Extended Abstract Track

Supervised Quadratic Feature Analysis: An information geometry approach to dimensionality reduction

Editors: List of editors' names

Abstract

Supervised dimensionality reduction seeks to map class-conditional data to a low-dimensional feature space while maximizing class discriminability. Although differences in class-conditional second-order statistics can often aid discriminability, most supervised dimensionality reduction methods focus on first-order statistics. Here, we present Supervised Quadratic Feature Analysis (SQFA), a dimensionality reduction technique that finds a set of features that preserves second-order differences between classes. For this, we exploit a relation between class discriminability and the Information geometry of second-moment (or covariance) matrices as points on the symmetric positive definite (SPD) manifold. We discuss the reasoning behind the approach, and demonstrate its utility in a simple vision task.

Keywords: Dimensionality reduction, Information geometry, Discriminative features

1. Introduction

Consider the random variable $\boldsymbol{x} \in \mathbb{R}^n$. The aim of dimensionality reduction is to find some mapping $\sigma : \mathbb{R}^n \to \mathbb{R}^m$, with $m < n$ such that the mapped variable $\boldsymbol{z} = \sigma(\boldsymbol{x})$ retains the maximum amount of information in $\boldsymbol{x}$ for the relevant purpose. In supervised dimensionality reduction, $\boldsymbol{x}$ has a corresponding class label $y \in \{1, ..., q\}$, the dataset is composed of pairs $\{\boldsymbol{x}_t, y_t\}_{t=1}^p$, and the goal is to find the mapping σ such that $\boldsymbol{z}$ supports the best possible estimation or discrimination of y.

In many real world problems, class-conditional second-order statistics $\boldsymbol{C}_i = \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^T | y = i\right]$ can support class discriminability. Few dimensionality reduction techniques, however, exploit second-order class differences for feature learning. Generalized eigenvectors have previously been used as feature sets for second-order discrimination (Karampatziakis and Mineiro, 2014). The generalized eigenvectors $\boldsymbol{v}_k^{\{i,j\}}$ for a pair of matrices $(\boldsymbol{C}_i, \boldsymbol{C}_j)$ are the local maximizers of the ratio of the expected class-conditional squared feature outputs

$$R_{ij}(\boldsymbol{w}) = \frac{\mathbb{E}[(\boldsymbol{w}^T\boldsymbol{x})^2 | y = i]}{\mathbb{E}[(\boldsymbol{w}^T\boldsymbol{x})^2 | y = j]} = \frac{\boldsymbol{w}^T\boldsymbol{C}_i\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{C}_j\boldsymbol{w}} \tag{1}$$

The ratio $R_{ij}(\boldsymbol{v}_k^{\{i,j\}})$ equals the generalized eigenvalue $\lambda_k^{\{i,j\}}$ associated with $\boldsymbol{v}_k^{\{i,j\}}$. The farther $\log(\lambda_k^{\{i,j\}})$ is from 0, the more different the expected squared feature outputs of $\boldsymbol{v}_k^{\{i,j\}}$ are for classes i, j, and the more discriminable the classes will be in the direction specified by $\boldsymbol{v}_k^{\{i,j\}}$. For each pair $(\boldsymbol{C}_i, \boldsymbol{C}_j)$, Karampatziakis and Mineiro (2014) use as features the m vectors $\boldsymbol{v}_k^{\{i,j\}}$ with log-eigenvalues farthest from 0. However, because a different set of features is obtained for every pair of classes, the number of feature sets scales quadratically with the number of classes. Across these multiple sets, features are often redundant.

Here, we introduce Supervised Quadratic Feature Analysis (SQFA), a method for learning a single low-dimensional feature set that maximizes second-order discriminability across classes. Its loss function exploits a relation between class discriminability and the Information geometry of the symmetric positive definite (SPD) manifold (see Appendix A).

2. Model

2.1. Setup of the problem

Our goal is to find the linear function $\sigma_{\boldsymbol{f}} : \mathbb{R}^n \to \mathbb{R}^m$ given by $\sigma_{\boldsymbol{f}}(\boldsymbol{x}) = \boldsymbol{f}^T\boldsymbol{x} = \boldsymbol{z}$ that maximizes class discriminability by the feature second-order statistics $\Psi_i = \mathbb{E}\left[\boldsymbol{z}\boldsymbol{z}^T | y = i\right]$. Given a set of q class conditional matrices $\{\boldsymbol{C}_i\}_{i=1}^q$, and noting that $\Psi_i = \boldsymbol{f}^T\boldsymbol{C}_i\boldsymbol{f}$, the general problem can be expressed as

$$\underset{\boldsymbol{f}\in\mathbb{R}^{n\times m}}{\arg\max}\, L\left(\{\boldsymbol{f}^T\boldsymbol{C}_i\boldsymbol{f}\}_{i=1}^q\right) \tag{2}$$

where $L\left(\{\boldsymbol{f}^T\boldsymbol{C}_i\boldsymbol{f}\}_{i=1}^q\right)$ measures second-order discriminability in the feature space. Intuitively, classes i, j are more discriminable if their matrices (Ψ_i, Ψ_j) are more different from one another, or, in geometric terms: more "far apart". This motivates a geometric loss using squared pairwise distances between matrices

$$L\left(\{\Psi_i\}_{i=1}^{i=q}\right) = \sum_{i=1}^{q}\sum_{j=1}^{i} d(\Psi_i, \Psi_j)^2 \tag{3}$$

Second moment matrices Ψ belong to the $m \times m$ SPD manifold $\mathcal{S}_+^m$, which has a well studied geometry and many distances that could be used in Equation (3) (Atkinson and Mitchell, 1981; Thanwerdas and Pennec, 2023). However, our goal requires a distance that relates to discriminability, as discussed next.

2.2. Affine Invariant distance and discriminability

The most commonly used distance in $\mathcal{S}_+^m$ is the Affine Invariant distance

$$d_{AI}(\Psi_i, \Psi_j) = \left\|\log(\Psi_i^{-1/2}\Psi_j\Psi_i^{-1/2})\right\|_F = \sqrt{\sum_{k=1}^{m}\left(\log\lambda_k^{\{i,j\}}\right)^2} \tag{4}$$

where log is the matrix logarithm, $\|\cdot\|_F$ is the Frobenius norm, and $\lambda_k^{\{i,j\}}$ are the generalized eigenvalues of (Ψ_i, Ψ_j). Some properties of d_{AI} make it a sensible distance for learning discriminative features.

First, Equation (4) summarizes, for an individual pair of class feature statistics (Ψ_i, Ψ_j), how different the log generalized eigenvalues $\log(\lambda_k^{\{i,j\}})$ are from 0. As the value of Equation (4) increases, so should the discriminability of the two classes (see Section 1, Equation (1)). Note that, unlike Karampatziakis and Mineiro (2014), we do not use a different set of m generalized eigenvectors $\boldsymbol{v}_k^{\{i,j\}}$ for each pair of class statistics $(\boldsymbol{C}_i, \boldsymbol{C}_j)$ as features. Rather, we find the single set of m features $\boldsymbol{f}$ that simultaneously discriminate between all pairs of classes. This set of features maximizes the sum of squared Affine Invariant distances between statistics of all pairs of classes in the feature space. The specific objective

$$\underset{\boldsymbol{f}\in\mathbb{R}^{n\times m}}{\arg\max}\sum_{i=1}^{q}\sum_{j=1}^{i} d_{AI}(\Psi_i, \Psi_j)^2 = \underset{\boldsymbol{f}\in\mathbb{R}^{n\times m}}{\arg\max}\sum_{i=1}^{q}\sum_{j=1}^{i}\sum_{k=1}^{m}(\log\lambda_k^{\{i,j\}})^2 \tag{5}$$

uses the generalized eigenvalues $\lambda_k^{\{i,j\}}$ of all pairwise statistics (Ψ_i, Ψ_j) in the feature space, and is obtained by combining equations Equations (2), (3) and (4).

Second, the Affine Invariant metric that induces this distance is very closely related to the Fisher Information in zero-mean Gaussians (Atkinson and Mitchell, 1981). Specifically, the squared norm under this metric for a vector (i.e. symmetric matrix) $\boldsymbol{A}$ in the tangent space of point Ψ, $T_\Psi \mathcal{S}_+^m$ is $\|\boldsymbol{A}\|_\Psi^2 = \langle \boldsymbol{A}, \boldsymbol{A} \rangle_\Psi = \mathrm{Tr}(\Psi^{-1}\boldsymbol{A}\Psi^{-1}\boldsymbol{A})$ which is twice the Fisher Information of $\mathcal{N}(0, \Psi)$ with respect to covariance matrix Ψ along direction $\boldsymbol{A}$. The distance $d_{AI}(\Psi_i, \Psi_j)$ is obtained by taking the geodesic (i.e. shortest) curve from Ψ_i to Ψ_j and integrating the norm of the velocity of the curve at each point. The norm at point Ψ along the geodesic is given by $\|\Psi'\|_\Psi$, where $\Psi' \in T_\Psi \mathcal{S}_+^m$ is the differential of the geodesic at point Ψ. Thus, the Affine Invariant distance can be thought of as the accumulated Fisher Information (i.e. discriminability) of the infinitesimal perturbations that convert $\mathcal{N}(0, \Psi_i)$ into $\mathcal{N}(0, \Psi_j)$ along the geodesic. For the interested reader, the geometry of probability distribution manifolds is the subject of information geometry (Amari, 2016).

2.3. Relation to gaPCA

SQFA bears similarities to geometry-aware PCA (gaPCA), a method for unsupervised dimensionality reduction in $\mathcal{S}_+^n$ (Harandi et al., 2014). gaPCA transforms data $\boldsymbol{\Sigma} \in \mathcal{S}_+^n$ using the function $\sigma_{\boldsymbol{f}} : \mathcal{S}_+^n \to \mathcal{S}_+^m$ of the form $\sigma_{\boldsymbol{f}}(\Sigma) = \boldsymbol{f}^T \Sigma \boldsymbol{f}$, with $m < n$. Analogous to PCA, gaPCA maximizes the Fréchet variance of the transformed data. Like SQFA, gaPCA transforms a set of matrices $\mathcal{S}_+^n \to \mathcal{S}_+^m$ and uses a geometric loss on $\mathcal{S}_+^m$. However, there are important differences between the methods. Conceptually, SQFA performs supervised dimensionality reduction of data in $\mathbb{R}^n$, whereas gaPCA performs unsupervised dimensionality reduction of data in $\mathcal{S}_+^n$. This conceptual difference has methodological consequences. First, SQFA maximizes squared distances between classes, which are related to discriminability, while gaPCA maximizes Fréchet variance. These two objectives are the same in Euclidean space but not in $\mathcal{S}_+^m$. Second, gaPCA does not require the distance between matrices to reflect discriminability in the underlying feature space, while SQFA does (Section 2.2).

3. Results

We tested SQFA on a dataset of naturalistic contrast video patches (1 deg) of moving surfaces that has previously been used to study the human visual system (Burge and Geisler, 2015; Chin and Burge, 2020). Each video patch $\boldsymbol{x} \in \mathbb{R}^{450}$ has 30 pixels (in space) and 15 frames (in time), and moves at one of 41 different speeds $y \in \{s_1, ..., s_{41} | s_i \in \mathbb{R}^n\}$. The vertical axis is averaged over, resulting in XT frames, rather than XYT frames. The task is to estimate the speed y of each video. We learn $\boldsymbol{f}$ by Equation (5) with gradient ascent using NAdam. The columns of $\boldsymbol{f}$ (features) are constrained to have unit norm.

This dataset is appealing for three reasons. First, finding features (receptive fields) that are useful for solving visual tasks is fundamental to systems neuroscience and perception science (Burge, 2020). Second, because $\boldsymbol{x}$ are contrast videos, the means $\mathbb{E}[\boldsymbol{x}|y = s_i]$ are zero (or nearly so), so that discrimination must rely on higher-order statistics. Third, an existing approach called AMA-Gauss that learns features from class-conditional random samples to optimize a probabilistic decoder, has been applied to the same task, and provides a useful benchmark for comparison (Jaini and Burge, 2017).

Table 1: Probabilistic decoder loss. MAP estimates are used to compute MSE and MAE.

Features	$\mathbb{E}\left[-\log P(y = y_t \mid \boldsymbol{z}_t)\right]$	**MSE**	**MAE**
PCA	2.21	1.60	0.72
FA	2.22	1.59	0.73
AMA-Gauss	**2.00**	1.46	**0.61**
SQFA	2.12	**1.30**	0.64

Table 1 shows the performance of a probabilistic decoder (see Jaini and Burge (2017) for details) that uses the features learned with PCA, Factor Analysis (FA) (both performed over the whole pooled dataset), AMA-Gauss, and SQFA. SQFA features outperform PCA and FA features for all the performance metrics analyzed. SQFA features are on par with AMA-Gauss features, despite the fact that AMA-Gauss learns features that optimize the decoder performance using as a loss the negative log posterior of the true class y_t (column 2 of Table 1). Furthermore, SQFA and AMA-Gauss features share some similarities, although there are also clear differences between the two (Figure 1). Importantly, SQFA features are learned 30 times faster than AMA-Gauss features for this dataset. Our model performs similarly well when tested on related datasets of disparity estimation (Jaini and Burge, 2017), 3D-motion estimation (Herrera-Esposito and Burge, 2024), and when using a K-nearest neighbors classifier (Appendix B).

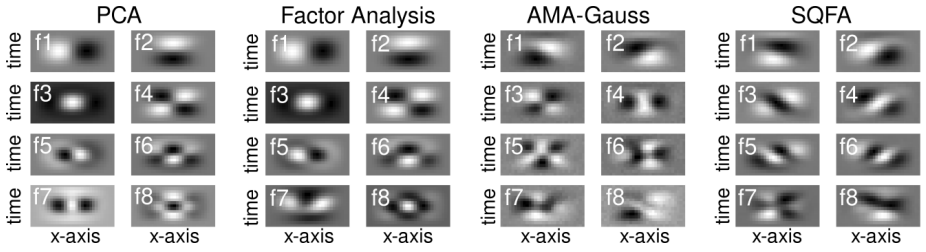

Figure 1: Features learned with different methods.

4. Conclusion

We present SQFA, a supervised dimensionality reduction method for learning features that maximize second-order differences between classes, using an information geometric approach. SQFA may be applied on its own or in combination with methods based on first-order statistics (e.g. LDA) to find a linear feature space where class distributions are most different. SQFA is also a principled approach to finding features when there is access to the data statistics but not to the data itself (e.g when statistics of unobserved classes are obtained through interpolation (Nejatbakhsh et al., 2023)). Future work will further develop the theoretical relationship between the Affine Invariant distance and discriminability, examine other distances in $\mathcal{S}_+^m$, and test SQFA on a wider range of datasets.

5. Citations and Bibliography

Acknowledgments

Acknowledgements go here.

References

Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, February 2016. ISBN 978-4-431-55978-8. Google-Books-ID: UkSFCwAAQBAJ.

Colin Atkinson and Ann F. S. Mitchell. Rao's Distance Measure. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 43(3):345–365, 1981. ISSN 0581-572X. URL https://www.jstor.org/stable/25050283. Publisher: Springer.

Johannes Burge. Image-Computable Ideal Observers for Tasks with Natural Stimuli. *Annual Review of Vision Science*, 6(1):491–517, 2020. doi: 10.1146/annurev-vision-030320-041134. URL https://doi.org/10.1146/annurev-vision-030320-041134. _eprint: https://doi.org/10.1146/annurev-vision-030320-041134.

Johannes Burge and Wilson S. Geisler. Optimal speed estimation in natural image movies predicts human performance. *Nature Communications*, 6(1):7900, August 2015. ISSN 2041-1723. doi: 10.1038/ncomms8900. URL https://www.nature.com/articles/ncomms8900. Publisher: Nature Publishing Group.

Benjamin M. Chin and Johannes Burge. Predicting the Partition of Behavioral Variability in Speed Perception with Naturalistic Stimuli. *The Journal of Neuroscience*, 40(4):864–879, January 2020. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1904-19.2019. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.1904-19.2019.

Mehrtash T. Harandi, Mathieu Salzmann, and Richard Hartley. From Manifold to Manifold: Geometry-Aware Dimensionality Reduction for SPD Matrices, November 2014. URL http://arxiv.org/abs/1407.1120. arXiv:1407.1120 [cs].

Daniel Herrera-Esposito and Johannes Burge. Optimal motion-in-depth estimation with natural stimuli, March 2024. URL https://www.biorxiv.org/content/10.1101/2024.03.14.585059v1. Pages: 2024.03.14.585059 Section: New Results.

Priyank Jaini and Johannes Burge. Linking normative models of natural tasks to descriptive models of neural response. *Journal of Vision*, 17(12):16, October 2017. ISSN 1534-7362. doi: 10.1167/17.12.16. URL https://doi.org/10.1167/17.12.16.

Nikos Karampatziakis and Paul Mineiro. Discriminative Features via Generalized Eigenvectors. *International conference on machine learning*, 2014.

Amin Nejatbakhsh, Isabel Garon, and Alex Williams. Estimating Noise Correlations Across Continuous Conditions With Wishart Processes. *Advances in Neural Information Processing Systems*, 36:54032–54045, December 2023.

URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a935ba2236c6ba0fb620f23354e789ff-Abstract-Conference.html.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.

Yann Thanwerdas and Xavier Pennec. O(n)-invariant Riemannian metrics on SPD matrices. *Linear Algebra and its Applications*, 661:163–201, March 2023. ISSN 0024-3795. doi: 10.1016/j.laa.2022.12.009. URL https://www.sciencedirect.com/science/article/pii/S0024379522004360.

Appendix A. Graphical illustration of SQFA

Figure 2 provides a graphical illustration of the geometric perspective of SQFA, and how the class-conditional statistics in the SQFA feature space might compare to those of the features obtained from Linear Discriminant Analysis (LDA).

In the left column, the second moment matrices of the data for the different classes $\boldsymbol{C}_i$ are shown as points in $\mathcal{S}_+^n$ (which is a cone in the vector space of symmetric matrices Thanwerdas and Pennec (2023)). In the center column of Figure 2 we see how these matrices are mapped to class-conditional second moment matrices of features $\Psi_i \in \mathcal{S}_+^m$. This mapping is shown for features learned with SQFA (top) and LDA (bottom). The objective maximized by SQFA is the squared distances between the Ψ_i, and thus, the points are farther apart for SQFA features than for the LDA features. This is because SQFA features explicitly maximize squared distances.

On the right column, we see how the distributions of the features learned by the two methods for a same dataset may differ. LDA maximizes the distances between class means normalized by within class variability. Thus, we see that the class means for LDA are farther away from one another, and within-class variance is also smaller than for SQFA. However, SQFA classes have very different second-order structures, which result from maximizing the distances between the class second moments. In this illustrative case, the classes are better separated by SQFA than by LDA.

Appendix B. Performance with KNN decoder

To verify that our results are not dependent on the decoder used, we tested the performance of a k-nearest neighbors (KNN) classifier using the same sets of features as in the main text. We also used features extracted with Kernel PCA (KPCA) with RBF kernel and default parameters from scikit-learn (Pedregosa et al.) (which we omitted for the probabilistic decoder because KPCA does not produce linear filters). Like for the probabilistic decoder, SQFA features outperform PCA and FA features for the speed estimation task.

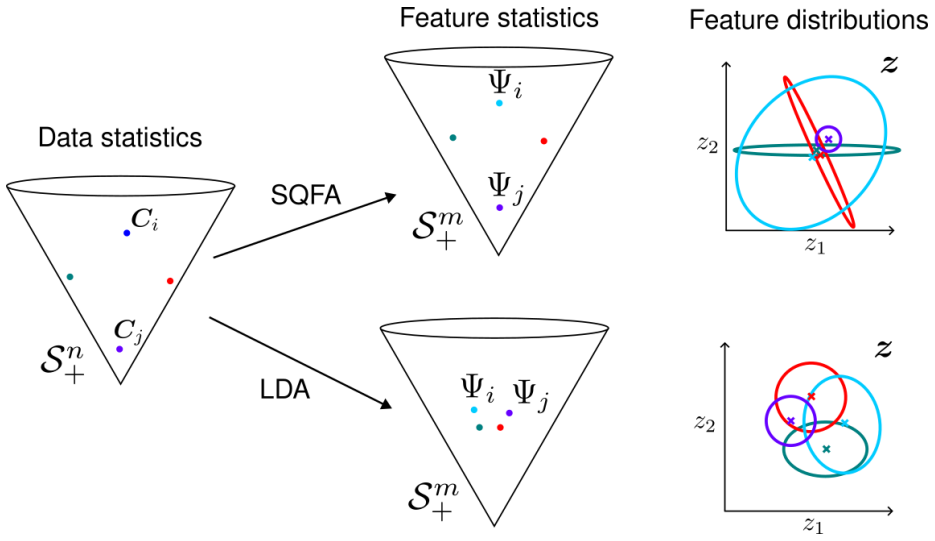

Figure 2: Illustration of SQFA as compared to LDA. The left column shows the second moment matrices $\boldsymbol{C}_i$ of the data for the different classes, as points in $\mathcal{S}_+^n$ (different classes are different points). The center column shows second moment matrices Ψ_i for the outputs of filters learned with two methods: SQFA (top) and LDA (bottom). The right column shows the class-conditional distributions in a feature space of reduced dimension. Crosses indicate the class means. Ellipses indicate the class covariances (color matched to the points on the manifold).

Table 2: KNN classifier loss with $k = 5$. KNN outputs are used to compute MSE and MAE.

Features	**% Correct**	**MSE**	**MAE**
PCA	22.0	1.45	0.68
FA	19.7	1.71	0.77
KPCA	22.0	1.45	0.68
AMA-Gauss	**33.4**	**0.96**	**0.45**
SQFA	30.0	0.99	0.51