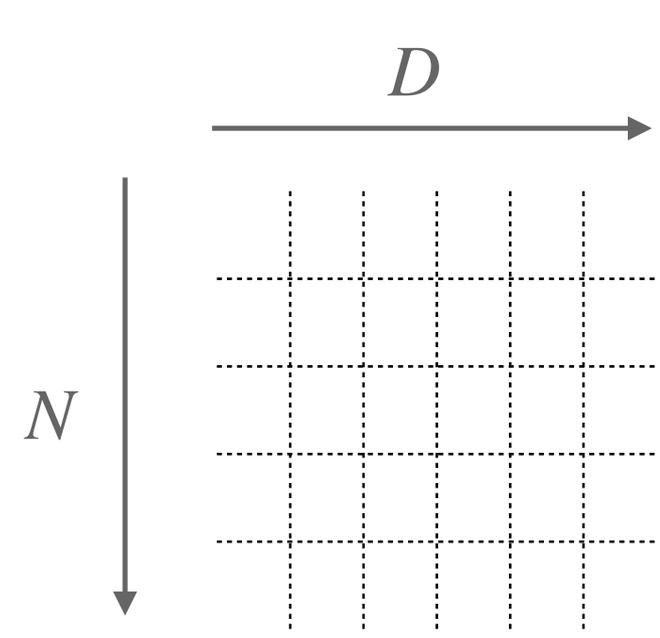

www.123rf.com/

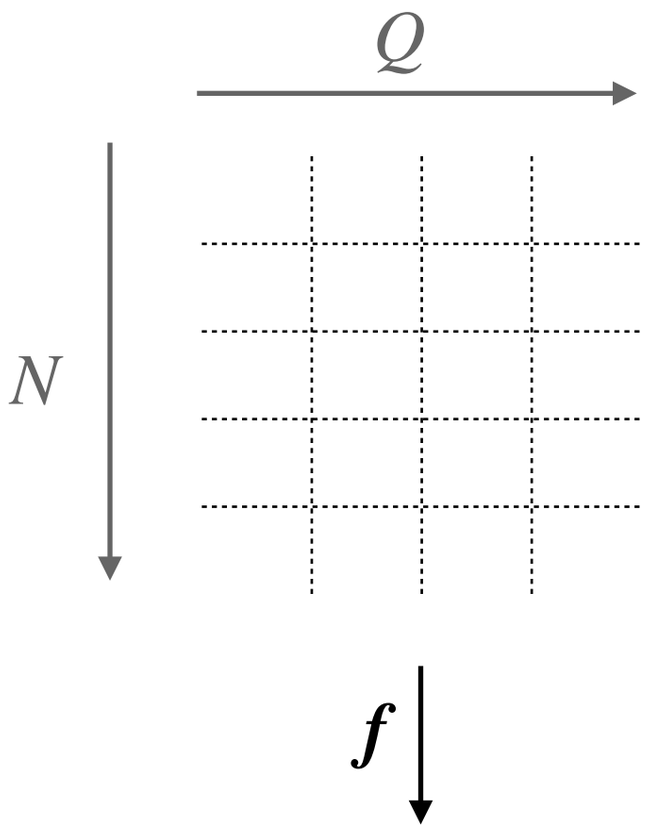

Q
N
f
D
N


www.123rf.com/



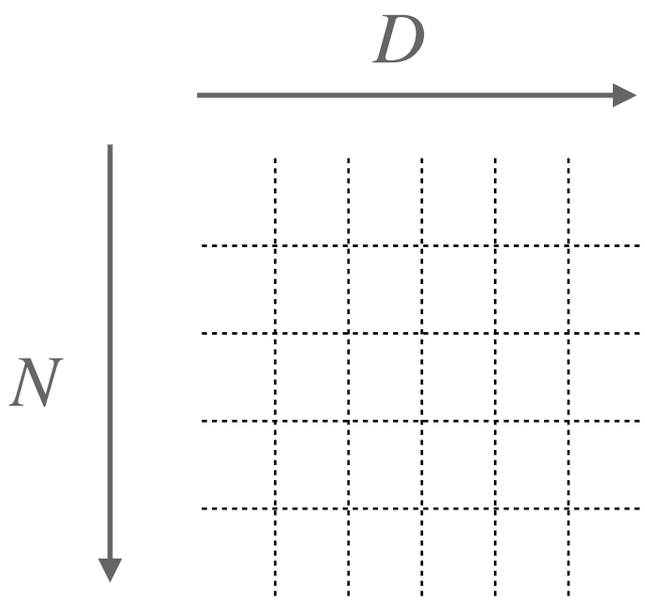
$$\begin{pmatrix} \textbf{math} \\ \textbf{physics} \\ \textbf{language} \end{pmatrix}$$

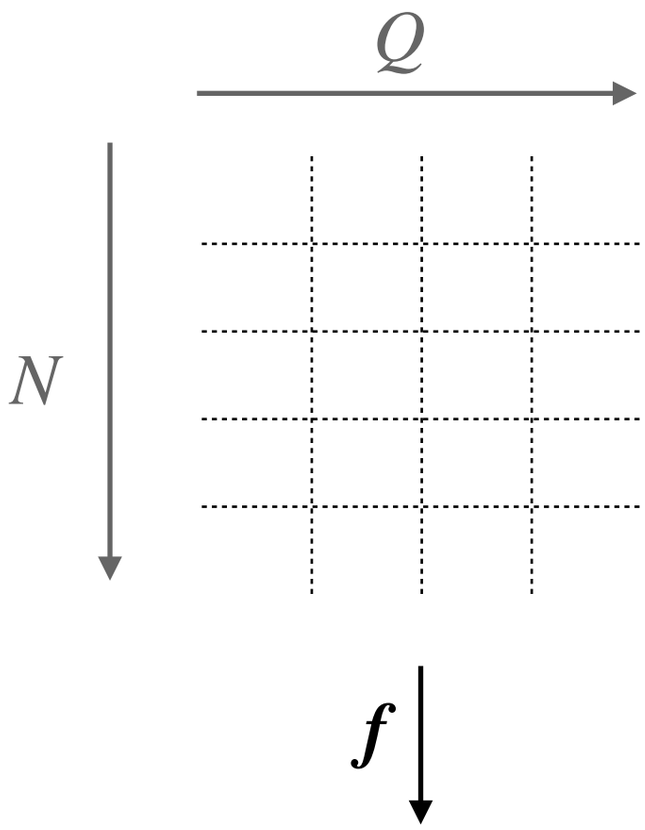




$$\begin{pmatrix} \textbf{math} \\ \textbf{physics} \\ \textbf{language} \end{pmatrix}$$

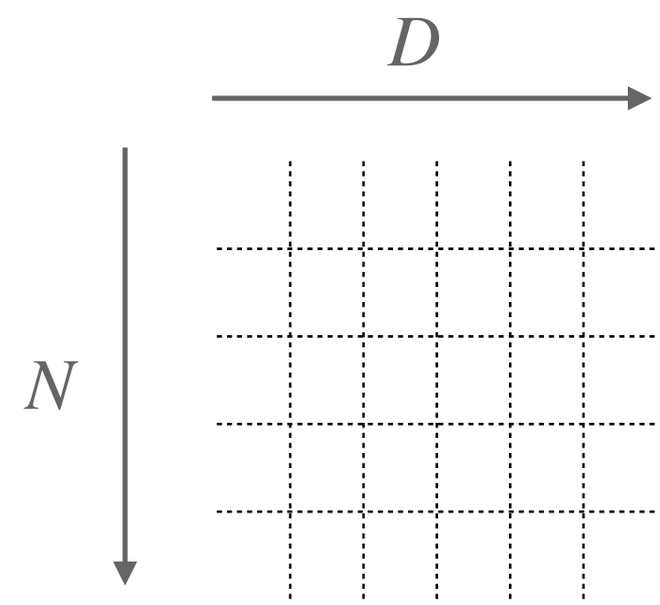
$$\begin{pmatrix} \textbf{math} \\ \textbf{physics} \\ \textbf{language} \end{pmatrix} = f(x_1, x_2)$$



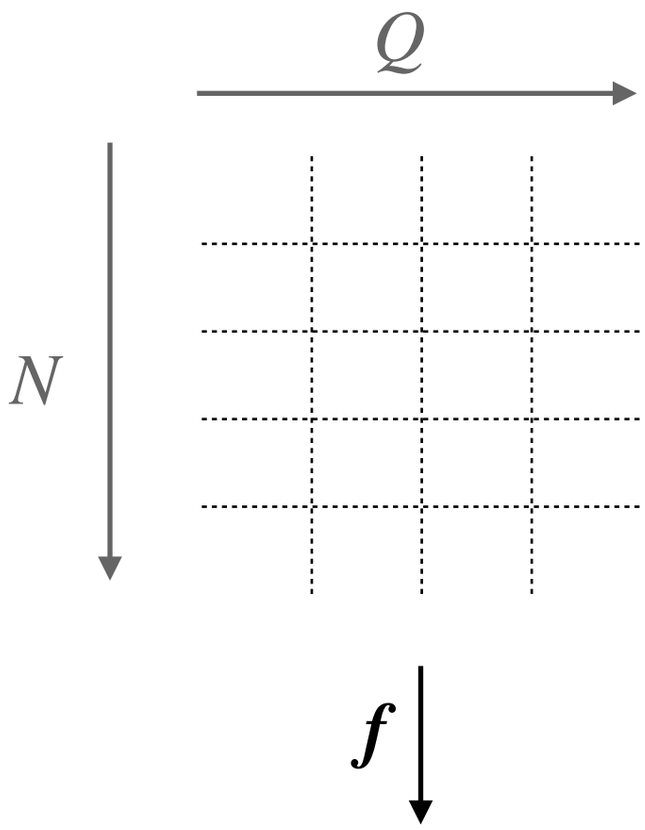


$$\begin{pmatrix} \textbf{math} \\ \textbf{physics} \\ \textbf{language} \end{pmatrix}$$

$$\begin{pmatrix} \textbf{math} \\ \textbf{physics} \\ \textbf{language} \end{pmatrix} = f(x_1, x_2)$$

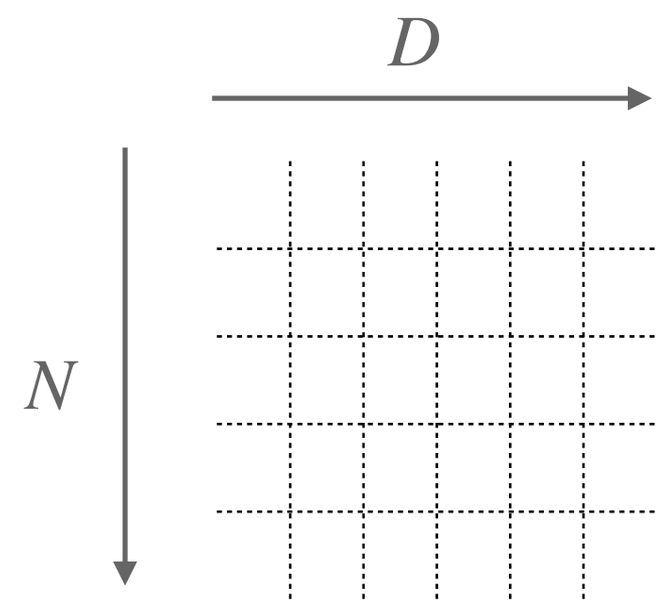


$$\begin{pmatrix} \textbf{math} \\ \textbf{physics} \\ \textbf{language} \end{pmatrix} = \begin{pmatrix} \uparrow & \downarrow \\ \uparrow & \downarrow \\ \downarrow & \uparrow \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$



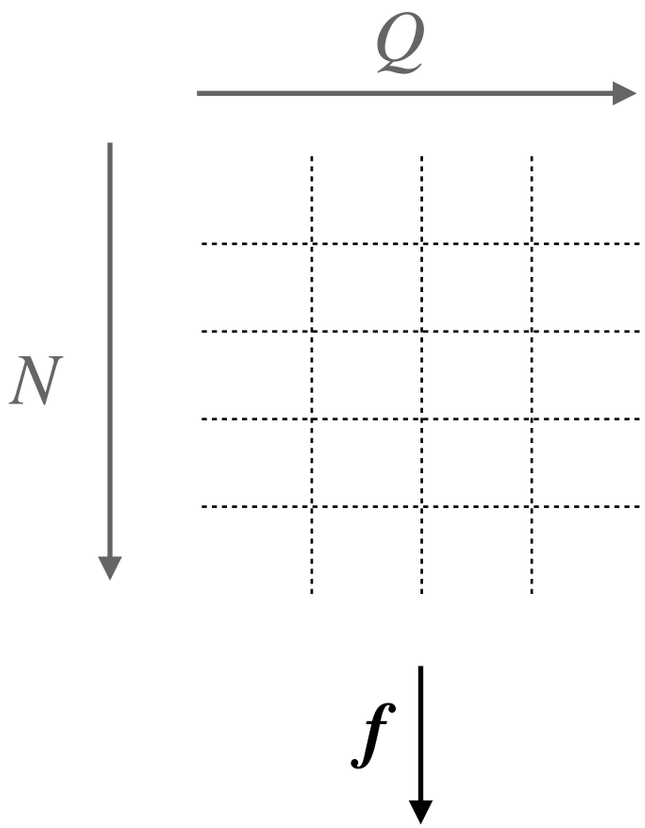
$$\begin{pmatrix} \textbf{math} \\ \textbf{physics} \\ \textbf{language} \end{pmatrix}$$

$$\begin{pmatrix} \textbf{math} \\ \textbf{physics} \\ \textbf{language} \end{pmatrix} = f(x_1, x_2)$$

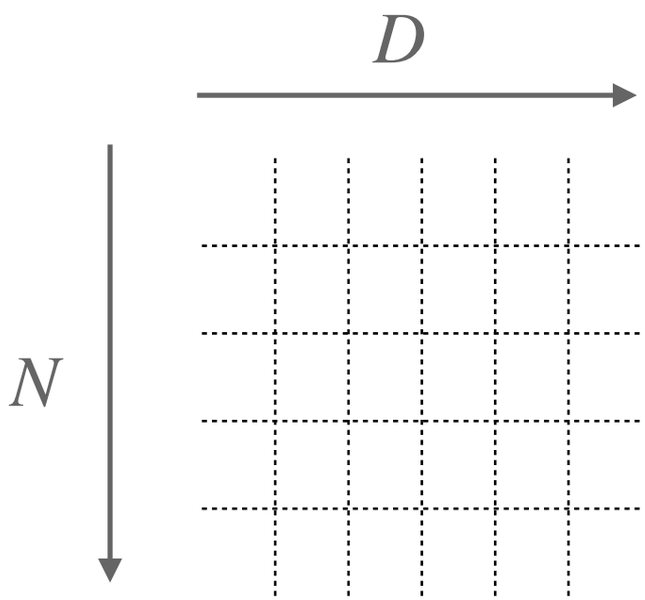


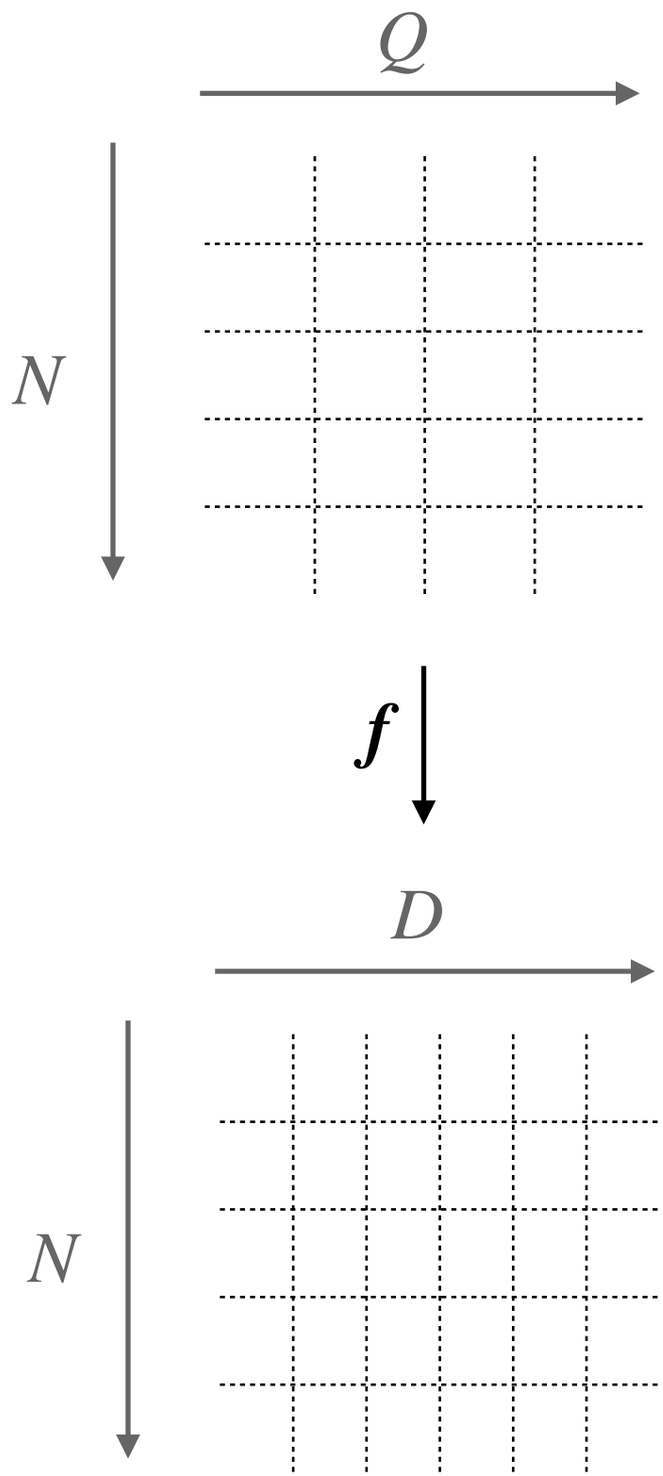
$$\begin{pmatrix} \textbf{math} \\ \textbf{physics} \\ \textbf{language} \end{pmatrix} = \begin{pmatrix} \uparrow & \downarrow \\ \uparrow & \downarrow \\ \downarrow & \uparrow \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$\begin{pmatrix} \textbf{math} \\ \textbf{physics} \\ \textbf{language} \end{pmatrix} = \begin{pmatrix} \uparrow & \downarrow \\ \uparrow & \downarrow \\ \downarrow & \uparrow \end{pmatrix} \begin{pmatrix} IQ \\ EQ \end{pmatrix}$$



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



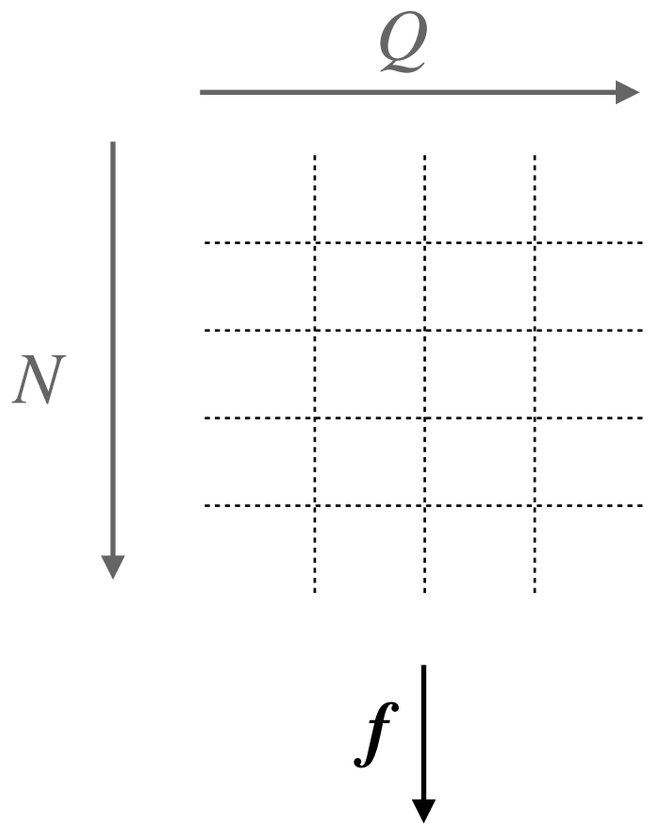


$$\boldsymbol{X} \in \mathbb{R}^{N \times Q} \quad \rightarrow \quad \boldsymbol{Y} \in \mathbb{R}^{N \times D}$$

Generative point of view:

$$\boldsymbol{x} \sim p_0(\boldsymbol{x})$$

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon}$$



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

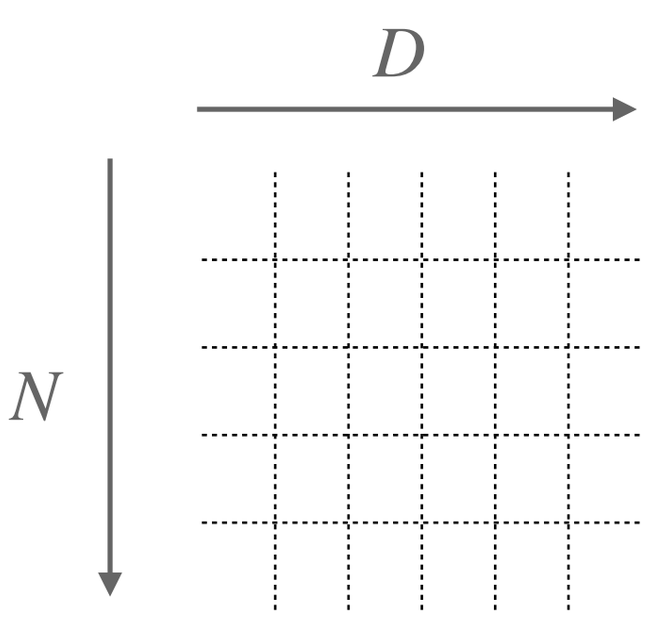
Generative point of view:

$$\boldsymbol{x} \sim p_0(\boldsymbol{x})$$

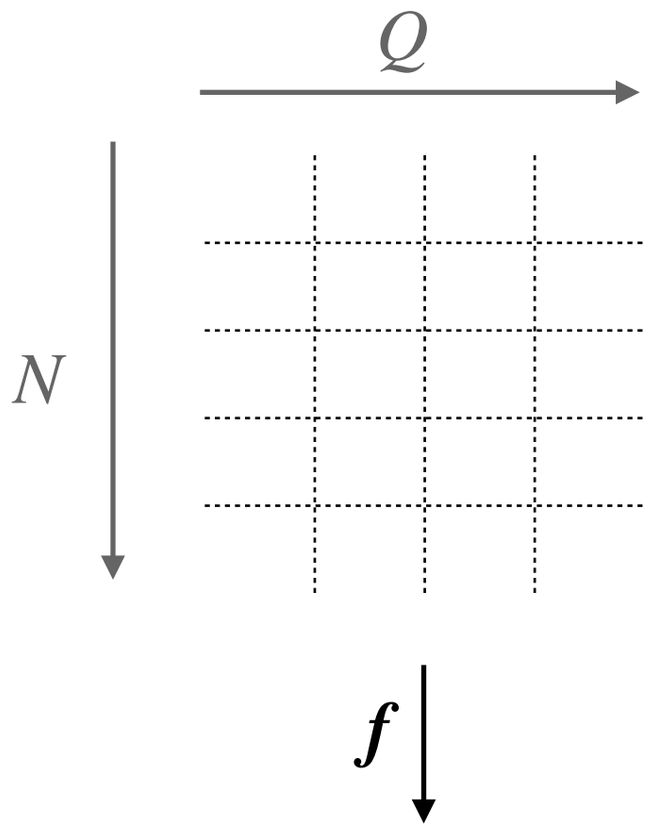
$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon}$$

Given $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$, what is $p(\boldsymbol{X} \mid \boldsymbol{Y})$ and $p(\boldsymbol{Y})$**?**

$$p(\boldsymbol{X} \mid \boldsymbol{Y}) = \frac{p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})}{p(\boldsymbol{Y})}$$



$$p(\boldsymbol{Y}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})\ d\boldsymbol{X}$$



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

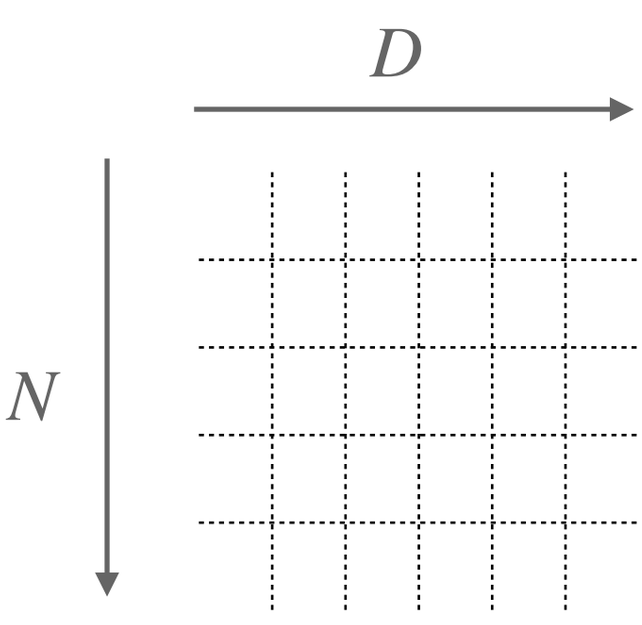
Generative point of view:

$$\boldsymbol{x} \sim p_0(\boldsymbol{x})$$

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon}$$

Given $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$, what is $p(\boldsymbol{X} \mid \boldsymbol{Y})$ and $p(\boldsymbol{Y})$**?**

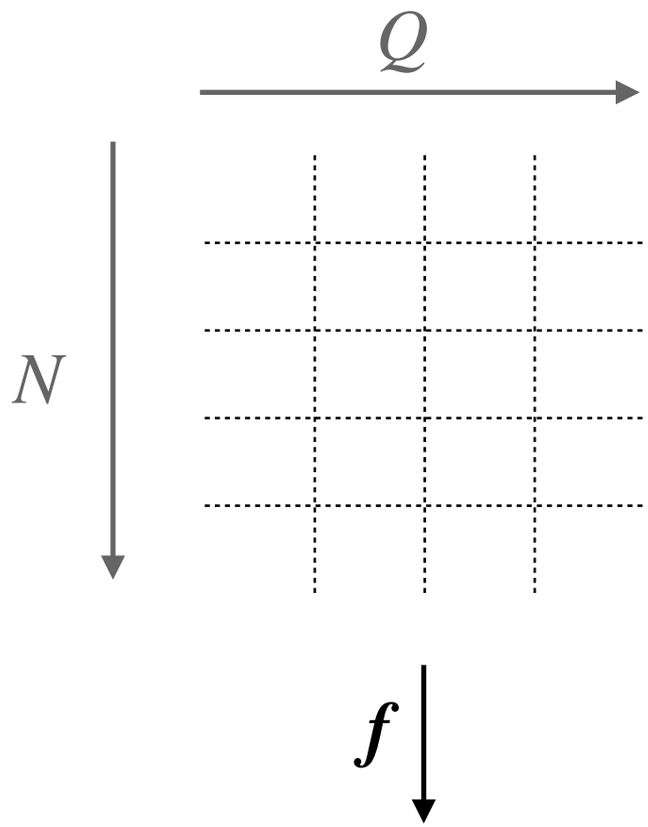
$$p(\boldsymbol{X} \mid \boldsymbol{Y}) = \frac{p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})}{p(\boldsymbol{Y})}$$



$$p(\boldsymbol{Y}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})\ d\boldsymbol{X}$$

PPCA

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\epsilon}$$



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

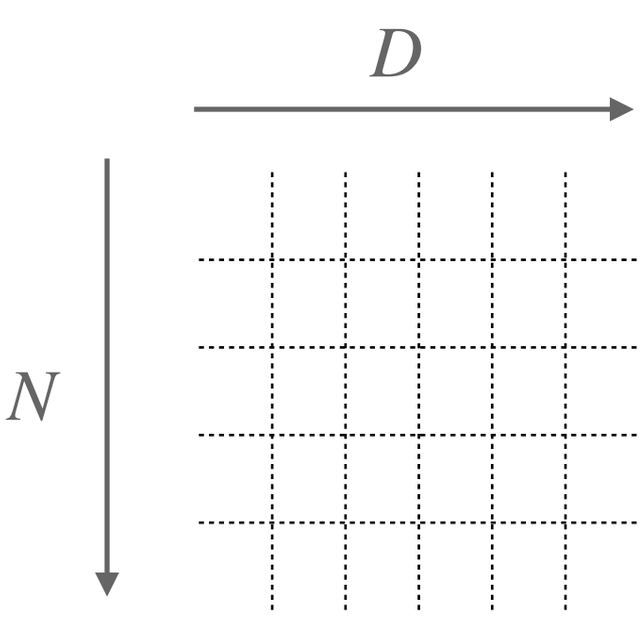
Generative point of view:

$$\boldsymbol{x} \sim p_0(\boldsymbol{x})$$

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon}$$

Given $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$, what is $p(\boldsymbol{X} \mid \boldsymbol{Y})$ and $p(\boldsymbol{Y})$**?**

$$p(\boldsymbol{X} \mid \boldsymbol{Y}) = \frac{p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})}{p(\boldsymbol{Y})}$$



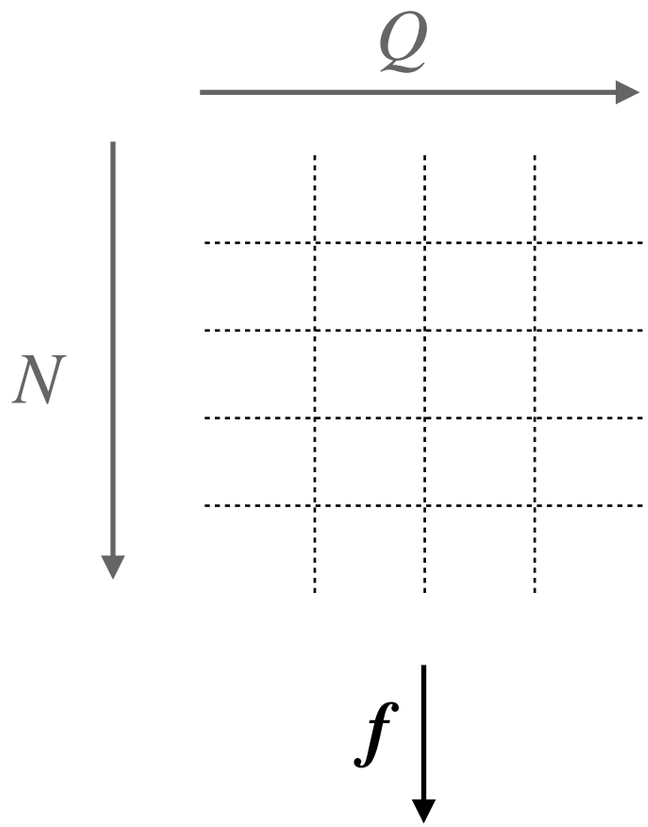
$$p(\boldsymbol{Y}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})\ d\boldsymbol{X}$$

PPCA

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\epsilon}$$

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W}) = \mathcal{N}(\boldsymbol{Y} \mid \boldsymbol{X}\boldsymbol{W}^T, \sigma^2\boldsymbol{I}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{W}\boldsymbol{x}_n, \sigma^2\boldsymbol{I})$$

$$p(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$



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

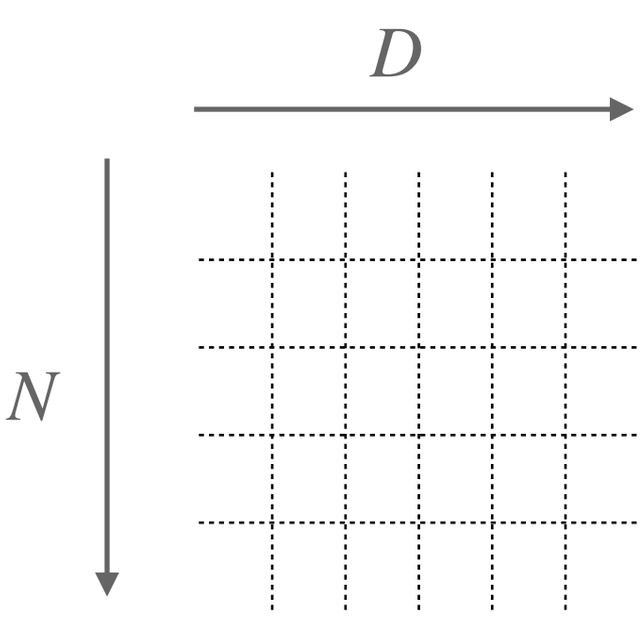
Generative point of view:

$$\boldsymbol{x} \sim p_0(\boldsymbol{x})$$

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon}$$

Given $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$, what is $p(\boldsymbol{X} \mid \boldsymbol{Y})$ and $p(\boldsymbol{Y})$**?**

$$p(\boldsymbol{X} \mid \boldsymbol{Y}) = \frac{p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})}{p(\boldsymbol{Y})}$$



$$p(\boldsymbol{Y}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})\ d\boldsymbol{X}$$

PPCA

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\epsilon}$$

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W}) = \mathcal{N}(\boldsymbol{Y} \mid \boldsymbol{X}\boldsymbol{W}^T, \sigma^2 \boldsymbol{I}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{W}\boldsymbol{x}_n, \sigma^2 \boldsymbol{I})$$

$$p(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

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

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

$$p(\boldsymbol{X} \mid \boldsymbol{Y}) = \frac{p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})}{p(\boldsymbol{Y})}$$

$$p(\boldsymbol{Y}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})\ d\boldsymbol{X}$$

Given $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$, what is $p(\boldsymbol{X} \mid \boldsymbol{Y})$ and $p(\boldsymbol{Y})$?

PPCA

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\epsilon}$$

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W}) = \mathcal{N}(\boldsymbol{Y} \mid \boldsymbol{X}\boldsymbol{W}^T, \sigma^2 \boldsymbol{I})$$

$$= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{W}\boldsymbol{x}_n, \sigma^2 \boldsymbol{I})$$

$$p(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

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

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

$$p(\boldsymbol{X} \mid \boldsymbol{Y}) = \frac{p(\boldsymbol{Y} \mid \boldsymbol{X})\, p(\boldsymbol{X})}{p(\boldsymbol{Y})}$$

$$p(\boldsymbol{Y}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X})\, p(\boldsymbol{X})\, d\boldsymbol{X}$$

Given $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$, what is $p(\boldsymbol{X} \mid \boldsymbol{Y})$ and $p(\boldsymbol{Y})$?

PPCA

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\epsilon}$$

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W}) = \mathcal{N}(\boldsymbol{Y} \mid \boldsymbol{X}\boldsymbol{W}^T, \sigma^2 \boldsymbol{I})$$

$$= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{W}\boldsymbol{x}_n, \sigma^2 \boldsymbol{I})$$

$$p(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

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

Variational Autoencoder

$$\boldsymbol{y} = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{\epsilon}$$

$\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$ is NN

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_n), \sigma^2 \boldsymbol{I}), \quad \boldsymbol{X} \sim p(\boldsymbol{X})$$

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

$$p(\boldsymbol{X} \mid \boldsymbol{Y}) = \frac{p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})}{p(\boldsymbol{Y})}$$

$$p(\boldsymbol{Y}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})\ d\boldsymbol{X}$$

Given $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$, what is $p(\boldsymbol{X} \mid \boldsymbol{Y})$ and $p(\boldsymbol{Y})$?

PPCA

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\epsilon}$$

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W}) = \mathcal{N}(\boldsymbol{Y} \mid \boldsymbol{X}\boldsymbol{W}^T, \sigma^2 \boldsymbol{I})$$

$$= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{W}\boldsymbol{x}_n, \sigma^2 \boldsymbol{I})$$

$$p(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

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

Variational Autoencoder

$$\boldsymbol{y} = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{\epsilon}$$

$\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$ is NN

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_n), \sigma^2 \boldsymbol{I}), \quad \boldsymbol{X} \sim p(\boldsymbol{X})$$

$$p(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\ p(\boldsymbol{X})\ d\boldsymbol{X}$$

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

$$p(\boldsymbol{X} \mid \boldsymbol{Y}) = \frac{p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})}{p(\boldsymbol{Y})}$$

$$p(\boldsymbol{Y}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})\ d\boldsymbol{X}$$

Given $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$, what is $p(\boldsymbol{X} \mid \boldsymbol{Y})$ and $p(\boldsymbol{Y})$?

PPCA

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\epsilon}$$

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W}) = \mathcal{N}(\boldsymbol{Y} \mid \boldsymbol{X}\boldsymbol{W}^T, \sigma^2 \boldsymbol{I})$$

$$= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{W}\boldsymbol{x}_n, \sigma^2 \boldsymbol{I})$$

$$p(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

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

Variational Autoencoder

$$\boldsymbol{y} = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{\epsilon}$$

$\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$ is NN

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_n), \sigma^2 \boldsymbol{I}), \quad \boldsymbol{X} \sim p(\boldsymbol{X})$$

$$p(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\ p(\boldsymbol{X})\ d\boldsymbol{X}$$

$$= \int q(\boldsymbol{X}) \frac{p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\ p(\boldsymbol{X})}{q(\boldsymbol{X})}\ d\boldsymbol{X}$$

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

$$p(\boldsymbol{X} \mid \boldsymbol{Y}) = \frac{p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})}{p(\boldsymbol{Y})}$$

$$p(\boldsymbol{Y}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})\ d\boldsymbol{X}$$

Given $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$, what is $p(\boldsymbol{X} \mid \boldsymbol{Y})$ and $p(\boldsymbol{Y})$?

PPCA

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\epsilon}$$

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W}) = \mathcal{N}(\boldsymbol{Y} \mid \boldsymbol{X}\boldsymbol{W}^T, \sigma^2 \boldsymbol{I})$$

$$= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{W}\boldsymbol{x}_n, \sigma^2 \boldsymbol{I})$$

$$p(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

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

Variational Autoencoder

$$\boldsymbol{y} = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{\epsilon}$$

$\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$ is NN

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_n), \sigma^2 \boldsymbol{I}), \quad \boldsymbol{X} \sim p(\boldsymbol{X})$$

$$p(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\ p(\boldsymbol{X})\ d\boldsymbol{X}$$

$$= \int q(\boldsymbol{X}) \frac{p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\ p(\boldsymbol{X})}{q(\boldsymbol{X})}\ d\boldsymbol{X}$$

Meaningful $q(\boldsymbol{X})$ would be the posterior but is intractable → Variational Approximation

$$q(\boldsymbol{X}) = q_{\boldsymbol{\phi}}(\boldsymbol{X} \mid \boldsymbol{Y}) = \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{f}_{\boldsymbol{\phi}}(\boldsymbol{y}_n), \sigma^2 \boldsymbol{I})$$

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

$$p(\boldsymbol{X} \mid \boldsymbol{Y}) = \frac{p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})}{p(\boldsymbol{Y})}$$

$$p(\boldsymbol{Y}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X})\ p(\boldsymbol{X})\ d\boldsymbol{X}$$

Given $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$, what is $p(\boldsymbol{X} \mid \boldsymbol{Y})$ and $p(\boldsymbol{Y})$?

PPCA

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\epsilon} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\epsilon}$$

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W}) = \mathcal{N}(\boldsymbol{Y} \mid \boldsymbol{X}\boldsymbol{W}^T, \sigma^2 \boldsymbol{I})$$

$$= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{W}\boldsymbol{x}_n, \sigma^2 \boldsymbol{I})$$

$$p(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

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

Variational Autoencoder

$$\boldsymbol{y} = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{\epsilon}$$

$\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$ is NN

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_n), \sigma^2 \boldsymbol{I}), \quad \boldsymbol{X} \sim p(\boldsymbol{X})$$

$$p(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\ p(\boldsymbol{X})\ d\boldsymbol{X}$$

$$= \int q(\boldsymbol{X}) \frac{p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\ p(\boldsymbol{X})}{q(\boldsymbol{X})}\ d\boldsymbol{X}$$

Meaningful $q(\boldsymbol{X})$ would be the posterior but is intractable → Variational Approximation

$$q(\boldsymbol{X}) = q_{\boldsymbol{\phi}}(\boldsymbol{X} \mid \boldsymbol{Y}) = \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{f}_{\boldsymbol{\phi}}(\boldsymbol{y}_n), \sigma^2 \boldsymbol{I})$$

$$\ln p(\boldsymbol{Y} \mid \boldsymbol{\theta}) - \mathbf{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{X} \mid \boldsymbol{Y}) \| p(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{\theta})\right] =$$

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{X}|\boldsymbol{Y})}\left[\ln p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\right] - \mathbf{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{X} \mid \boldsymbol{Y}) \| p(\boldsymbol{X})\right]$$

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***

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

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

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

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

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

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***

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

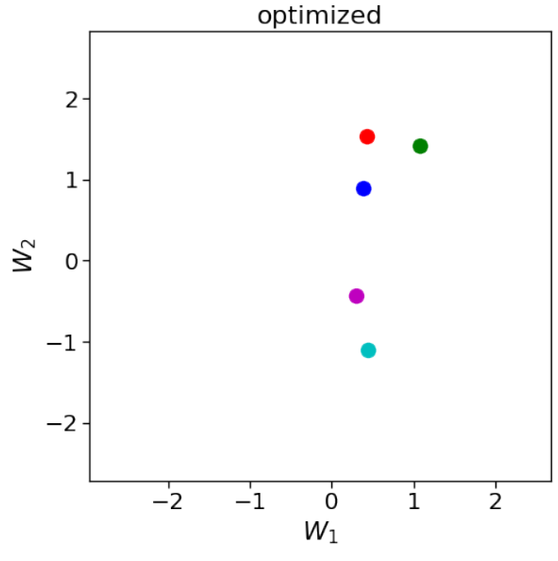
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$$

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

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

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

- 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***

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

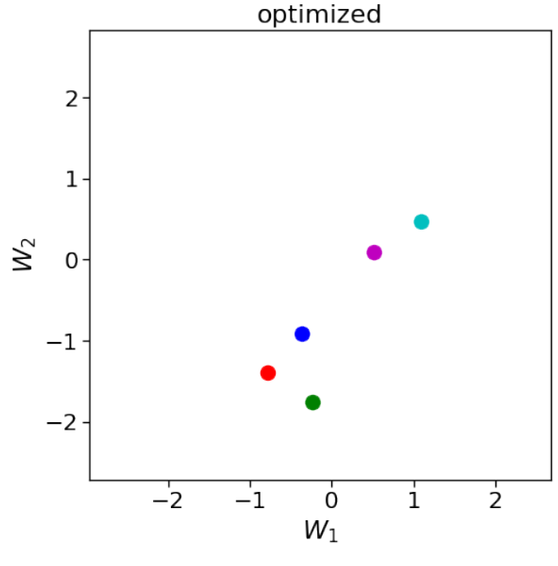
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$$

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

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

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

- 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***

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

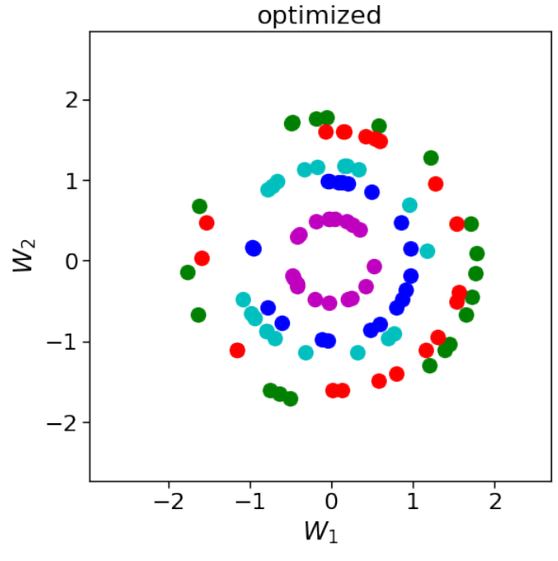
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$$

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

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

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

- 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***

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

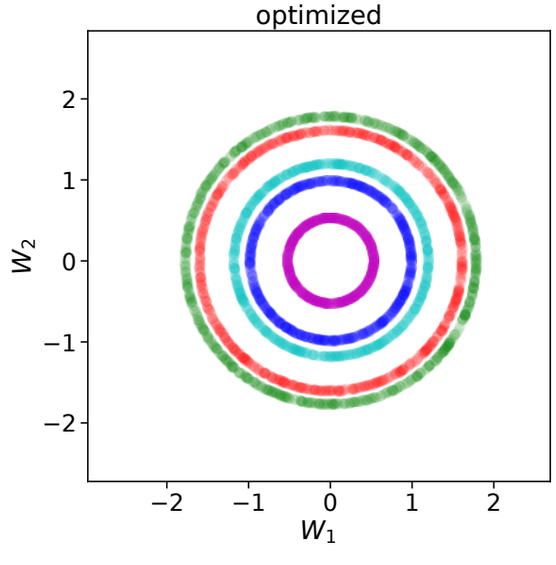
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^T + \boldsymbol{\epsilon}$$

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

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

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

  - Rotation invariant likelihood



Bayesian approach to PPCA

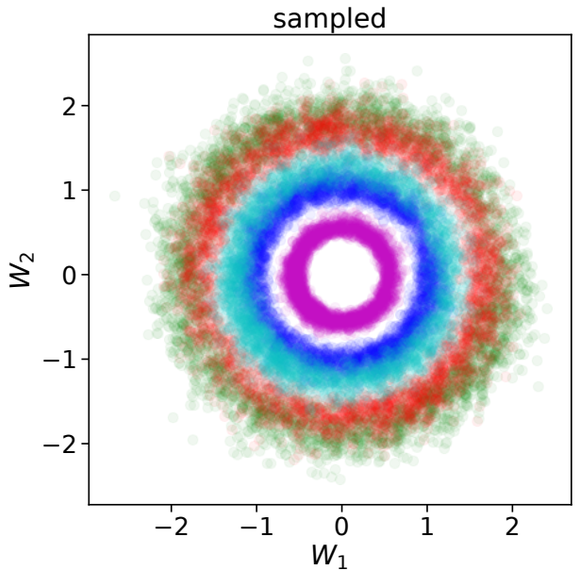
$$p\left(\boldsymbol{W}\middle|\boldsymbol{Y}\right) = \frac{p\left(\boldsymbol{Y}\middle|\boldsymbol{W}\right)p\left(\boldsymbol{W}\right)}{p(\boldsymbol{Y})}$$

- 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

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



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

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 $\boldsymbol{W}$ and $\boldsymbol{U}$, $\boldsymbol{\Sigma}^2$ is eigenvalue decomposition of $\boldsymbol{W}\boldsymbol{W}^\mathsf{T}$ → $\boldsymbol{U}$ is eigenvector matrix

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

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

→ U is uniformly distributed on the Stiefel manifold

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

Theory

- Since $\boldsymbol{U}$, $\boldsymbol{\Sigma}$ is SVD of $\boldsymbol{W}$ and $\boldsymbol{U}$, $\boldsymbol{\Sigma}^2$ is eigenvalue decomposition of $\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}$ → $\boldsymbol{U}$ is eigenvector matrix

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

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

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

---

- Square of ordered eigenvalue matrix $\boldsymbol{\Sigma}$ 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:

$U \sim$ **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** $\mathbb{S}^{n-1}$

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

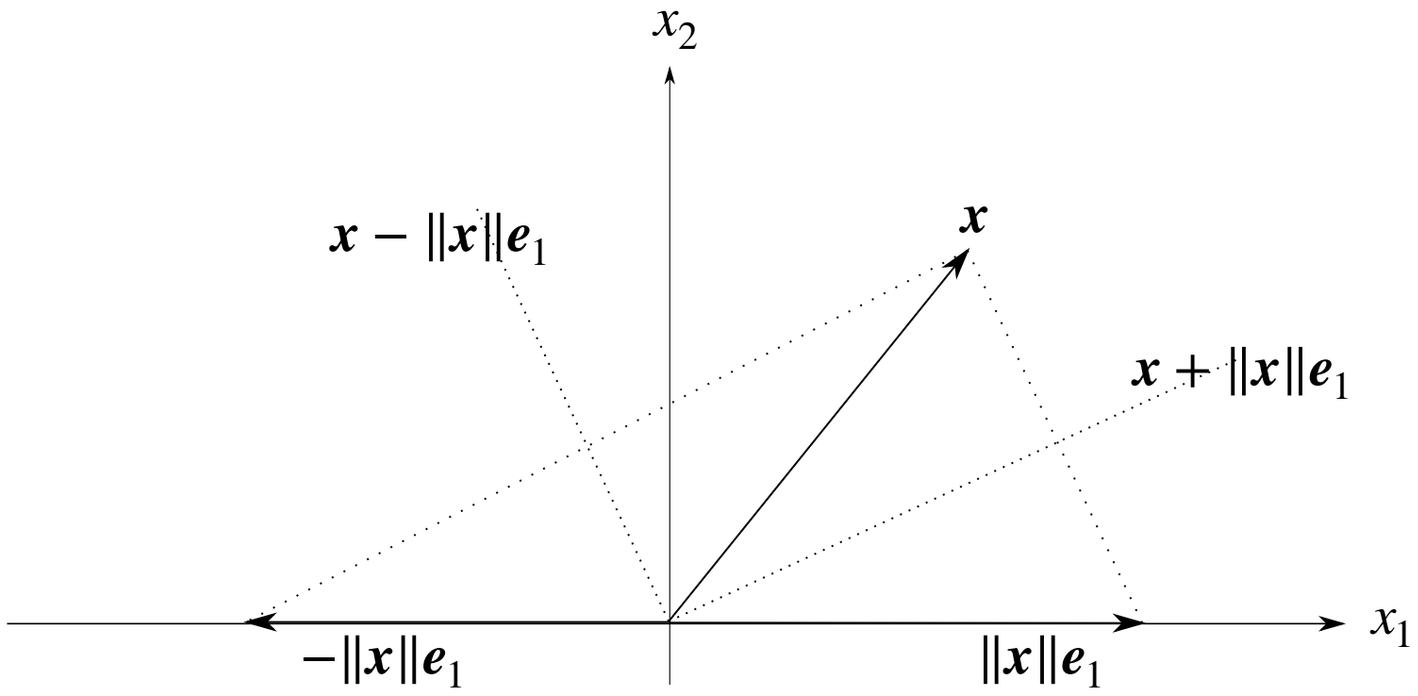
$$\tilde{\boldsymbol{H}}_{\boldsymbol{n}}(\boldsymbol{v}_n) = -\operatorname{sgn}(\boldsymbol{v}_{n1})\left(\boldsymbol{I} - 2\boldsymbol{u}_n\boldsymbol{u}_n^T\right)$$

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

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

Householder Transformations

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



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

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

Householder Transformations

Example for $(D, Q) = (2,2)$

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

for $n = D : 1$

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

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

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

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

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

Householder Transformations

Example for $(D, Q) = (2,2)$

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

Construction of $\boldsymbol{H}_1$

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

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

for $n = D : 1$

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

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

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

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

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

Householder Transformations

Example for $(D, Q) = (2,2)$

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

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 $\boldsymbol{H}_2$

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

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

for $n = D : 1$

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

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

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

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

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

Householder Transformations

Example for $(D, Q) = (2,2)$

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

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 $\boldsymbol{H}_2$

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

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

for $n = D : 1$

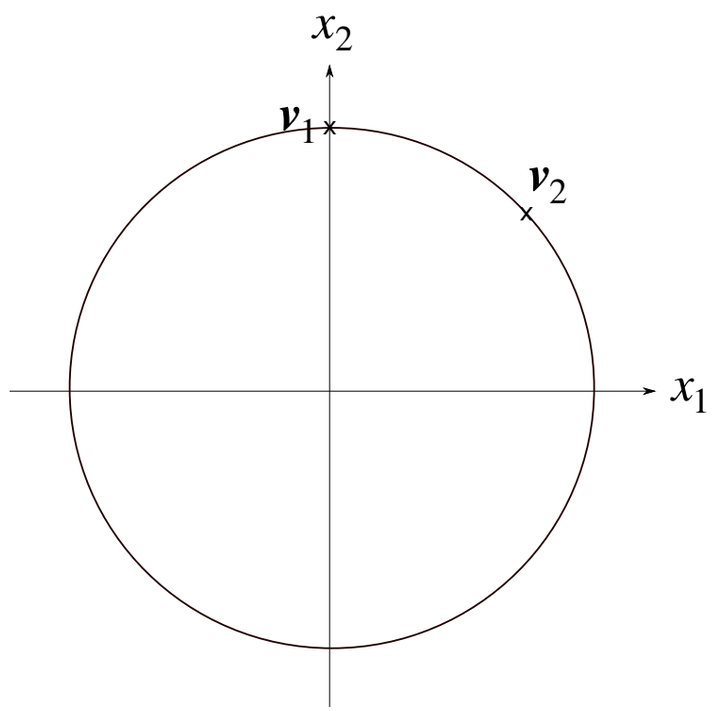
$\boldsymbol{v}_n \sim$ **uniform on** $\mathbb{S}^{n-1}$

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

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

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

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



Householder Transformations

Example for $(D, Q) = (2,2)$

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

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 $\boldsymbol{H}_2$

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

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

for $n = D : 1$

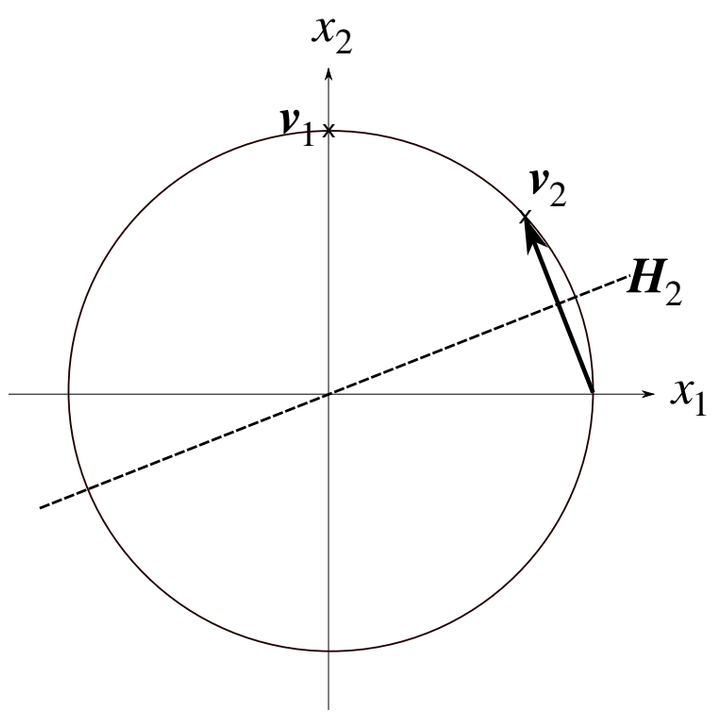
$$\boldsymbol{v}_n \sim \textbf{uniform on } \mathbb{S}^{n-1}$$

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

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

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

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



Householder Transformations

Example for $(D, Q) = (2,2)$

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

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 $\boldsymbol{H}_2$

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

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

for $n = D : 1$

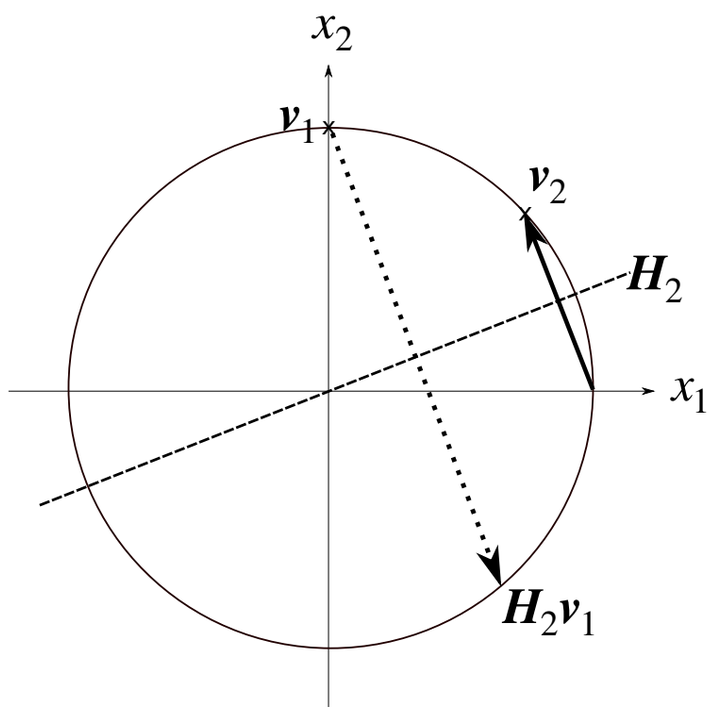
$\boldsymbol{v}_n \sim$ **uniform on** $\mathbb{S}^{n-1}$

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

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

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

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



Householder Transformations

Example for $(D, Q) = (2,2)$

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

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 $\boldsymbol{H}_2$

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

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

for $n = D : 1$

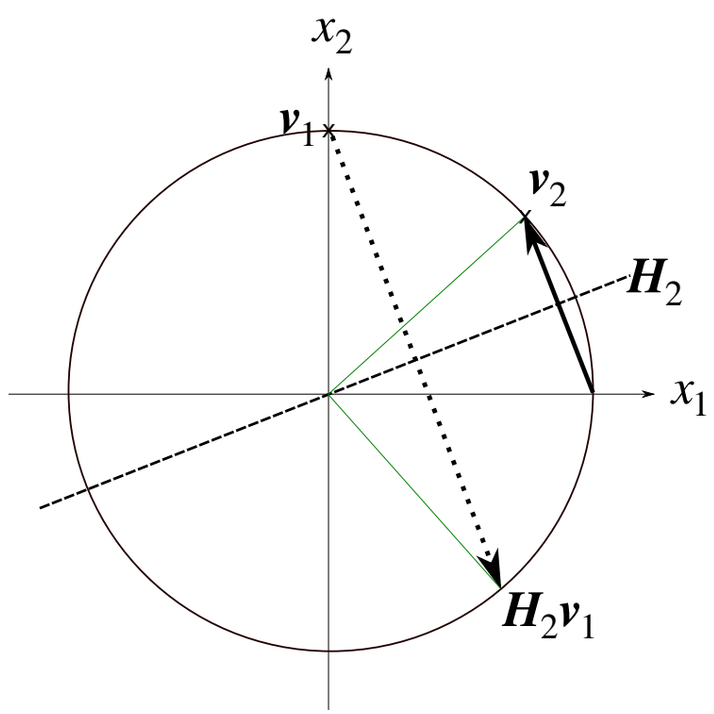
$$\boldsymbol{v}_n \sim \textbf{uniform on } \mathbb{S}^{n-1}$$

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

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

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

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



Householder Transformations

Example for $(D, Q) = (2,2)$

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

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 $n = D : 1$

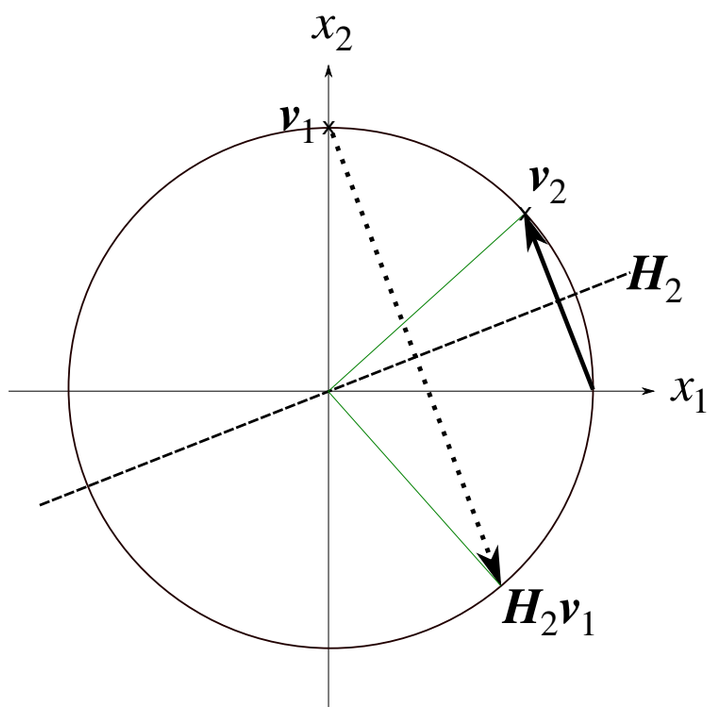
$\boldsymbol{v}_n \sim$ **uniform on** $\mathbb{S}^{n-1}$

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

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

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

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



Implementation

The full generative model for Bayesian PPCA:

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

Results

Synthetic Dataset

- Construction

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

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

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

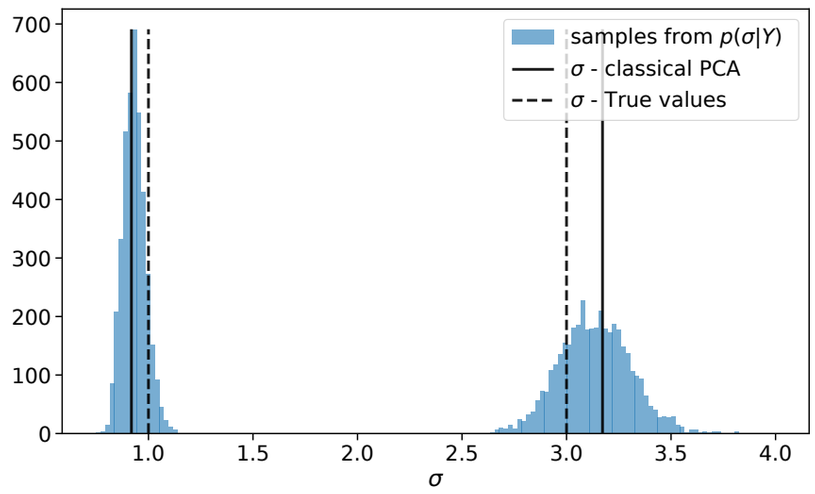
$$\boldsymbol{\epsilon} \sim \mathcal{N}(0,\ 0.01) \in \mathbb{R}^{N \times D}$$

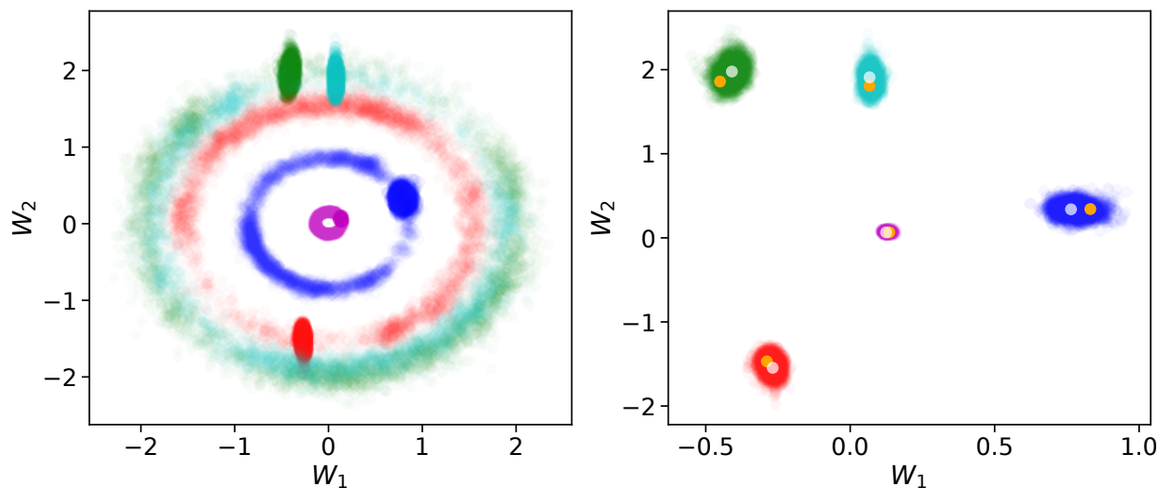
$$\boldsymbol{\Sigma} = \operatorname{diag}\left(\sigma_1, \sigma_2\right) = \operatorname{diag}\left(3.0,\ 1.0\right)$$

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

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

- Inference

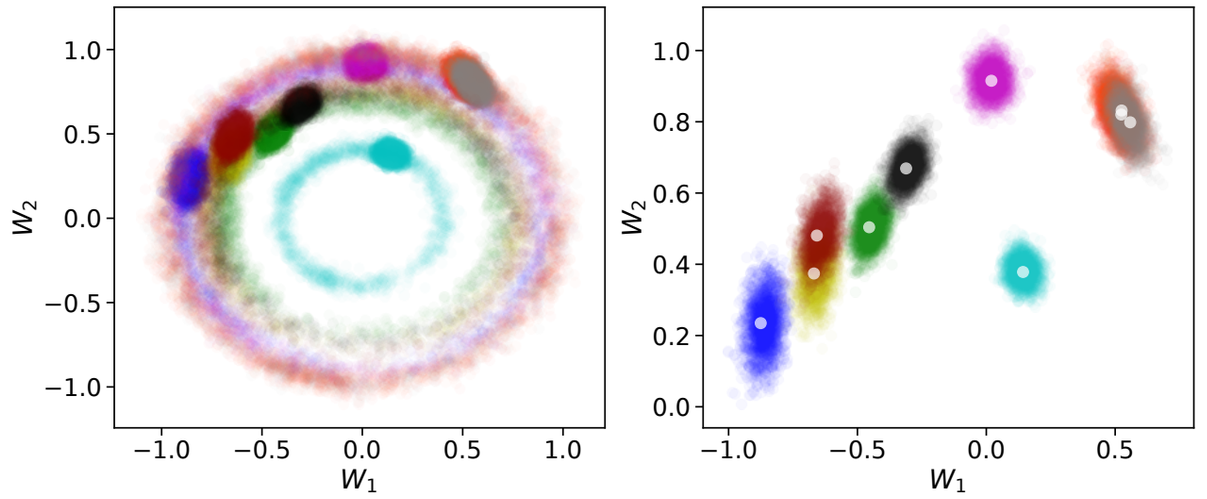




Results

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

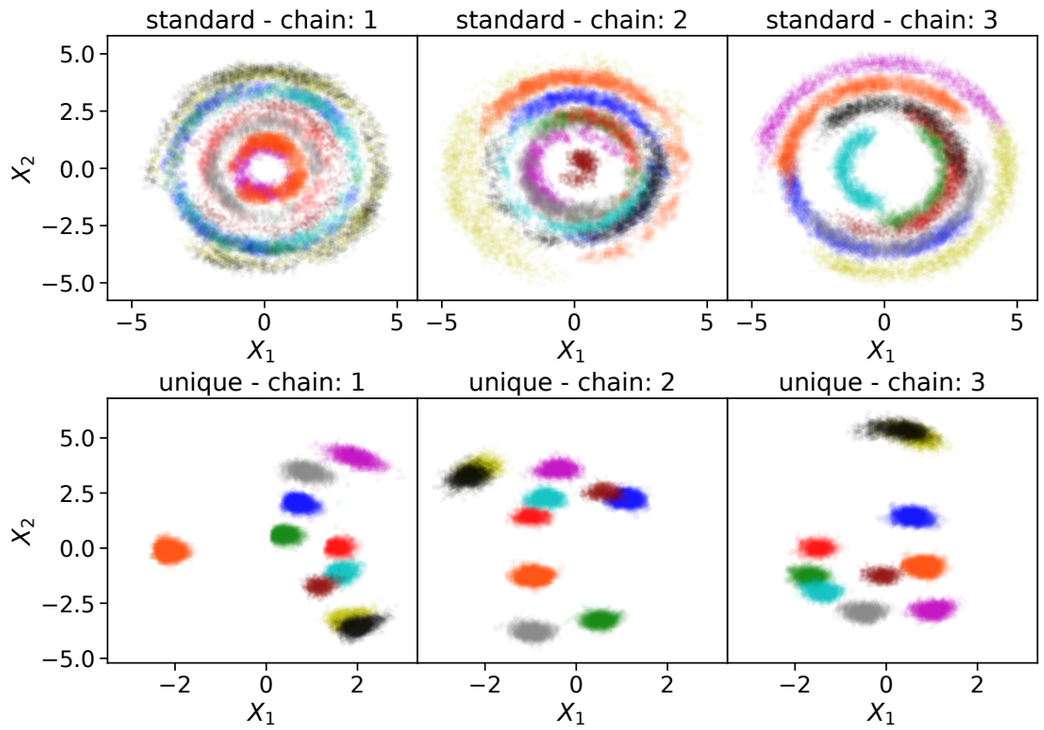
Extension to non-linear models

- GPLVM with the same rotation invariant problem

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

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

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



- 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 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 $\boldsymbol{W}$

- 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

Thanks for your attention!

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