JMLR: Workshop and Conference Proceedings 63:302–317, 2016 ACML 2016

Multitask Principal Component Analysis

Ikko Yamane YAMANE@MS.K.U-TOKYO.AC.JP
Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8561, Japan.

Florian Yger FLORIAN.YGER@DAUPHINE.FR
LAMSADE Université Paris-Dauphine, 75775 Paris Cedex 16, France.

Maxime Berar MAXIME.BERAR@UNIV-ROUEN.FR
LITIS Université de Rouen, 76800 Saint-Etienne du Rouvray, France.

Masashi Sugiyama SUGI@K.U-TOKYO.AC.JP
RIKEN / Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8561, Japan.

Editors: Robert J. Durrant and Kee-Eung Kim

Abstract

Principal Component Analysis (PCA) is a canonical and well-studied tool for dimensionality reduction. However, when few data are available, the poor quality of the covariance estimator at its core may compromise its performance. We leverage this issue by casting the PCA into a multitask framework, and doing so, we show how to solve simultaneously several related PCA problems. Hence, we propose a novel formulation of the PCA problem relying on a novel regularization. This regularization is based on a distance between subspaces, and the whole problem is solved as an optimization problem over a Riemannian manifold. We experimentally demonstrate the usefulness of our approach as pre-processing for EEG signals.

Keywords: dimensionality reduction, multitask learning, Riemannian geometry, Grassmann manifold

1. Introduction

Principal Component Analysis (PCA) (Pearson, 1901; Hotelling, 1933; Joliffe, 1986) is a preprocessing technique widely used in data processing and is a prominent dimensionality reduction technique in machine learning.

In few words, PCA seeks an accurate low-dimensional approximation to high-dimensional data. To do so, PCA finds an orthogonal projection of the data to a low-dimensional subspace while preserving as much variance as possible, or equivalently while minimizing the projection error (see Bishop (2006, Chap 12)).

In practice, this boils down to an eigenproblem involving the covariance estimator of the input variables. Hence, its simplicity and efficiency allowed the proposition of several variants of PCA over the course of time, ranging from non-linear extensions (Schölkopf et al., 1997; Vincent et al., 2010) to sparse (Zou et al., 2006) or supervised extensions (De Bie et al., 2005). In order to efficiently cope with non-stationary data streams, PCA has been studied from the point of view of subspace tracking (Badeau et al., 2008; Balzano et al., 2010), where the emphasis is put on efficiently updating the principal subspace while maintaining the orthonormality constraint. In a related setup, it has also been studied from the online

learning point of view (Warmuth and Kuzmin, 2007) in order to derive bounds on the projection error.

In one-dimensional cases, the PCA can be solved by extracting the dominant eigenvector of the covariance matrix of the input data. Then multidimensional PCA can be iteratively performed by solving a one-dimensional problem and followed by a deflation scheme. However, this problem can also be solved at once by optimizing a generalization of the one-dimensional cost under orthonormality constraints or by optimizing on a Riemannian manifold[1] involving orthogonal matrices (Edelman et al., 1998; Absil et al., 2009). When such a cost is optimized, the solution may not exactly diagonalize the covariance matrix but will have the same span as the leading eigenvectors.

As any machine learning method, the quality of the solution obtained by the PCA is greatly affected by the quality of the covariance estimator used in practice. Being based on the minimization of a least-squares cost, the quality of the covariance estimator is particularly affected by outliers. Hence, in order to overcome this situation, several robust versions of the PCA have been proposed for dealing with noisy data and outliers. Those approaches either rely on multivariate trimming of the samples (Devlin et al., 1981) or on a cost function giving less influence to outliers (Candès et al., 2011).

However, in a context where few data are available, the covariance matrix may not be accurately estimated and the robust approaches are not adapted. If such a situation happens to several related datasets, one straightforward approach consists in finding a common principal subspace to all the datasets. As studied in Wang et al. (2011), it boils down to applying a single PCA over all the data or to finding a subspace approximating all the covariance matrices. This latter formulation makes the problem close to the Approximate Joint Diagonalization (AJD) encountered in the Signal Processing community (Flury and Gautschi, 1986; Cardoso and Souloumiac, 1996).

Obviously, in this context of data scarcity, as the covariance estimator is not reliable, independently solving a PCA for every dataset would fail. Hence, we need a trade-off between the flexible straightforward approach (independent PCAs) and the easy-to-use single PCA. To do so, we propose to cast the PCA into the Multitask Learning (MTL) framework (Evgeniou and Pontil, 2004; Argyriou et al., 2008; Zhang and Yeung, 2011). In this setup, every task amounts to finding a low-rank transformation of each dataset (maximizing the retained information), and while solving those tasks, a regularization term is introduced to make those transformations similar to each other. As we focus on the multidimensional case, we formulate our problem as an optimization problem over a Riemannian matrix manifold, and using this geometry, we propose a novel regularization term.

Eigenproblems being a classical tool of machine learning (De Bie et al., 2005), their study in the multitask framework has naturally been proposed. Recently, such an approach has been developed in Wang et al. (2016). This approach studied the generalized eigenproblems and only extracted the leading eigenvector by casting the problem into a multitask dictionary learning problem. In essence, our contribution in this paper is very different from this previous work. Indeed, instead of studying the generalized eigenproblems, we focused on the PCA problem, but we are able to extract directly a dominant subspace (i.e. the

1. A Riemannian manifold is a smoothly curved non-Euclidean space with additional structures such as a set of linear local approximations, i.e. the tangent spaces, that are equipped with an inner product (Absil et al., 2009).

span of a set of leading eigenvectors) without having to resort to any deflation scheme. As exposed in this paper, we propose a simple and elegant MTL formulation relying on a novel regularization.

The use of a multitask methodology has been advocated in challenging applications such as Brain Computer Interfaces (BCI) where it is difficult to collect data from each task but the tasks are related (Devlaminck et al., 2011; Samek et al., 2013). In this paper, we provide some promising results on this difficult application. In order to analyze the behavior of our approach, we also apply it on synthetic data.

To summarize, the key contributions of our paper are twofold: First and foremost, we formulate on a matrix manifold the problem of dominant subspace extraction for multitask variance maximization. As a result, it makes it possible to solve at once several related PCA problems of fixed dimensionality. Secondly, we propose a relevant regularization (having an interpretation from the point of view of Riemannian geometry) for this multitask problem. Then, the problem is naturally formulated as an optimization problem over a Riemannian matrix manifold. Through experiments on synthetic data and a signal processing application, we demonstrate the efficacy of our proposed dimensionality reduction method.

2. Multitask Variance Maximization

In this section, we define the problem of multitask variance maximization and then present our proposed method.

2.1. Problem Setup

This problem being defined as a collection of instances of single-task variance maximization, we first start by introducing the single-task version of variance maximization.

For any random data variable $\boldsymbol{x} \in \mathbb{R}^d$ following a distribution $p(\boldsymbol{x})$, the goal of variance maximization is to estimate from i.i.d. samples $\{\boldsymbol{x}_i\}_{i=1}^n$ drawn from $p(\boldsymbol{x})$ the k-dimensional subspace ($k < d$; we assume k is known and fixed) on which the projected point of $\boldsymbol{x}$ has the maximum variance.

For any matrix $\boldsymbol{M} \in \mathbb{R}^{d\times k}$, we denote the span of the columns of $\boldsymbol{M}$ by $\mathrm{Span}(\boldsymbol{M})$. For any k-dimensional subspace S and any orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{d\times k}$, we say that $\boldsymbol{U}$ is an *orthogonal basis matrix* of S if $\mathrm{Span}(\boldsymbol{U}) = S$. Any d-by-k orthogonal matrix determines a unique subspace as an orthogonal basis matrix while there are infinitely many orthogonal basis matrices for any given subspace with dimensionality $d \geq 2$. Since the orthogonal projection onto any subspace S is given as $\mathbb{R}^d \ni \boldsymbol{x} \mapsto \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x} \in S$ using any basis matrix $\boldsymbol{U}$ of S, an orthogonal basis matrix $\boldsymbol{U}^*$ of the optimal subspace S^* is obtained as a solution to the following problem:

$$\boldsymbol{U}^* = \operatorname*{argmax}_{\boldsymbol{U}\in\mathbb{R}^{d\times k}:\ \boldsymbol{U}^\top\boldsymbol{U}=\boldsymbol{I}_k} \mathop{\mathbf{E}}_{\boldsymbol{x}\sim p(\boldsymbol{x})} \left[\left\|\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x} - \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{\mu}\right\|^2\right] \tag{1}$$

$$= \operatorname*{argmax}_{\boldsymbol{U}\in\mathbb{R}^{d\times k}:\ \boldsymbol{U}^\top\boldsymbol{U}=\boldsymbol{I}_k} \mathrm{Tr}(\boldsymbol{U}^\top\boldsymbol{C}\boldsymbol{U}), \tag{2}$$

where $\boldsymbol{\mu} = \mathbf{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}[\boldsymbol{x}]$ is the population mean of $\boldsymbol{x}$, $\boldsymbol{C} = \mathbf{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top]$ is the population covariance of $\boldsymbol{x}$, and $\boldsymbol{I}_k$ is the k-by-k identity matrix. Again, $\boldsymbol{U}^*$ is not uniquely

determined because the objective function is invariant under orthogonal transformations since $\mathrm{Tr}((\boldsymbol{U}\boldsymbol{O})^\top \boldsymbol{C}(\boldsymbol{U}\boldsymbol{O})) = \mathrm{Tr}(\boldsymbol{U}^\top \boldsymbol{C}\boldsymbol{U}\boldsymbol{O}\boldsymbol{O}^\top) = \mathrm{Tr}(\boldsymbol{U}^\top \boldsymbol{C}\boldsymbol{U})$ for any orthogonal matrix $\boldsymbol{O} \in \mathbb{R}^{k\times k}$. As made clear later in this paper, in order to deal with the orthogonality constraint as well as with this invariance to rotations, we will use Grassmann manifolds for the formulation of the problem (Edelman et al., 1998).

In *multitask variance maximization*, which is the main subject of this paper, we have multiple different instances of variance maximization. We call such instances as *tasks*. More specifically, given T sets of i.i.d. samples $\{\boldsymbol{x}_{t,i}\}_{i=1}^{n_t}$, $t = 1, \ldots, T$, following underlying probability distributions $p_1(\boldsymbol{x}_1), \ldots, p_T(\boldsymbol{x}_T)$ respectively, we are required to estimate the optimal k-dimensional subspaces, whose basis matrices $\boldsymbol{U}_t^*$, $t = 1, \ldots, T$, are given by

$$\boldsymbol{U}_t^* = \underset{\boldsymbol{U}_t \in \mathbb{R}^{d\times k}:\ \boldsymbol{U}_t^\top \boldsymbol{U}_t = \boldsymbol{I}_k}{\operatorname{argmax}} \mathrm{Tr}(\boldsymbol{U}_t^\top \boldsymbol{C}_t \boldsymbol{U}_t), \tag{3}$$

where $\boldsymbol{\mu}_t = \mathbf{E}_{\boldsymbol{x}_t \sim p_t(\boldsymbol{x}_t)}[\boldsymbol{x}_t]$, and $\boldsymbol{C}_t = \mathbf{E}_{\boldsymbol{x}_t \sim p_t(\boldsymbol{x}_t)}[(\boldsymbol{x}_t - \boldsymbol{\mu}_t)(\boldsymbol{x}_t - \boldsymbol{\mu}_t)^\top]$.

2.2. Principal Component Analysis

In many applications, the population covariance matrix $\boldsymbol{C}_t$ is often unknown, and the objective function of Eq. (3) cannot be directly evaluated. A common way to alleviate this is to resort to the *sample covariance matrix* defined by $\widehat{\boldsymbol{C}}_t = \frac{1}{n_t - 1}\sum_{i=1}^{n_t}(\boldsymbol{x}_{t,i} - \widehat{\boldsymbol{\mu}}_t)(\boldsymbol{x}_{t,i} - \widehat{\boldsymbol{\mu}}_t)^\top$ with $\widehat{\boldsymbol{\mu}}_t = \frac{1}{n_t}\sum_{i=1}^{n_t}\boldsymbol{x}_{t,i}$ to approximate the objective function as $\mathrm{Tr}(\boldsymbol{U}_t^\top \boldsymbol{C}_t \boldsymbol{U}_t) \approx \mathrm{Tr}(\boldsymbol{U}_t^\top \widehat{\boldsymbol{C}}_t \boldsymbol{U}_t)$.

In the case of the single task learning setup (i.e. $T = 1$), the method of solving such an approximated problem is widely known as Principal Component Analysis (PCA) (see, e.g., Joliffe (1986)) and can be solved by taking the leading k orthonormal eigenvectors of $\widehat{\boldsymbol{C}}_t$. PCA and its variants have been proven to be useful in many applications such as model reduction in control theory (Moore, 1981) and denoising for image processing (Zhang et al., 2010). In our multitask setting, we refer to the method of applying PCA to every task independently as *Independent PCA (I-PCA)* and the method of applying it to the union of the datasets from all the tasks as *Common PCA (C-PCA)*. The notable difference between these two methods is that C-PCA gives the same subspace for all the tasks whereas I-PCA could give completely different subspaces for different tasks.

I-PCA may provide good estimates of the optimal subspaces when sufficiently many data samples are available, but when we have only scarce data samples, the solutions $\widehat{\boldsymbol{U}}_t$ to the problem in Eq. (3) may be badly affected by unreliable covariance estimation resulting in poor performance on unseen data. In fact, the solution is undetermined when the sample size n_t is less than the dimensionality k of the subspace.

A straightforward countermeasure to this data-scarcity problem is to adopt C-PCA in order to simply increase the sample size. However, this corresponds to assuming that all the tasks share the identical optimal solution, which may be unreasonable when the tasks have considerable heterogeneity.

The objective of this paper is to improve the performance over both I-PCA and C-PCA when the tasks are different but related to each other in the sense that their optimal subspaces are similar to each other. In such a case, solving all the tasks simultaneously while sharing information with each other may improve performance. This strategy of jointly

learning multiple tasks with taking the advantage of their relatedness is called *multitask learning* and has been shown to work well in many other applications (Caruana, 1998; Evgeniou and Pontil, 2004; Argyriou et al., 2008; Zhang and Yeung, 2011; Jacob et al., 2009).

2.3. Regularized Multitask PCA

One of the most successful approaches to multitask learning is the regularization approach (Evgeniou and Pontil, 2004; Argyriou et al., 2008; Jacob et al., 2009). In this approach, the tasks maintain different learning parameters but they are simultaneously optimized with an appropriately designed regularization term which, e.g., makes the parameters close to each other or imposes similar sparsity patterns on them.

In this paper, we propose a method based on this approach for solving multitask variance maximization. In the proposed method, we directly search the space of subspaces instead of searching the space of orthogonal skinny matrices. More specifically, we solve the following optimization problem:

$$(\widehat{\boldsymbol{S}}_1, \ldots, \widehat{\boldsymbol{S}}_T) = \underset{\substack{(\boldsymbol{U}_1,\ldots,\boldsymbol{U}_T)\\ \in \mathrm{Gr}(d,k)^{\otimes T}}}{\operatorname{argmax}} \underbrace{\left[\frac{1}{2} \sum_{t\in[T]} \operatorname{Tr}(\boldsymbol{U}_t^\top \widehat{\boldsymbol{C}}_t \boldsymbol{U}_t) + \frac{\lambda}{4} \sum_{s,t\in[T]:\, s\neq t} \operatorname{Tr}(\boldsymbol{U}_s \boldsymbol{U}_s^\top \boldsymbol{U}_t \boldsymbol{U}_t^\top) \right]}_{J(\boldsymbol{U}_1,\ldots,\boldsymbol{U}_T)}, \tag{4}$$

where $\lambda > 0$ is a regularization parameter, $[T] = \{1, \ldots, T\}$, and $\mathrm{Gr}(d,k)^{\otimes T}$ denotes the product manifold consisting of T *Grassmann manifolds*. Each of those manifolds consists of all the k-dimensional linear subspaces of the d-dimensional Euclidean space $\mathbb{R}^d$[2], and $\widehat{\boldsymbol{S}}_t$ is the estimate of the optimal subspace for task t. We call this method *Regularized MultiTask Principal Component Analysis (RMT-PCA)*. As we will see later, the objective function does not depend on the choice of the orthogonal basis matrices $\boldsymbol{U}_t$, $t = 1, \ldots, T$, and thus the optimization problem is well-defined on $\mathrm{Gr}(d,k)^{\otimes T}$.

Intuitively, we try to maximize the PCA objective function $\operatorname{Tr}(\boldsymbol{U}_t^\top \widehat{\boldsymbol{C}} \boldsymbol{U}_t)$ for every task t simultaneously while maximizing the similarity between the subspaces $\mathrm{Span}(\boldsymbol{U}_s)$ and $\mathrm{Span}(\boldsymbol{U}_t)$ quantified by $\operatorname{Tr}(\boldsymbol{U}_s \boldsymbol{U}_s^\top \boldsymbol{U}_t \boldsymbol{U}_t^\top)$ for every task pair (s,t) at the same time.

Fig. 1 illustrates the idea of our multitask PCA approach. In this example, three datasets of three-dimensional examples are observed. Those three datasets share similar (but different) behaviors as their two-dimensional principal subspaces are close to be parallel. Hence, the overall objective is to find similar subspaces (i.e. having similar angles) expressing most of the variance of each dataset. This example shows the flexibility of our approach as it is immune to the choice of bases representing the subspaces.

Maximizing the term $\operatorname{Tr}(\boldsymbol{U}_s \boldsymbol{U}_s^\top \boldsymbol{U}_t \boldsymbol{U}_t^\top)$ in the regularization can be interpreted as minimizing the *projection F-norm distance* which is defined and denoted for any subspaces S and S' by $\delta_{\mathrm{pF}}(S, S') = \|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}'\boldsymbol{U}'^\top\|_{\mathrm{F}}$, where $\|\boldsymbol{M}\|_{\mathrm{F}} = \sqrt{\operatorname{Tr}(\boldsymbol{M}^\top \boldsymbol{M})}$, and $\boldsymbol{U}$ and $\boldsymbol{U}'$ are d-by-k orthogonal basis matrices of S and S' respectively. This follows from the equality $\delta_{\mathrm{pF}}^2(S, S') = 2d - 2\operatorname{Tr}(\boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{U}'\boldsymbol{U}'^\top)$. $\delta_{\mathrm{pF}}(S, S')$, and thus the regularization term in Eq. (4),

2. Note that a point X on this manifold can be represented by any orthonormal basis of $\mathbb{R}^{k\times d}$. The chosen orthonormal basis is called a representative of its subspace $\mathrm{Span}(X)$.

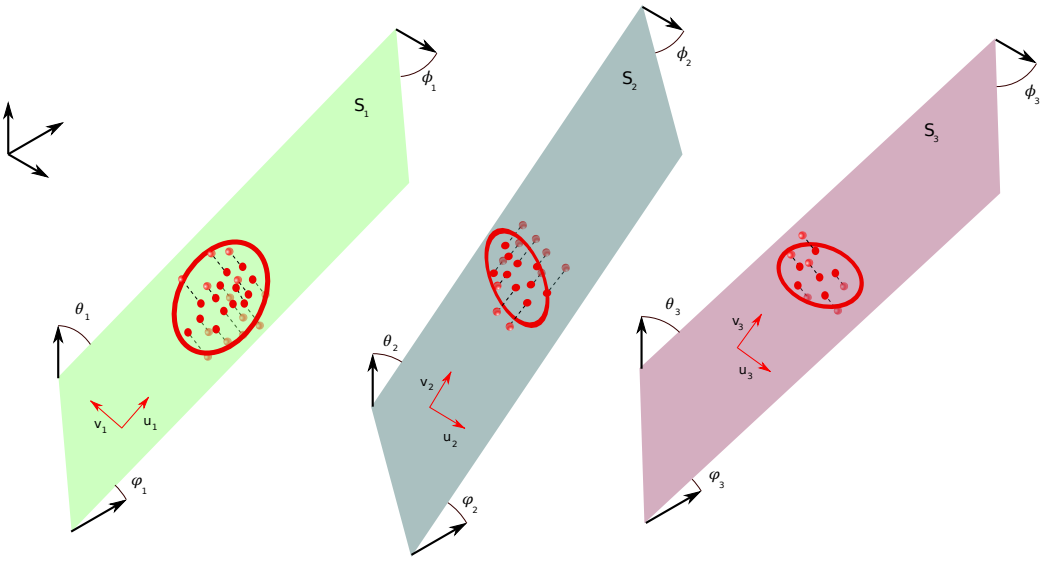


Figure 1: Illustration of the multitask setup for the PCA problem. Few observations are available for every task of PCA, and we aim at extracting similar subspaces (hence being oriented according to similar angles). In this example, each subspace S_t is represented by a basis of two vectors u_t, v_t and the angles between the canonical basis and the subspaces are ϕ_t, θ_t, ψ_t.

are invariant up to the choice of $\boldsymbol{U}_s$ and $\boldsymbol{U}_t$. A nice property of the projection F-norm distance is that for subspaces with small geodesic distance, it is asymptotically equivalent to other several important measures including that induced by the intrinsic geometry of the Grassmann manifold (Edelman et al., 1998; Chevallier et al., 2013).

As already mentioned, the multidimensional PCA loss function is invariant under the group action $\boldsymbol{U} \mapsto \boldsymbol{U}\boldsymbol{O}$ for all orthogonal matrices $\boldsymbol{O}$ of size $k \times k$. Hence optimizing on the space of orthogonal skinny matrices (i.e. the Stiefel manifold) without taking into account this invariance would be inefficient as the critical points of the cost function are not isolated on the Stiefel manifold. Then, such a property should be taken into account for defining a multitask regularization. It can be easily shown that this is the case for the proposed regularization of this paper as for any orthogonal matrices $\boldsymbol{O}$ and $\boldsymbol{O}'$ of size $k \times k$, we have

$$\mathrm{Tr}(\boldsymbol{U}\underbrace{\boldsymbol{O}\boldsymbol{O}^\top}_{\boldsymbol{I}_k}\boldsymbol{U}^\top\boldsymbol{U}'\underbrace{\boldsymbol{O}'\boldsymbol{O}'^\top}_{\boldsymbol{I}_k}\boldsymbol{U}'^\top) = \mathrm{Tr}(\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{U}'\boldsymbol{U}'^\top).$$

It would have been tempting to use a simpler regularization such as the matrix scalar product $\mathrm{Tr}(\boldsymbol{U}^\top\boldsymbol{U}')$. However, this regularizer is not invariant under the group action over the product of Grassmann manifolds and this has some very bad consequences. In cases where the top k eigenvalues of a covariance matrix are close, it can happen that the value of those eigenvalues are different (and hence their order changed) for the estimated covariance. Such a situation in two dimensions is illustrated in Fig. 2. In this case, the subspaces S

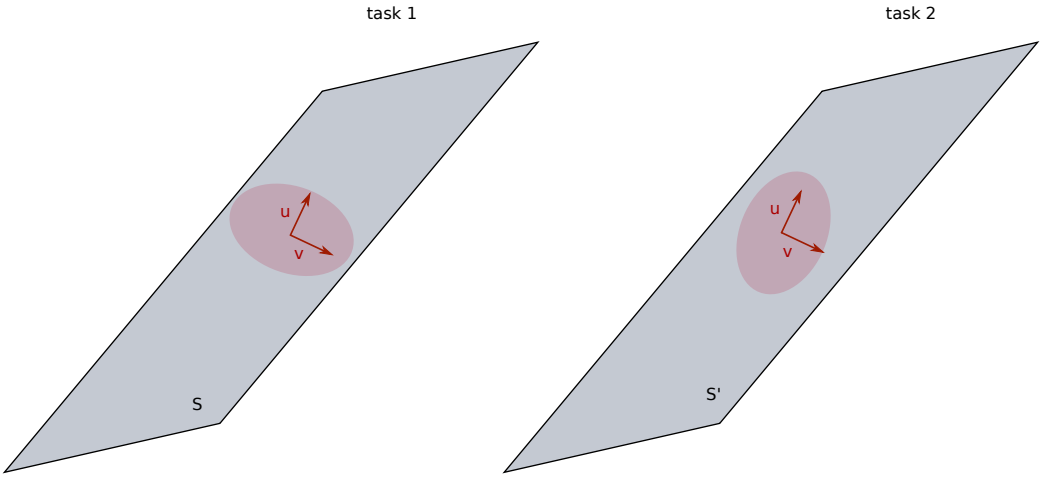


Figure 2: Illustration of the invariance of subspaces to the choice of basis. In this 3-dimensional example, the two tasks are generated from the same distribution, but due to sampling, the order of the two main eigenvectors is changed (even though the subspaces are the same). Hence, if we are interested in comparing subspaces, our regularizer should be immune to the choice of bases.

and S' are identical and respectively represented by the basis matrices $\boldsymbol{U} = \left[\begin{array}{c|c} \boldsymbol{u} & \boldsymbol{v} \end{array} \right]$ and $\boldsymbol{U}' = \left[\begin{array}{c|c} \boldsymbol{v} & \boldsymbol{u} \end{array} \right]$, with $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ such that $\boldsymbol{u}^\top \boldsymbol{u} = 1$, $\boldsymbol{v}^\top \boldsymbol{v} = 1$ and $\boldsymbol{u}^\top \boldsymbol{v} = 0$. Then, it naturally follows that: $\mathrm{Tr}(\boldsymbol{U}^\top \boldsymbol{U}') = 0 \quad \text{and} \quad \mathrm{Tr}(\boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{U}'\boldsymbol{U}'^\top) = 2$.

When dealing with covariance matrices estimated from few samples, it can happen that the order of the principal eigenvectors is changed compared to the principal eigenvectors of the population covariance. Compared to the naive regularization, our regularization is robust to such a practical problem.

2.4. Optimization on Product of Grassmann Manifolds

The Grassmann manifold is a powerful mathematical tool for modeling low-rank transformations, and as noted in Edelman et al. (1998), it is usually involved for solving eigenproblems. As it directly models fixed dimensionality subspaces, it is independent of the bases chosen to represent the subspaces. Hence, as described in Absil et al. (2009, Sec. 3.4.4), a Grassmann manifold is a quotient manifold and the group structure enables us to encode the invariance properties. In few words, if two representations have the same span, they are said to be equivalent. For a comprehensive tour on this topic, the reader should refer to Absil et al. (2009); Edelman et al. (1998).

In this work, instead of modeling our dimensionality reduction problem as an optimization problem under a set of orthonormality constraints, we write it as an unconstrained optimization on Grassmann manifolds. Hence, our approach consists in finding several lower-dimensional subspaces by optimizing several transformations (parameterized by $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_T$) that maximize the variance on each dataset meanwhile being similar. As each parameter

$\boldsymbol{U}_t$ lies in a Grassmann manifold $\mathrm{Gr}(d,k)$ (Absil et al., 2009; Edelman et al., 1998), we solve the optimization problem on the product of these manifolds.

In Ma et al. (2001), the authors proved that the geodesics in the product manifold are the products of the geodesics in the factor manifolds. This helpful property enables us to compute the gradients on each of the factor manifolds separately and hence to apply easily the machinery of the field of optimization on Riemannian manifolds.

Optimization on Riemannian matrix manifolds is a mature field and by now most of the classical optimization algorithms have been extended to this setting (Absil et al., 2009). In this setting, descent directions are not straight lines but rather curves on the manifold. For a function $f(\boldsymbol{U})$, applying a Riemannian gradient descent can be expressed by the following steps:

1. At any iteration, at the point $\boldsymbol{U}$, transform a Euclidean gradient $D_{\boldsymbol{U}}f$ into a Riemannian gradient $\nabla_{\boldsymbol{U}}f$. In our case, $\nabla_{\boldsymbol{U}}f = D_{\boldsymbol{U}}f - \boldsymbol{U}\boldsymbol{U}^\top D_{\boldsymbol{U}}f$ (Absil et al., 2009).

2. Perform a line search along geodesics at $\boldsymbol{U}$ in the direction $H = \nabla_{\boldsymbol{U}}f$. In our case, on the geodesic going from a point U in direction H (with a step-size t), a new iterate is obtained as $\boldsymbol{U}(t) = \boldsymbol{U}\boldsymbol{V}\cos(\boldsymbol{\Sigma}t)\boldsymbol{V}^\top + \boldsymbol{W}\sin(\boldsymbol{\Sigma}t)\boldsymbol{V}^\top$, where $\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is the compact singular value decomposition of H.

Our cost function being defined in Eq. (4), its Euclidean gradient (w.r.t. a given task t) can be written as:

$$D_{\boldsymbol{U}_t}J = \widehat{\boldsymbol{C}}_t\boldsymbol{U}_t + \lambda \sum_{s\in[T]\setminus\{t\}} \boldsymbol{U}_s\boldsymbol{U}_s^\top\boldsymbol{U}_t. \tag{5}$$

In practice, we employ a Riemannian trust-region method described in Absil et al. (2009) and efficiently implemented in Boumal et al. (2014).

3. Experiments

In this section, we present numerical experiments on synthetic and real-life data in order to study the effect of the proposed regularization. We run the proposed method with various regularization parameter values and in various conditions to understand how the performance of the proposed method shifts as the regularization level changes. In this experiment, we compare the performance of the proposed method to the performances of independently applying the PCA to each task (noted as I-PCA and corresponding to the case of $\lambda = 0$) and applying a single PCA over all the datasets (noted as C-PCA and corresponding to the case of $\lambda = \infty$)[3].

In our *scarce setup*, every task has only scarce data, and the goal is to estimate the optimal subspaces accurately for all the tasks.

3. Note that the method of Wang et al. (2016) being fundamentally a rank-1 method, and as it relies on several hyper-parameters (the number of dictionary atoms and the sparsity level). For these reasons, we decided not to include it in our comparisons.

3.1. Setup

In the scarce setup, we estimate the optimal subspaces with the proposed method using a small number of training samples under several configurations, and then evaluate the quality of the obtained estimates using a large number of test samples. The specific numbers of training and test samples differ from dataset to dataset. We will provide the information in Sec. 3.2.

In the evaluation phase, we measure how much ratio of the variance is preserved when the test sample points are projected onto the estimated subspaces. We refer to this ratio as the *retained variance ratio (RVR)*. We calculate the RVR for every subject t by

$$r_t = \frac{\mathrm{Tr}(\widehat{\boldsymbol{U}}_t^\top \widehat{\boldsymbol{C}}'_t \widehat{\boldsymbol{U}}_t)}{\mathrm{Tr}(\widehat{\boldsymbol{C}}'_t)}, \tag{6}$$

where $\widehat{\boldsymbol{U}}_t$ denotes an arbitrary basis matrix of the estimated subspace, $\widehat{\boldsymbol{C}}'_t$ is the sample covariance matrix calculated using test samples. Then, we average $r_1, \ldots, r_T$ to obtain the overall score: $r = \frac{1}{T}\sum_{t=1}^{T} r_t$. In regularization parameter selection by cross-validation, we also use this score but calculated with hold-out samples in place of the test samples.

For statistically reliable evaluation, we run several trials of this experiment with different data realizations[4]. The specific numbers of trials will be provided in Sec. 3.2.

3.2. Data

We tested the method on the following synthetic data and BCI data.

Synthetic Data Sample points for each task t are drawn from the 6-dimensional Gaussian distribution with mean zero and covariance matrix $\boldsymbol{C}_t$ generated in the following way. First, we prepare the 'core' covariance matrix $\boldsymbol{C}_0$ as $\boldsymbol{C}_0 = \boldsymbol{O}_0 \boldsymbol{\Sigma}_0 \boldsymbol{O}_0^\top$, where $\boldsymbol{O}_0 \in \mathbb{R}^{d\times d}$ is a random orthogonal matrix, and $\boldsymbol{\Sigma}_0$ is the diagonal matrix whose diagonal elements are $1, 1, 2, 2, 3, 3$. Second, for each task t, we slightly 'tilt' $\boldsymbol{C}_0$ in order to obtain the task specific covariance $\boldsymbol{C}_t$: $\boldsymbol{C}_t = \boldsymbol{O}_t \boldsymbol{C}_0 \boldsymbol{O}_t^\top$, where $\boldsymbol{O}_t \in \mathbb{R}^{d\times d}$ is an orthogonal matrix nearly equal to $\boldsymbol{I}_d$. We generate $\boldsymbol{O}_t$ as the projected point of $\boldsymbol{I}_d + \boldsymbol{N}$ onto the space of orthogonal matrices[5], where $\boldsymbol{N}$ is the noise matrix whose elements are i.i.d. samples drawn from the Gaussian distribution with mean zero and variance 0.3.

We conduct the experiment for $k = 1, \ldots, 5$ on this dataset. In the scarce setup, the training sample size is 10 for every task. The test sample size is 10000.

BCI Data This dataset consists of *electroencephalogram (EEG)* signals made available in the context of the *BCI competition IV dataset IIa* (Naeem et al., 2006). This data set is made of EEG signals (recorded from 22 electrodes) from 9 subjects who performed left-hand, right-hand, foot and tongue imaginary movements. As in Yger et al. (2015), we focus on the hand signals (72 trials for each class). This classical paradigm of motor imagination is used for building BCI so that a patient can send commands to a computer by performing imaginary actions.

4. By "data realization", we indicate data instances generated with a pseudo random generator in the case of the synthetic dataset, and re-sampled data points from the dataset in the case of the BCI data.
5. The projection on the space of orthogonal matrices is defined by $\boldsymbol{X} \mapsto \mathrm{argmin}_{\boldsymbol{O}\in\mathbb{R}^{d\times d}:\, \boldsymbol{O}\boldsymbol{O}^\top = \boldsymbol{I}_d} \|\boldsymbol{X} - \boldsymbol{O}\|_{\mathrm{F}}$.

Then the challenge remains for the computer to accurately detect the correct signal pattern. Nowadays, covariance matrices of EEG signals are commonly used as features for training BCIs (Yger, 2013). In this area, it is time consuming to gather data for a given subject but the data of several subjects are available.

Hence, in this context, our first task will be to investigate the performance of the proposed method in principal subspace extraction of the signals of all the subjects given only the covariance matrix of 1 epoch per subject (Sec. 3.3.1).

Furthermore, we tackle the second task, regularization parameter selection by 2-fold cross-validation, under the setup where two covariance matrices are available (Sec. 3.3.2).

We conduct these experiments for $k = 1, 4, 7, 10$. We sequentially pick one/two epoch(s) (as described above) for each task for subspace estimation, and then the rest of the epochs are used for evaluation of the estimates. We run 72 iterations in the experiment in Sec. 3.3.1 and 36 iterations in the experiment in Sec. 3.3.2.

3.3. Results

We show the results of the experiments below.

3.3.1. Performance Transition over Regularization-Level Change

First, we investigate the performance transition of the proposed method when the regularization level is varied.

The results on the synthetic data in the scarce setup are summarized in Fig. 3. Fig. 3 shows that the best λ value is somewhere in the middle between 0 and Inf (which denotes infinity) for all of $k = 1, \ldots, 5$, meaning that the proposed method with an appropriate λ value outperforms I-PCA and C-PCA. We can also see the tendency that the performance improves more when we have more tasks.

The results on the BCI data in the scarce setup are shown in Fig. 4. Similarly to the case of the synthetic data, the performance was improved for all k with appropriate λ values.

The BCI data have most of their variance in a few principal components; the test RVR score for $k = 4$ was more than 96% in all the trials of our experiments, which means that the largest possible RVR gain is less than 4%. Hence, there is less room for improvement for larger k. Nevertheless, the proposed method significantly improved the performance even in such challenging cases.

These results show that the proposed method is useful as long as the regularization level is in a moderate range.

3.3.2. Regularization Parameter Selection by Cross-Validation

In Sec. 3.3.1, the experiments in the scarce setup showed that there exists a regularization parameter such that our method achieves better results than those of the baseline methods. In order to select such a parameter, we apply a cross-validation method and provide some experimental results on BCI data. On the BCI data, the proposed method outperformed the other two methods for all of $k = 1, 4, 7, 10$ on average (see Tab. 1). The box plots in Fig. 5 detail the results, showing that RMT-PCA scored larger RVRs compared to I-PCA and C-PCA in most of the trials in every setting.

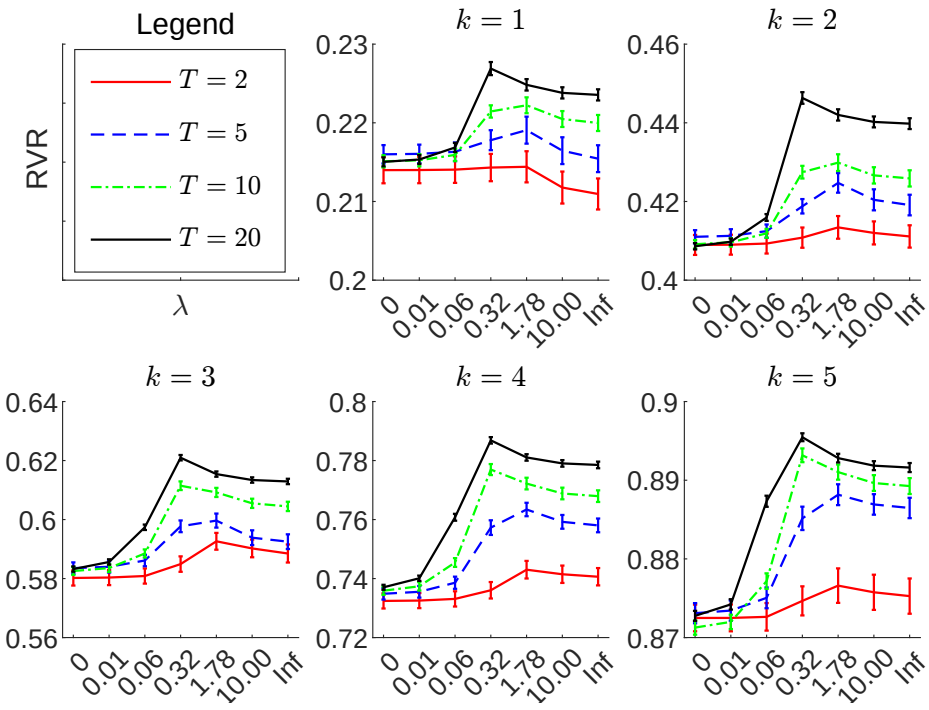
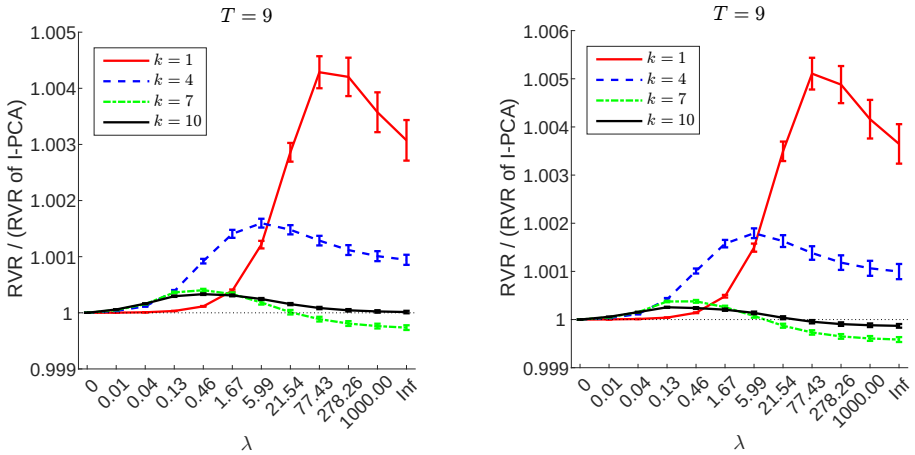
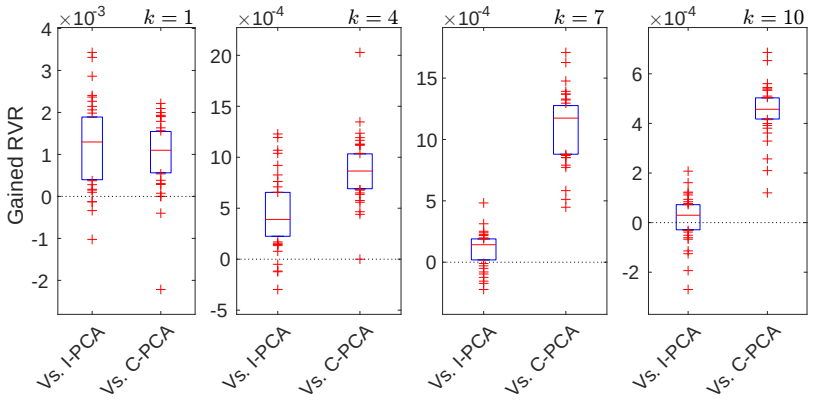


Figure 3: The transition of the RVR score over the level of regularization on synthetic data. Each plot corresponds to a different dimensionality k, and each curve corresponds to a different number of tasks T. 'Inf' denotes infinity. The error bars show the mean scores and their standard errors over 100 trials of the experiment.

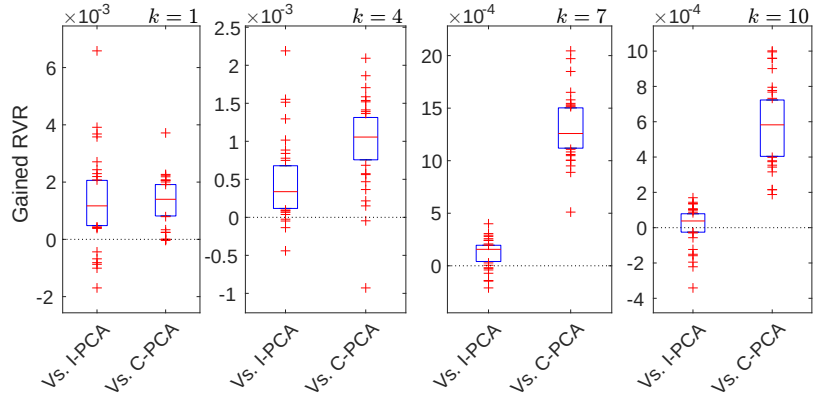


(a) Class 1. (b) Class 2.

Figure 4: Transition of the RVR score of the proposed method divided by the score of I-PCA over the the level of regularization on BCI data ('Inf' denotes infinity). Each plot corresponds to a different class, and each curve corresponds to a different dimensionality k. The black dotted lines indicate ratio of 1. The error bars show the mean scores and their standard errors over 72 trials.



(*a*) Class 1



(*b*) Class 2

Figure 5: The RVR of the proposed method using a cross-validated regularization parameter substracted by the RVRs of its competitors (I-PCA and C-PCA) on BCI data. The samples between the 25% and the 75% quantiles are summarized as a blue box and the rest are shown as red + symbols in each plot.

From these experiments, it is demonstrated that the proposed method with a regularization parameter automatically selected by cross-validation performs significantly better than I-PCA and C-PCA.

4. Conclusion

In this paper, we introduced a novel regularization term for orthogonal skinny matrices. Based on this regularization term, we provided a novel and elegant formulation of the multitask PCA problem. Using tools from the field of optimization on manifolds, we solved this problem, applied our method to synthetic and real-world data, and demonstrated its usefulness.

Table 1: Averages and standard errors of the RVRs on BCI data. The best and comparable to the best scores by the paired t-test (5% significance level) are shown in bold face.

		CV-MTL	Independent	Common
(Class 1)	$k = 1$	**0.7997(0.0001)**	0.7985(0.0002)	0.7987(0.0001)
	$k = 4$	**0.9670(0.0001)**	0.9666(0.0001)	0.9662(0.0001)
	$k = 7$	**0.9877(0.0000)**	0.9876(0.0000)	0.9866(0.0000)
	$k = 10$	**0.9945(0.0000)**	**0.9945(0.0000)**	0.9941(0.0000)
(Class 2)	$k = 1$	**0.7857(0.0001)**	0.7844(0.0003)	0.7844(0.0001)
	$k = 4$	**0.9655(0.0001)**	0.9651(0.0001)	0.9646(0.0000)
	$k = 7$	**0.9872(0.0000)**	0.9871(0.0000)	0.9859(0.0000)
	$k = 10$	**0.9943(0.0000)**	**0.9943(0.0000)**	0.9938(0.0000)

We only considered multitask learning in the scarce setting, but the proposed regularization can be applied to transfer learning and adaptation problems, whose goals are to improve the performance for a single target task utilizing the information from other similar tasks. Real-world examples where multitask principal component analysis plays important roles include analysis of multi-country government bond returns (Pérignon et al., 2007) and preprocessing for learning biometric verification systems (Delac and Grgic, 2004). The particular usefulness of principal component analysis in face image processing with scarce samples is argued in Jafri and Arabnia (2009).

In future work, we consider several extensions of our method. We may cast our multitask dimensionality reduction to a supervised setup. Such an approach may be particularly useful for BCI applications. Other subspace methods such as locality preserving projections (He and Niyogi, 2004), Fisher's discriminant analysis (Fisher, 1936), and canonical correlation analysis (Hotelling, 1936) can be extended to multitask scenarios using the proposed regularization by replacing the sample covariance $\widehat{C}_t$ in Eq. (4) with appropriate symmetric matrices. In addition, it would be interesting to use our approach with different criteria in the spirit of Harandi et al. (2014); Horev et al. (2015), leading to a multitask Riemannian dimensionality reduction.

Acknowledgments

IY acknowledges KAKENHI 16J07970 and MS acknowledges KAKENHI 26280054.

References

Pierre-Antoine Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

Roland Badeau, Gaël Richard, and Bertrand David. Fast and stable YAST algorithm for principal and minor subspace tracking. *IEEE Transactions on Signal Processing*, 56(8): 3437–3446, 2008.

Laura Balzano, Robert Nowak, and Benjamin Recht. Online identification and tracking of subspaces from highly incomplete information. In *48th Annual Allerton Conference on Communication, Control, and Computing*, pages 704–711. IEEE, 2010.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Nicolas Boumal, Bamdev Mishra, Pierre-Antoine Absil, and Rodolphe Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014. URL http://www.manopt.org.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.

Jean-Francois Cardoso and Antoine Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 17(1):161–164, 1996.

Rich Caruana. Multitask learning. In *Learning to Learn*, pages 95–133. Springer, 1998.

Sylvain Chevallier, Quentin Barthélemy, and Jamal Atif. Metrics for multivariate dictionaries. *arXiv preprint arXiv:1302.4242*, 2013.

Tijl De Bie, Nello Cristianini, and Roman Rosipal. Eigenproblems in pattern recognition. In *Handbook of Geometric Computing*, pages 129–167. Springer, 2005.

Kresimir Delac and Mislav Grgic. A survey of biometric recognition methods. In *46th International Symposium on Electronics in Marine*, pages 184–193, June 2004.

Dieter Devlaminck, Bart Wyns, Moritz Grosse-Wentrup, Georges Otte, and Patrick Santens. Multisubject learning for common spatial patterns in motor-imagery BCI. *Computational Intelligence and Neuroscience*, 2011:8, 2011.

Susan J. Devlin, Ramanathan Gnanadesikan, and Jon R. Kettenring. Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76(374):354–362, 1981.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2): 303–353, 1998.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multi-task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117. ACM, 2004.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

Bernhard N. Flury and Walter Gautschi. An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM Journal on Scientific and Statistical Computing*, 7(1):169–184, 1986.

Mehrtash T. Harandi, Mathieu Salzmann, and Richard Hartley. From manifold to manifold: Geometry-aware dimensionality reduction for SPD matrices. In *European Conference on Computer Vision*, pages 17–32. Springer International Publishing, 2014.

Xiaofei He and Partha Niyogi. Locality preserving projections. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 153–160. MIT Press, 2004.

Inbal Horev, Florian Yger, and Masashi Sugiyama. Geometry-aware principal component analysis for symmetric positive definite matrices. In *Proceedings of The 7th Asian Conference on Machine Learning*, pages 1–16, 2015.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417, 1933.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Laurent Jacob, Jean-Philippe Vert, and Francis R. Bach. Clustered multi-task learning: A convex formulation. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 745–752. Curran Associates, Inc., 2009.

Rabia Jafri and Hamid R. Arabnia. A survey of face recognition techniques. *Journal of Information Processing Systems*, 5(2):41–68, 2009.

Ian Joliffe. *Principal Component Analysis*. Springer, 1986.

Yi Ma, Jana Košecká, and Shankar Sastry. Optimization criteria and geometric algorithms for motion and structure estimation. *International Journal of Computer Vision*, 44(3): 219–249, 2001.

Bruce Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26(1):17–32, 1981.

Muhammad Naeem, Clemens Brunner, Robert Leeb, Bernhard Graimann, and Gert Pfurtscheller. Seperability of four-class motor imagery data using independent components analysis. *Journal of Neural Engineering*, 3(3):208, 2006.

Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

Christophe Pérignon, Daniel R. Smith, and Christophe Villa. Why common factors in international bond returns are not so common. *Journal of International Money and Finance*, 26(2):284–304, 2007.

Wojciech Samek, Frank C. Meinecke, and Klaus-Robert Müller. Transferring subspaces between subjects in brain-computer interfacing. *IEEE Transactions on Biomedical Engineering*, 60(8):2289–2298, 2013.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pages 583–588. Springer, 1997.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

Boyu Wang, Joelle Pineau, and Borja Balle. Multitask generalized eigenvalue program. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2115–2121, 2016.

Huahua Wang, Arindam Banerjee, and Daniel Boley. Common component analysis for multiple covariance matrices. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 956–964. ACM, 2011.

Manfred K Warmuth and Dima Kuzmin. Randomized PCA algorithms with regret bounds that are logarithmic in the dimension. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1481–1488. MIT Press, 2007.

Florian Yger. A review of kernels on covariance matrices for BCI applications. In *2013 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2013.

Florian Yger, Fabien Lotte, and Masashi Sugiyama. Averaging covariance matrices for EEG signal classification based on the CSP: an empirical study. In *23rd European Signal Processing Conference*, pages 2721–2725. IEEE, 2015.

Lei Zhang, Weisheng Dong, David Zhang, and Guangming Shi. Two-stage image denoising by principal component analysis with local pixel grouping. *Pattern Recognition*, 43(4): 1531–1549, 2010.

Yu Zhang and Dit-Yan Yeung. Multi-task learning in heterogeneous feature spaces. In *Proceedings of the National Conference on Artificial Intelligence*, 2011.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.