Exploring patterns enriched in a dataset with **contrastive principal component analysis**

Abubakar Abid, Martin J. Zhang, Vivek K. Bagaria & James Zou

Nature Communications, 2018

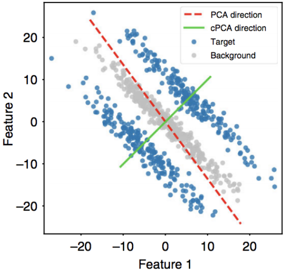
18th Feb, 2019

Background

- Dimensionality reduction is a fundamental tool for exploratory data analysis and visualization.
- While there are many dimensionality reduction methods these methods typically assume a **single** dataset.
- However, it is often the case we have multiple datasets and wish to find **projections which exhibit interesting differences between the datasets**.

Background

- Dimensionality reduction is a fundamental tool for exploratory data analysis and visualization.
- While there are many dimensionality reduction methods these methods typically assume a **single** dataset.
- However, it is often the case we have multiple datasets and wish to find **projections which exhibit interesting differences between the datasets**.



Contrastive PCA

- We observe target data $\{\mathbf{x}_i \in \mathbb{R}^d\}$ and background data $\{\mathbf{y}_i \in \mathbb{R}^d\}$ with sample covariances C_X and C_Y.
- For any unit vector $\mathbf{v}$, define:

$$\lambda_X(\mathbf{v}) = \mathbf{v}^T C_X \mathbf{v}$$
$$\lambda_Y(\mathbf{v}) = \mathbf{v}^T C_Y \mathbf{v}$$

- Standard PCA simply maximizes $\lambda_X(\mathbf{v}) \Rightarrow$ problematic if leading eigenvectors are shared in C_X and C_Y.

Contrastive PCA

- We observe target data $\{\mathbf{x}_i \in \mathbb{R}^d\}$ and background data $\{\mathbf{y}_i \in \mathbb{R}^d\}$ with sample covariances C_X and C_Y.
- For any unit vector $\mathbf{v}$, define:

$$\lambda_X(\mathbf{v}) = \mathbf{v}^T C_X \mathbf{v}$$
$$\lambda_Y(\mathbf{v}) = \mathbf{v}^T C_Y \mathbf{v}$$

- Standard PCA simply maximizes $\lambda_X(\mathbf{v}) \Rightarrow$ problematic if leading eigenvectors are shared in C_X and C_Y.
- For a fixed $\alpha \in \mathbb{R}_+$, contrastive PCA solves the following optimization:

$$\begin{aligned}\mathbf{v}^* &= \underset{\mathbf{v}}{\text{argmax}} \ \{\lambda_X(\mathbf{v}) - \alpha\lambda_Y(\mathbf{v})\} \\ &= \underset{\mathbf{v}}{\text{argmax}} \ \left\{\mathbf{v}^T(C_X - \alpha C_Y)\mathbf{v}\right\}\end{aligned}$$

Special case: simultaneously diagonalizable system

- We assume C_X and C_Y have shared eigen-structure such that:
$$C_X = Q\Lambda_X Q^T \text{ and } C_Y = Q\Lambda_Y Q^T,$$
for $\Lambda_X = \text{diag}(\lambda_{X,1}, \dots, \lambda_{X,d})$ and where $\mathbf{q}_1, \dots, \mathbf{q}_d$ are eigenvectors.
- Then we can write any unit vector in terms of the basis defined by Q as: $\mathbf{v} = \sum_{i=1}^{d} \sqrt{c_i}\mathbf{q}_i$ where $\sum_{i=1}^{d} c_i = 1$.

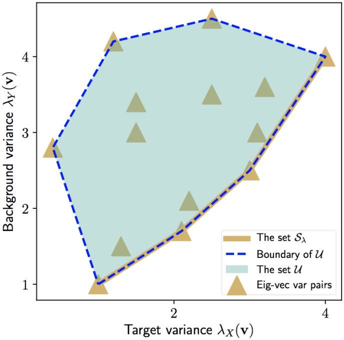
Special case: simultaneously diagonalizable system

- We assume C_X and C_Y have shared eigen-structure such that:

$$C_X = Q\Lambda_X Q^T \quad \text{and} \quad C_Y = Q\Lambda_Y Q^T,$$

  for $\Lambda_X = \text{diag}(\lambda_{X,1}, \ldots, \lambda_{X,d})$ and where $\mathbf{q}_1, \ldots, \mathbf{q}_d$ are eigenvectors.
- Then we can write any unit vector in terms of the basis defined by Q as: $\mathbf{v} = \sum_{i=1}^{d} \sqrt{c_i}\mathbf{q}_i$ where $\sum_{i=1}^{d} c_i = 1$.

- Thus $\lambda_X(\mathbf{v}) = \sum_{i=1}^{d} c_i \lambda_{X,i}$ and similarly for $\lambda_Y(\mathbf{v})$.
- $\mathbf{v}^*$ will be along bottom right of figure $\Rightarrow$ convex hull of eigenvalues, will be piecewise linear



Final example