Rotation Invariant Householder Parameterization for Bayesian PCA

Rajbir-Singh Nirwan, Nils Bertschinger
June 11, 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 $Y$ to a lower dimensional latent space $X$
  
  $Y \in \mathbb{R}^{N \times D} \rightarrow X \in \mathbb{R}^{N \times Q}$

  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 \( X \)
  \[
  Y \in \mathbb{R}^{N \times D} \quad \rightarrow \quad X \in \mathbb{R}^{N \times Q}
  \]
  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 \( Y \)
  \[
  X \in \mathbb{R}^{N \times Q} \quad \rightarrow \quad Y \in \mathbb{R}^{N \times D}
  \]
  \[
  Y = XW^T + \epsilon
  \]
  \[
  X \sim \mathcal{N}(0, I), \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I)
  \]
  \[
  p(Y|W) = \prod_{n=1}^{N} \mathcal{N}(Y_{n,:}|0, WW^T + \sigma^2 I)
  \]
  \[
  WRR^T W^T = WW^T \quad \forall \; RR^T = I
  \]
Probabilistic PCA

• Classical PCA

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

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

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 $Y$

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

$$Y = XW^T + \epsilon$$

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

$$p(Y|W) = \prod_{n=1}^{N} \mathcal{N}(Y_n, |0, WW^T + \sigma^2 I)$$

$$WRR^TW^T = WW^T \quad \forall RR^T = I$$

• Optimization for $D=5$, $Q=2$
Probabilistic PCA

• Classical PCA
  
  Formulated as a projection from data space $Y$ to a lower dimensional latent space $X$
  
  $$Y \in \mathbb{R}^{N \times D} \rightarrow X \in \mathbb{R}^{N \times Q}$$

  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 $Y$

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

  $$Y = XW^T + \epsilon$$

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

  $$p(Y|W) = \prod_{n=1}^{N} \mathcal{N}(Y_n; |0, WW^T + \sigma^2 I)$$

  $$WRR^TW^T = WW^T \quad \forall RR^T = I$$

• Optimization for D=5, Q=2
Probabilistic PCA

• Classical PCA
  Formulated as a projection from data space $Y$ to a lower dimensional latent space $X$
  \[ Y \in \mathbb{R}^{N \times D} \rightarrow X \in \mathbb{R}^{N \times Q} \]
  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 $Y$
  \[ X \in \mathbb{R}^{N \times Q} \rightarrow Y \in \mathbb{R}^{N \times D} \]
  \[ Y = XW^T + \epsilon \]
  \[ X \sim \mathcal{N}(0, I), \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I) \]
  \[ p(Y|W) = \prod_{n=1}^{N} \mathcal{N}(Y_{n,:}|0, WW^T + \sigma^2 I) \]
  \[ WRR^TW^T = WW^T \quad \forall RR^T = I \]

• Optimization for $D=5$, $Q=2$
Probabilistic PCA

• Classical PCA

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

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

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 $Y$

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

$$Y = XW^T + \epsilon$$

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

$$p(Y|W) = \prod_{n=1}^{N} \mathcal{N}(Y_n; |0, WW^T + \sigma^2 I)$$

$WRR^T W^T = WW^T \quad \forall \ RR^T = I$

• Rotation invariant likelihood
Bayesian approach to PPCA

\[ p(W|Y) = \frac{p(Y|W)p(W)}{p(Y)} \]

• 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

\[ p(W|Y) = \frac{p(Y|W)p(W)}{p(Y)} \]

- 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 $W$
  \[
  WW^T = U \Sigma V^T (U \Sigma V^T)^T = U \Sigma^2 U^T
  \]

• Fix coordinate system
  \[V = I\]

• Specify correct prior
  \[p(U, \Sigma)\]

• Sample from
  \[p(U, \Sigma | Y)\]
Solution

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

Outline of procedure

• SVD of $W$
  \[ WW^T = U\Sigma V^T (U\Sigma V^T)^T = U\Sigma^2 U^T \]
  \[ V = I \]
• Specify correct prior
  \[ p(U, \Sigma) \]
• Sample from
  \[ p(U, \Sigma | Y) \]

\[ W \sim \mathcal{N}(0, I) \quad \rightarrow \quad WW^T \text{ is Wishart distributed} \]
\[ U \sim ? \quad \rightarrow \quad U\Sigma\Sigma^TU^T \text{ is Wishart distributed} \]
\[ \Sigma \sim ? \]
Theory

• Since $U, \Sigma$ is SVD of $W$ and $U, \Sigma^2$ is eigenvalue decomposition of $WW^T \rightarrow U$ is eigenvector matrix

$$U \in \mathcal{V}_{Q,D} \quad \text{Stiefel manifold} \quad \mathcal{V}_{Q,D} = \{ U \in \mathbb{R}^{D \times Q} | U^T U = I \}$$

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

$\rightarrow U$ is uniformly distributed on the Stiefel manifold
Theory

• Since $U$, $\Sigma$ is SVD of $W$ and $U$, $\Sigma^2$ is eigenvalue decomposition of $WW^T \rightarrow U$ is eigenvector matrix

$$U \in \mathcal{V}_{Q,D} \quad \text{Stiefel manifold} \quad \mathcal{V}_{Q,D} = \{ U \in \mathbb{R}^{D \times Q} | U^T U = I \}$$

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

$\rightarrow U$ is uniformly distributed on the Stiefel manifold

• Square of ordered eigenvalue matrix $\Sigma$ is distributed as (James & Lee (2014))

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

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

- Need:
  \[ U \sim \text{uniform on Stiefel } \mathcal{V}_{Q,D} \]
  \[ \Sigma \sim p(\Sigma) \leftarrow \text{easy, since we know the analytic exp for density} \]

**Theorem 2** Let \( v_D, v_{D-1}, \ldots, v_1 \) be uniformly distributed on the unit spheres \( S^{D-1}, \ldots, S^0 \) respectively, where \( S^{n-1} \) is the unit sphere in \( \mathbb{R}^n \). Furthermore, let \( \tilde{H}_n(v_n) \) be the \( n \)-th Householder transformation as defined in equation (2.20)

The product

\[ Q = H_D(v_D)H_{D-1}(v_{D-1}) \ldots H_1(v_1) \] (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** \( \mathcal{V}_{Q,D} \)

**for** \( n = D : 1 \)

\[ v_n \sim \text{uniform on } S^{n-1} \]

\[ u_n = \frac{v_n + \text{sgn}(v_{n1}) \parallel v_n \parallel e_1}{\parallel v_n + \text{sgn}(v_{n1}) \parallel v_n \parallel e_1 \parallel} \]

\[ \tilde{H}_n(v_n) = -\text{sgn}(v_{n1}) (I - 2u_nu_n^T) \]

\[ H_n = \begin{pmatrix} I & 0 \\ 0 & H_n \end{pmatrix} \]

\[ U = H_D(v_D)H_{D-1}(v_{D-1}) \ldots H_1(v_1) \]
Implementation

The full generative model for Bayesian PPCA:

\[ v_D, \ldots, v_{D-Q+1} \sim \mathcal{N}(0, I) \]
\[ \sigma \sim p(\sigma) \]
\[ \mu \sim p(\mu) \]
\[ U = \prod_{q=1}^{Q} H_{D-q+1}(v_{D-q+1}) \]
\[ \Sigma = \text{diag}(\sigma) \]
\[ W = U \Sigma \]
\[ \sigma_{\text{noise}} \sim p(\sigma_{\text{noise}}) \]
\[ Y \sim \prod_{n=1}^{N} \mathcal{N}(Y_n; |\mu, WW^T + \sigma_{\text{noise}}^2 I) \]
Results

Synthetic Dataset

• Construction

\[(N, D, Q) = (150, 5, 2)\]

\[X \sim \mathcal{N}(0, I) \in \mathbb{R}^{N \times Q}\]

\[U \sim \text{uniform on Stiefel } \mathcal{V}_{Q,D}\]

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

\[\Sigma = \text{diag } (\sigma_1, \sigma_2) = \text{diag } (3.0, 1.0)\]

\[W = U\Sigma \in \mathbb{R}^{D \times Q}\]

\[Y = XW^T + \epsilon\]

• 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

\[ p(Y|X) = \prod_{d=1}^{D} \mathcal{N}(Y_{:,d}|\mu, K + \sigma^2 I) \]

\[ K = XX^T, \quad K_{ij} = X_{i,:}^T X_{j,:} = k(X_{i,:}, X_{j,:}) \]

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

• No rotation symmetry in the posterior for the suggested parameterization

• Different chains converge to different solutions due to increased model complexity
Conclusion

• Suggested new parameterization for $W$ in PPCA, which uniquely identifies principle components even though the likelihood and the posterior are rotationally symmetric

• Showed how to set the prior on the new parameters such that the model is not changed compared to a standard Gaussian prior on $W$

• Provided an efficient implementation via Householder transformations (no Jacobian correction needed)

• New parameterization allows for other interpretable priors on rotation and principle variances

• Extended to non-linear models and successfully solved the rotation problem there as well
Thanks for your attention!

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