User-steering Interpretable Visualization with Probabilistic PCA

Viet Minh Vu and Benoît Frénay

NADI Institute - PReCISE Research Center

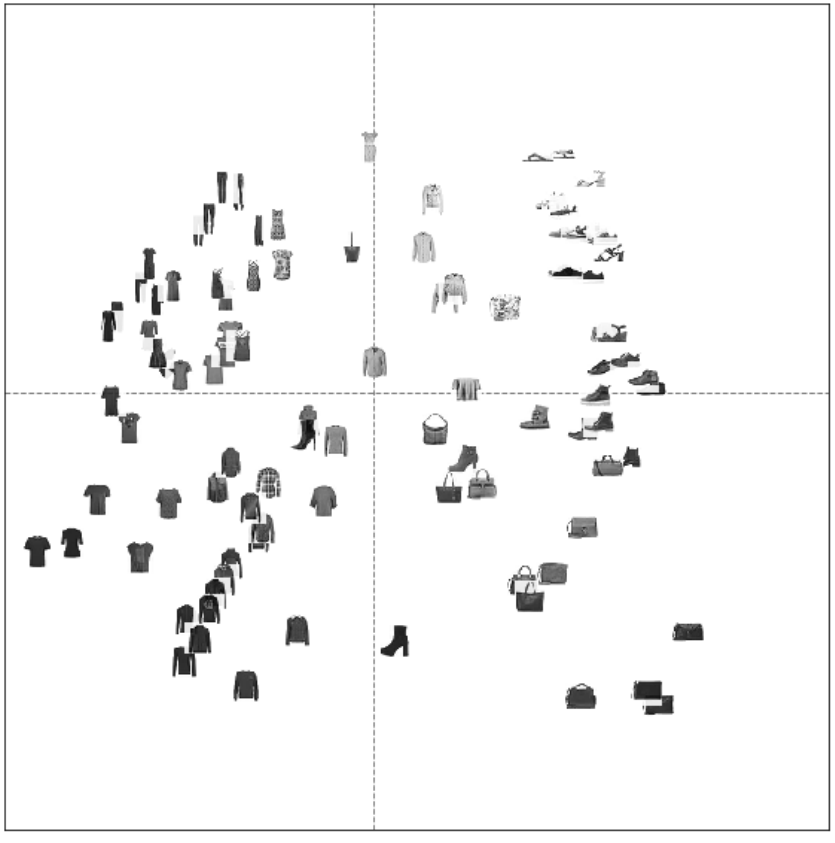
University of Namur, Belgium

25/04/2019

Problem: Dimensionality Reduction (DR)

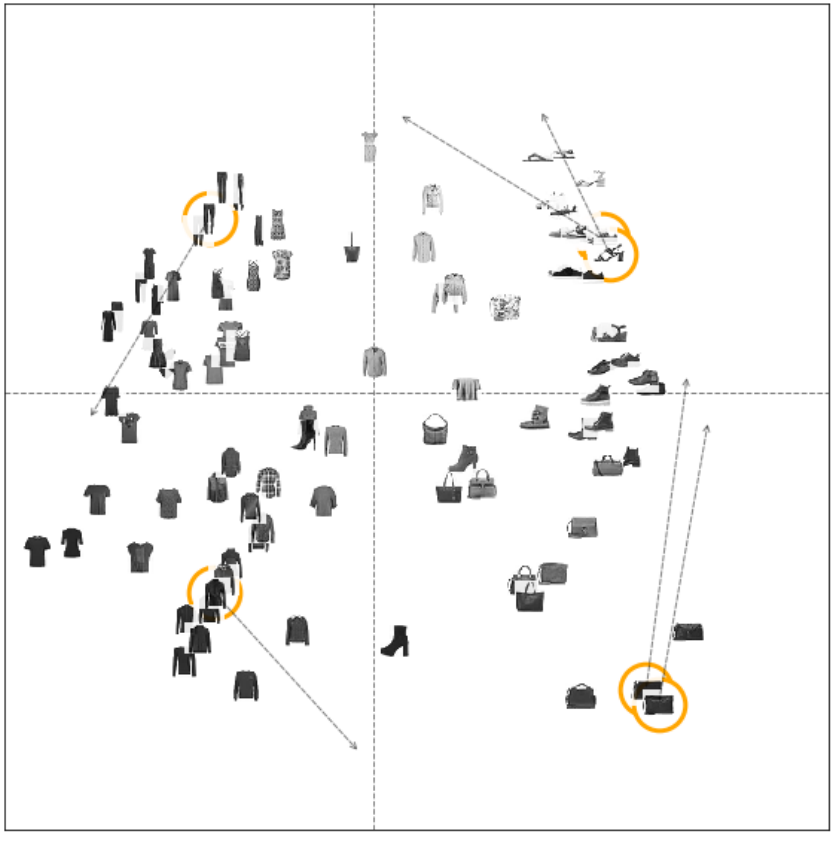
Samples from the Fashion-MNIST dataset

Visualization of high dimensional data



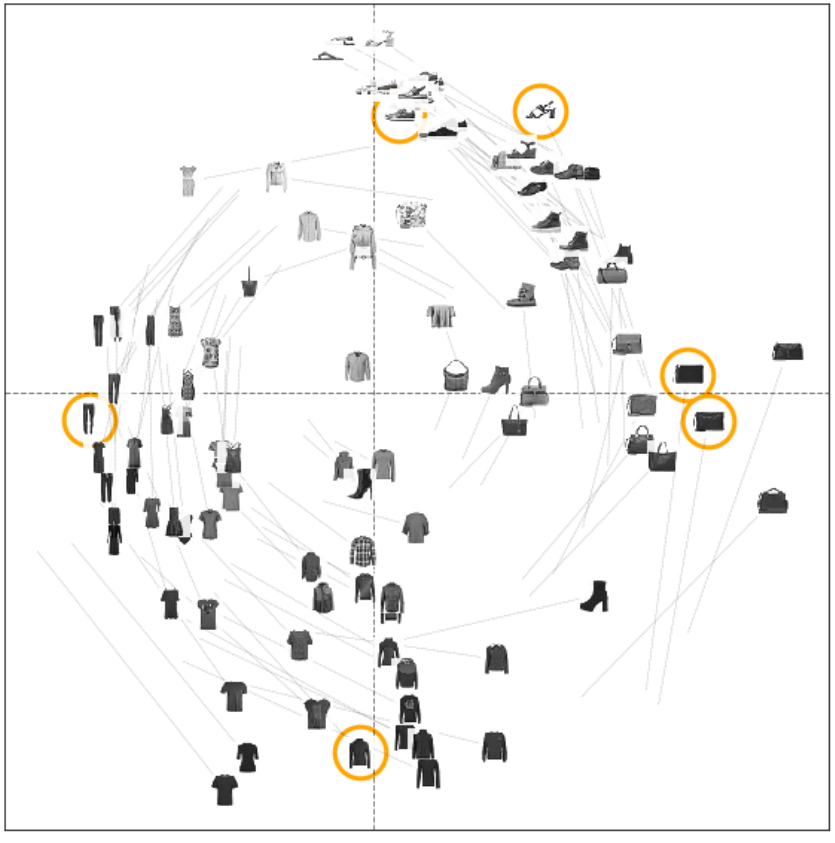
Having an initial visualization with the Probabilistic Principle Component Analysis (PPCA) model ...

Proposed interactive PPCA model (iPPCA)



The user can manipulate the visualization by moving some points.

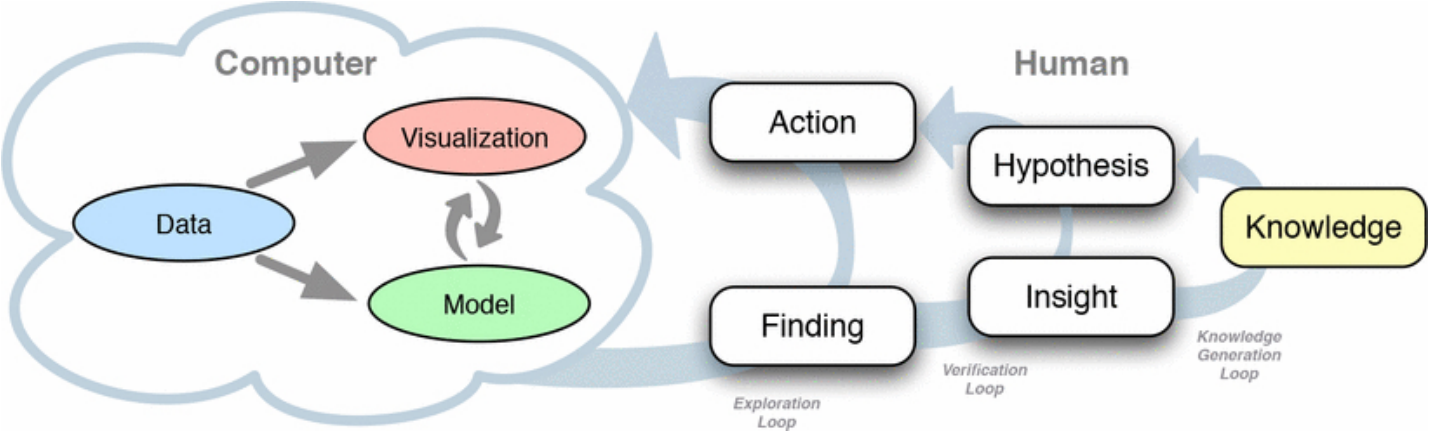
iPPCA result



The result of the interactive model is explainable to the users.

Motivation

User interaction in model design and analysis



Visual analytic method with Human-in-the-loop [1]

- The user can interact directly with the visualization to give their feedbacks.
- The model can update itself to take into account these feedbacks and produce a new visualization.

---

[1]Sacha, Dominik, et al. "Knowledge generation model for visual analytics." IEEE TVCG 2014

Existing approaches

Integrating user's feedbacks into existing Dimensionality Reduction (DR) methods

- Weighted MDS with the some fixed points to modify the weights ω_F:

$$\mathbf{Y} = \text{argmin}_{\mathbf{Y}} \sum_{i<j\leq n} \rho\Big|d_{\omega}(i,j) - d_Y(i,j)\Big| + (1-\rho)\Big|d_{\omega_F}(i,j) - d_Y(i,j)\Big|$$

- Semi-supervised PCA with sets of Must-links (ML) and Cannot-links (CL):

$$J(\mathbf{W}) = \frac{1}{2n^2}\sum_{i,j}|\mathbf{x}_i - \mathbf{y}_j|^2 + \frac{\alpha}{2n_{CL}}\sum_{CL}|\mathbf{x}_i - \mathbf{y}_j|^2 - \frac{\beta}{2n_{ML}}\sum_{ML}|\mathbf{x}_i - \mathbf{y}_j|^2$$

- Constrained Locality Preserving Projections with ML and CL:

$$\mathbf{W} = \text{argmin}_{\mathbf{W}} \frac{1}{2}\Big(\sum_{i,j}(\mathbf{y}_i - \mathbf{y}_j)^2 \widetilde{M}_{ij} + \sum_{ML'}(\mathbf{y}_i - \mathbf{y}_j)^2 - \sum_{CL'}(\mathbf{y}_i - \mathbf{y}_j)^2\Big)$$

  - $\mathbf{y}_j = \mathbf{W}^T\mathbf{x}_j$, $\mathbf{W}$ is projection matrix, $\mathbf{M}$ is weights matrix
  - ML', CL' are the extended set of Must-links and Cannot-links constraints

Existing approaches

Integrating user's feedbacks into existing Dimensionality Reduction (DR) methods

- Weighted MDS with the some fixed points to modify the weights ω_F:

$$\mathbf{Y} = \text{argmin}_{\mathbf{Y}} \sum_{i<j\leq n} \rho\left|d_{\omega}(i,j) - d_Y(i,j)\right| + (1-\rho)\left|d_{\omega_F}(i,j) - d_Y(i,j)\right|$$

- Semi-supervised PCA with sets of Must-links (ML) and Cannot-links (CL):

$$J(\mathbf{W}) = \frac{1}{2n^2}\sum_{i,j}\left|\mathbf{x}_i - \mathbf{y}_j\right|^2 + \frac{\alpha}{2n_{CL}}\sum_{CL}\left|\mathbf{x}_i - \mathbf{y}_j\right|^2 - \frac{\beta}{2n_{ML}}\sum_{ML}\left|\mathbf{x}_i - \mathbf{y}_j\right|^2$$

- Constrained Locality Preserving Projections with ML and CL:

$$\mathbf{W} = \text{argmin}_{\mathbf{W}}\frac{1}{2}\left(\sum_{i,j}(\mathbf{y}_i - \mathbf{y}_j)^2\widetilde{M}_{ij} + \sum_{ML'}(\mathbf{y}_i - \mathbf{y}_j)^2 - \sum_{CL'}(\mathbf{y}_i - \mathbf{y}_j)^2\right)$$

  - $\mathbf{y}_j = \mathbf{W}^T\mathbf{x}_j$, $\mathbf{W}$ is projection matrix, $\mathbf{M}$ is weights matrix
  - ML', CL' are the extended set of Must-links and Cannot-links constraints

Existing approaches

Integrating user's feedbacks into existing DR methods

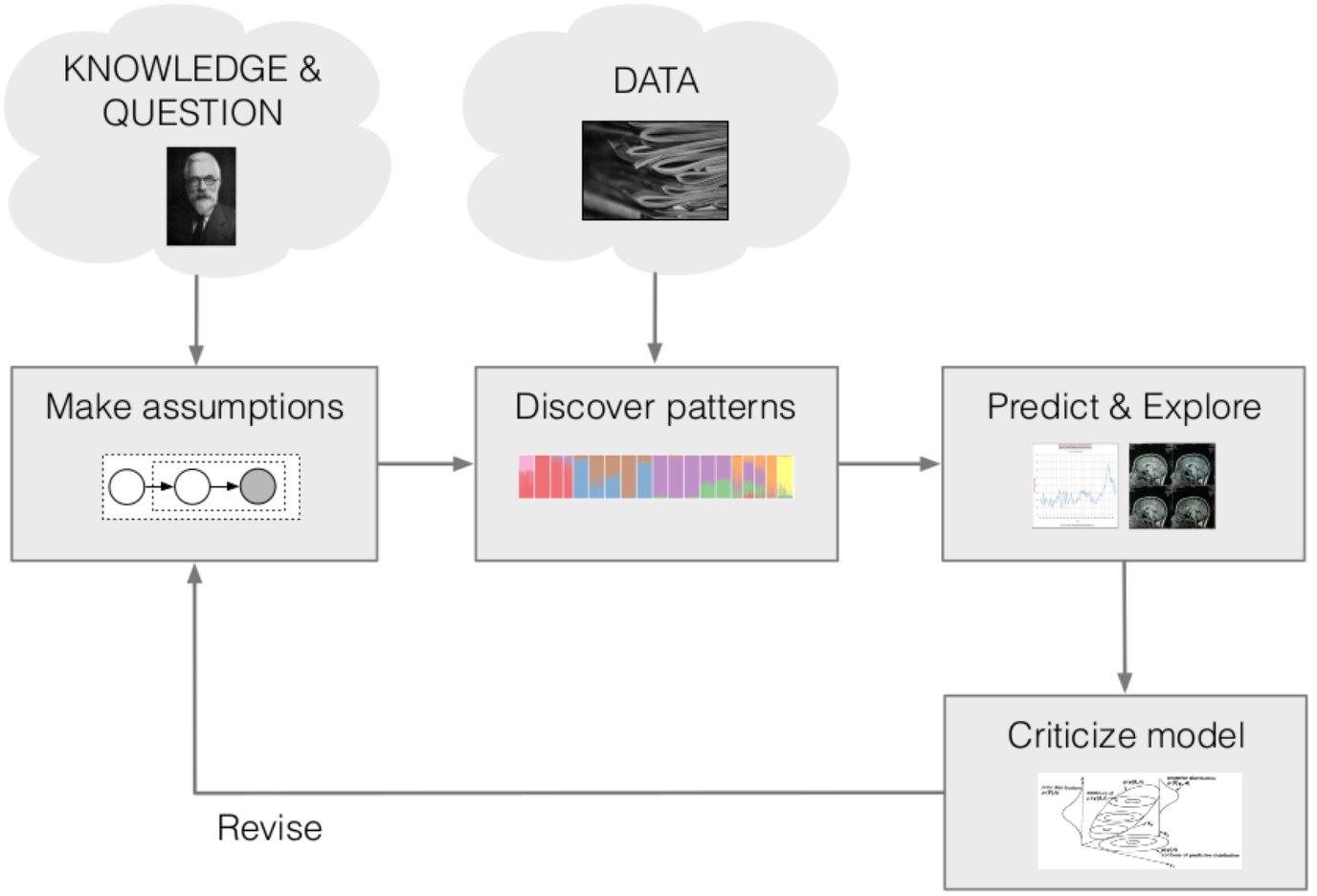
- User's feedbacks $\Longrightarrow$ Explicit regularization term
- Jointly optimized with the objective function of the basic DR methods.

Problems?

- Many discrete methods
- Manually design the regularization term explicitly

$\Longrightarrow$ Can we find another approach?

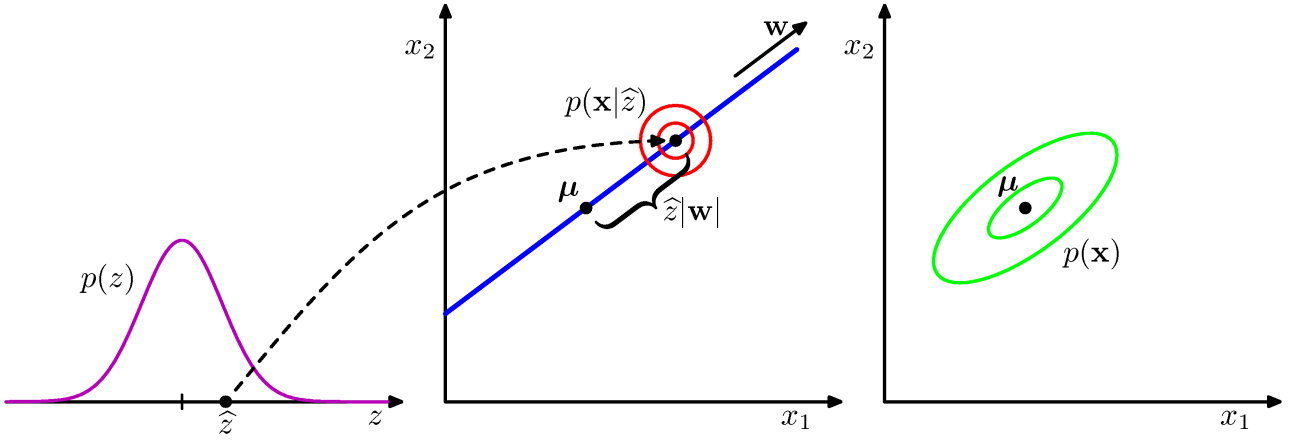
Probabilistic approach



David Blei, et al. "Variational Inference: Foundations and Modern Methods". NIPS 2016 Tutorial

Probabilistic PCA

- Probabilistic reformulation as the basic for a Bayesian treatment of PCA [1]
- Illustration for the generative process in PPCA model [2]
  - generate 2-dimensional data $p(\mathbf{x})$ from 1-dimensional latent variable $p(z)$



[1] Bishop, Christopher M. "Bayesian pca." Advances in neural information processing systems. 1999.
[2] Bishop's PRML book, Figure 12.9

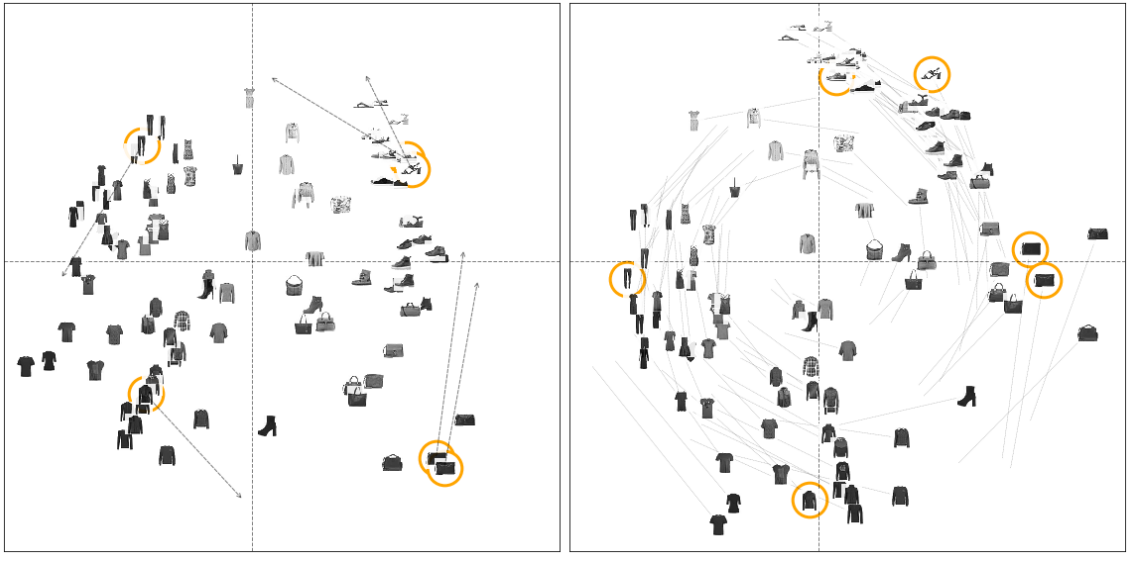
Probabilistic PCA

- $\mathbf{X} = \{\mathbf{x}_n\}$: N observations of D-dimensions.
- The embedded points in the 2D visualization are the latent variables $\mathbf{Z} = \{\mathbf{z}_n\}$.
- Likelihood $\mathbf{x}_n \mid \mathbf{z}_n \sim \mathcal{N}(\mathbf{x}_n \mid \mathbf{W}\mathbf{z}_n,\ \sigma^2\mathbf{I}_D)$
- The inference problem: $\theta_{MAP} = \text{argmax}_\theta \log p(\theta \mid \mathbf{X})$ where θ represents all the model's parameters (including $\mathbf{Z}$).
- The MAP estimate of the latent variables $\mathbf{Z}$ is found by following the partial gradient $\nabla_\mathbf{Z} \log p(\theta, \mathbf{X})$ to its local optima.

Proposed Interactive PPCA model

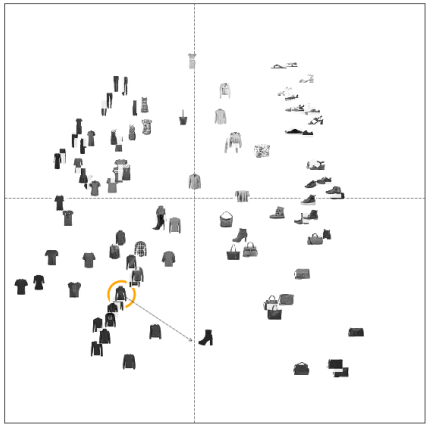
- iPPCA: The user-indicated position of the selected points is modelled directly in the prior distribution of the PPCA model.

$$\boldsymbol{z}_n \sim \begin{cases} \mathcal{N}(\boldsymbol{z}_n \mid \boldsymbol{\mu}_n,\ \sigma^2_{\text{fix}}) & \text{if } \boldsymbol{z}_n \text{ is fixed by user,} \\ \mathcal{N}(\boldsymbol{z}_n \mid \mathbf{0}, \mathbf{1}) & \text{otherwise.} \end{cases}$$

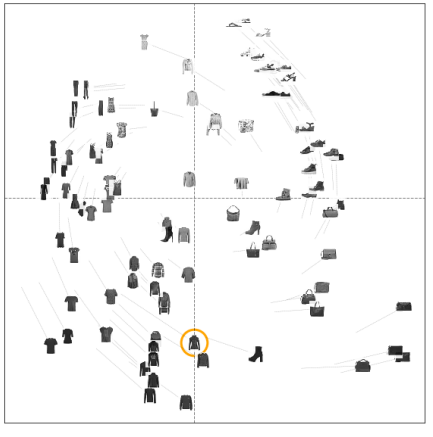


How user's constraints are handled?

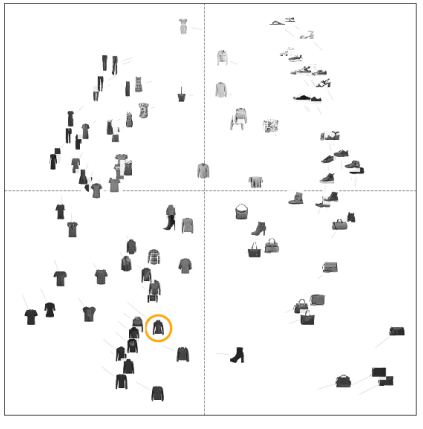
- The user can fix the position of several interested points, with some level of uncertainty (σ^2_{fix})
- A very small variance $\Longrightarrow$ the user is very certain.
- A large variance $\Longrightarrow$ the user is not sure.



user's uncertainty σ^2_{fix}



$\sigma^2_{fix} = 1e-4$*:*
very sure



$\sigma^2_{fix} = 0.2$*:*
very uncertain

Evaluation of the iPPCA model

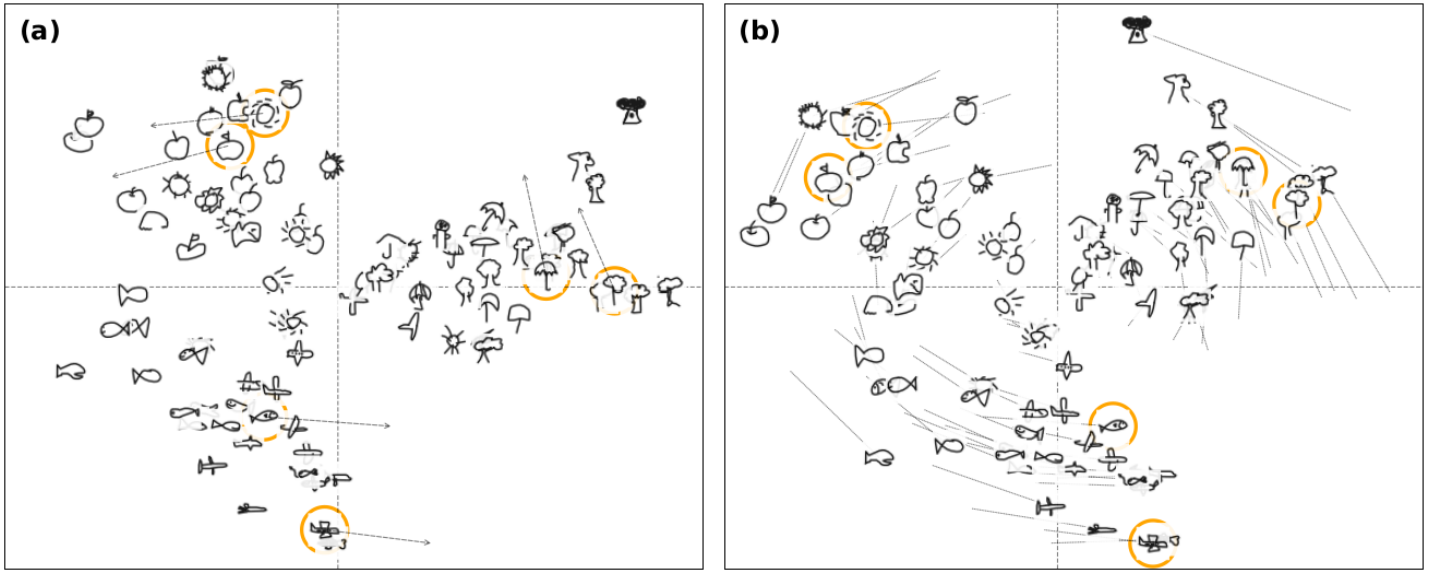
The workflow:

- Show the initial visualization of the (original) PPCA model
- The user selects and moves some anchor points
- Reconstruct the iPPCA model to create a new visualization.
  - The uncertainty of the feedbacks (σ_{fix}^2) is small
  - Hyper parameters of the optimization process are chosen to be the best

How to evaluate:

- Show how to explain the new visualization
  - The level on which we can understand / explain the visualization is considered as a qualitative measure

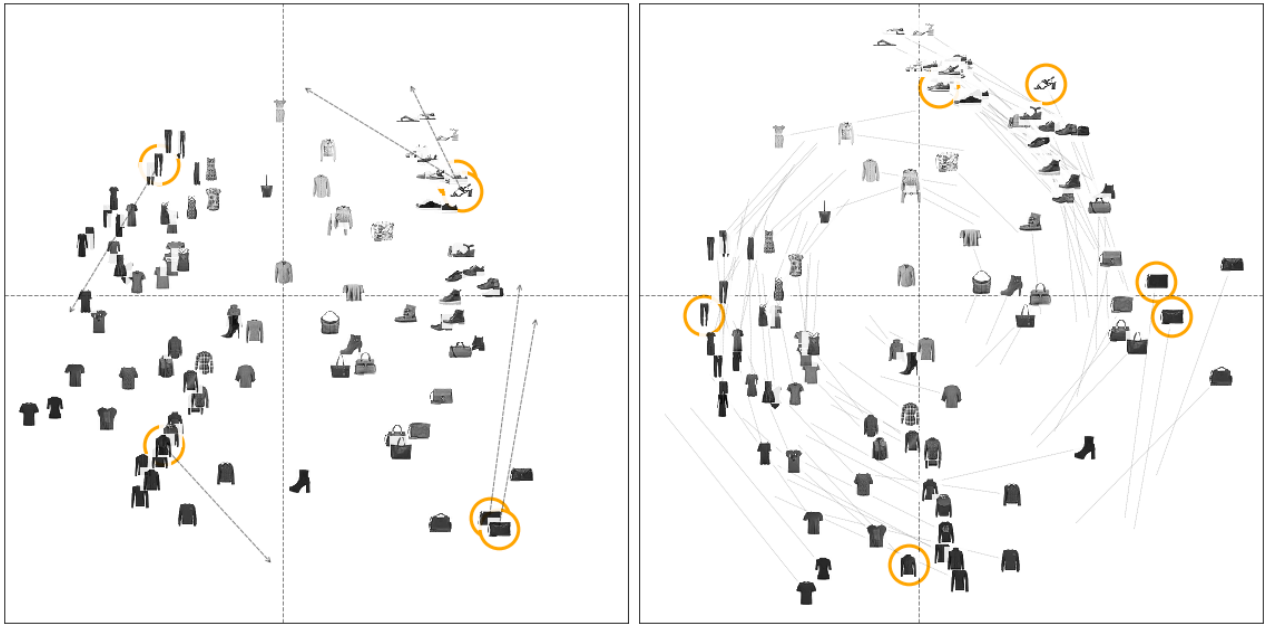
Quickdraw dataset



90 sample images from Quickdraw dataset

- Move 6 different points of different groups
- The global structure of the embedding is preserved

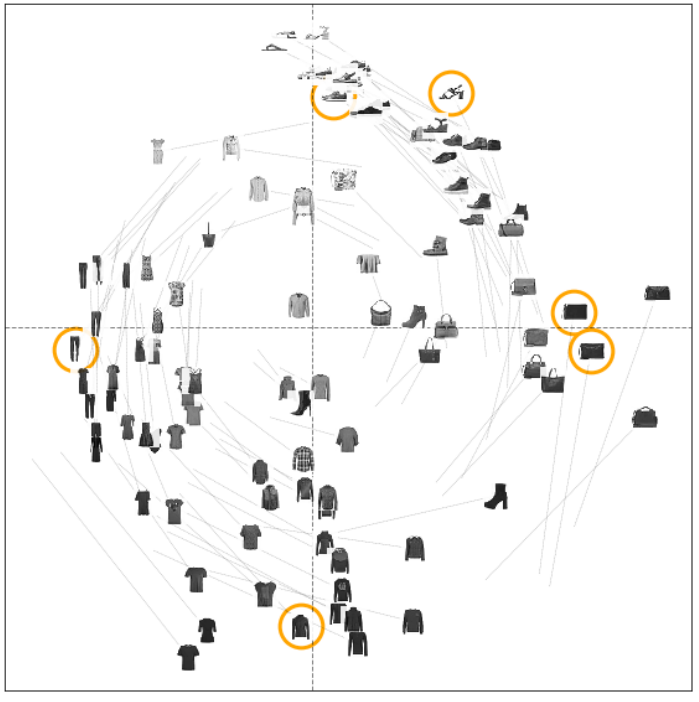
Fashion dataset

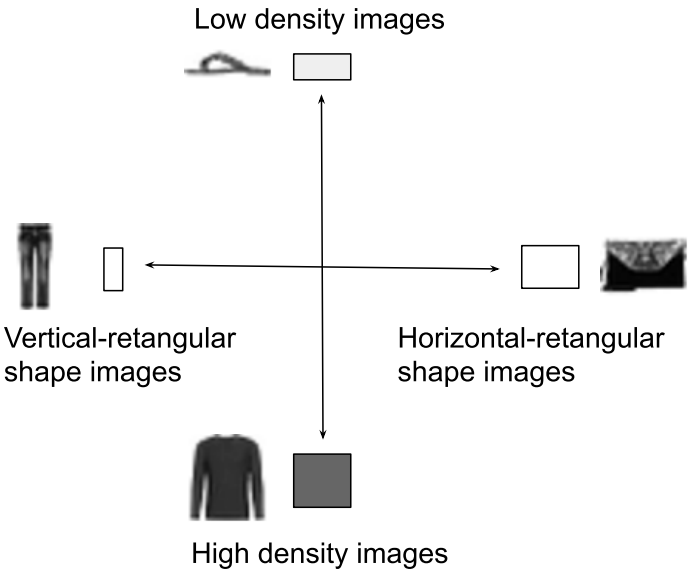


100 sample images from Fashion dataset

- Moves 6 points towards the coordinate axes
- The goal of this interaction is to re-define the axes in the visualization

Fashion dataset

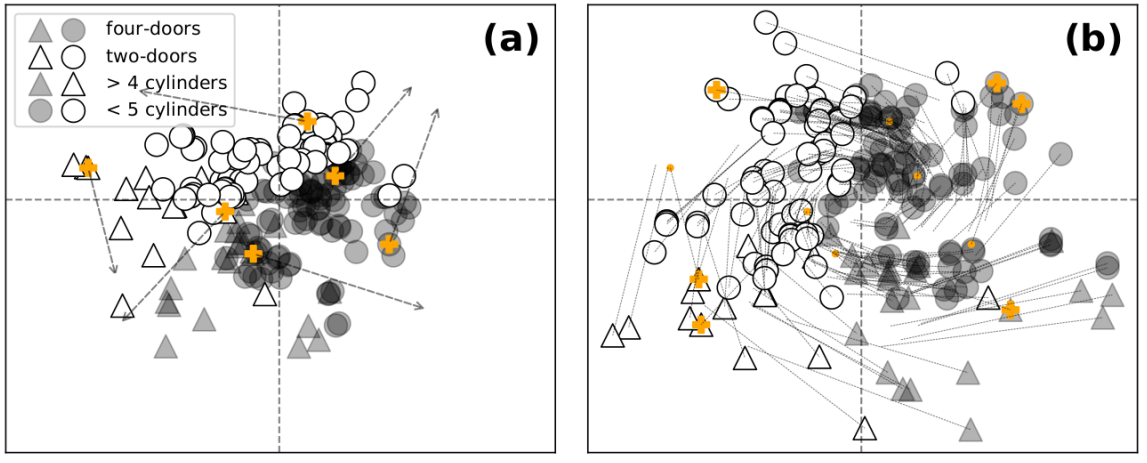




How to explain the new axes?

- Horizontal axis represents shape
- Vertical axis represents color density

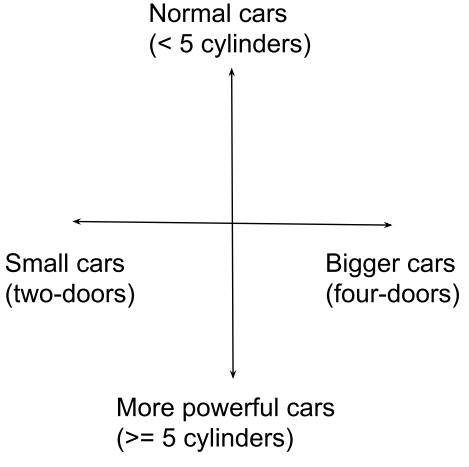
Automobile dataset



203 data points of the Automobile dataset

How to explain the new axes?

- Horizontal axis: cars' size
- Vertical axis: cars' power



Advantage of probabilistic approach

Combination of solid theoretical models and modern powerful inference toolboxes

- Take any old-class model or modern generative model
- Plug into a probability framework [1] which support modern inference methods like Stochastic Variational Inference (SVI)

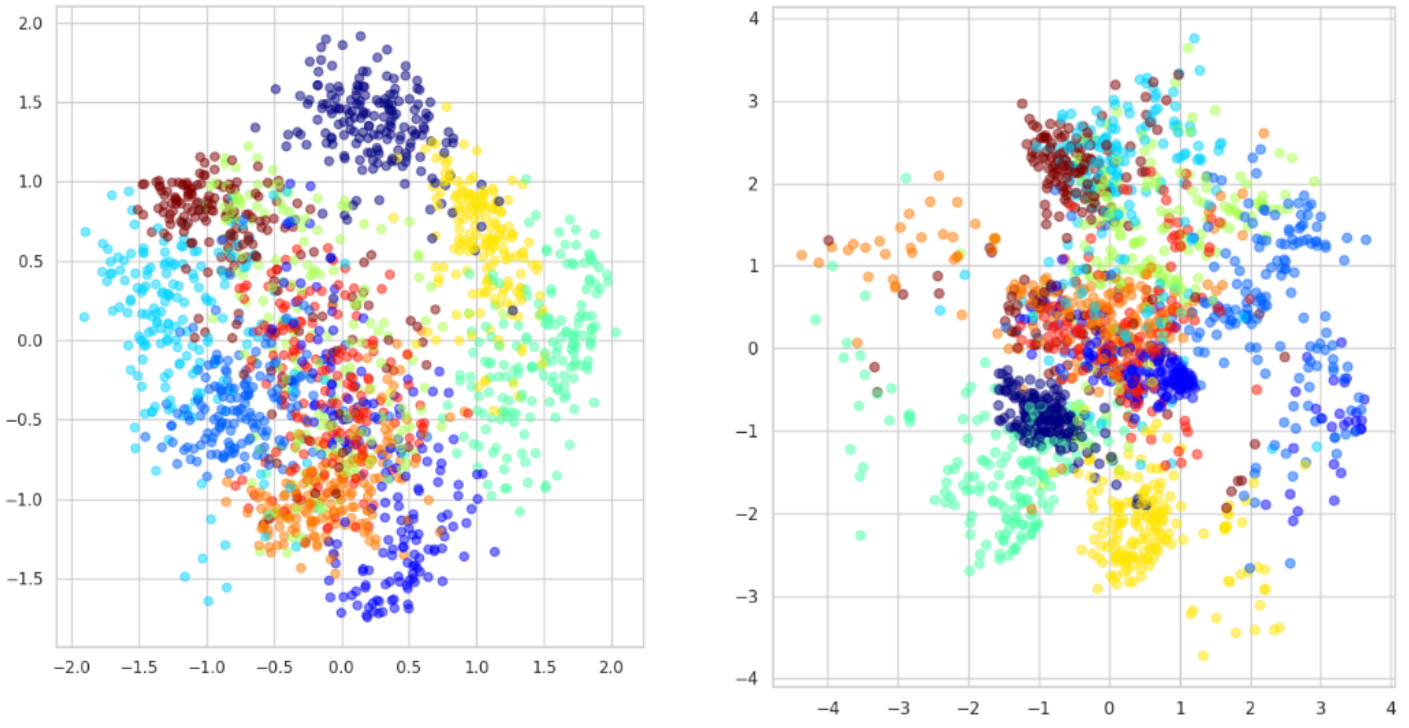
Can easily extend the generative process

$$\mathbf{x}_n \mid \mathbf{z}_n \sim \mathcal{N}(f(\mathbf{z}_n), \sigma^2 \mathbf{I})$$

- in PPCA model, $f(\mathbf{z}_n) = \mathbf{W}\mathbf{z}_n$
- $f(\mathbf{z}_n)$ can be any high-capacity representation function (a neural net)

---

[1] Stan, PyMC3, Pyro, TensorFlow Probability



Embedding of DIGITS dataset with the original PCA and the modified PPCA model

- The decoder $f(\mathbf{z})$ of PPCA is a simple neural network with one hidden layer of 50 units and a sigmoid activation function.
- The inference is done by the Pyro's built-in SVI optimizer [1].

[1] Pyro, Deep Universal Probabilistic Programming, http://pyro.ai/

Recap

Propose the interactive PPCA model allowing the user to control the visualization

- [Why] To communicate the analytical result (e.g., create an explainable visualization) and to explore the visualizations ("what-if" analysis)
- [How] The user's feedbacks can be efficiently integrated into a probabilistic model via prior distributions of latent variables.
- [Potential] The probabilistic model is flexible to extend and can be easily optimized by the black-box inference methods.
- [Future work] Focus on the user's feedback modeling problem without worrying about the complex optimization procedure.

User-steering Interpretable Visualization with Probabilistic PCA

Viet Minh Vu and Benoît Frénay

NADI Institute - PReCISE Research Center

University of Namur, Belgium

-