Supervised Quadratic Feature Analysis: An Information Geometry approach to dimensionality reduction

Daniel Herrera-Esposito   Johannes Burge

University of Pennsylvania

CNI Computational Neuroscience Initiative

Introduction: Supervised Dimensionality Reduction

Supervised dimensionality reduction involves labeled data $\{\mathbf{x}_t, y_t\}_{t=1}^N$, where $\mathbf{x}_t \in \mathbb{R}^n$ is observation t and $y_t \in \{1, \dots, C\}$ is the class label.

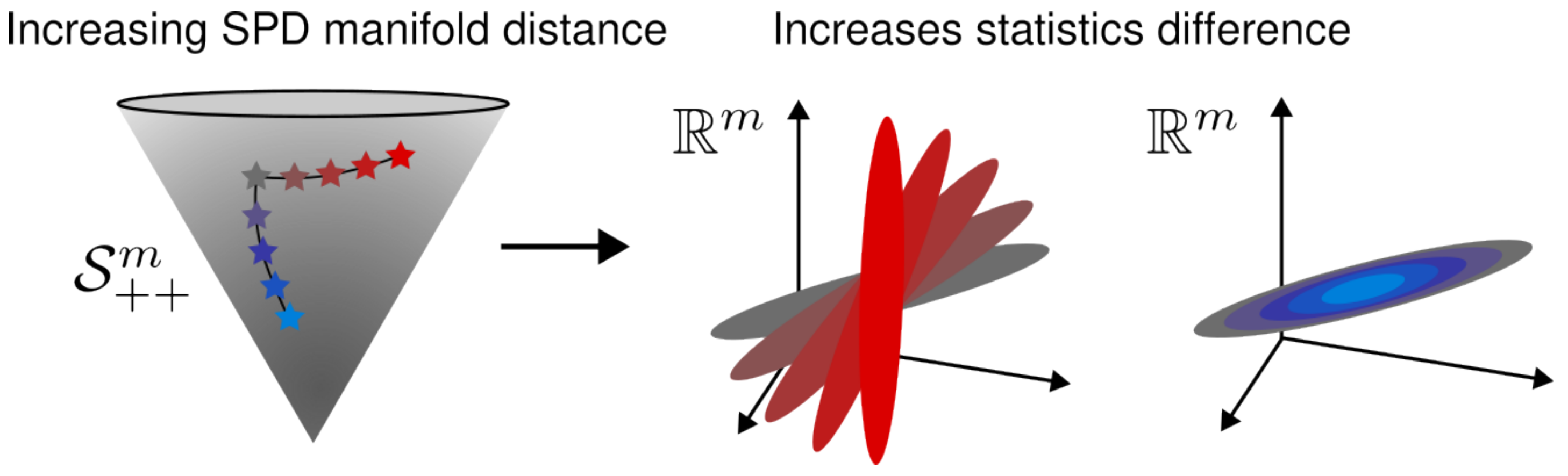
The goal is to map the variable $\mathbf{x}$ to a lower-dimensional variable $\mathbf{z} \in \mathbb{R}^m$ that maximizes information about y.

We introduce a novel supervised dimensionality reduction method called Supervised Quadratic Feature Analysis (SQFA):

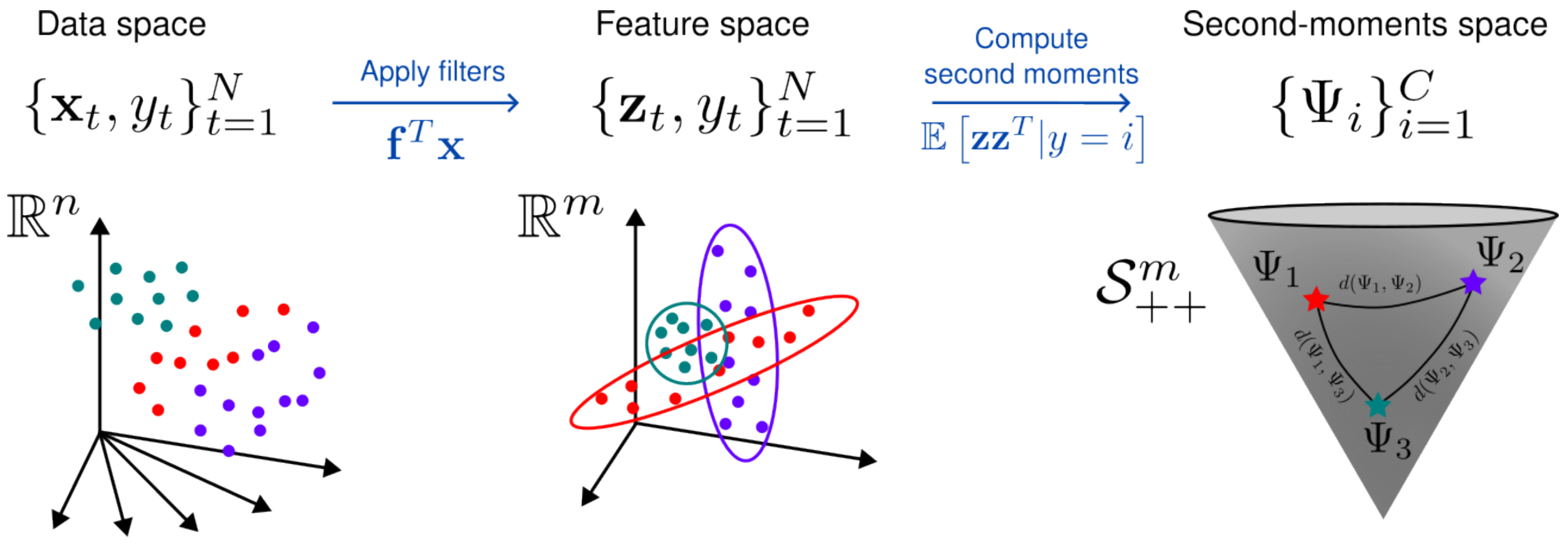
- SQFA learns a set of filters $\mathbf{f} \in \mathbb{R}^{n \times m}$ to obtain the linear projection $\mathbf{z} = \mathbf{f}^\intercal \mathbf{x}$ that maximizes second-order differences between classes
- Quadratic decoders (e.g. probabilistic Gaussian decoders, Quadratic Discriminant Analysis) are sensitive to second-order differences

Information Geometry objective

GEOMETRY: Class-conditional second-moment matrices $\Psi_i = \mathbb{E}\left[\mathbf{z}\mathbf{z}^T | y = i\right]$ in feature space are in Symmetric Positive Definite (SPD) manifold $\mathcal{S}_{++}^m$.


Increasing SPD manifold distance
Increases statistics difference
$\mathcal{S}_{++}^m$
$\mathbb{R}^m$
$\mathbb{R}^m$

The class-conditional second-moment matrices $\{\Psi_i\}_{i=1}^C$ define C points in $\mathcal{S}_{++}^m$. SQFA uses Riemannian distances in $\mathcal{S}_{++}^m$ to characterize class differences:



Data space   Feature space   Second-moments space

$$\{\mathbf{x}_t, y_t\}_{t=1}^N \xrightarrow[\mathbf{f}^T\mathbf{x}]{\text{Apply filters}} \{\mathbf{z}_t, y_t\}_{t=1}^N \xrightarrow[\mathbb{E}\left[\mathbf{z}\mathbf{z}^T|y=i\right]]{\text{Compute second moments}} \{\Psi_i\}_{i=1}^C$$

OBJECTIVE FUNCTION: SQFA maximizes the sum of all pairwise Riemannian distances between the class-conditional feature second-moment matrices:

$$\max_{\mathbf{f}} \sum_{i=2}^{C} \sum_{j \neq i} d_{AI}(\Psi_i, \Psi_j) \tag{1}$$

using the Affine-Invariant Riemannian distance

$$d_{AI}(\Psi_i, \Psi_j) = \left\| \log\left(\Psi_i^{-1/2} \Psi_j \Psi_i^{-1/2} \right) \right\|_F = \sqrt{\sum_{k=1}^{m} \log^2 \lambda_k} \tag{2}$$

where λ_k is the k-th generalized eigenvalue of (Ψ_i, Ψ_j), and $\mathbf{v}_k$ its eigenvector.

CHOICE OF METRIC: $d_{AI}(\Psi_i, \Psi_j)$ reflects discriminability between classes i, j:

- $\log^2 \lambda_k$ relates to quadratic discriminability of classes i, j along $\mathbf{v}_k$. Discriminability along all generalized eigenvectors in feature space is summarized by $d_{AI}(\Psi_i, \Psi_j) = \sqrt{\sum_{k=1}^m \log^2 \lambda_k}$
- d_{AI} reflects Fisher information ($\frac{1}{2} d_{AI}$ is Fisher distance for 0-mean Gaussians):
  - Let curve $\Sigma : [0,1] \to \mathcal{S}_{++}^m$ be the Affine-Invariant geodesic from Ψ_i to Ψ_j, with $\Sigma(0) = \Psi_i$ and $\Sigma(1) = \Psi_j$.
  - Fisher information of $\mathcal{N}\left(0, \Sigma(\theta)\right)$ along the curve is $\mathcal{I}(\theta) = \frac{1}{2} \mathrm{Tr}\left(\Sigma(\theta)^{-1} \Sigma'(\theta) \Sigma(\theta)^{-1} \Sigma'(\theta)\right)$, where $\Sigma'(\theta)$ is the velocity of the curve at $\Sigma(\theta)$.
  - $\mathcal{I}(\theta)$ measures how discriminable are local changes along the curve defined by $\mathcal{N}\left(0, \Sigma(\theta)\right)$
  - $d_{AI}(\Psi_i, \Psi_j) = 2 \int_0^1 \sqrt{\mathcal{I}(\theta)} d\theta$. In words, $d_{AI}(\Psi_i, \Psi_j)$ is the accumulated discriminability of transforming $\mathcal{N}\left(0, \Psi_i\right)$ into $\mathcal{N}\left(0, \Psi_j\right)$.

SQFA Python package

See the code for reproducing these results in our SQFA package tutorials: https://sqfa.readthedocs.io/en/latest/ or **scan the QR**:

sqfa

Toy example: SQFA vs LDA and PCA

We compare SQFA to LDA and PCA using a toy 6D dataset with 3 classes. The dataset has 3 distinct subspaces, each favored by one method:

- Dimensions 1-2 have highly discriminative covariances, but have low variance and no differences in class means.
- Dimensions 3-4 have small differences in class means, but low discriminability.
- Dimensions 5-6 have high variance and no discriminability.

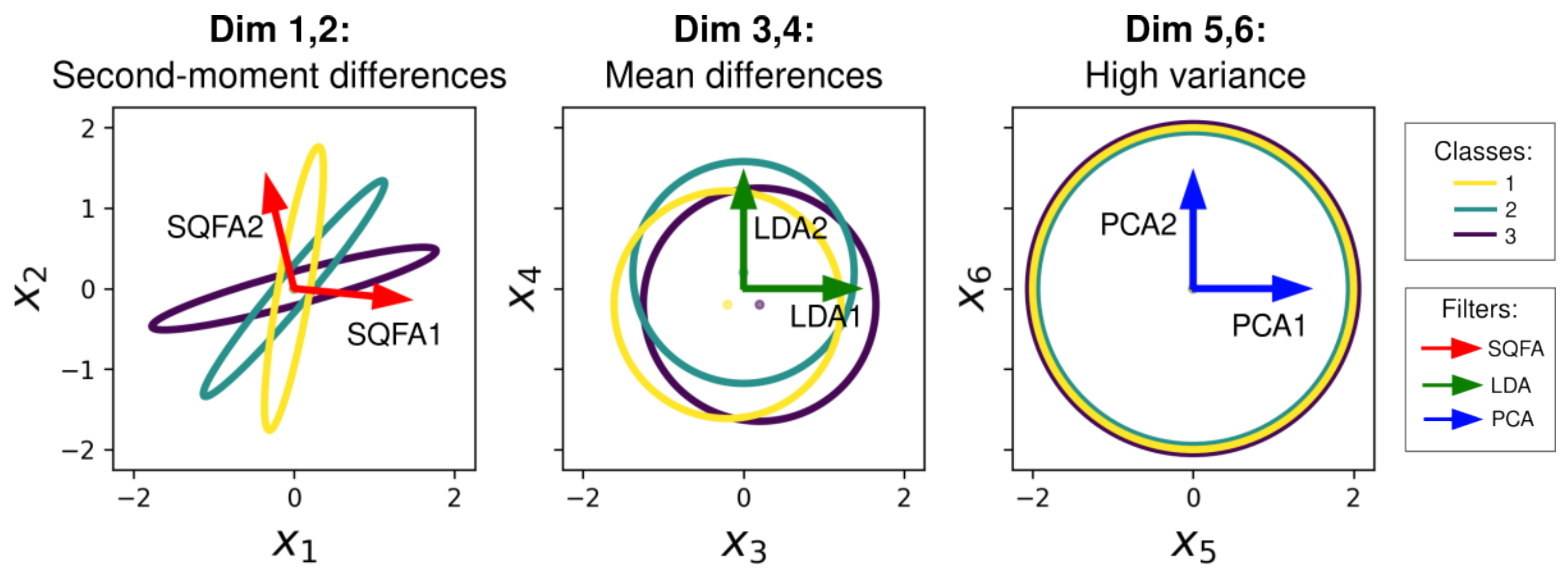
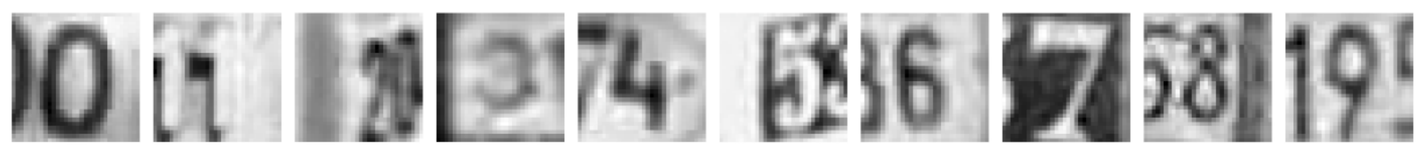


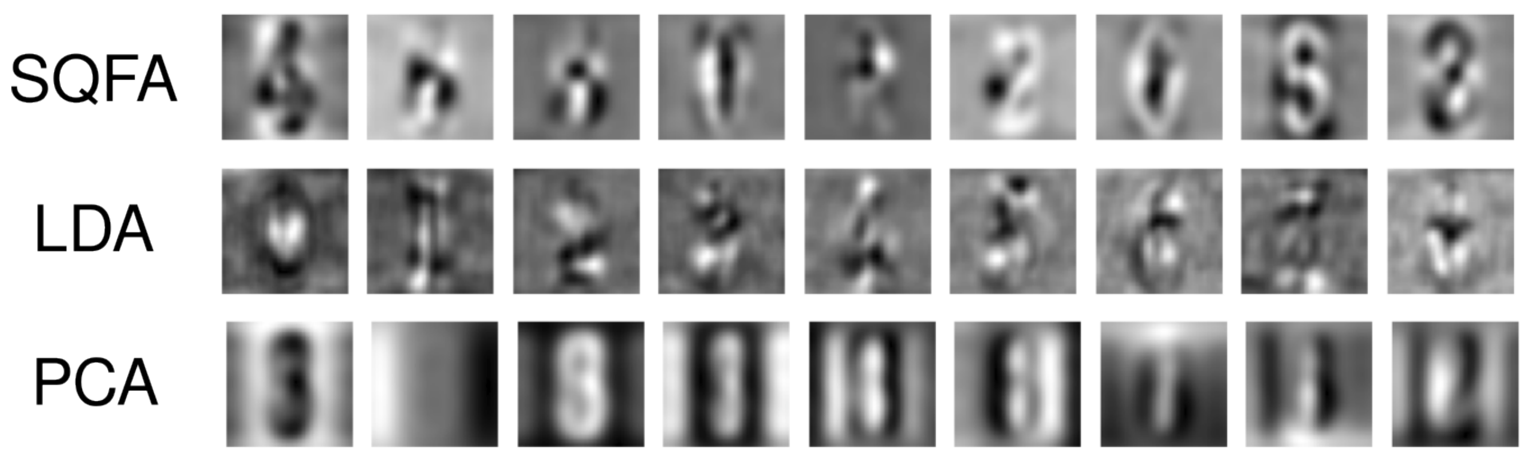
Figure 1. The 3 subspaces given by dimension pairs (1,2), (3,4) and (5,6) are shown. Colored ellipses show the mean and covariance of each class. **We learn 2 filters in the same 6D data space with each of SQFA (red), LDA (green) and PCA (blue).** The filters learned by each method are shown as arrows overlayed on the data space. Each method learns filters in a different subspace. SQFA captures the most discriminative subspace, with second-order class differences.

SQFA for digit classification

Street View House Numbers is a challenging classification dataset.



We learn 9 filters with each SQFA, LDA and PCA. SQFA filters look more digit-like.



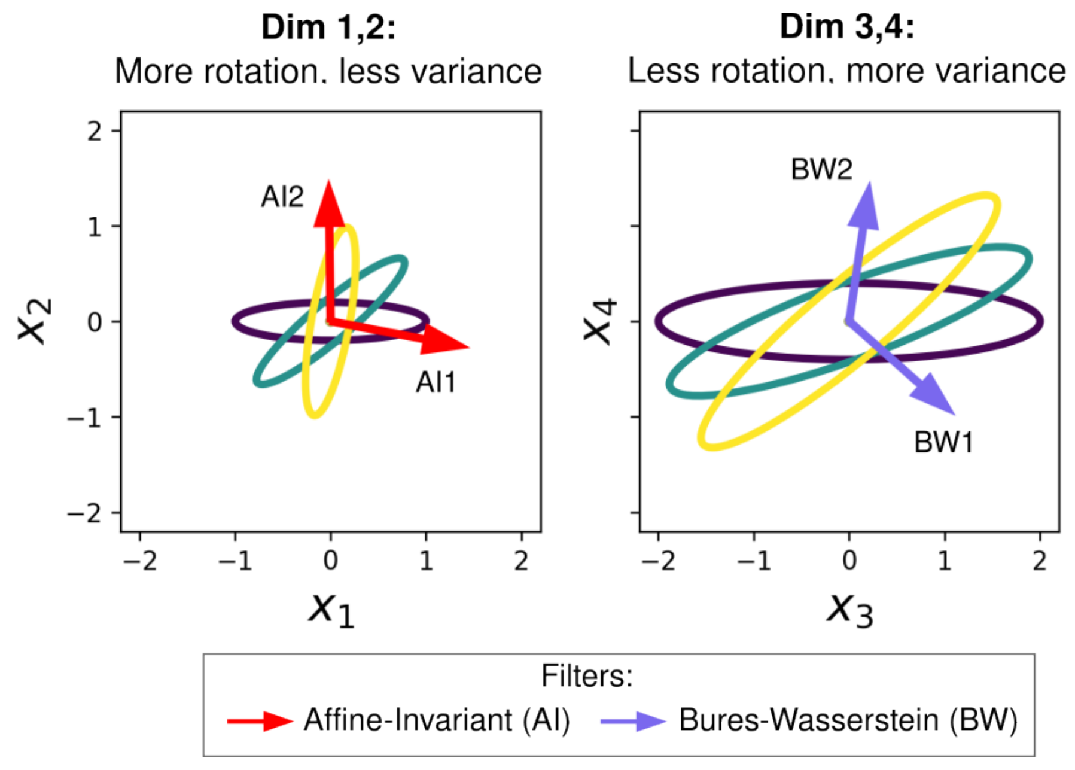
A quadratic classifier (QDA) has higher classification accuracy when using features learned with SQFA than when using features learned with LDA, PCA, ICA or Factor Analysis.

Features	QDA Accuracy (%)
SQFA	**67.5**
LDA	37.4
PCA	38.6
ICA	38.6
Factor Analysis	33.4

Choice of metric is important

Other Riemannian distances (defined by different metrics) can be used in $\mathcal{S}_{++}^m$, but not all of them reflect discriminability.
We compare the filters learned by maximizing the Affine-Invariant distance and the Bures-Wasserstein (optimal transport) distance in a toy 4D example.

Figure 2. Dimensions (1,2) have less variance than (3,4), but are more discriminative because classes are more rotated. Filters learned with the Affine Invariant distance favor the most discriminative subspace. Filters learned with the Wasserstein distance are not invariant to scale (this can be intuited form the earth movers analogy of optimal transport), and they favor the subspace with the most variance. The Wasserstein distance does not capture the most discriminative subspace.



Conclusion

- SQFA is a method for dimensionality reduction that leverages the information geometry of SPD matrices
- SQFA learns features that are different from those learned by other common methods. The features are sensitive to second-order class differences
- SQFA can help tackle problems with high-dimensional covariance matrices
- Information geometry can be a powerful tool for machine learning

NeurReps @ NEURIPS 2024

dherrera1911@gmail.com