Supervised Quadratic Feature Analysis: Information Geometry for Dimensionality Reduction

Anonymous authors
Paper under double-blind review

Abstract

Supervised dimensionality reduction maps labeled data into a low-dimensional feature space while preserving class discriminability. A common approach is to maximize a statistical measure of dissimilarity between classes in the feature space. Information geometry provides an alternative framework for measuring class dissimilarity, with the potential for improved insights and novel applications. Information geometry, which is grounded in Riemannian geometry, uses the Fisher information metric, a local measure of discriminability that induces the Fisher-Rao distance. Here, we present Supervised Quadratic Feature Analysis (SQFA), a linear dimensionality reduction method that maximizes Fisher-Rao distances between class-conditional distributions, under Gaussian assumptions. We motivate the Fisher-Rao distance as a good proxy for discriminability. We show that SQFA features support good classification performance with Quadratic Discriminant Analysis (QDA) on three real-world datasets. SQFA provides a novel framework for supervised dimensionality reduction, motivating future research in applying information geometry to machine learning and neuroscience.

1 Introduction

Consider a random vector $\mathbf{x} \in \mathbb{R}^n$ with label $y \in \{1, \ldots, c\}$, where c is the number of classes. Supervised dimensionality reduction aims to map the high-dimensional variable $\mathbf{x}$ to a lower-dimensional variable $\mathbf{z} \in \mathbb{R}^m$ that best supports classification performance, or discriminability. There are many methods for learning nonlinear features, like deep neural networks (LeCun et al., 2015) and manifold learning (Sainburg et al., 2021). However, methods for learning linear features of the form $\mathbf{z} = \mathbf{F}^T\mathbf{x}$ are still in great demand because of their simplicity, interpretability, and data efficiency (Cunningham & Ghahramani, 2015).

Directly maximizing class discriminability in the reduced-dimensional feature space is often impractical (Fukunaga, 1990). A common approach is to maximize a measure of dissimilarity between class distributions, as a proxy for discriminability. Linear Discriminant Analysis (LDA), for instance, maximizes the squared Mahalanobis distances between the classes, assuming a shared within-class covariance. Other methods use different dissimilarity measures, typically derived from statistics or information theory (Choi & Lee, 2003; Dwivedi et al., 2022).

Information geometry provides an alternative approach for measuring the dissimilarity between probability distributions, by considering them as points in a statistical manifold. It uses the Fisher information metric, a principled measure of local discriminability, to measure distances between the distributions. This metric is the unique invariant metric under reparameterizations (Nielsen, 2020), and is widely used in statistics, machine learning, and neuroscience (Dayan & Abbott, 2005). The distance induced by integrating this metric along the geodesics of the statistical manifold is the Fisher-Rao distance.

Despite its theoretical and practical appeal, Fisher-Rao distances between class-conditional distributions have not been used for dimensionality reduction. One potential reason is that closed-form expressions for the Fisher-Rao distance are not available for many important statistical distributions, such as the multivariate Gaussian (Miyamoto et al., 2024; Nielsen, 2023). Another is that alternative measures of dissimilarity from statistics and information theory (e.g. Bhattacharyya distance, Kullback-Leibler divergence) are more directly linked to classification error, making them a more obvious choice (Fukunaga, 1990).

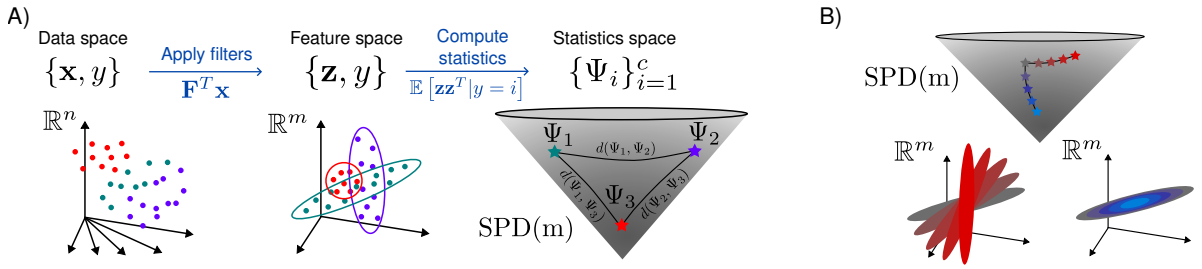

Figure 1: SQFA learns features using information geometry. **A)** SQFA and smSQFA map the n-dimensional data into an m-dimensional feature space using the linear filters $\mathbf{F}$. In smSQFA, the class-specific second-moment matrices of the features are represented as points in the SPD(m) manifold (which is an open cone). Fisher-Rao distances in SPD(m) are used for learning. **B)** Each point in SPD(m) (top) corresponds to a second-moment ellipse (below). As the distance in SPD(m) increases, the second-order statistics become more different and more discriminable.

However, maximizing Fisher-Rao distances between classes is an interesting learning objective for multiple reasons. First, there is a growing interest in studying representation geometry, an important area of research in both machine learning and neuroscience (Kriegeskorte & Wei, 2021; Chung & Abbott, 2021). The Fisher information metric is an important tool for characterizing the local geometry of neural representations in terms of the local discriminability, but it is limited to infinitesimally close distributions. (Wang & Ponce, 2021; Ding et al., 2023; Feather et al., 2024; Ye & Wessel, 2024; Zhou et al., 2024). The Fisher-Rao distance extends this local metric into a non-local geometry, providing a richer tool for characterizing representation geometry (Kriegeskorte & Wei, 2021). Additionally, it has been hypothesized that such geodesic distances that integrate local discriminability along a path may be related to suprathreshold (i.e. non-local) similarity judgments in both neuroscience and perceptual psychology (Fechner, 1860; Vacher & Mamassian, 2024; Zhou et al., 2024; Hong et al., 2025).

Second, information geometry provides a rich set of tools with a growing number of applications (Amari, 2016; Miyamoto et al., 2024). A dimensionality reduction method based on Fisher-Rao distances could provide low-dimensional features that support the use of these tools, for example clustering of classes by their similarity. Third, given the broad interest in information geometry, it is of standalone theoretical interest to understand how maximizing Fisher-Rao distances compares to other common objective functions.

Here, we present Supervised Quadratic Feature Analysis (SQFA), a dimensionality reduction method that learns linear features by maximizing Fisher-Rao distances between class-conditional distributions, under Gaussian assumptions (but this assumption can be relaxed). We show that SQFA features support excellent quadratic discriminability, using a Quadratic Discriminant Analysis (QDA) classifier, on real-world datasets. We use an exact expression for the Fisher-Rao distance for zero-mean Gaussians, and the Calvo-Oller lower bound (Calvo & Oller, 1990) as a closed-form approximation for the Fisher-Rao distance in the general case; we show that the Calvo-Oller bound closely approximates the true distance for the analyzed datasets. We also analyze variants of SQFA using well-known statistical measures dissimilarity (i.e. Bhattacharyya and Hellinger distances), and compare their results.

2 Fisher-Rao distance as a discriminability proxy

Notation. The class-conditional means and covariances of the data variable $\mathbf{x}$ are denoted by $\boldsymbol{\gamma}_i = \mathbb{E}\left[\mathbf{x}|y=i\right]$ and $\boldsymbol{\Phi}_i = \mathbb{E}\left[(\mathbf{x}-\boldsymbol{\gamma}_i)(\mathbf{x}-\boldsymbol{\gamma}_i)^T|y=i\right]$, respectively. The low-dimensional projection of $\mathbf{x}$ is given by $\mathbf{z} = \mathbf{F}^T\mathbf{x}$, where $\mathbf{F} \in \mathbb{R}^{n\times m}$ is a matrix of filters, and the respective class-conditional means, second moments, and covariances are given by $\boldsymbol{\mu}_i = \mathbb{E}\left[\mathbf{z}|y=i\right]$, $\boldsymbol{\Psi}_i = \mathbb{E}\left[\mathbf{z}\mathbf{z}^T|y=i\right]$, and $\boldsymbol{\Sigma}_i = \boldsymbol{\Psi}_i - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T$. We also denote by $\boldsymbol{\theta}_i$ the parameters of $p(\mathbf{z}|y=i) = p(\mathbf{z}|\boldsymbol{\theta}_i)$, which in the Gaussian cases is $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

2.1 Supervised dimensionality reduction via dissimilarity maximization

The standard goal of linear supervised dimensionality reduction is to find filters $\mathbf{F}$ such that the classes are as discriminable as possible in the low-dimensional space of the variable $\mathbf{z} = \mathbf{F}^T\mathbf{x}$. A common way to achieve this is to maximize a dissimilarity measure between the class-conditional distributions. In the case of Gaussian distributions $p(\mathbf{z}|y=i) = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, dissimilarity measures depend on the parameters $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Given a dissimilarity measure $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ between classes, the optimization problem reduces to

$$\underset{\mathbf{F}\in\mathbb{R}^{n\times m}}{\arg\max} \sum_{i=1}^{c}\sum_{j=1}^{c} d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \tag{1}$$

If $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ is a good proxy for discriminability, the learned features should support high classification accuracy (Figure 1). We argue that the Fisher-Rao distance, $d_{FR}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$, is a sensible proxy for discriminability.

2.2 Fisher-Rao distance as accumulated local discriminability

For a distribution $p(\mathbf{z}|\boldsymbol{\theta})$, the Fisher information in the direction $\boldsymbol{\theta}'$ in the parameter space is defined as

$$\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta}') = \mathbb{E}\left[(s(\mathbf{z},\boldsymbol{\theta})\cdot\boldsymbol{\theta}')^2\right] = \boldsymbol{\theta}'^T\mathbf{J}(\boldsymbol{\theta})\boldsymbol{\theta}' \tag{2}$$

where the expectation is over $\mathbf{z}$, $s(\mathbf{z},\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\log p(\mathbf{z}|\boldsymbol{\theta})$ is the score function, and $\mathbf{J}(\boldsymbol{\theta}) = \mathbb{E}[s(\mathbf{z},\boldsymbol{\theta})s(\mathbf{z},\boldsymbol{\theta})^T]$ is the Fisher information matrix. The quantity $\sqrt{\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta}')}$ measures the discriminability between $p(\mathbf{z}|\boldsymbol{\theta})$ and $p(\mathbf{z}|\boldsymbol{\theta}+\epsilon\boldsymbol{\theta}')$, where $\epsilon\boldsymbol{\theta}'$ is a small perturbation.

Let $\boldsymbol{\theta}(t) = (\boldsymbol{\mu}(t), \boldsymbol{\Sigma}(t))$ be the Fisher-Rao geodesic (i.e. the shortest curve) connecting $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Psi}_i)$ to $\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Psi}_j)$ along the manifold of Gaussian distributions, where $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_i$, $\boldsymbol{\theta}(1) = \boldsymbol{\theta}_j$, and $\boldsymbol{\theta}'(t)$ is the velocity. The Fisher-Rao distance is obtained by integrating the speed along the geodesic, $d_{FR}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \int_0^1 \|\boldsymbol{\theta}'(t)\|dt$. The Fisher information metric defines the speed as $\|\boldsymbol{\theta}'\| = \sqrt{\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta}')}$, which for the Gaussian case is

$$\|\boldsymbol{\theta}'(t)\| = \sqrt{\mathcal{I}_{\boldsymbol{\theta}(t)}(\boldsymbol{\theta}'(t))} = \left[\boldsymbol{\mu}'(t)^T\boldsymbol{\Sigma}(t)^{-1}\boldsymbol{\mu}'(t) + \tfrac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Sigma}(t)^{-1}\boldsymbol{\Sigma}'(t)\boldsymbol{\Sigma}(t)^{-1}\boldsymbol{\Sigma}'(t)\right)\right]^{1/2} \tag{3}$$

Then, $\|\boldsymbol{\theta}'(t)\|$ is a measure of the discriminability between $\mathcal{N}(\boldsymbol{\mu}(t), \boldsymbol{\Sigma}(t))$ and $\mathcal{N}(\boldsymbol{\mu}(t+\mathrm{d}t), \boldsymbol{\Sigma}(t+\mathrm{d}t))$. Thus, the Fisher-Rao distance can be expressed as $d_{FR}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \int_0^1 \sqrt{\mathcal{I}_{\boldsymbol{\theta}(t)}(\boldsymbol{\theta}'(t))}dt$, and it can be conceptualized as the accumulated discriminability of the infinitesimal perturbations transforming $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ into $\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ along the geodesic, making it a sensible proxy for discriminability.

2.3 Closed-form expressions for the Fisher-Rao distance

There is no closed-form expression for the Fisher-Rao distance between arbitrary Gaussians. While numerical methods exist to compute it (Nielsen, 2023; Nielsen & Soen, 2024), they are too costly for dimensionality reduction. To overcome this problem, we use two complementary approaches: 1) we consider the special case of zero-mean Gaussians, and 2) we use a closed-form lower bound for the general case.

The manifold of m-dimensional zero-mean Gaussians is equivalent to the manifold of m-by-m symmetric positive definite matrices, SPD(m), where each point corresponds to a covariance matrix. The Fisher-Rao distance in this case, denoted $d_{FR}(\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j)$, is proportional to the affine-invariant distance d_{AI} in SPD(m) (Atkinson & Mitchell, 1981)

$$d_{AI}(\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j) = \sqrt{\sum_{k=1}^{m}\log^2\lambda_k} = d_{FR}(\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j)\sqrt{2} \tag{4}$$

where λ_k is the k-th generalized eigenvalue of the pair $(\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j)$. Zero-mean Gaussians are relevant for some applications, and having an expression for the true distance in this special case is useful for validating the method. We use a variant of SQFA that assumes zero-mean Gaussians, and that maximizes the distances $d_{AI}(\boldsymbol{\Psi}_i, \boldsymbol{\Psi}_j)$ where $\boldsymbol{\Psi}_i$ is the second-moment matrix for class i in the feature space. We call this variant

smSQFA (for 'second-moment SQFA'). In the Appendix B we further discuss the link between Fisher-Rao distance and discriminability for the zero-mean case.

For arbitrary Gaussians, we use the Calvo-Oller bound (Calvo & Oller, 1990). It is obtained by embedding $\boldsymbol{\theta}$ into $\mathrm{SPD}(\mathrm{m}+1)$ as

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{bmatrix} \tag{5}$$

and the bound is $d_{FR}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \geq d_{AI}(\boldsymbol{\Omega}_i, \boldsymbol{\Omega}_j)/\sqrt{2}$. This bound has desirable properties: it is a true distance, it matches d_{FR} when $\boldsymbol{\mu}_i = \boldsymbol{\mu}_j$, it is invariant to invertible linear transformations of $\mathbf{z}$, and it is often a good approximation to the true distance in simple cases (Nielsen, 2023). The Calvo-Oller bound is also a good approximation for the real-world datasets used here (Appendix C), highlighting its potential for real-world applications of information geometry. Importantly, the Calvo-Oller bound can be easily extended to other elliptical distributions (e.g. multivariate t-Student and Cauchy distributions) (Calvo & Oller, 2002; Nielsen, 2023), allowing to relax the Gaussian assumption. SQFA learns filters by maximizing the Calvo-Oller bound $d_{AI}(\boldsymbol{\Omega}_i, \boldsymbol{\Omega}_j)/\sqrt{2}$ between classes.

3 Related work

Maximizing a measure of dissimilarity between classes is a well-established approach for supervised dimensionality reduction. This is exemplified by the canonical example of LDA, which maximizes pairwise squared Mahalanobis distances (see proof in Appendix D). However, the Mahalanobis distance is limited by the assumption that all class-conditional distributions have identical covariance. Other measures, from statistics and information theory, are more flexible in this regard.

The most widely used dissimilarity measure between Gaussians with different covariances for this purpose is likely the Chernoff distance and its special case, the Bhattacharyya distance[1] (Duin & Loog, 2004; Rueda & Herrera, 2008; Choi & Lee, 2003). One reason for their success is that they provide an upper bound for the Bayes classification. This bound is a stronger link to classification error than that provided by the Kullback-Leibler (KL) divergence (Kailath, 1967; Fukunaga, 1990). Another reason is that they have closed-form expressions for the Gaussian case unlike, for example, the Jensen-Shannon divergence.

The Bhattacharyya distance between two distributions $p(\mathbf{z})$ and $q(\mathbf{z})$ is given by $d_B(p,q) = -\log BC(p,q)$, where $BC(p,q) = \int \sqrt{p(\mathbf{z})}\sqrt{q(\mathbf{z})}d\mathbf{z}$ is the Bhattacharyya coefficient. Note that $BC(p,q)$ is the dot product between the square-root densities. For two Gaussians with parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, and defining $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j)/2$, the distance is given by

$$d_B(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \frac{1}{8}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{1}{2}\log \frac{\det(\boldsymbol{\Sigma})}{\sqrt{\det(\boldsymbol{\Sigma}_i)\det(\boldsymbol{\Sigma}_j)}} \tag{6}$$

Another important dissimilarity measure is the Hellinger distance, defined as the (scaled) Euclidean distance between square-root densities, $d_H(p,q) = \sqrt{\frac{1}{2}\int(\sqrt{p(\mathbf{z})} - \sqrt{q(\mathbf{z})})^2 d\mathbf{z}}$. It is a true distance, bounded between 0 and 1, and it converges to the Fisher-Rao distance for infinitesimally close distributions (Amari, 2016). It relates to $BC(p,q)$ via $d_H(p,q) = \sqrt{1 - BC(p,q)}$, meaning that a closed-form expression is available for Gaussians, and that it provides a bound on the Bayes classification error. Notably, d_H has not been used for dimensionality reduction in the multiclass Gaussian case (see Carter et al. (2009) for a non-parametric example, and Dwivedi et al. (2022) for a two-class Gaussian example). This is likely because it is monotonically related to d_B, making them equivalent in the two-class case.

Unlike the Fisher-Rao distance, d_{FR}, d_B, and d_H are not geodesic distances along a statistical manifold. While not essential for maximizing discriminability, the geometric perspective is appealing. On the other hand, the distances d_B and d_H have direct links to Bayes error in the two-class case, unlike d_{FR}. This link, however, is less clear in the multiclass case, since pairwise errors do not determine overall error (Loog

[1] Neither is a true distance, since they do not satisfy the triangle inequality.

et al., 2001; Thangavelu & Raich, 2008). It is therefore an empirical question which objective is better for maximizing discriminability in the multiclass case for specific real-world datasets, as we explore here.

Wasserstein Discriminant Analysis (WDA), is another geometry-based method, that maximizes the regularized Wasserstein distance between the class-conditional empirical distributions (Flamary et al., 2018). The Wasserstein distance is another geodesic statistical distance on the manifold of probability distributions. WDA is a non-parametric method that can learn complex class boundaries, maximizing the performance of classifiers such as k-nearest neighbors (kNN). However, like other non-parametric methods, it is computationally costly, scaling poorly with the number of data points per class. Moreover, unlike the Fisher-Rao distance, the Wasserstein distance is not directly linked to local discriminability, and it is not invariant to invertible linear transformations.

4 Methods

Optimization. Filters $\mathbf{F}$ were optimized by maximizing Equation 1 using L-BFGS, until the change in the objective was $\leq 10^{-6}$. The columns of $\mathbf{F}$ were initialized to random unit vectors and constrained to have unit norm. Adding an orthogonality constraint made optimization slower without improving performance. $\mathbf{F}$ can be orthogonalized after learning without affecting QDA accuracy. Constraints were implemented by parametrizing the filters in terms of an unconstrained variable (see Lezcano Casado (2019)). Optimization is non-convex, but the performance across 20 different initializations was very similar.

The class-conditional statistics of $\mathbf{z}$ were computed by transforming the means and covariances of $\mathbf{x}$ (i.e. $\boldsymbol{\gamma}_i$ and $\boldsymbol{\Phi}_i$). Namely, $\boldsymbol{\mu}_i = \mathbf{F}^T\boldsymbol{\gamma}_i$ and $\boldsymbol{\Sigma}_i = \mathbf{F}^T\boldsymbol{\Phi}_i\mathbf{F}$. Each optimization step has complexity $O(m^3c^2 + cmn^2)$, where c is the number of classes, m is the number of filters, and n is the dimensionality of the data (Appendix E). The expensive computations involve only low-dimensional statistics, making optimization efficient (training takes seconds on a consumer laptop).

Regularization and invariance. To prevent rank-deficient or ill-conditioned covariances, a regularization term was added as $\boldsymbol{\Sigma}_i = \mathbf{F}^T\boldsymbol{\Phi}_i\mathbf{F} + \mathbf{I}_m\sigma^2$, where $\mathbf{I}_m$ is the identity matrix, and σ^2 is the regularization parameter. For the digit datasets, σ^2 was selected via grid-search, using 15% of the training set for validation. For the speed-estimation dataset, we set $\sigma^2 = 0.001$ to match the biologically-informed value of the original work (Burge & Geisler, 2015).

Fisher-Rao distances are invariant to invertible linear transformations of the data, implying that learned filters are unique only up to the subspace they span. When regularization is used, however, the invariance no longer holds, and obtained filters are more consistent (see Appendix G). To obtain filters rank-ordered by usefulness, the filters can be learned sequentially in pairs: two filters are learned first, then this pair can be fixed and two more can be learned, and so on. The resulting filter pairs are ordered by how well they separate the classes.

Evaluation. To evaluate the models, we use the accuracy of the QDA classifier trained on the low-dimensional features, which is the optimal Bayes classifier under Gaussian assumptions (see Appendix H for a discussion of the effect of non-Gaussianity on the results). We compared SQFA to other linear dimensionality reduction techniques, focusing on supervised methods. We tested the canonical methods PCA and LDA, as well as the popular method Supervised PCA (SPCA) (Barshan et al., 2011). Because no efficient implementation of WDA was found, we used the similar method Large Margin Nearest Neighbors (LMNN) (Weinberger & Saul, 2009), that performed similar to WDA in the original work (Flamary et al., 2018) and is implemented in Python (Vazelhes et al., 2020)[2]. We learned filters by maximizing the Bhattacharyya and Hellinger distances, using the same optimization procedure as for SQFA (denoted SQFA-B and SQFA-H, respectively). We used scikit-learn implementations of PCA, LDA, and the QDA classifier (Pedregosa et al., 2011). For LDA, covariance shrinkage was used to improve performance.

Code. A Python package with an efficient implementation of SQFA and its variants will be made available.

[2]To train LMNN we used the first 100 principal components of the data and 10% of the samples, because of the high computational cost. Results for the other methods were similar when using the same reduced dataset.

5 Results

5.1 Toy problem: SQFA vs. LDA vs. PCA

First, we illustrate the differences between SQFA and the canonical dimensionality reduction methods, LDA and PCA. For this, we designed a toy dataset with a six-dimensional variable $\mathbf{x}$ and three classes (Figure 2A). The six-dimensional space contains three different 2D subspaces, each represented by a panel in Figure 2A. The statistics of the dataset are built such that each subspace is preferred by one of the three techniques, i.e. SQFA, LDA, or PCA. We trained the three models on the same synthetic six-dimensional dataset, learning two filters with each. The filters $\mathbf{F}$ learned by each model are shown as arrows in the data space.

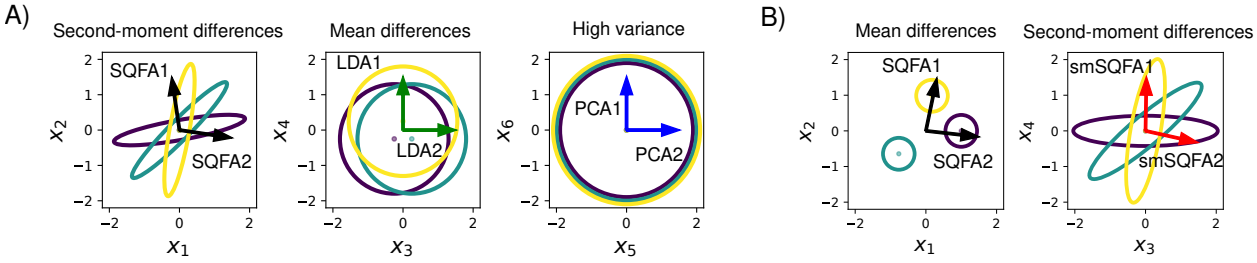

Figure 2: **A)** SQFA vs. LDA vs. PCA. Each of the three panels depicts two dimensions of a 6D data space. Ellipses show the conditional probability distributions of the data vector $\mathbf{x}$ for three classes (colors) in a 6D toy dataset. The classes are separated by different statistical properties. Classes are distinguished by large differences in the covariances (dimensions 1-2), small differences in the means (dimensions 3-4), or neither (dimensions 5-6). Two filters were learned with each of SQFA, LDA, and PCA. The learned filters are shown as arrows in the data space, indicating the axis of the data selected by each filter. SQFA prefers the most discriminative subspace. **B)** SQFA vs. smSQFA. Each of the two panels depicts two dimensions of a 6D data space. Ellipses show the conditional probability distributions of the data vector $\mathbf{x}$ for three classes (colors) in a 4D toy dataset. Classes are distinguished by large differences in the means (dimensions 1-2), and by large differences in the covariances (dimensions 3-4). We learned two filters with SQFA and smSQFA, shown as arrows in the data space. The SQFA filters select for the most discriminative subspace (dimensions 1-2).

Dimensions 1-2 (Figure 2A, left) have no differences between the class means, but have very different–and hence highly discriminative–class-specific covariances. This subspace is selected for by SQFA (black arrows), because it is where the classes are most distant in terms of Fisher-Rao distance. Dimensions 3-4 (Figure 2A, center) contain slight differences between the class means, but these are not very discriminative because the differences are small relative to the covariances. This subspace is selected by LDA (green arrows) because it is the only subspace containing differences in the means. Dimensions 5-6 (Figure 2A, right) contain large covariances, but the class-specific means and covariances are identical. Because this subspace contains the largest overall variances, it is favored by PCA.

Filters that are learned with each of the three methods–SQFA, LDA, and PCA–select for the expected subspace. PCA selects for the subspace with the highest variance, LDA selects for the subspace with the largest differences in the class means, and SQFA selects for the subspace that best supports discrimination.

The previous toy problem shows that SQFA can capture the differences between class-specific covariances that support classification. However, SQFA can also capture differences in the class means. Importantly, it can flexibly select one or the other depending on which one is most informative. To illustrate this, we designed a second toy dataset with three classes and a four-dimensional variable $\mathbf{x}$, composed of two 2D subspaces (like the previous example).

Dimensions 1-2 (Figure 2A) contain large differences in the class-specific means, supporting strong discrimination. Dimensions 3-4 (Figure 2B) have different class-specific covariances but identical means, supporting weaker discrimination than dimensions 1-2. We designed the classes such that their second-moment matrices

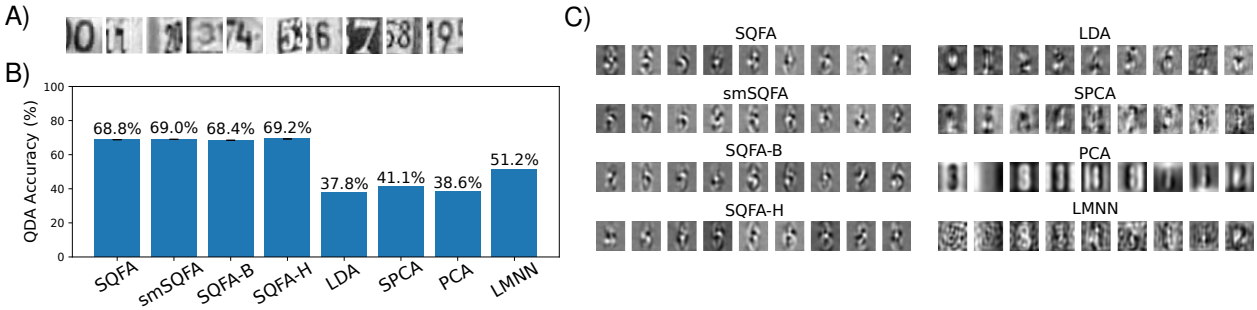

Figure 3: SQFA extracts useful features using class-conditional second-order statistics. **A)** Example images from SVHN. **B)** QDA accuracy using the features learned by the different methods. For SQFA variants, the median and interquartile range of 20 different initializations are shown. **C)** Filters learned by the methods.

(i.e. $\mathbf{\Psi}_i$) are more different for dimensions 3-4 than for dimensions 1-2. We trained both SQFA and smSQFA on the full four-dimensional dataset, and show the learned filters as arrows in the data space.

Because SQFA can select for either first-order or second-order class differences, its filters select for dimensions 1-2, where the classes are more separated (black arrows). On the other hand, smSQFA is only sensitive to class-specific second-moment matrices, so its filters select for dimensions 3-4. This shows that SQFA can flexibly select differences in either the means or the covariances, and that smSQFA should be used when class-specific means are not informative.

5.2 SQFA for digit classification with poor first-order information

To examine how SQFA performs on high-dimensional real-world data, we compared it to the other techniques using the Street View House Numbers (SVHN) dataset, composed of 1024-dimensional images. For each of the eight methods, we learned nine filters and computed QDA classification accuracy in the low-dimensional feature space. (Figure 3). (LDA can learn only $c-1$ filters where c is the number of classes; hence, nine filters were learned.)

All versions of SQFA perform well, with higher performance than the remaining methods. Different versions of SQFA also perform similar, with SQFA-H slightly better, and SQFA-B slightly worse than SQFA and smSQFA (Choi & Lee, 2003; Duin & Loog, 2004). SQFA and smSQFA were also among the fastest to learn the features (Appendix E). The many sources of variation in the images (e.g. digits with mixed contrast polarity, variation in background intensity), make the first-order differences between classes a poor signal for discriminating between the digits. Second-order differences support much better discrimination performance.

5.3 SQFA for digit classification with useful first-order information

To examine how SQFA performs when both first- and second-order statistics are informative, we compared the same eight methods using the MNIST dataset (Figure 4). Because MNIST digits are white digits on a black background, the class-conditional means are quite different, making them useful for discrimination. Nine filters were learned with each method.

SQFA-H had the highest performance, followed by SQFA, SQFA-B and LMNN, which performed similarly. Interestingly, SQFA outperformed both smSQFA and LDA. smSQFA uses the within-class second-moments but not the means; LDA uses the within-class means but not the second moments. The pattern of results indicates that both first- and second-order statistics carry useful information for the task, and that SQFA exploits these two information sources to learn discriminative features.

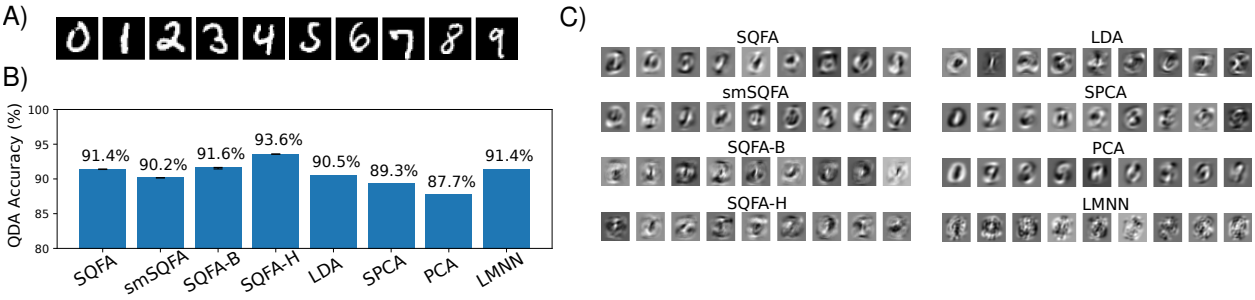

Figure 4: SQFA can exploit class-conditional first- and second-order information. **A)** Example MNIST images. **B)** QDA accuracy using the features learned by the different methods. For SQFA variants, the median and interquartile range of 20 initializations are shown. **C)** Filters learned by the methods.

5.4 Naturalistic speed-estimation task

Next, we examine how SQFA features compare to optimal features for quadratic discrimination. To answer this question, we tested SQFA with a dataset used to investigate the neural computations underlying speed-estimation (Burge & Geisler, 2015; Chin & Burge, 2020; Herrera-Esposito & Burge, 2024). This video-based dataset is interesting for several reasons. First, finding features (receptive fields) that are useful for solving visual tasks is essential for systems neuroscience and perception science (Burge, 2020), and is a potential application for SQFA. Second, the within-class means approximately equal zero, meaning that the exact Fisher-Rao distance computed by smSQFA is accurate. Third, SQFA results can be directly compared to results obtained with a method called AMA-Gauss, which directly maximizes the performance of a Bayesian quadratic decoder (Jaini & Burge, 2017), providing a principled benchmark for comparison (see Appendix F).

Each video consists of 30 horizontal pixels and 15 frames. The vertical dimension was averaged out. Hence, the videos can be represented as 2D space-time plots (Figure 5A). Each video shows a naturally textured surface moving with one of 41 different speeds (i.e. classes). SQFA filters were learned in sequential pairs to aid interpretability (see Section 4). We learned 8 filters (4 pairs) with each method, following the original work (Burge & Geisler, 2015).

All different versions of SQFA, as well as AMA-Gauss learn filters that are similar to typical motion-sensitive receptive fields in visual cortex, selecting for a range of spatio-temporal frequencies (Movshon et al., 1978; Rust et al., 2005; Priebe et al., 2006). The filters learned by the other methods, in contrast, either lack clear motion sensitivity, or do not cover a range of spatio-temporal frequencies (Figure 5C). As expected, AMA-Gauss performed best, since it directly optimizes the quadratic decodability of speed. Notably, SQFA-H performance was very close, followed closely by SQFA and smSQFA. A performance gap was expected for two reasons. First, the dissimilarity measures are not perfect measures of discriminability. Second, maximizing pairwise discriminabilities, as in Equation 1, is not guaranteed to maximize multiclass discriminability (Loog et al., 2001; Thangavelu & Raich, 2008). AMA-Gauss, on the other hand, directly optimizes multiclass quadratic decodability. The fact that SQFA and SQFA-H perform so close to AMA-Gauss is a noteworthy result, highlighting the utility of Fisher-Rao and Hellinger distances as objectives for learning features.

6 Discussion

We have introduced SQFA, a supervised dimensionality reduction method that maximizes the Fisher-Rao distances between class-conditional distributions under Gaussian assumptions. SQFA is a computationally efficient method that learns features supporting excellent quadratic discriminability, i.e. QDA accuracy. The same optimization procedure can be used to define variants of SQFA, that maximize other dissimilarity measures from statistics and information theory.

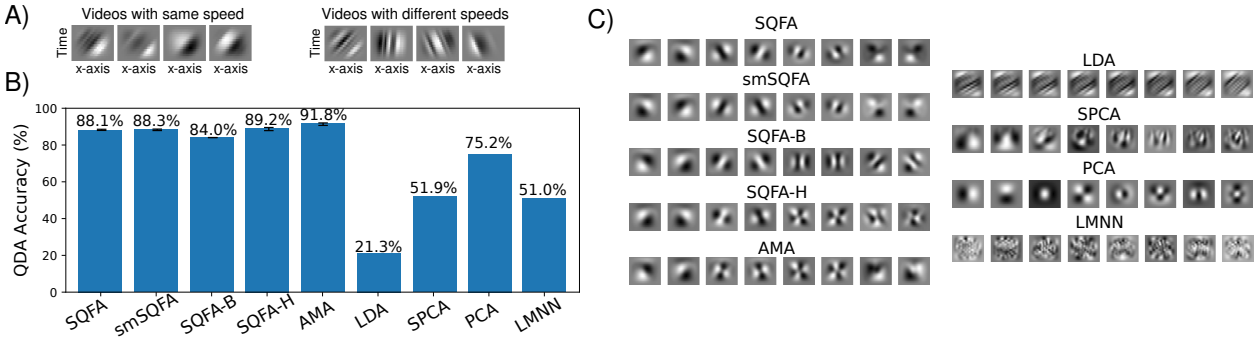

Figure 5: SQFA features are close to optimal for quadratic discrimination. **A)** 4 example videos with the same speed (left) and 4 example videos with different speeds (right). Each video is shown as a 2D space-time plot where the vertical axis is time and the horizontal axis is space. **B)** Performance of a QDA classifier using the filters learned by the different methods. For SQFA variants, the median and interquartile range of 20 different initializations are shown. **C)** Filters learned by the methods (each image shows a 2D space-time plot).

The QDA classification results exhibit noteworthy patterns. First, maximizing the Fisher-Rao distance, which is derived in a geometric framework, leads to features with similar performance to dissimilarity measures that are directly linked to classification error: the Bhattacharyya and Hellinger distances. While the three measures are all linked to local discriminability, and are closely related at an infinitesimal scale (d_H and $\sqrt{d_B}$ converge to d_{FR}), they differ substantially at the larger distances at which SQFA operates. This shows that, besides the geometric properties that make the Fisher-Rao distance useful in other applications (Arvanitidis et al., 2022; Miyamoto et al., 2024), the distance is a good proxy for discriminability.

A second interesting pattern is that, despite d_H and d_B being equivalent objectives for the two-class case (i.e. they are monotonic transformations of each other), SQFA-H consistently outperformed SQFA-B in the multiclass problems tested here. It is known that pairwise discriminabilities do not necessarily translate to multiclass discriminability (Loog et al., 2001; Tao et al., 2007; Thangavelu & Raich, 2008). Two dissimilarity measures that scale differently will assign different relative weights of pairwise discriminabilities in Equation 1. For example, d_H is bounded between 0 and 1, so classes that are already well separated will not contribute much to the loss gradient, as opposed to the unbounded d_B, where the loss gradient can be dominated by a few well-separated classes. To our knowledge, this advantage of d_H over d_B has not been reported before, and it should be of interest to practitioners, since d_B is a popular choice for dimensionality reduction. Notably, the Fisher-Rao distance is unbounded, but it often weights well-separated classes less than d_B (Appendix C), making it an interesting compromise to explore.

A novel methodological aspect of this work is the use of the Calvo-Oller bound as a surrogate for the Fisher-Rao distance in learning (Calvo & Oller, 1990; Nielsen, 2023). The fact that closed-form expressions for the Fisher-Rao distance between multivariate Gaussians are unavailable has limited its use in practice. Here we show that the Calvo-Oller bound is close to the true distance for the real-world datasets tested. Importantly, with minor modifications, the Calvo-Oller bound can be used for other elliptical distributions, such as multivariate Student-t and Cauchy distributions (Calvo & Oller, 2002; Nielsen, 2023), which should allow SQFA to be extended to the non-Gaussian case.

Information geometry is a promising tool for studying neural representations in neuroscience and machine learning (Kriegeskorte & Wei, 2021; Wang & Ponce, 2021; Arvanitidis et al., 2022; Duong et al., 2023; Feather et al., 2024). Finding features that maximize Fisher-Rao distances between classes, or conditions, is a potentially useful tool in this context. In particular, our results should be relevant to research in neuroscience and psychology that proposes a relation between geodesic distances in perceptual space and suprathreshold perception (Fechner, 1860; Bujack et al., 2022; Zhou et al., 2024; Vacher & Mamassian, 2024; Hong et al., 2025), since SQFA shows that maximizing such distances can be a useful learning objective. One concrete

example of a potential application of SQFA in neuroscience is to analyze the magnitude and relevance of the condition-dependent covariability of neural responses, which is a topic of much debate, and for which the Fisher information has been used extensively (Moreno-Bote et al., 2014; Kohn et al., 2016; Ding et al., 2023). SQFA can find the modes of neural activity where different conditions differ most in their first- and second-order neural responses, potentially providing insights into the relevance of condition-dependent covariability. SQFA can also leverage modern tools from neuroscience that estimate the response statistics in conditions where raw data itself is not available (Nejatbakhsh et al., 2023; Ding et al., 2023; Maheswaranathan et al., 2023).

SQFA is a first step towards using information geometry for dimensionality reduction. Several questions and extensions remain open. For example, extending SQFA to non-Gaussian elliptic distributions, or even to the non-parametric case, using the non-parametric Fisher-Rao distance (Srivastava et al., 2007) are promising directions. Future work should also study under what circumstances maximizing the Fisher-Rao distance might be preferable to maximizing other measures of dissimilarity, such as the Hellinger distance. Finally, it should be straightforward to extend this framework to nonlinear feature learning, which should be explored in future work.

References

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

Georgios Arvanitidis, Miguel González-Duque, Alison Pouplin, Dimitris Kalatzis, and Søren Hauberg. Pulling back information geometry. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, Valencia, Spain, April 2022. PMLR. doi: 10.48550/arXiv.2106.05367. URL http://arxiv.org/abs/2106.05367. arXiv:2106.05367 [cs].

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.

Elnaz Barshan, Ali Ghodsi, Zohreh Azimifar, and Mansoor Zolghadri Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7):1357–1371, July 2011. ISSN 0031-3203. doi: 10.1016/j.patcog.2010.12.015. URL https://www.sciencedirect.com/science/article/pii/S0031320310005819.

Roxana Bujack, Emily Teti, Jonah Miller, Elektra Caffrey, and Terece L. Turton. The non-Riemannian nature of perceptual color space. *Proceedings of the National Academy of Sciences*, 119(18):e2119753119, May 2022. doi: 10.1073/pnas.2119753119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2119753119. Publisher: Proceedings of the National Academy of Sciences.

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.

Miquel Calvo and Josep M. Oller. A distance between multivariate normal distributions based in an embedding into the siegel group. *Journal of Multivariate Analysis*, 35(2):223–242, November 1990. ISSN 0047-259X. doi: 10.1016/0047-259X(90)90026-E. URL https://www.sciencedirect.com/science/article/pii/0047259X9090026E.

Miquel Calvo and Josep M. Oller. A distance between elliptical distributions based in an embedding into the Siegel group. *Journal of Computational and Applied Mathematics*, 145(2):319–334, August 2002. ISSN 0377-0427. doi: 10.1016/S0377-0427(01)00584-2. URL https://www.sciencedirect.com/science/article/pii/S0377042701005842.

Kevin M. Carter, Raviv Raich, and Alfred O. Hero. An information geometric approach to supervised dimensionality reduction. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1829–1832, Taipei, Taiwan, April 2009. IEEE. ISBN 978-1-4244-2353-8. doi: 10.1109/ICASSP.2009.4959962. URL http://ieeexplore.ieee.org/document/4959962/.

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.

Euisun Choi and Chulhee Lee. Feature extraction based on the Bhattacharyya distance. *Pattern Recognition*, 36(8):1703–1709, August 2003. ISSN 0031-3203. doi: 10.1016/S0031-3203(03)00035-9. URL https://www.sciencedirect.com/science/article/pii/S0031320303000359.

SueYeon Chung and L. F. Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology*, 70:137–144, October 2021. ISSN 0959-4388. doi: 10.1016/j.conb.2021.10.010. URL https://www.sciencedirect.com/science/article/pii/S0959438821001227.

John P Cunningham and Zoubin Ghahramani. Linear Dimensionality Reduction: Survey, Insights, and Generalizations. *Journal of Machine Learning Research*, 16:2859–2900, 2015.

Peter Dayan and Laurence F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, August 2005. ISBN 978-0-262-54185-5. Google-Books-ID: fLT4DwAAQBAJ.

Xuehao Ding, Dongsoo Lee, Joshua Melander, George Sivulka, Surya Ganguli, and Stephen Baccus. Information Geometry of the Retinal Representation Manifold. *Advances in Neural Information Processing Systems*, 36:44310–44322, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8a267516a7a697965c6ae4f48b908605-Abstract-Conference.html.

R.P.W. Duin and M. Loog. Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):732–739, June 2004. ISSN 1939-3539. doi: 10.1109/TPAMI.2004.13. URL https://ieeexplore.ieee.org/document/1288523/?arnumber=1288523. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Lyndon R. Duong, Jingyang Zhou, Josue Nassar, Jules Berman, Jeroen Olieslagers, and Alex H. Williams. Representational dissimilarity metric spaces for stochastic neural networks, February 2023. URL http://arxiv.org/abs/2211.11665. arXiv:2211.11665 [cs, q-bio].

Anmol Dwivedi, Sihui Wang, and Ali Tajer. Discriminant Analysis under f-Divergence Measures. *Entropy*, 24(2):188, February 2022. ISSN 1099-4300. doi: 10.3390/e24020188. URL https://www.mdpi.com/1099-4300/24/2/188. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Jenelle Feather, David Lipshutz, Sarah E. Harvey, Alex H. Williams, and Eero P. Simoncelli. Discriminating image representations with principal distortions, October 2024. URL http://arxiv.org/abs/2410.15433. arXiv:2410.15433 [q-bio].

Gustav Theodor Fechner. *Elemente der psychophysik*, volume 2. Breitkopf u. Härtel, 1860.

Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein Discriminant Analysis. *Machine Learning*, 107(12):1923–1945, December 2018. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-018-5717-1. URL http://arxiv.org/abs/1608.08063. arXiv:1608.08063 [cs, stat].

Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, second edition, 1990. ISBN 978-0-08-047865-4. Google-Books-ID: BIJZTGjTxBgC.

Daniel Herrera-Esposito and Johannes Burge. Optimal estimation of local motion-in-depth with naturalistic stimuli. *Journal of Neuroscience*, November 2024. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0490-24.2024. URL https://www.jneurosci.org/content/early/2024/11/25/JNEUROSCI.0490-24.2024. Publisher: Society for Neuroscience Section: Research Articles.

Fangfang Hong, Ruby Bouhassira, Jason Chow, Craig Sanders, Michael Shvartsman, Phillip Guan, Alex H. Williams, and David H. Brainard. Comprehensive characterization of human color discrimination thresholds, August 2025. URL https://www.biorxiv.org/content/10.1101/2025.07.16.665219v2. ISSN: 2692-8205 Pages: 2025.07.16.665219 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.

T. Kailath. The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Transactions on Communications*, 15(1):52–60, February 1967. ISSN 0096-2244. doi: 10.1109/TCOM.1967.1089532. URL http://ieeexplore.ieee.org/document/1089532/.

Adam Kohn, Ruben Coen-Cagli, Ingmar Kanitscheider, and Alexandre Pouget. Correlations and Neuronal Population Information. *Annual Review of Neuroscience*, 39(1):237–256, 2016. doi: 10.1146/annurev-neuro-070815-013851. URL https://doi.org/10.1146/annurev-neuro-070815-013851. _eprint: https://doi.org/10.1146/annurev-neuro-070815-013851.

Nikolaus Kriegeskorte and Xue-Xin Wei. Neural tuning and representational geometry. *Nature Reviews Neuroscience*, 22(11):703–718, November 2021. ISSN 1471-0048. doi: 10.1038/s41583-021-00502-3. URL https://www.nature.com/articles/s41583-021-00502-3. Publisher: Nature Publishing Group.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 1476-4687. doi: 10.1038/nature14539. URL https://www.nature.com/articles/nature14539. Publisher: Nature Publishing Group.

Mario Lezcano Casado. Trivializations for Gradient-Based Optimization on Manifolds. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

M. Loog, R.P.W. Duin, and R. Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(7):762–766, July 2001. ISSN 1939-3539. doi: 10.1109/34.935849. URL https://ieeexplore.ieee.org/abstract/document/935849. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Niru Maheswaranathan, Lane T. McIntosh, Hidenori Tanaka, Satchel Grant, David B. Kastner, Joshua B. Melander, Aran Nayebi, Luke E. Brezovec, Julia H. Wang, Surya Ganguli, and Stephen A. Baccus. Interpreting the retinal neural code for natural scenes: From computations to neurons. *Neuron*, 111 (17):2742–2755.e4, September 2023. ISSN 08966273. doi: 10.1016/j.neuron.2023.06.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0896627323004671.

Henrique K. Miyamoto, Fábio C. C. Meneghetti, Julianna Pinele, and Sueli I. R. Costa. On closed-form expressions for the Fisher–Rao distance. *Information Geometry*, 7(2):311–354, November 2024. ISSN 2511-249X. doi: 10.1007/s41884-024-00143-2. URL https://doi.org/10.1007/s41884-024-00143-2.

Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature Neuroscience*, 17(10):1410–1417, October 2014. ISSN 1546-1726. doi: 10.1038/nn.3807. URL https://www.nature.com/articles/nn.3807. Publisher: Nature Publishing Group.

J A Movshon, I D Thompson, and D J Tolhurst. Spatial and temporal contrast sensitivity of neurones in areas 17 and 18 of the cat's visual cortex. *The Journal of Physiology*, 283(1):101–120, 1978. ISSN 1469-7793. doi: 10.1113/jphysiol.1978.sp012490. URL https://onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1978.sp012490.

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.

Frank Nielsen. An Elementary Introduction to Information Geometry. *Entropy*, 22(10):1100, October 2020. ISSN 1099-4300. doi: 10.3390/e22101100. URL https://www.mdpi.com/1099-4300/22/10/1100. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

Frank Nielsen. A Simple Approximation Method for the Fisher–Rao Distance between Multivariate Normal Distributions. *Entropy*, 25(4):654, April 2023. ISSN 1099-4300. doi: 10.3390/e25040654. URL https://www.mdpi.com/1099-4300/25/4/654. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Frank Nielsen and Alexander Soen. pyBregMan: A Python library for Bregman Manifolds, August 2024. URL http://arxiv.org/abs/2408.04175. arXiv:2408.04175 [cs].

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. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Nicholas J. Priebe, Stephen G. Lisberger, and J. Anthony Movshon. Tuning for Spatiotemporal Frequency and Speed in Directionally Selective Neurons of Macaque Striate Cortex. *Journal of Neuroscience*, 26(11):2941–2950, March 2006. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3936-05.2006. URL https://www.jneurosci.org/content/26/11/2941. Publisher: Society for Neuroscience Section: Articles.

Luis Rueda and Myriam Herrera. Linear dimensionality reduction by maximizing the Chernoff distance in the transformed space. *Pattern Recognition*, 41(10):3138–3152, October 2008. ISSN 0031-3203. doi: 10.1016/j.patcog.2008.01.016. URL https://www.sciencedirect.com/science/article/pii/S003132030800037X.

Nicole C. Rust, Odelia Schwartz, J. Anthony Movshon, and Eero P. Simoncelli. Spatiotemporal Elements of Macaque V1 Receptive Fields. *Neuron*, 46(6):945–956, June 2005. ISSN 0896-6273. doi: 10.1016/j.neuron.2005.05.021. URL https://www.sciencedirect.com/science/article/pii/S089662730500468X.

Tim Sainburg, Leland McInnes, and Timothy Q. Gentner. Parametric UMAP Embeddings for Representation and Semisupervised Learning. *Neural Computation*, pp. 1–27, August 2021. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco_a_01434. URL https://direct.mit.edu/neco/article/doi/10.1162/neco_a_01434/107068/Parametric-UMAP-Embeddings-for-Representation-and.

Anuj Srivastava, Ian Jermyn, and Shantanu Joshi. Riemannian Analysis of Probability Density Functions with Applications in Vision. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Minneapolis, MN, USA, June 2007. IEEE. ISBN 978-1-4244-1179-5 978-1-4244-1180-1. doi: 10.1109/CVPR.2007.383188. URL http://ieeexplore.ieee.org/document/4270213/.

Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J. Maybank. General Averaged Divergence Analysis. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pp. 302–311, October 2007. doi: 10.1109/ICDM.2007.105. URL https://ieeexplore.ieee.org/document/4470254/. ISSN: 2374-8486.

Madan Thangavelu and Raviv Raich. Multiclass linear dimension reduction via a generalized Chernoff bound. In *2008 IEEE Workshop on Machine Learning for Signal Processing*, pp. 350–355, October 2008. doi: 10.1109/MLSP.2008.4685505. URL https://ieeexplore.ieee.org/abstract/document/4685505. ISSN: 2378-928X.

Jonathan Vacher and Pascal Mamassian. Perceptual Scales Predicted by Fisher Information Metrics, March 2024. URL http://arxiv.org/abs/2310.11759. arXiv:2310.11759 [q-bio].

William de Vazelhes, C. J. Carey, Yuan Tang, Nathalie Vauquier, and Aurélien Bellet. metric-learn: Metric Learning Algorithms in Python. *Journal of Machine Learning Research*, 21(138):1–6, 2020. ISSN 1533-7928. URL http://jmlr.org/papers/v21/19-678.html.

Binxu Wang and Carlos R. Ponce. The Geometry of Deep Generative Image Models and its Applications, March 2021. URL http://arxiv.org/abs/2101.06006. arXiv:2101.06006 [cs, math].

Kilian Q Weinberger and Lawrence K Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

Zeyuan Ye and Ralf Wessel. Speed modulations in grid cell information geometry, September 2024. URL https://www.biorxiv.org/content/10.1101/2024.09.18.613797v1. Pages: 2024.09.18.613797 Section: New Results.

Jingyang Zhou, Lyndon R. Duong, and Eero P. Simoncelli. A unified framework for perceived magnitude and discriminability of sensory stimuli. *Proceedings of the National Academy of Sciences*, 121(25): e2312293121, June 2024. doi: 10.1073/pnas.2312293121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2312293121. Publisher: Proceedings of the National Academy of Sciences.

A Appendix

B Fisher-Rao distance and discriminability in the zero-mean case

B.1 Generalized eigenvalues reflect quadratic differences between classes

As described in the main text, the Fisher-Rao distance between two zero-mean Gaussian distributions $\mathcal{N}(0, \mathbf{\Psi}_i)$ and $\mathcal{N}(0, \mathbf{\Psi}_j)$ is given (up to a factor of $\sqrt{2}$) by the affine-invariant distance in the manifold of symmetric positive definite matrices, SPD(m)

$$d_{FR}(\mathbf{\Psi}_i, \mathbf{\Psi}_j) = \sqrt{\frac{1}{2}\sum_{k=1}^{m} \log^2(\lambda_k)} \tag{7}$$

where λ_k are the generalized eigenvalues of the pair of matrices $(\mathbf{\Psi}_i, \mathbf{\Psi}_j)$.

The generalized eigenvalues λ_k and generalized eigenvectors $\mathbf{v}_k$ of the pair of matrices $\mathbf{A} \in \text{SPD(m)}$ and $\mathbf{B} \in \text{SPD(m)}$ are the solutions to the generalized eigenvalue problem $\mathbf{A}\mathbf{v}_k = \lambda_k \mathbf{B}\mathbf{v}_k$. The solution to the problem is given by the eigenvalues and eigenvectors of $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$, where $\mathbf{B}^{-1/2}$ is the inverse square root of $\mathbf{B}$. If $\mathbf{A}$ and $\mathbf{B}$ are identical, $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ is the identity matrix, all the eigenvalues are 1, and $d_{FR}(\mathbf{A}, \mathbf{B}) = 0$. The farther the λ_k are from 1, the more different the matrices $\mathbf{A}$ and $\mathbf{B}$ are (i.e. the more different $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ is from the identity matrix).

Consider a random variable $\mathbf{z} \in \mathbb{R}^m$ that belongs to one of two classes i, j, with second moment matrices $\mathbf{\Psi}_i = \mathbb{E}\left[\mathbf{z}\mathbf{z}^T | y = i\right]$ and $\mathbf{\Psi}_j = \mathbb{E}\left[\mathbf{z}\mathbf{z}^T | y = j\right]$. Next, consider a vector $\mathbf{w} \in \mathbb{R}^m$ and the squared projection of $\mathbf{z}$ onto $\mathbf{w}$, $(\mathbf{w}^T\mathbf{z})^2$. The following ratio relates to how different the squared projections are for the two classes, which is a useful proxy for quadratic discriminability:

$$R(\mathbf{w}) = \frac{\mathbb{E}[(\mathbf{w}^T\mathbf{z})^2 | y = i]}{\mathbb{E}[(\mathbf{w}^T\mathbf{z})^2 | y = j]} = \frac{\mathbf{w}^T\mathbf{\Psi}_i\mathbf{w}}{\mathbf{w}^T\mathbf{\Psi}_j\mathbf{w}} \tag{8}$$

The local extrema of the ratio $R(\mathbf{w})$ are obtained at the generalized eigenvectors $\mathbf{w} = \mathbf{v}_k$ of $(\mathbf{\Psi}_i, \mathbf{\Psi}_j)$, where the ratio $R(\mathbf{v}_k) = \lambda_k$ (Fukunaga, 1990).

The more different λ_k is from 1, the more different are the expected squared projections $(\mathbf{v}^T\mathbf{z})^2$ for the two classes. The magnitude of $\log^2 \lambda_k$ indicates how different the ratio in Equation 8 is from 1, in proportional terms. The set of $\mathbf{v}_k$'s spans the space of $\mathbf{z}$, so $d_{FR}(\mathbf{\Psi}_i, \mathbf{\Psi}_j)$ summarizes the quadratic differences between

the classes i, j along all directions in the feature space, thus relating to the quadratic discriminability of the classes.

Of course, the quadratic discriminability between the classes depends on factors other than the ratio of the expected values of the squared projections, and larger differences in this ratio do not strictly indicate higher discriminability. However, empirical studies show that the generalized eigenvalues tend to be a good indicator of quadratic discriminability in real world datasets (**?**).

B.2 Fisher-Rao distance and Bayes error

Here, we explicitly compare the Fisher-Rao distance to the Bayes error for zero-mean Gaussians in the 1D and 2D cases.

Two symmetric positive definite (SPD) matrices $\mathbf{\Psi}_i$ and $\mathbf{\Psi}_j$ can be simultaneously diagonalized by a linear transformation of the data space. Given the affine-invariance of the Fisher-Rao distance and of the Bayes error, we can reduce the analysis of these quantities to the analysis of pairs of Gaussians of the form $\mathcal{N}(0, \mathbf{D})$ and $\mathcal{N}(0, \mathbf{I})$, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{I}$ is the identity matrix. Hence, we can analyze the Fisher-Rao distance and the Bayes error as a function of the diagonal elements of $\mathbf{D}$. The diagonal elements of $\mathbf{D}$ are the generalized eigenvalues of the pair $(\mathbf{D}, \mathbf{I})$. Thus, the Fisher-Rao distance between $\mathcal{N}(0, \mathbf{D})$ and $\mathcal{N}(0, \mathbf{I})$ is given by

$$d_{FR}(\mathbf{D}, \mathbf{I}) = \sqrt{\frac{1}{2}\sum_{k=1}^{m} \log^2(\mathbf{D}_{kk})} \tag{9}$$

We computed the Bayes error (e_B) for each value of $\mathbf{D}$ by simulating 100,000 samples from each $\mathcal{N}(0, \mathbf{D})$ and $\mathcal{N}(0, \mathbf{I})$ distribution, and obtaining the error rate of the optimal probabilistic classifier[3]. We compute the Bayes accuracy (a_B) as $a_B = 1 - e_B$. Because the accuracy is bounded to be between 0 and 1, we also computed the log-odds of correctly classifying a sample, given by $\log\left(\frac{a_B}{e_B}\right)$. The log-odds is a common way to map probabilities onto the real line.

Zero-mean Gaussian, 1D case. In the 1D case, the matrix $\mathbf{D}$ is a single positive number σ^2. Figure 6 shows the Bayes error, the log-odds of correct classification, and the Fisher-Rao distance as a function of $\log \sigma^2$.

It is easy to see in Equation 9 and in Figure 6 that the Fisher-Rao distance grows linearly with $|\log_{10} \sigma^2|$ in the 1D case. Interestingly, the Bayes accuracy is approximately linear for small values of σ^2, although as expected, it begins to saturate for larger values of σ^2 (since it is bounded by 1). However, the log-odds of a correct classification also looks like a linear function of $|\log_{10} \sigma^2|$.

The above results suggest that the Fisher-Rao distance is a good proxy for discriminability in the zero-mean 1D Gaussian case, in particular, it is a good proxy for the log-odds of a correct classification.

Zero-mean Gaussian, 2D case. Next we analyze the 2D case. Here, there are two parameters, σ_1^2 and σ_2^2. The top row of Figure 7 shows the level sets of accuracy, the log-odds correct, and the Fisher-Rao distance as a function of $\log_{10} \sigma_1^2$ and $\log_{10} \sigma_2^2$. The second row shows the values of the same quantities as a function of $\log_{10} \sigma_1^2$ for fixed values of $\log_{10} \sigma_2^2$, indicated by dashed lines in the top row. The bottom row shows examples of two zero-mean distributions for different values of $\log_{10} \sigma_1^2$ and $\log_{10} \sigma_2^2$.

Interestingly, we see that the level sets of the Fisher-Rao distance have a different shape than the level sets of the accuracy and log-odds correct. The contours of the accuracy and log-odds correct have more circular shapes close to the origin, but acquire a tilted hexagonal shape farther from the origin. As expected from Equation 7, the Fisher-Rao contour sets are circles.

These contour shapes show that the Fisher-Rao distance is not a perfect proxy for discriminability, since it does not capture the interactions between the two parameters $\log_{10} \sigma_1^2$ and $\log_{10} \sigma_2^2$ that lead to the

[3]For the univariate case the Bayes error can be computed in closed-form using the error function. Since the results are the same, we present the simulation results

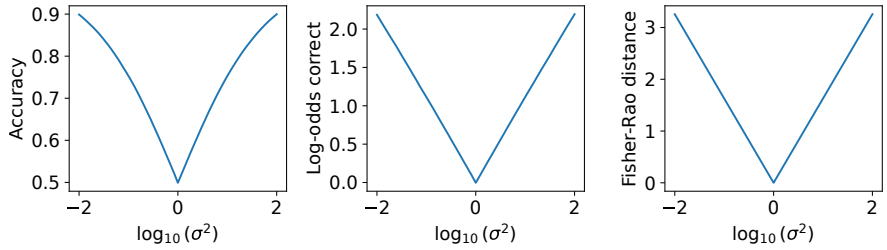

Figure 6: Distances and Bayes error for 1D Gaussian distributions. From left to right, the three panels show, as a function of $\log_{10}(\sigma^2)$, accuracy of the Bayes classifier, the log odds-ratio of correct classification, and the Fisher-Rao distance.

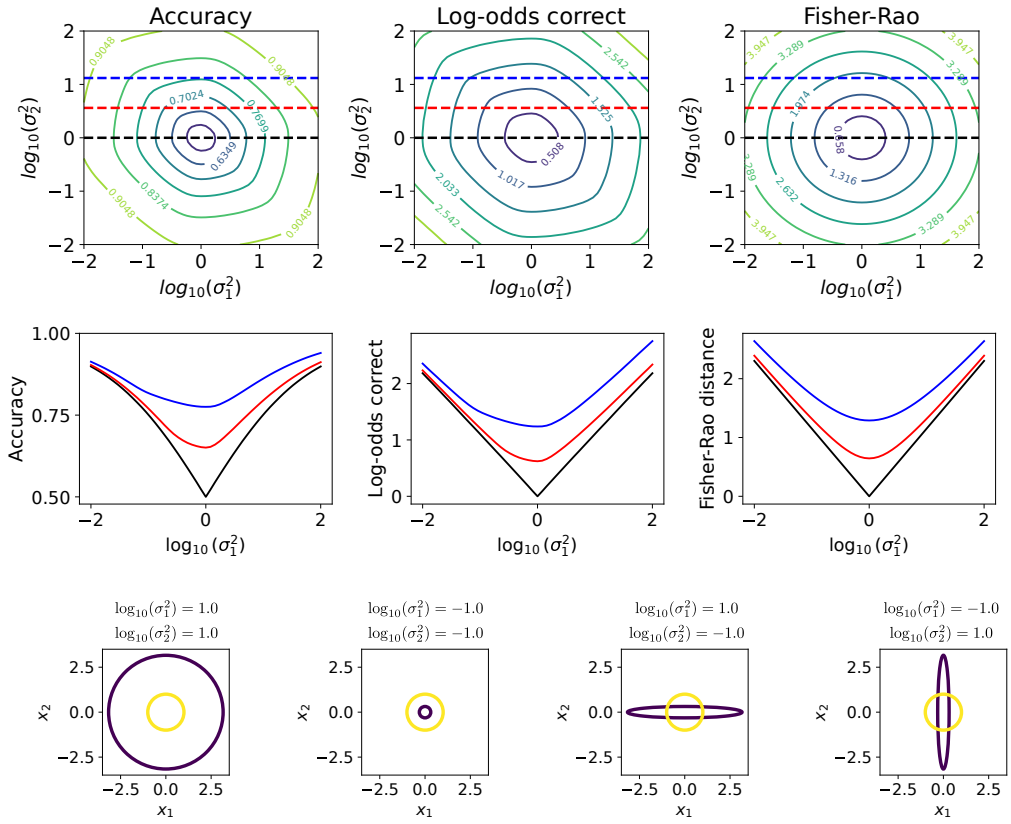

Figure 7: Distances and Bayes error for 2D Gaussian distributions. From left to right, the panels show for two Gaussian classes, the accuracy of the Bayes classifier, the log-odds of a correct classification, and the Fisher-Rao distance. **Top**. Contour plots of the quantities as a function of $\log_{10}(\sigma_1^2)$ and $\log_{10}(\sigma_2^2)$. The values for the contour lines are shown as a color map. The horizontal dashed lines indicate the 1D slices that are shown in the middle row. **Middle**. The same quantities as the top row, shown as a function of $\log_{10}(\sigma_1^2)$, for a fixed values of $\log_{10}(\sigma_2^2)$. The values used are 0.0 (black), 0.56 (red), and 1.12 (blue). **Bottom**. Examples of two zero-mean distributions for different values of $\log_{10}\sigma_1^2$ and $\log_{10}\sigma_2^2$.

distinctive hexagonal shape of the accuracy and log-odds correct plots. The Fisher-Rao distance provides a good approximation of the log-odds correct, however, at smaller values of $\log_{10}\sigma_1^2$ and $\log_{10}\sigma_2^2$.

To visualize this further, we plotted the values of the distances and accuracies for three different slices of the 2D space, with fixed values of $\log_{10}\sigma_2^2$ (0.0, 0.56, 1.12). (Figure 7, middle row). For the first two slices (black and red), the Fisher-Rao distance and the log-odds correct are very similar across the range of values of $\log_{10}\sigma_1^2$, but they become less similar for the third slice (blue), where the asymmetry of the log-odds correct is more pronounced.

In sum, this analysis shows that the Fisher-Rao distance broadly captures the dependence of discriminability on the distribution parameters, particularly in the 1D case, and in the 2D case when the generalized eigenvalues are closer to 1 (i.e. when $\log_{10}\sigma^2$ is close to 0), but that it also fails to capture some patterns of the dependence of discriminability on the parameters.

C Comparing Fisher-Rao to the Calvo-Oller and other distances

C.1 Quality of the Calvo-Oller bound

The tightness of the Calvo-Oller bound has been analyzed for some simple cases (Nielsen, 2023), where it has been shown to be a good approximation of the true Fisher-Rao distance. Two special cases that are of interest are the of distributions with equal mean, and the case of distributions with equal covariance, both of which have closed-form expressions for the Fisher-Rao distance.

In the case of Gaussian distributions with equal mean, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_i)$ and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_j)$, the Calvo-Oller bound is equal to the true Fisher-Rao distance. To see this, we can use the affine-invariant property of the Calvo-Oller bound (Nielsen, 2023) to subtract the common mean from the data without changing the distance. Then, the Calvo-Oller embedding for distribution i becomes $\boldsymbol{\Omega}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix}$, where $\mathbf{0}$ is a vector of zeros. It is then easy to see that $\boldsymbol{\Omega}_i^{-1}\boldsymbol{\Omega}_j = \begin{bmatrix} \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix}$, and that the generalized eigenvalues of $(\boldsymbol{\Omega}_i, \boldsymbol{\Omega}_j)$ are the eigenvalues of $\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j$ plus an additional generalized eigenvalue equal to 1. Because the the generalized eigenvalue equal to 1 does not contribute to the distance (see Equation 4), the Calvo-Oller bound equals the exact Fisher-Rao distance formula available for the zero-mean case (Equation 7).

In the case of Gaussian distributions with equal covariance, there is an exact expression for the Fisher-Rao distance. First, we denote the squared Mahalanobis distance between the distributions $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ and $\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ as

$$d_M(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)^2 = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{10}$$

Then, the Fisher-Rao distance is given by (Nielsen, 2023)

$$d_{FR}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) = \sqrt{2}\operatorname{arccosh}\left(1 + \frac{1}{4} d_M(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)^2\right) \tag{11}$$

We used this formula to compare the Calvo-Oller bound to the true Fisher-Rao distance for a range of $d_M(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ values going from 0 to 20, which is a large range in terms of discriminability. The Calvo-Oller bound is, in general, a good approximation to the Fisher-Rao distance for this range of Mahalanobis distances (Figure 8).

Finally, we examine how tight the bound is for the real-world datasets used in this work. Specifically, we computed the Fisher-Rao distance between the Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Psi}_i)$ and $\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Psi}_j)$ for all pairs of classes i, j in the SVHN, MNIST and motion datasets, and compared the true distance to the Calvo-Oller bound. To compute the true Fisher-Rao distance, we used the numerical method of (Nielsen, 2023) for computing the true Fisher-Rao distance, which is implemented in the Python package 'pyBregMan' (Nielsen & Soen, 2024).

For all datasets, the Calvo-Oller bound was, in general, close to the true Fisher-Rao distance (Figure 7). For the SVHN and the speed estimation datasets in particular, the match was almost perfect. This is not surprising, since for these datasets the means of the classes are close to each other (see the Results section), and the Calvo-Oller bound is exact when the means are equal. For the MNIST dataset, the relation between

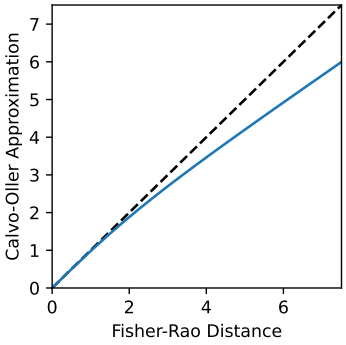

Figure 8: Calvo-Oller bound vs. true Fisher-Rao distance for the equal-covariance case. The Fisher-Rao distance and the Calvo-Oller distance were computed for a range of Mahalanobis distances ranging from 0 to 20. The dashed line shows the identity line.

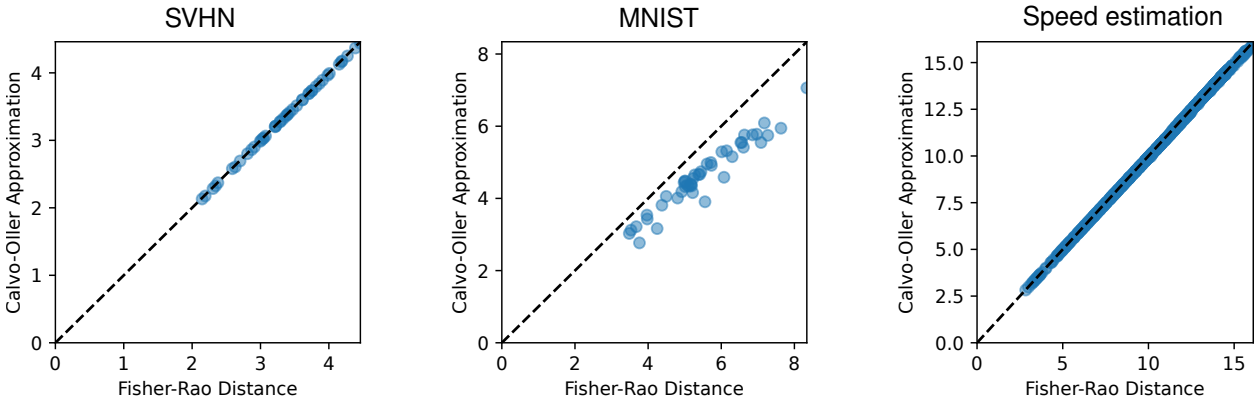

Figure 9: Calvo-Oller bound vs. true Fisher-Rao distance in real-world datasets. Each panel shows the Calvo-Oller bound and the Fisher-Rao distance for the class pairs of a given dataset. From left to right, the SVHN, MNIST and the speed estimation dataset are shown. Each point shows a different pair of classes. The dashed line shows the identity line.

the two distances tends to follow a similar pattern to the equal-covariance case above (Figure 8), with some relatively small deviations. Overall, these results show that the Calvo-Oller bound is a good approximation to the true Fisher-Rao distance in the real-world datasets used in this work.

C.2 Fisher-Rao distance vs other distances

In this section, we compare how the different distances (i.e. Fisher-Rao, Hellinger, Bhattacharyya) behave as a function of differences in the parameters of two Gaussian distributions.

First, we consider the case of two Gaussians with equal covariance $\mathbf{\Sigma}_i = \mathbf{\Sigma}_j = \mathbf{I}$, and different means. All three distances have closed form expressions as a function of the Mahalanobis distance. Some differences between the distances can be observed (Figure 10). The Hellinger distance is bounded between 0 and 1, whereas the Bhattacharyya and Fisher-Rao distances are unbounded. However, the growth rates of the Fisher-Rao distance and the Bhattacharyya distance are different. As the Mahalanobis distance grows, the Bhattacharyya distance grows faster. On the other hand, as the Mahalanobis distance grows, the Fisher-Rao distance grows slower. In the context of supervised dimensionality reduction, this means that the Bhattacharyya distance

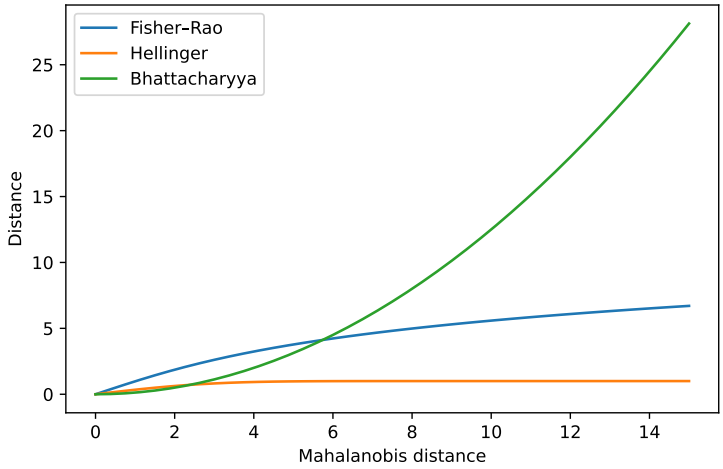

Figure 10: Distances as a function of Mahalanobis distance. The three distances (Fisher-Rao, Hellinger, Bhattacharyya) are shown as a function of the Mahalanobis distance between two Gaussian distributions $\mathcal{N}(0, \mathbf{I})$ and $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$.

will prefer to further separate already well-separated classes, while the Fisher-Rao distance will prefer to separate less separated classes. The implication of this is that in some problems where some classes are already well-separated, the Fisher-Rao distance might lead to better overall discriminability, by focusing on the less separated classes.

Next, we consider the case of two Gaussians with zero mean and different covariances. Like in Section B.2, we examine 2D distributions with diagonal covariance matrices, that is, $\boldsymbol{\Sigma}_i = \mathbf{D} = \text{diag}(\sigma_1^2, \sigma_2^2)$ and $\boldsymbol{\Sigma}_j = \mathbf{I}$ (note that the results for the Fisher-Rao distance are the same as in Figure 7).

Again, the three distances behave differently (Figure 11). The Hellinger distance is bounded between 0 and 1, while the Bhattacharyya and Fisher-Rao distances are unbounded. One interesting pattern is that whereas the contours of the Fisher-Rao distance are circles (as expected from its formula), the contours of the Hellinger and Bhattacharyya distances are more circular closer to the origin, but grow more square farther from the origin.

Other interesting patterns can be observed in the curves at the slices of the 2D space (indicated by colored dashed lines in the first row of Figure 11). For the Fisher-Rao distance, the curve at the slice through the origin $\log_{10}(\sigma_2^2) = 0$ (black line) is shaped like the absolute value function. But the slices at different values of $\log_{10}(\sigma_2^2)$ have different shapes, becoming shallower at the origin. On the other hand, for the Bhattacharyya distance the slices are just shifted versions of each other. This observation can be explained by the following alternaive expression for the Bhattacharyya distance in the zero-mean Gaussian case

$$d_B(\boldsymbol{\Psi}_i, \boldsymbol{\Psi}_j) = \sum_{k=1}^{m} \left[\log\left(\frac{1+\lambda_k}{2}\right) - \frac{1}{2}\log\lambda_k \right] \tag{12}$$

where λ_k are the generalized eigenvalues of $(\boldsymbol{\Psi}_i, \boldsymbol{\Psi}_j)$, and in this case, $\lambda_k = \sigma_k^2$. It is easy to see from this formula that the contribution of each σ_k^2 to the distance is independent of the value of the other $\sigma_{k'}^2$'s, explaining why the slices are just shifted versions of each other. It is interesting to note that in Figure 7, the curves at the different slices for both the accuracy and the log-odds correct show an interaction between the two parameters, closer to the behavior of the Fisher-Rao distance.

These results illustrate some of the differences between the Fisher-Rao distance and other distances. Particularly, in the multi-class case, where the sum of the pairwise distances is not necessarily a good proxy for multi-class discriminability, it is important to consider how different distance functions weight the different

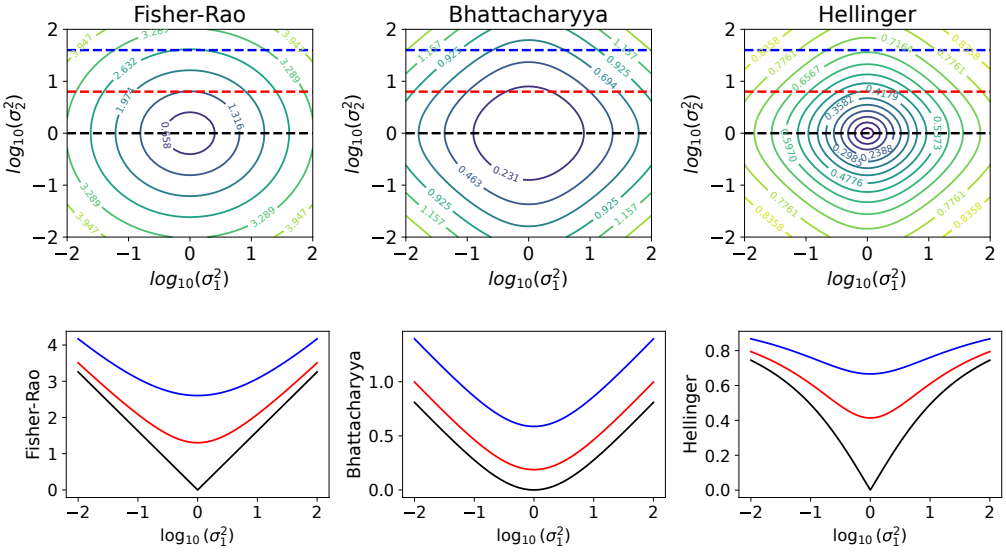

Figure 11: Distances as a function of covariance differences. The three distances (Fisher-Rao, Hellinger, Bhattacharyya) are shown as a function of $\log_{10}(\sigma_1^2)$ and $\log_{10}(\sigma_2^2)$, where the two Gaussian distributions are $\mathcal{N}(0, \mathbf{D})$ and $\mathcal{N}(0, \mathbf{I})$. The top row shows contour plots of the distances, while the bottom row shows 1D slices of the distances. The values used for the slices are shown as dashed lines in the top row.

pairwise distances. The Fisher-Rao distance seems to have some desirable properties in this regard. Ultimately, it might be an empirical question what distance is best suited for a particular application.

D LDA features maximize squared Mahalanobis distances

In this section, we prove that Linear Discriminant Analysis (LDA) maximizes the squared Mahalanobis distances between classes when the classes are homoscedastic Gaussians. Because the Mahalanobis distance equals the Fisher-Rao distance along the submanifold of Gaussians with equal covariance[4], this also means that LDA maximizes the pairwise Fisher-Rao (squared) distances between classes, and that it is thus closely related to SQFA.

For a labeled random variable $\mathbf{x} \in \mathbb{R}^n$ with class labels $y \in \{1, \ldots, c\}$, the goal of LDA is to find the filters $\mathbf{F} \in \mathbb{R}^{n \times m}$ such that the variable $\mathbf{z} = \mathbf{F}^T \mathbf{x}$ maximizes the between-class scatter relative to the within-class scatter. This is typically formulated as maximizing the Fisher criterion

$$\operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1} \mathbf{S}_{\mathbf{z}}\right) \tag{13}$$

where $\boldsymbol{\Sigma}$ is the residual within-class covariance matrix of the data (i.e. the covariance matrix of the data after subtracting the class mean from each data point), and $\mathbf{S}_{\mathbf{z}}$ is the between-class scatter matrix of $\mathbf{z}$, defined as

$$\mathbf{S}_{\mathbf{z}} = \frac{1}{c} \sum_{i=1}^{c} (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$

where $\boldsymbol{\mu}_i$ is the mean of class i in the feature space and $\boldsymbol{\mu} = \sum_{i=1}^{c} \boldsymbol{\mu}_i$.

Denoting the squared Mahalanobis distance as in Equation 10, the sum of pairwise squared Mahalanobis distances between the classes is given by

$$\frac{1}{2} \sum_{i=1}^{c} \sum_{j=1}^{c} d_M(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)^2 = \frac{1}{2} \sum_{i=1}^{c} \sum_{j=1}^{c} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{14}$$

[4]This is not the same as the Fisher-Rao distance along the general manifold of Gaussian distributions.

where the factor of 1/2 is included to control for double-counting. Thus, we must prove that the objectives in Equation 13 and Equation 14 are equivalent.

First, we assume without loss of generality that the overall mean $\boldsymbol{\mu} = \sum_{i=1}^{c} \boldsymbol{\mu}_i = 0$. (Note that Equation 14 is invariant to translations.) Then, we have

$$\begin{aligned}
\frac{1}{2}\sum_{i=1}^{c}\sum_{j=1}^{c} d_M(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)^2 &= \frac{1}{2}\sum_{i=1}^{c}\sum_{j=1}^{c}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \\
&= \frac{1}{2}\sum_{i=1}^{c}\sum_{j=1}^{c}\left(\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j - 2\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j\right) \\
&= \sum_{i=1}^{c}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i - \sum_{i=1}^{c}\sum_{j=1}^{c}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j \\
&= \sum_{i=1}^{c}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i - \left(\sum_{i=1}^{c}\boldsymbol{\mu}_i^T\right)\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{c}\boldsymbol{\mu}_j\right) \\
&= \sum_{i=1}^{c}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \\
&= \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{c}\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T\right) \\
&= c\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}_\mathbf{z}\right)
\end{aligned}$$

In the first five lines we expanded the squared Mahalanobis distance, used the linearity of the dot product and the fact that $\boldsymbol{\mu} = 0$. In the last two lines we used the linearity and the cyclic property of the trace, and that $\sum_{i=1}^{c}\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T = c\mathbf{S}_\mathbf{z}$ because $\boldsymbol{\mu} = 0$. This shows that the maximizing the pairwise squared Mahalanobis distances between classes is equivalent to maximizing the Fisher criterion in Equation 13.

E Complexity analysis and training times

E.1 Computational complexity of SQFA

Here we analyze the computational complexity of SQFA. Specifically, we analyze the cost of taking a gradient step using the SQFA objective. For simplicity, we analyze the case of smSQFA (assuming zero means), but this is equivalent to the case of SQFA, since the operations are the same except substituting the $m \times m$ matrices with $(m+1) \times (m+1)$ matrices.

The first step is to compute the class covariances in the feature space. As mentioned in the Methods section, we achieve this by transforming the covariances of the raw data using the formula $\boldsymbol{\Sigma}_i = \mathbf{F}^T\boldsymbol{\Phi}_i\mathbf{F}$, where $\boldsymbol{\Phi}_i$ is the covariance of the data for class i. The matrix $\boldsymbol{\Phi}_i$ is $n \times n$, and the matrix $\mathbf{F}$ is $n \times m$, where n is the dimensionality of the data and m the number of filters. Thus, the cost of computing $\boldsymbol{\Sigma}_i$ for all c classes is $O(cmn^2)$.

Next, we compute the generalized eigenvalues of the pair $(\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j)$, where the matrices are $m \times m$. For this, the cost of computing $\boldsymbol{\Sigma}_i^{-1}$ is $O(m^3)$. Then, we compute the product matrix $\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j$, which also costs $O(m^3)$. Finally, the cost of computing the eigenvalues of $\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j$ is also $O(m^3)$. Performing this for all pairs of classes has a complexity of $O(c^2m^3)$.

The cost of computing the gradient for the three operations described above is the same as the cost of computing the operations themselves. Thus, the overall cost of taking a gradient step in SQFA is $O(c^2m^3 + cmn^2)$.

It is important to note that the most expensive operations occur in the feature space, and the dimensionality of the feature space, m, is typically much smaller than the dimensionality of the data, n.

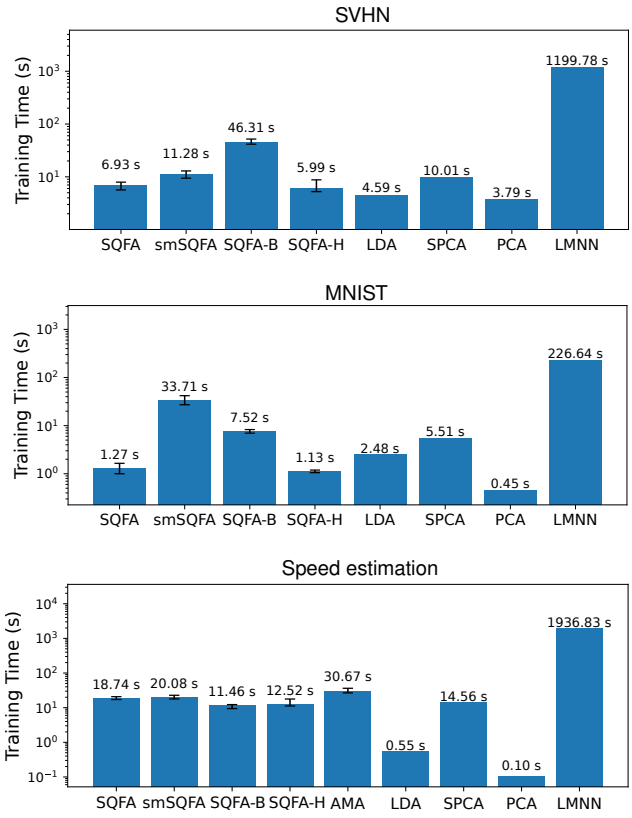

Figure 12: Training times for the different models on the different datasets. Times are in seconds, and indicated by the numbers on top of each bar. The y-axis is logarithmic. **Top**: SVHN. **Center**:MNIST. **Bottom**: Speed estimation dataset. We note that LMNN was trained on reduced datasets compared to the other methods (see Section 4), so its training time would be higher if trained on the full datasets. The bars and the numbers for the SQFA variants indicate the median training time across different seeds, and the error bars indicate the interquartile range.

E.2 Training times

For each dataset, we recorded the time it took to train the models on a consumer laptop with an 12th Gen Intel Core i7-1270P with 32 GB of RAM.

The training times of SQFA and its variants were in the order of seconds for all datasets (Figure 12). Remarkably, the training time for SQFA was comparable to that of LDA and SPCA. We note that SQFA training times were higher for the speed estimation dataset compared to the other datasets. This is partly because the filters were learned sequentially for this task, which is more computationally expensive than learning all filters at once.

F Details of the motion estimation task

F.1 Dataset synthesis

The motion estimation dataset consists of synthetic naturalistic videos of surfaces moving at different frontoparallel speeds, that were synthesized following the procedure described by Burge & Geisler (2015).

Briefly, each video is initially 60-by-60 pixels and 15 frames long. In this synthesis procedure, 60 pixels correspond to 1 degree of visual field (roughly equivalent to the foveal sampling of photoreceptors) and the videos have a duration of 250 ms (roughly equivalent to the duration of a fixation in natural viewing).

A video is synthesized by taking a random patch from an image of a natural scene and moving it horizontally at a given speed behind a 60-by-60 pixel aperture, for a duration of 15 frames. The resulting video is then filtered spatiotemporally to simulate the response of retinal photoreceptors to the video, following the procedure described by Herrera-Esposito & Burge (2024). Then, the resulting video is downsampled spatially by a factor of 2, to 30-by-30 pixel frames, and each frame is averaged vertically, leading to 30 pixel by 15 frame videos (thus, each video can be represented as a 450 dimensional vector). Vertically averaging the movies is equivalent to only considering filters that are vertically-oriented in the original 2D frame videos.

Then, to simulate further retinal processing, the video is converted to contrast, by subtracting and dividing by its mean intensity across pixels and frames. Denoting the resulting contrast video by $\mathbf{c}$, we apply divisive normalization to the video, dividing it by $\sqrt{(\|\mathbf{c}\|^2 + nc_{50}^2)}$ where $c_{50}^2 = 0.045$ is a constant and $n = 450$ is the number of pixels in the video multiplied by the number of frames. The resulting video simulates the retinal output in response to a naturalistic motion, as discussed in Burge & Geisler (2015); Herrera-Esposito & Burge (2024).

Above, we described the process by which a video is synthesized for a given speed. In the dataset, 41 retinal speeds (i.e. classes) were used, ranging from -6.0 to 6.0 deg/s with 0.3 deg/s intervals. For each of the 41 retinal speeds, we synthesized 800 different naturalistic videos, by using a different randomly sampled patch from a natural scene for each video. This adds nuisance naturalistic variability to the dataset. We used 500 videos per speed for training (a total of 20500 videos) and 300 videos per speed for testing (a total of 12300 videos).

Note that because the videos are generated with small patches randomly sampled from natural images, the expected intensity value at each pixel and frame is approximately the same (because natural image patches are approximately stationary), independent of speed. Then, because the videos were constrained to be contrast videos–i.e. they are formed by subtracting off and dividing by the intensity mean, consistent with early operations in the human visual system (Burge & Geisler, 2015)–the mean across patches equals zero, independent of speed. This results in a dataset where the classes are all approximately zero-mean.

F.2 AMA-Gauss training

The AMA-Gauss model was implemented following the description in Jaini & Burge (2017); Herrera-Esposito & Burge (2024). First, a set of linear filters $\mathbf{F}$ is applied to each pre-processed video, and a sample of independent noise is added to each filter output. This results in a noisy response vector $\mathbf{R} = \mathbf{F}^T\mathbf{c} + \lambda$, where $\lambda \sim \mathcal{N}(0, \mathbf{I}\sigma^2)$ is the added noise. Then, a QDA-like decoder is used to classify the videos based on the noisy response vectors, assuming that the noisy response vectors are Gaussian distributed conditional on the speed (i.e. the class). The decoder computes the likelihood of the response given each class, that is, $p(\mathbf{R}|y = i)$, and using the priors (which we set as flat) and Bayes rule, it computes the posterior distribution $p(y|\mathbf{R})$ for the different speeds. For a given video $\mathbf{c}$ corresponding to true class $y = j$, the AMA-Gauss training loss is the negative log-posterior at the correct class, that is, $-\log p(y = j|\mathbf{R})$. The filters $\mathbf{F}$ that minimize the loss are obtained using gradient descent. The filters (columns of $\mathbf{F}$) are constrained to have unit norm. We use the same regularization parameter σ^2 for AMA-Gauss and for SQFA. AMA-Gauss filters were learned in a pairwise fashion.

G Regularization and invariance

G.1 Invariance to invertible linear transformations

To better understand SQFA, it is important to consider the invariance properties of the Fisher-Rao and the affine-invariant distances.

First, we consider the affine-invariant distance and the case of smSQFA. If $\mathbf{G} \in GL(m)$ where $GL(m)$ is the General Linear group, composed of non-singular m-by-m matrices, then

$$d_{AI}(\mathbf{\Psi}_i, \mathbf{\Psi}_j) = d_{AI}(\mathbf{G}^T\mathbf{\Psi}_i\mathbf{G}, \mathbf{G}^T\mathbf{\Psi}_j\mathbf{G}) \tag{15}$$

In words, the affine-invariant distance is invariant to the action by congruence of $GL(m)$.

For a variable $\mathbf{z} \in \mathbb{R}^m$ with second-moment matrix $\mathbf{\Psi}_i$, the transformed matrix $\mathbf{\Psi}'_i = \mathbf{G}^T\mathbf{\Psi}_i\mathbf{G}$ corresponds to the second moment matrix of the transformed variable $\mathbf{z}' = \mathbf{G}^T\mathbf{z}$. Thus, the distance is invariant to invertible linear transformations of the underlying variable $\mathbf{z}$. Importantly, for the case of SQFA, where the variable $\mathbf{z}$ is obtained as $\mathbf{z} = \mathbf{F}^T\mathbf{x}$, this is also equivalent to a transformation of the filters. Specifically, if $\mathbf{z} = \mathbf{F}^T\mathbf{x}$, then $\mathbf{z}' = \mathbf{F}'^T\mathbf{x}$, where $\mathbf{F}' = \mathbf{F}\mathbf{G}$.

In the context of smSQFA, and in the absence of regularization (see Methods section), if a set of filters are transformed as $\mathbf{F}' = \mathbf{F}\mathbf{G}$, then the second-moment matrices for all the classes will be transformed as $\mathbf{\Psi}'_i = \mathbf{G}^T\mathbf{\Psi}_i\mathbf{G}$. Therefore, according to Equation 15, the pairwise distances between classes will be the same for both sets of filters. In other words, the objective function of smSQFA is invariant to invertible linear transformations of the filters.

The Fisher-Rao distance and the Calvo-Oller bound are also invariant to affine transformations of the data space Nielsen (2020). This makes the filters learned by SQFA (like smSQFA above) non-unique up to invertible linear transformations (again, in the absence of regularization). To see this, let $\mathbf{H} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}$, where $\mathbf{0}$ is a m-by-1 vector of zeros, and $\mathbf{H} \in GL(m+1)$. We denote the moments of the transformed variable $\mathbf{z}' = \mathbf{G}^T\mathbf{z}$ as $\boldsymbol{\mu}' = \mathbf{G}^T\boldsymbol{\mu}$ and $\mathbf{\Sigma}'_i = \mathbf{G}^T\mathbf{\Sigma}_i\mathbf{G}$. Then we have the following relation

$$\mathbf{H}^T\mathbf{\Omega}_i\mathbf{H} = \mathbf{H}^T \begin{bmatrix} \mathbf{\Sigma}_i + \boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{bmatrix} \mathbf{H} = \begin{bmatrix} \mathbf{G}^T(\mathbf{\Sigma}_i + \boldsymbol{\mu}\boldsymbol{\mu}^T)\mathbf{G} & \mathbf{G}^T\boldsymbol{\mu} \\ \mathbf{G}\boldsymbol{\mu}^T & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{\Sigma}'_i + \boldsymbol{\mu}'\boldsymbol{\mu}'^T & \boldsymbol{\mu}' \\ \boldsymbol{\mu}'^T & 1 \end{bmatrix} = \mathbf{\Omega}'_i \tag{16}$$

Following the same reasoning as above, if the filters are transformed as $\mathbf{F}' = \mathbf{F}\mathbf{G}$, this amounts to transforming the Calvo-Oller embedding of each class as $\mathbf{\Omega}'_i = \mathbf{H}^T\mathbf{\Omega}_i\mathbf{H}$, meaning that the pairwise distances between classes remain the same.

In sum, in absence of regularizing noise, there is an equivalent set of solutions, given by the sets of filters that span the same subspace. This can be a useful property, for example because there is no need to worry about scalings in the data space. But the lack of a unique solution also makes for less interpretable features (the interpretable object is the subspace spanned by the filters).

G.2 Regularization breaks invariance

The equivalence of solutions above, however, is eliminated when we introduce regularization as an additive term $\mathbf{I}\sigma^2$ to the covariance matrices. We show this for the simpler case of smSQFA, but the same reasoning applies to SQFA.

As described in the Methods section, regularization is introduced by adding $\mathbf{I}\sigma^2$ to each second-moment matrix (equal to the covariance matrix in the zero-mean case of smSQFA) in the feature space. That is

$$\mathbf{\Psi}_i = \mathbf{F}^T\mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right]\mathbf{F} + \mathbf{I}\sigma^2 \tag{17}$$

This means that if the filters are transformed as $\mathbf{F}' = \mathbf{F}\mathbf{G}$, it is no longer true that the regularized second-moment matrices are related by $\mathbf{\Psi}'_i = \mathbf{G}^T\mathbf{\Psi}_i\mathbf{G}$. Rather, there will be a different matrix $\mathbf{G}_i$ satisfying $\mathbf{\Psi}'_i = \mathbf{G}_i^T\mathbf{\Psi}_i\mathbf{G}_i$ for each class i. Therefore, the pairwise distances between classes will change, and the objective function will not be invariant to invertible linear transformations of the filters.

One consequence of adding regularization is that the solution is no longer invariant to the scale of the filters. For filters with small norms, matrices $\mathbf{\Psi}_i = \mathbf{F}^T\mathbf{\Phi}_i\mathbf{F} + \mathbf{I}\sigma^2$ will be dominated by the regularization term, and thus more similar (i.e. closer) to each other. Then, the solution will tend towards filters with infinite norm that make the contribution of the regularization term negligible. In our case the filters (i.e. each column of $\mathbf{F}$) are constrained to have unit norm, so this effect is avoided.

Following the same reasoning, the regularization term will penalize filters that lead to small second-moment matrices, since these will be dominated by the regularization term, which is identical across classes. Thus, regularization will favor the directions in the data space that lead to larger second-moment matrices. This might be useful in that it makes the filters more robust to estimation noise along dimensions with small variance, but it might also mask important information in directions with low squared values. Also, this makes the results dependent on the choice of the regularization parameter σ^2. Future work should explore the effect of regularization on the features learned by SQFA, and examine how to choose the regularization parameter.

The breaking of invariance by regularization has the desirable effect of making the specific filters learned more reproducible and therefore more interpretable.

H Effect of Gaussianity on model evaluation

SQFA maximizes the Fisher-Rao distances between classes, under Gaussian assumptions. We also evaluate the performance of the different learning methods using Quadratic Discriminant Analysis (QDA), which that the data is Gaussian. We are interested in the Gaussian case for several reasons. First, the Gaussian assumption makes the SQFA objective computationally tractable, and is a good first step to applying information geometry for dimensionality reduction. Second, the Gaussian assumption is a good approximation to the data in many problems. Third, many scientific models assume that the data is Gaussian, and SQFA can be a useful dimensionality reduction method in these cases.

Nonetheless, the datasets used in this work are not Gaussian, and it is natural to ask to what extent the non-Gaussianity of the data affects SQFA. However, this is not an easy question to answer. First, non-Gaussianity can affect both the filter learning process and the evaluation of the learned filters, and it is not easy to disentangle the two effects. We need to ask how non-Gaussianity affects each of these processes separately. Second, there are many ways in which data can be non-Gaussian. It is thus difficult to make general statements about the effect of non-Gaussianity on the performance of SQFA. Despite these difficulties, here we present an analysis of how non-Gaussianity affects the evaluation of the learned filters.

First, we tested how the non-Gaussianity of the data affects the performance of QDA. That is, how does QDA performance with the used datasets differ from performance with Gaussian data? To answer this, after learning the filters with each method, we simulated testing data from Gaussian distributions with the same mean and covariance as the testing data for each class. We then obtained the QDA accuracy using the simulated Gaussian data.

The results are qualitatively similar for all Gaussian datasets and all the original datasets. In most cases, the SQFA variants continue to outperform the other methods (except AMA for the speed estimation dataset). The relative performance of the different SQFA variants is also maintained. Probably the most notable difference is that for the MNIST dataset the gap between SQFA and SQFA-B got larger for the Gaussian data. Additionally, for MNIST the performance of LDA features improved, slightly outperforming SQFA.

Then, we also asked how using a non-Gaussian decoder affects the results. That is, how does SQFA and its variants compare to the other methods when using a decoder that does not assume Gaussianity? To answer this, we used a k-Nearest Neighbors (kNN) decoder instead of QDA.

The results are again qualitatively similar for all datasets (Figure 14). One of the differences with the QDA results is that LMNN performs better than most SQFA variants for the MNIST dataset, although it is still outperformed by SQFA-H. Then, SQFA-H continues to be the best performing method across datasets (except for AMA in the speed estimation dataset), often followed by SQFA. Notably, performance for the speed estimation dataset is quite smaller than for QDA. This makes sense considering that the class-conditional data in this dataset has been shown to be close to Gaussian (Burge & Geisler, 2015), so a Gaussian decoder is more appropriate. Additionally, there is a large number of classes (41) with considerable overlap between them, and a somewhat small number of samples per class (500), which might make it difficult for kNN to work well.

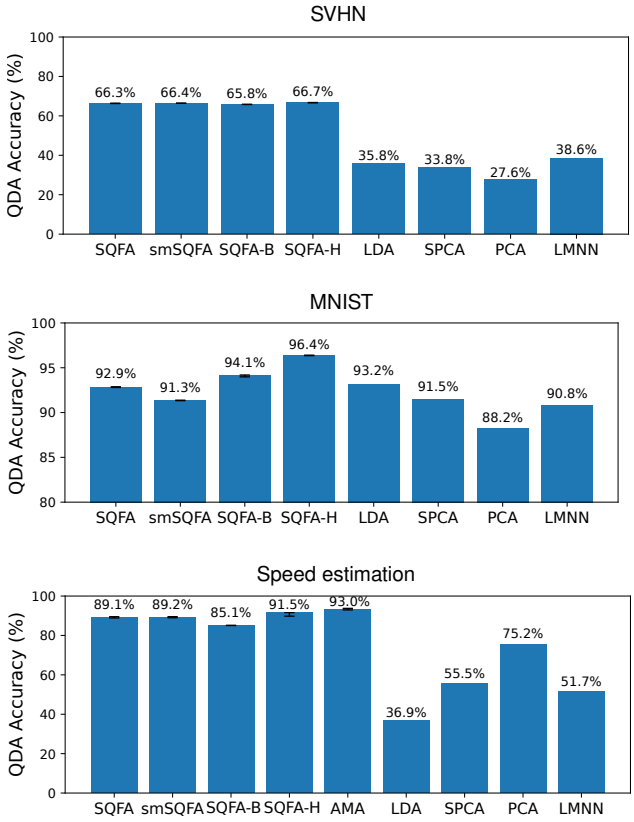

Figure 13: QDA accuracy with simulated Gaussian data. After learning the filters with each method, we simulated testing data from Gaussian distribution, using the mean and covariance of the testing data for each class. Then we evaluated the QDA accuracy using the simulated Gaussian data. From top to bottom, the datasets are SVHN, MNIST, and the speed estimation dataset.

In conclusion, here we analyzed the effect of non-Gaussianity on the evaluation of the learned filters. We found that the main results of the paper are maintained both when using actually Gaussian data, and when using a non-Gaussian decoder.

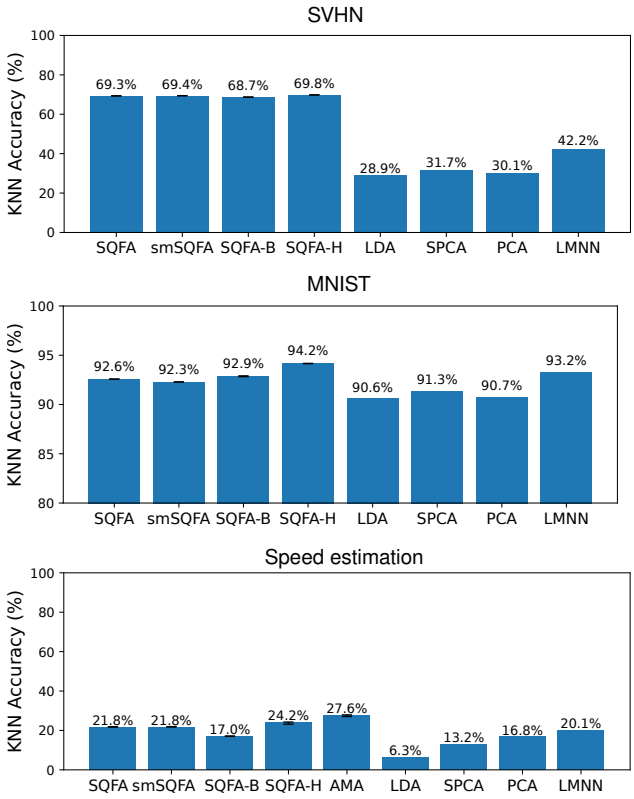

Figure 14: KNN accuracy for each of the datasets. For each dataset, the filters were learned as described in the main text. Then, a kNN decoder with k=3 was used to evaluate class discriminability. From top to bottom, the datasets are SVHN, MNIST, and the speed estimation dataset.