Unsupervised Learning with Contrastive Latent Variable Models

Kristen Severson, Soumya Ghosh, and Kenney Ng

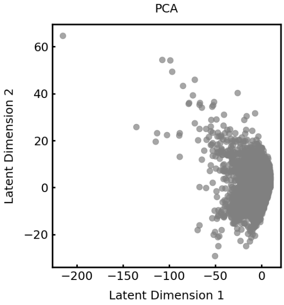
IBM Research
MIT-IBM Watson AI Lab

kristen.severson@ibm.com

January 2019

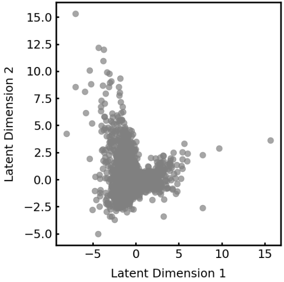
Data visualization with PCA

- When working with high dimensional data, principal component analysis (PCA) is often the first tool used to explore the dataset
- PCA maximizes the retained variance in a lower dimensional space
- Ideally, the lower dimensional representation will reveal structure that can be related to the application

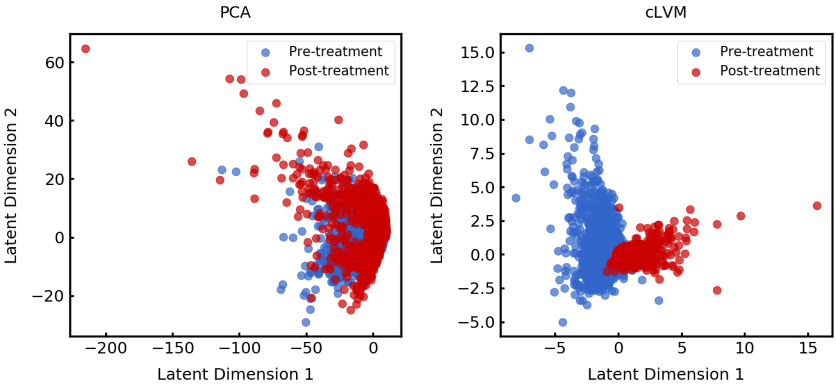


Data visualization with PCA

- When working with high dimensional data, principal component analysis (PCA) is often the first tool used to explore the dataset
- PCA maximizes the retained variance in a lower dimensional space
- Ideally, the lower dimensional representation will reveal structure that can be related to the application

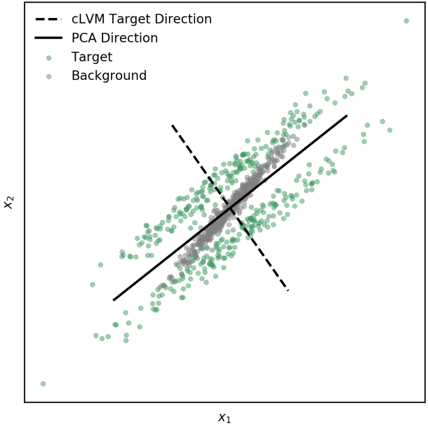


Data visualization with PCA



Contrastive dimensionality reduction

- High variance directions aren't necessarily relevant to the domain application
- If we can characterize the expected baseline variation, we can search for variance directions that differentiate the dataset of interest



Target and background datasets

There are many natural settings where pairs of datasets occur:

- Control vs. study populations
- Pre- vs. post-intervention groups
- Signal vs. signal-free measurements

Contrastive PCA

- Despite this natural setting of contrasting dataset, there are few methods that leverage this type of structure
- Abid *et al.*[1] proposed contrastive PCA as a way to achieve this:

$$C = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\mathsf{T}} - \alpha \frac{1}{m}\sum_{j=1}^{m} y_i y_i^{\mathsf{T}}$$

where x_i are samples from the dataset of interest (*target*) and y_i are samples from the *background* dataset

1. A Abid, MJ Zhang, VK Bagaria and J Zou (2018) Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications*.

A probabilistic approach

Probabilistic models have several advantages as compared to deterministic approaches:

- Possibility to incorporate prior information
- Natural handling of noisy and missing data
- Ability to perform model and feature selection
- Incorporation into larger probabilistic systems

Contrastive Latent Variable Models

Given a *target* dataset, $\{x_i\}_{i=1}^n$, and a *background* dataset, $\{y_j\}_{j=1}^m$, the model is specified

$$x_i = Sz_i + Wt_i + \mu_t + \epsilon_i, \quad i = 1 \dots n$$
$$y_j = Sz_j + \mu_b + \epsilon_j, \quad j = 1 \dots m$$

where $x_i, y_j \in \mathbb{R}^d$ are the observed data, $z_i, z_j \in \mathbb{R}^k$ and $t_i \in \mathbb{R}^t$ are the latent variables, $S \in \mathbb{R}^{d \times k}$ and $W \in \mathbb{R}^{d \times t}$ are the corresponding factor loadings, μ_t, μ_b are the dataset specific means and ϵ_i, ϵ_j are the noise.

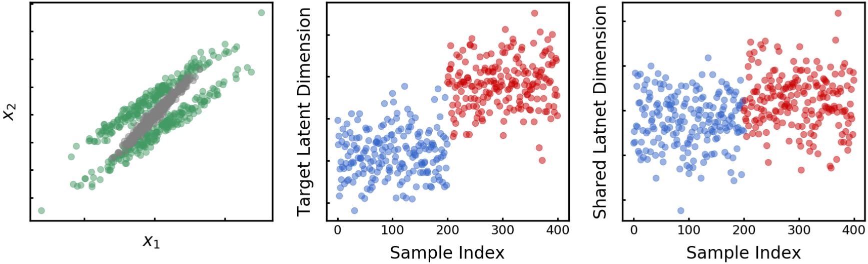
Contrastive Latent Variable Models

$$p(\mathcal{D}, \{z_i, t_i\}_{i=1}^n, \{z_j\}_{j=1}^m; \Theta) = p(\Theta) \prod_{i=1}^n p(x_i | z_i, t_i; W, S, \mu_x, \sigma^2) p(z_i) p(t_i) \prod_{j=1}^m p(y_j | z_j; S, \mu_y, \sigma^2) p(z_j)$$

- The primary modeling decisions are to select the likelihood and priors on the loading matrices, W and S
- Many choices of prior will lead to posterior distributions that are not tractable, therefore we use black-box variational inference to solve for the parameters

Gaussian likelihoods and priors

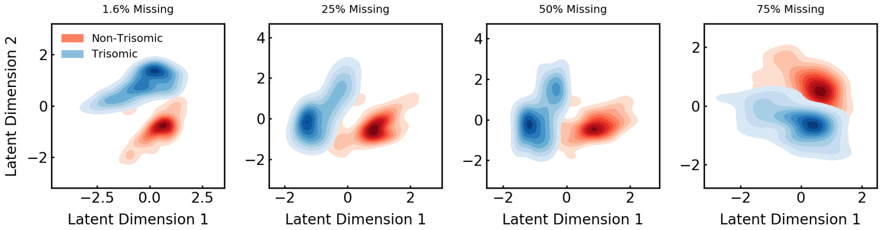
- The cLVM is to similar probabilistic PCA when standard Gaussian variables are used to model the latent representations, t_i, z_i, z_j and the likelihood is Gaussian
- The main difference is the part of the data that is captured by the shared space is projected away before updating the target space, and vice versa



Robustness and missing data

Probabilistic formulation allows for handling of noisy and missing data

- Prior: $p(\sigma) \sim \mathrm{IG}(a, b)$
- Likelihood: Student's t
- Variational approximation: $q(\ln \sigma^2) \sim \mathcal{N}(\cdot, \cdot)$



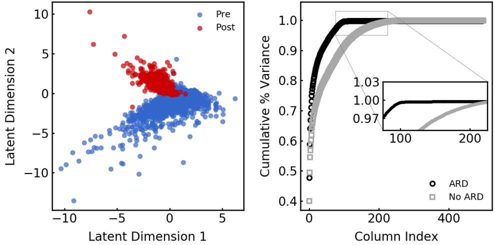
Application: subgroup discovery using mouse protein expression data

Target samples: 270
Background samples: 135
Observation dimensionality: 77

Model selection

Probabilistic formulation allows for automatic relevance detection

- Prior: $p(S) = \prod_{i=1}^{d} \mathcal{N}(S_{\cdot j} | 0, \alpha_j) \mathrm{IG}(\alpha_j | a, b)$
- Likelihood: Gaussian
- Variational approximation: $q(S) = \mathcal{N}(\cdot, \cdot), \quad q(\ln \alpha) = \mathcal{N}(\cdot, \cdot)$



Application: subgroup discovery using single cell RNA-Seq data

Target samples: 7898

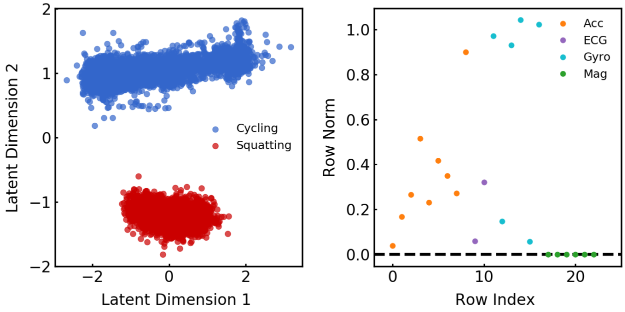
Background samples: 1985

Observation dimensionality: 500

Feature selection

Probabilistic formulation enables feature selection

- Penalty: $r(W) = \rho \sum_{i=1}^{d} \sqrt{p_i} \|W_{:i}\|_2$



Application: feature selection using heterogeneous sensor data

Target samples: 6451

Background samples: 3072

Observation dimensionality: 23

Nonlinear extension: Contrastive variational autoencoders

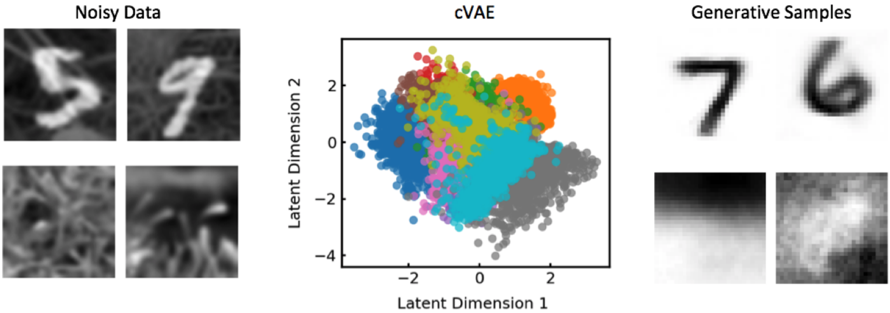
Given a *target* dataset, $\{x_i\}_{i=1}^{n}$, and a *background* dataset, $\{y_j\}_{j=1}^{m}$, the model is specified

$$x_i = f_{\theta_s}(z_i) + f_{\theta_t}(t_i) + \epsilon_i, \quad i = 1 \dots n$$
$$y_j = f_{\theta_s}(z_j) + \epsilon_j, \quad j = 1 \dots m$$

where f_{θ_s} and f_{θ_t} are non-linear transformations parameterized by neural networks.

Contrastive variational autoencoders

cVAE recovers meaningful structure from noisy data



Application: generative de-noising using image data

Target samples: 5000
Background samples: 5000
Observation dimensionality: 784

Conclusions

- Dimensionality reduction has important applications in data exploration, visualization, and pre-processing
- The design of the cLVM allows the model to learn structure that is enriched in one dataset relative to another
- The probabilistic formulation enables robust, sparse, and nonlinear variations of the model
- Future extensions will include larger numbers of datasets and data types (e.g. count, categorical)