Stratified principal component analysis

Tom Szwagier, Xavier Pennec

Stratified principal component analysis

Tom Szwagier and Xavier Pennec

Université Côte d'Azur and Inria, Sophia Antipolis, France

Emails: tom.szwagier@inria.fr, xavier.pennec@inria.fr

Abstract

This paper investigates a general family of covariance models with repeated eigenvalues extending probabilistic principal component analysis (PPCA). A geometric interpretation shows that these models are parameterised by flag manifolds and stratify the space of covariance matrices according to the sequence of eigenvalue multiplicities. The subsequent analysis sheds light on PPCA and answers an important question on the practical identifiability of individual eigenvectors. It notably shows that one rarely has enough samples to fit a covariance model with distinct eigenvalues and that block-averaging the adjacent sample eigenvalues with small gaps achieves a better complexity/goodness-of-fit tradeoff.

Key words: covariance model; eigenvalue multiplicity; flag manifold; parsimony; probabilistic principal component analysis; stratification

1. Introduction

Principal component analysis (PCA) [Jolliffe, 2002] is a well-known dimension reduction method that is based on the eigenvalue decomposition of the sample covariance matrix. Usually, after the decomposition, one plots the eigenvalue profile in decreasing order and decomposes it into two parts: the signal on the left and the noise on the right. The position of the separation relates to the so-called *intrinsic dimension* of the dataset [Shepard, 1962]. Such a decomposition can be done with simple rules relying on the shape of the profile, like the elbow method or the percentage of explained variance [Jolliffe, 2002]. Another large family of dimension selection methods relies on a generative modelling formulation of PCA, called *probabilistic principal component analysis (PPCA)* [Tipping and Bishop, 1999], which can be interpreted as a low-dimensional Gaussian model of the data up to an isotropic Gaussian noise. In such a framework, the choice of the intrinsic dimension is based on the *principle of parsimony* (also known as *Occam's razor*): the selected model is the one that has the lowest number of parameters, while still well representing the data distribution. Such a tradeoff can be achieved with model selection criteria such as the *Bayesian information criterion (BIC)* [Schwarz, 1978], which depends on the dataset size and promotes *low-complexity* beyond *goodness-of-fit* when the number of available samples is limited (*small-data*).

Due to the isotropic noise assumption, PPCA can be reinterpreted as a covariance model where the lowest eigenvalues are constrained to be all equal (cf. Section 2). This constraint greatly reduces the number of parameters with respect to the full covariance model, while not excessively lowering the approximation quality of the sample

covariance matrix, whose eigenvalues are almost-surely all distinct (see discussion in Appendix A.2). One may wonder however if such a complexity drop is enough, especially in the small-data regime. The eigenvalue-equality constraint could indeed naturally be extended to the signal space by equalising adjacent sample eigenvalues with small gaps, achieving a better complexity/goodness-of-fit tradeoff.

This motivates us to investigate a more general family of covariance models with repeated eigenvalues, which contains in particular PPCA. Those models, coined *stratified principal component analysis (SPCA)*, enjoy an explicit maximum likelihood estimate and a unifying geometric characterisation relying on flag manifolds. Such a geometric interpretation of SPCA enables us to answer a first key question on the inference of two adjacent eigenvalues and their associated eigenvectors. Among the outcomes, we get that a pair of adjacent eigenvalues with a relative gap lower than 21% needs at least 1000 data points to be distinguished. More precisely, if this condition is not met (which is often the case in real datasets), then a model with two equal eigenvalues and a two-dimensional eigenspace is more optimal in terms of BIC. To extend this result to more than two eigenvalues, we must perform model selection among the whole family of SPCA models. Since the number of candidate models grows exponentially with the data dimension, we are encouraged to design efficient model selection heuristics. This leads us to study the structure of the SPCA models, which are shown to have a particular hierarchy that is the one of a stratified family [Geiger et al., 2001], according to the multiplicities of the eigenvalues [Arnold, 1995, Groisser et al., 2017, Breiding et al., 2018]. The partial order induced by such a stratification enables us to design computationally efficient model selection heuristics, whose asymptotic consistency is moreover proven. The application of our model to synthetic and real datasets successfully shows that equalising groups of adjacent eigenvalues with small gaps is indeed relevant, especially when the number of available samples is limited. The experiments notably show that SPCA models achieve a better complexity/goodness-of-fit tradeoff than PPCA.

The paper is organised in the following way. In Section 2, we present the PPCA model, its maximum likelihood estimate and number of free parameters, as well as a parsimonious version of PPCA called *isotropic probabilistic principal component analysis (IPPCA)* [Bouveyron et al., 2011]. In Section 3, we introduce the SPCA model. We derive an explicit maximum likelihood estimate that boils down to an eigenvalue decomposition of the sample covariance matrix followed by a block-averaging of adjacent eigenvalues. We show that SPCA extends PPCA and IPPCA and comes with an insightful geometric unification relying on flag manifolds. This enables the accurate computation of the number of free parameters. In Section 4, we develop a model selection framework for SPCA, taking advantage of the partial order on the family of models induced by the stratification. This notably allows us to answer a key question on the distinguishability of adjacent sample eigenvalues. In Section 5, we compare PPCA and SPCA models on synthetic and real datasets and show the improvement brought by equalising adjacent eigenvalues with small gaps.

2. Probabilistic Principal Component Analysis

PCA is a ubiquitous tool in statistics, which however used to lack a statistical model formulation. Tipping and Bishop [1999] circumvented this issue by introducing PPCA that we describe in this section.

2.1. Model

Let $(\boldsymbol{x}_i)_{i=1}^n$ be a p-dimensional dataset and $q \in [0\,..\,p-1]$ a lower dimension. In PPCA, the observed data is assumed to stem from a q-dimensional latent variable via a linear-Gaussian model

$$\boldsymbol{x} = W\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{1}$$

with $\boldsymbol{z} \sim \mathcal{N}(0, I_q)$, $W \in \mathbb{R}^{p\times q}$, $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_p)$ and $\sigma^2 > 0$.

Through classical probability theory, one can show that the observed data is modeled as following a multivariate Gaussian distribution:

$$\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{\mu},\, WW^\top + \sigma^2 I_p\right). \tag{2}$$

An analysis of the covariance matrix reveals that the distribution is actually anisotropic on the first q dimensions and isotropic on the remaining $p-q$ ones. Hence there is an implicit constraint on the covariance model of the data, which is that the lowest $p-q$ eigenvalues are assumed to be all equal.

2.2. Maximum Likelihood

The PPCA model parameters are the shift $\boldsymbol{\mu}$, the linear map W and the noise factor σ^2. Let some observed dataset $(\boldsymbol{x}_i)_{i=1}^n$, $\overline{\boldsymbol{x}} := \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ its mean and $S := \sum_{j=1}^p \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j{}^\top$ its sample covariance matrix, with $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ its eigenvalues and $\boldsymbol{v}_1 \perp \cdots \perp \boldsymbol{v}_p$ some associated eigenvectors. One can explicitly infer the parameters that are the most likely to have generated these data using maximum likelihood estimation. Tipping and Bishop [1999] show that the most likely shift is the empirical mean, the most likely linear map is the composition of a scaling by the q highest eigenvalues $\Lambda_q := \operatorname{diag}(\lambda_1, \ldots, \lambda_q)$ (up to the noise) and an orthogonal transformation by the associated q eigenvectors $V_q := [\boldsymbol{v}_1 | \ldots | \boldsymbol{v}_q]$, and finally the most likely noise factor is the average of the $p-q$ discarded eigenvalues

$$\hat{\boldsymbol{\mu}} = \overline{\boldsymbol{x}}, \qquad \hat{W} = V_q\left(\Lambda_q - \hat{\sigma}^2 I_q\right)^{\frac{1}{2}}, \qquad \hat{\sigma}^2 = \frac{1}{p-q}\sum_{j=q+1}^{p} \lambda_j. \tag{3}$$

One can then easily express the maximum log-likelihood

$$\ln \hat{\mathcal{L}} := -\frac{n}{2}\left(p\ln(2\pi) + \sum_{j=1}^{q}\ln\lambda_j + (p-q)\ln\left(\frac{1}{p-q}\sum_{j=q+1}^{p}\lambda_j\right) + p\right). \tag{4}$$

2.3. Parsimony and model selection

The previously described PPCA is already a somewhat parsimonious statistical model. Indeed, it not only makes the assumption that the observed data follows

a multivariate Gaussian distribution, which is the entropy-maximising distribution at a fixed mean and covariance, but it also reduces the number of covariance parameters by constraining the last $p-q$ eigenvalues to be equal. The covariance matrix $\Sigma := WW^\top + \sigma^2 I_p$ is parameterised by $W \in \mathbb{R}^{p\times q}$ and σ^2. It is shown in Tipping and Bishop [1999] to have $\kappa := pq - \frac{q(q-1)}{2} + 1$ free parameters—the removal of $\frac{q(q-1)}{2}$ parameters being due to the invariance of the latent variable distribution to a rotation. Although not evident at first sight with this expression of κ, we have a drop of complexity—with respect to the full covariance model which is of dimension $\frac{p(p+1)}{2}$—due to the equality constraint on the low eigenvalues, and the number of parameters decreases along with q. As shown in Subsection 3.4, we can give an insightful geometric interpretation to the number of free parameters in the PPCA model using Stiefel manifolds [Edelman et al., 1998].

For a given data dimension p, a PPCA model is indexed by its latent variable dimension $q \in [0\,..\,p-1]$. The process of model selection then consists in comparing different PPCA models and choosing the one that optimises a criterion, like the BIC [Schwarz, 1978] or more PPCA-oriented ones [Bishop, 1998, Minka, 2000]. They often rely on a tradeoff between goodness-of-fit (via maximum likelihood) and complexity (via the number of parameters), ponderated by the number of samples.

2.4. Isotropic Probabilistic Principal Component Analysis

IPPCA [Bouveyron et al., 2011] is an even more constrained covariance model with only two distinct eigenvalues. For $a > b$ and $U \in \mathbb{R}^{p\times q}$ such that $U^\top U = I_q$, one defines it as

$$\Sigma := (a-b)\, UU^\top + bI_p. \tag{5}$$

Such a parsimonious model is shown to be efficient in high-dimensional classification problems [Bouveyron and Girard, 2009]. The authors derive the maximum likelihood of such a model, which is highly related to the one of PPCA, where this time the q first sample covariance eigenvalues are also averaged to fit the model. They also show that the maximum likelihood criterion alone is surprisingly asymptotically consistent for selecting the true intrinsic dimension under the assumptions of IPPCA.

3. Stratified Principal Component Analysis

Inspired by the complexity drop induced by the isotropy in the noise space in PPCA, we aim at investigating more general isotropy constraints on the full data space. In this section, we introduce SPCA, a covariance model with a general constraint on the sequence of eigenvalue multiplicities. SPCA generalises PPCA and IPPCA and unifies them in a new family of models parameterised by flag manifolds [Monk, 1959]. Flag manifolds are themselves generalisations of Stiefel manifolds and Grassmannians [Edelman et al., 1998], hence the link between PPCA, IPPCA and SPCA that is detailed in this section.

3.1. Model

We recall that in combinatorics, a *composition* of an integer p is an ordered sequence of positive integers that sums up to p. It has to be distinguished from a *partition* of an integer, which doesn't take into account the ordering of the parts.

Let $\boldsymbol{\gamma} := (\gamma_1, \gamma_2, \ldots, \gamma_d) \in \mathcal{C}(p)$ be a composition of a positive integer p. We define the SPCA model of *type* $\boldsymbol{\gamma}$, noted $\boldsymbol{\gamma}$-SPCA, as

$$\boldsymbol{x} = \sum_{k=1}^{d-1} \sigma_k \, U_k \boldsymbol{z}_k + \boldsymbol{\mu} + \boldsymbol{\epsilon}. \tag{6}$$

In this formula, $\sigma_1 > \cdots > \sigma_{d-1} > 0$ are decreasing scaling factors, $U_k \in \mathbb{R}^{p \times \gamma_k}$ verify ${U_k}^\top U_{k'} = \delta_{kk'} I$ (in Kronecker notation) and $\boldsymbol{z}_k \sim \mathcal{N}(0, I_{\gamma_k})$ are independent latent variables. $\boldsymbol{\mu} \in \mathbb{R}^p$, $\sigma^2 > 0$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_p)$ are the classical shift, variance and isotropic noise present in PPCA.

Similarly as for PPCA, we can compute the population density

$$\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \sum_{k=1}^{d-1} {\sigma_k}^2 \, U_k \, {U_k}^\top + \sigma^2 I_p\right). \tag{7}$$

The expression of the covariance matrix $\Sigma := \sum_k {\sigma_k}^2 \, U_k \, {U_k}^\top + \sigma^2 I_p \in \mathbb{R}^{p \times p}$ can be simplified by gathering all the orthonormal frames into one orthogonal matrix $Q := [U_1 | \ldots | U_{d-1} | U_d] \in \mathcal{O}(p)$ where $U_d \in \mathbb{R}^{p \times \gamma_d}$ is an orthogonal completion of the previous frames. Writing $L := \operatorname{diag}(\ell_1 I_{\gamma_1}, \ldots, \ell_d I_{\gamma_d})$, with $\ell_k := {\sigma_k}^2 + \sigma^2$ for $k \in [1 \,..\, d-1]$ and $\ell_d := \sigma^2$, one gets

$$\Sigma = Q L Q^\top. \tag{8}$$

Hence, the fitted density of $\boldsymbol{\gamma}$-SPCA is a multivariate Gaussian with repeated eigenvalues $\ell_1 > \cdots > \ell_d > 0$ of respective multiplicity $\gamma_1, \ldots, \gamma_d$. An illustration of the generative model is provided in Figure 1. Therefore, PPCA and IPPCA can be seen as SPCA models, with respective types $\boldsymbol{\gamma} = (1, \ldots, 1, p-q)$ and $\boldsymbol{\gamma} = (q, p-q)$. From a geometric point of view, the fitted density is isotropic on the eigenspaces of Σ, which constitute a sequence of mutually orthogonal subspaces of respective dimension $\gamma_1, \ldots, \gamma_d$, whose direct sum generates the data space. Such a sequence is called a *flag* of linear subspaces of *type* $\boldsymbol{\gamma}$ [Monk, 1959]. Hence flags are natural objects to geometrically interpret SPCA, and so a fortiori PPCA and IPPCA. We detail this point later in Subsection 3.4.

3.2. Type

Just like the latent variable dimension $q \in [1 \,..\, p-1]$ is a central notion in PPCA, the type $\boldsymbol{\gamma} \in \mathcal{C}(p)$ is a central notion in SPCA. In this subsection, we introduce the concepts of *refinement* and $\boldsymbol{\gamma}$-*composition* to make its analysis more convenient.

Let $\boldsymbol{\gamma} := (\gamma_1, \gamma_2, \ldots, \gamma_d) \in \mathcal{C}(p)$. We say that $\boldsymbol{\gamma}' \in \mathcal{C}(p)$ is a *refinement* of $\boldsymbol{\gamma}$, and note $\boldsymbol{\gamma} \preceq \boldsymbol{\gamma}'$, if we can write $\boldsymbol{\gamma}' := (\boldsymbol{\gamma}'_1, \boldsymbol{\gamma}'_2, \ldots, \boldsymbol{\gamma}'_d)$, with $\boldsymbol{\gamma}'_k \in \mathcal{C}(\gamma_k), \forall k \in [1 \,..\, d]$. For instance, one has $(2,3) \preceq (1,1,2,1)$, while one does *not* have $(2,3) \preceq (3,2)$.

Let $\boldsymbol{\gamma} := (\gamma_1, \gamma_2, \ldots, \gamma_d) \in \mathcal{C}(p)$. Then each integer between 1 and p can be uniquely assigned a *part* of the composition, indexed between 1 and d. We

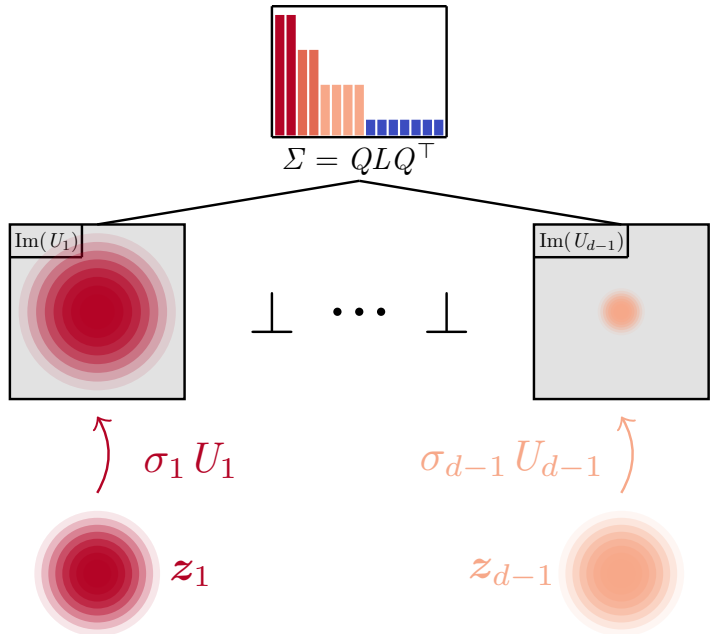

Fig. 1: SPCA generative model (6), assuming that the observed data was first sampled from a sequence of independent lower dimensional normal latent variables, then linearly mapped to mutually orthogonal subspaces and finally shifted and added an isotropic Gaussian noise. The resulting density is a multivariate Gaussian with repeated eigenvalues, whose multiplicities are given by the type $\boldsymbol{\gamma} = (2, 2, 4, 7)$.

define the γ-*composition function* $\phi_{\boldsymbol{\gamma}} \colon [1, p] \to [1, d]$ to be this surjective map, such that $\phi_{\boldsymbol{\gamma}}(j)$ is the index k of the part the integer j belongs to. For instance, one has $\phi_{(2,3)}(1) = \phi_{(2,3)}(2) = 1$ and $\phi_{(2,3)}(3) = \phi_{(2,3)}(4) = \phi_{(2,3)}(5) = 2$. Then, intuitively and with slight abuse of notation, each object of size p can be partitioned into d sub-objects of respective size γ_k, for $k \in [1\,..\,d]$. We call it the $\boldsymbol{\gamma}$-*composition* of an object. We give two examples. Let $Q \in \mathcal{O}(p)$. The $\boldsymbol{\gamma}$-composition of Q is the sequence $Q^{\boldsymbol{\gamma}} := (Q_1, \dots, Q_d)$ such that $Q_k \in \mathbb{R}^{p \times \gamma_k}, \forall k \in [1\,..\,d]$ and $Q = [Q_1 | \dots | Q_d]$. Let $\boldsymbol{\lambda} := (\lambda_1, \dots, \lambda_p)$ be a sequence of decreasing eigenvalues. The $\boldsymbol{\gamma}$-composition of $\boldsymbol{\lambda}$ is the sequence $\boldsymbol{\lambda}^{\boldsymbol{\gamma}} := \left(\boldsymbol{\lambda}^1, \dots, \boldsymbol{\lambda}^d\right)$ such that $\boldsymbol{\lambda}^k \in \mathbb{R}^{\gamma_k}$, $\forall k \in [1\,..\,d]$ and $\boldsymbol{\lambda} = \left[\boldsymbol{\lambda}^1 | \dots | \boldsymbol{\lambda}^d\right]$. We call $\boldsymbol{\gamma}$-*averaging* of $\boldsymbol{\lambda}$ the sequence $\overline{\boldsymbol{\lambda}^{\boldsymbol{\gamma}}} := \left(\overline{\boldsymbol{\lambda}^1}, \dots, \overline{\boldsymbol{\lambda}^d}\right) \in \mathbb{R}^d$ of average eigenvalues in $\boldsymbol{\lambda}^{\boldsymbol{\gamma}}$.

3.3. Maximum Likelihood

Similarly as for PPCA, the log-likelihood of the model can be easily computed

$$\ln \mathcal{L}\left(\boldsymbol{\mu}, \Sigma\right) = -\frac{n}{2}\left(p \ln(2\pi) + \ln|\Sigma| + \operatorname{tr}\left(\Sigma^{-1} C\right)\right), \tag{9}$$

with $C = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top$. We now show that the maximum likelihood estimate for $\boldsymbol{\gamma}$-SPCA consists in the eigenvalue decomposition of the sample covariance matrix followed by a block-averaging of adjacent eigenvalues such that the imposed type $\boldsymbol{\gamma}$ is respected; in other words, a $\boldsymbol{\gamma}$-averaging of the eigenvalues. Just before, we naturally extend the notion of *type* to symmetric matrices, as the sequence of multiplicities of its ordered-descending-eigenvalues.

Theorem 1 *Let $(x_i)_{i=1}^n$ be a p-dimensional dataset, $\overline{\boldsymbol{x}} := \frac{1}{n}\sum_{i=1}^n x_i$ its mean and $S := \sum_{j=1}^p \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j{}^\top$ its sample covariance matrix, with $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ its eigenvalues and $[\boldsymbol{v}_1|\ldots|\boldsymbol{v}_p] := V \in \mathcal{O}(p)$ some associated eigenvectors.*
The maximum likelihood parameters of γ-SPCA are

$$\hat{\boldsymbol{\mu}} = \overline{\boldsymbol{x}}, \qquad \hat{Q} = V, \qquad \left(\hat{\ell}_1, \ldots, \hat{\ell}_d\right) = \overline{\boldsymbol{\lambda}^{\boldsymbol{\gamma}}}. \tag{10}$$

The parameters $\hat{\boldsymbol{\mu}}$ and $\hat{\ell}_1, \ldots, \hat{\ell}_d$ are unique. $\hat{Q}$ is not unique but the flag of linear subspaces generated by its $\boldsymbol{\gamma}$-composition almost surely is—more precisely, the flag is unique if and only if the type of S is a refinement of γ, which is almost sure.

Proof – The proof is given in Appendix A. It relies on optimisation and linear algebra. We emphasize that the almost-sure uniqueness of the solution comes from the null Lebesgue measure of the set of symmetric matrices with repeated eigenvalues. □

One can then easily express the maximum log-likelihood of $\boldsymbol{\gamma}$-SPCA

$$\ln \hat{\mathcal{L}} = -\frac{n}{2}\left(p\ln(2\pi) + \sum_{k=1}^{d}\gamma_k \ln \overline{\boldsymbol{\lambda}^k} + p\right). \tag{11}$$

3.4. Geometric interpretation with flag manifolds

As intuited in Subsection 3.1 and then proven in Theorem 1, the accurate parameter space for Q in $\boldsymbol{\gamma}$-SPCA is the space of flags of type $\boldsymbol{\gamma}$, noted $\mathrm{Flag}(\boldsymbol{\gamma})$. The geometry of such a set is well known [Monk, 1959]. $\mathrm{Flag}(\boldsymbol{\gamma})$ is a smooth quotient manifold, consisting in equivalence classes of orthogonal matrices

$$\mathrm{Flag}(\boldsymbol{\gamma}) \cong \mathcal{O}(p)/\left(\mathcal{O}(\gamma_1) \times \cdots \times \mathcal{O}(\gamma_d)\right). \tag{12}$$

This result enables the accurate computation of the number of parameters in SPCA. Let us just before note that the other parameters are $\boldsymbol{\mu} \in \mathbb{R}^p$ and $L \in \mathrm{D}(\boldsymbol{\gamma}) := \left\{\mathrm{diag}\left(\ell_1 I_{\gamma_1}, \ldots, \ell_d I_{\gamma_d}\right) \in \mathbb{R}^{p\times p} : \ell_1 > \cdots > \ell_d > 0\right\}$, which can be seen as a convex cone of $\mathbb{R}^d$.

Proposition 1 *The number of free parameters in $\boldsymbol{\gamma}$-SPCA is*

$$\kappa := p + d + \frac{p(p-1)}{2} - \sum_{k=1}^{d}\frac{\gamma_k(\gamma_k - 1)}{2}. \tag{13}$$

This geometric interpretation sheds light on PPCA, which—we remind—is a special case of SPCA with $\boldsymbol{\gamma} = (1, \dots, 1, p-q)$. First, as flags of type $(1, \dots, 1, p-q)$ belong to Stiefel manifolds (up to changes of signs), we can naturally parameterise PPCA models with those spaces, which is already commonly done in the literature [Minka, 2000]. Second, we can now see PPCA as removing $(p-q-1) + \frac{(p-q)(p-q-1)}{2}$ parameters with respect to the full covariance model by imposing an isotropy constraint on the noise space. SPCA then goes beyond the noise space and results in even more parsimonious models.

We can extend this analysis to the IPPCA model, which—we remind—is a special case of SPCA with $\boldsymbol{\gamma} = (q, p-q)$. Hence we can parameterise it with flags of type $(q, p-q)$, which belong to Grassmannians. With that in mind, we notice that our formula (13) differs from the one given in Bouveyron et al. [2011]. We think that they overestimated the number of free parameters by implicitly assuming eigenvectors living on Stiefel manifolds like in PPCA, whereas the isotropy in the signal space yields an additional rotational invariance which makes them actually live on Grassmannians. Therefore IPPCA is even more parsimonious than originally considered.

4. Model selection

As discussed in Appendix A.2, sample covariance matrices almost surely have distinct eigenvalues. This makes the full covariance model the most likely to have generated some observed data. However, it does not mean that the true parameters—that are the eigenvectors and the eigenvalues—can be individually precisely inferred, especially in the small-data regime. Hence, one can wonder if a covariance model with repeated eigenvalues and multidimensional eigenspaces would not be more robust. The results of the previous section enable us to provide a possible answer, through SPCA model selection. First, we study the inference of two adjacent eigenvalues and their associated eigenvectors. We show that when the eigenvalue gap is small and the number of samples is limited, one should rather equalise the eigenvalues and gather the associated eigenvectors in a multidimensional eigenspace. Second, to extend this result to more than two eigenvalues, we develop a general model selection framework based on the stratified structure of SPCA.

4.1. Bayesian information criterion

In this work, we focus on one simple model selection criterion to set up the ideas. The Bayesian information criterion is defined as

$$\mathrm{BIC} := \kappa \ln n - 2 \ln \hat{\mathcal{L}}, \tag{14}$$

where κ is the number of free parameters—computed in Proposition 1—and $\ln \hat{\mathcal{L}}$ is the maximum log-likelihood (11). By removing the constant variables within model selection (like p and n), we get the following proposition.

Proposition 2 *The SPCA model minimising the BIC is*

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma} \in \mathcal{C}(p)}{\arg\min} \left(d - \sum_{k=1}^{d} \frac{\gamma_k(\gamma_k - 1)}{2} \right) \frac{\ln n}{n} + \sum_{k=1}^{d} \gamma_k \ln \overline{\boldsymbol{\lambda}^k}. \tag{15}$$

From now on, we remove the shift parameter $\boldsymbol{\mu} \in \mathbb{R}^p$ because it has the same complexity across models, and rather consider SPCA as a covariance model, like done in Tipping and Bishop [1999].

4.2. Eigenvalue equalisation

Willing to better apprehend the dynamics of SPCA model selection, we lead the experiment of quantifying the BIC variation induced by the equalisation of two adjacent eigenvalues. More precisely and without loss of generality, we compare the BIC of a *full covariance model* $\boldsymbol{\gamma} = (1, \dots, 1)$ to the one of an *equalised covariance model* $\boldsymbol{\gamma}' = (1 \dots 1, 2, 1 \dots 1)$, where the eigenvalues λ_j and λ_{j+1} have been equalised.

Proposition 3 *Let $(x_i)_{i=1}^n$ be a p-dimensional dataset with n samples, $\lambda_j \geq \lambda_{j+1}$ two adjacent sample eigenvalues and $\delta_j := \frac{\lambda_j - \lambda_{j+1}}{\lambda_j}$ their* relative eigengap. *If*

$$\delta_j < 2 - 2\exp\left(2\frac{\ln n}{n}\right) + 2\sqrt{\exp\left(4\frac{\ln n}{n}\right) - \exp\left(2\frac{\ln n}{n}\right)} := \delta(n), \tag{16}$$

then the equalised covariance model has a lower BIC than the full one.

Proof – The proof is given in Appendix B.1. □

The *threshold function* $\delta(n)$ is represented in Figure 2. One can deduce for instance that if a pair of adjacent sample eigenvalues has a relative eigengap lower than 21%, then one needs at least 1000 samples to justify the use of a model with distinct eigenvalues. This is an important result since many real datasets do not fulfill this condition, as we will see in the next section. As far as we know, this is the first time that a study on the parsimony induced by the equalisation of two adjacent sample eigenvalues is performed. This is enabled by the very design of SPCA and the geometric interpretation of its parameter space, involving flag manifolds. We could extend this study to the equalisation of more than two eigenvalues, but it would not necessarily yield a condition as simple as the one of Proposition 3. Hence, in the following, we establish a general framework for SPCA model selection. We study the structure of the family of models and design efficient model selection heuristics.

4.3. Structure of the SPCA family

Given a dimension p, PPCA has p models, ranging from the isotropic Gaussian ($q = 0$) to the full covariance model ($q = p - 1$). We can naturally equip the set of PPCA models with the *less-than-or-equal* relation $\leq$ on the latent variable dimension q, which makes it a totally ordered set. The complexity of the model then increases with q (cf. Subsection 3.4). The characterisation of the SPCA family structure is a bit more technical, as it requires to study the hierarchy of types, involving the concept of integer composition. Fortunately, the structure of such sets has already been well studied in combinatorics [Bergeron et al., 1995]. Moreover, several works have shown and exploited the stratification of symmetric matrices according to the multiplicities of the eigenvalues [Arnold, 1995, Groisser et al., 2017, Breiding et al., 2018]. Hence, without proof, we can state the following result.

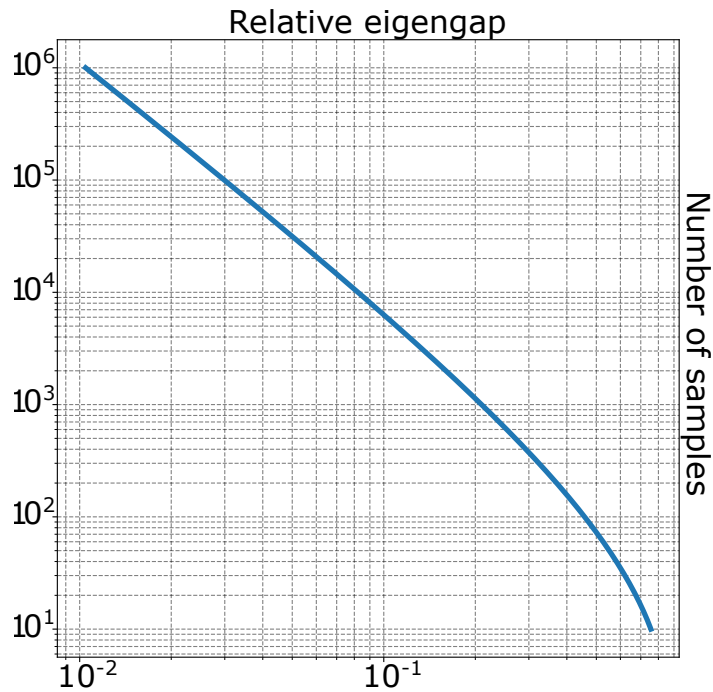

Fig. 2: Plot of the inverse threshold function δ^{-1} of Proposition 3, corresponding to the minimal number of samples needed to distinguish two adjacent eigenvalues separated by a given relative eigengap.

Proposition 4 *The family of p-dimensional SPCA models induces a stratification of the space of full-rank $p \times p$ covariance matrices according to the type $\boldsymbol{\gamma}$.*
The refinement relation $\preceq$ (3.2) makes it a partially ordered set of cardinal 2^{p-1}.

Hence the set of SPCA models at a given data dimension can be represented using a Hasse diagram, as done in Figure 3. We can see that SPCA contains PPCA, IPPCA, and many new models. SPCA therefore has the advantage of possibly providing more adapted models than PPCA and IPPCA, but also the drawback of requiring more comparisons for model selection. In high dimension this becomes quickly computationally heavy, so we need to define heuristics for selecting only a few number of models to compare. The previously derived partial order $\preceq$ on the set of SPCA models allows simple efficient heuristics for model selection.

4.4. Heuristics

In this subsection, we develop two simple heuristics for model selection. Their common idea is to a priori choose a subfamily of candidate models based on the shape of the eigenvalue profile, and then restrict the model selection process to this smaller subset.

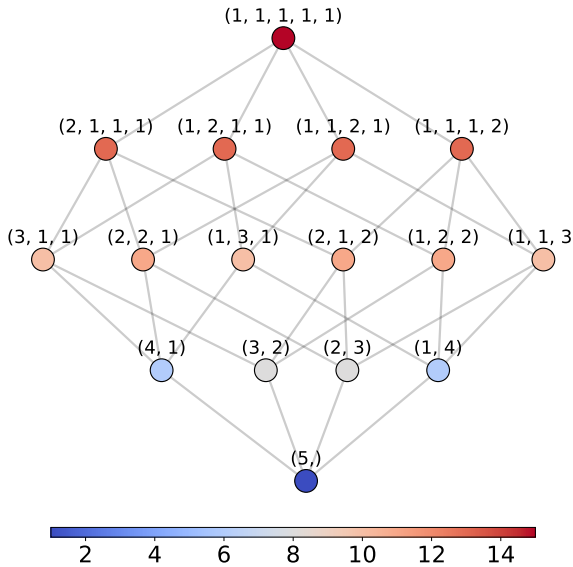

Fig. 3: Hasse diagram of 5-dimensional SPCA models. Each node represents a model. The associated label and color represent respectively the model type and its number of free parameters. The family contains 16 models: the isotropic Gaussian is the bottom node, the full covariance model is the top node, the five PPCA models are on the right side and the four IPPCA models are on the first floor.

4.4.1. Hierarchical clustering of eigenvalues

In this heuristic, the subset of candidate models is generated by the *hierarchical clustering* [Ward, 1963] of the sample eigenvalues. The general principle of hierarchical clustering is to agglomerate one by one the eigenvalues into clusters, thanks to a so-called *cluster-linkage criterion*, which is a measure of dissimilarity between clusters. Here, given two clusters of sample eigenvalues A, B and any *continuous* distance Δ (such as the relative eigengap defined in Proposition 3), we take as a cluster-linkage criterion the distance between the average eigenvalue in each cluster, $\Delta\left(\overline{A}, \overline{B}\right)$. The method is detailed in Algorithm 1 and illustrated in Figure 4. The hierarchical clustering heuristic creates a *trajectory* $(\boldsymbol{\gamma}^t)_{t=1}^p$ in the Hasse diagram of SPCA models. The sequence starts from $\boldsymbol{\gamma}^1 = (1, \ldots, 1)$, the full covariance model, in which each eigenvalue is in its own cluster. Then, one by one, the eigenvalues that are the closest in terms of distance Δ are agglomerated, and the inter-cluster distances are updated. The algorithm ends when one reaches the isotropic covariance model, $\boldsymbol{\gamma}^p = (p)$, in which all the eigenvalues are in the same cluster. This corresponds to an *agglomerative* approach in the hierarchical clustering vocabulary, in opposition to a *divisive* approach, that we could similarly develop for this heuristic.

The hierarchical clustering heuristic hence generates a subfamily of p models that can be then compared within a classical model selection framework. In order to assess the quality of such a heuristic, we show the following consistency result.

Algorithm 1 Hierarchical clustering heuristic for SPCA model selection

Require: $\lambda_1 \geq \cdots \geq \lambda_p, \Delta$ ▷ sample eigenvalues and distance
Ensure: $(\boldsymbol{\gamma}^t)_{t=1}^p$ ▷ subfamily of SPCA models
  $\boldsymbol{\gamma}^1 \leftarrow (1, \dots, 1)\,, \quad \boldsymbol{\lambda}^1 \leftarrow (\lambda_1, \dots, \lambda_p) := \boldsymbol{\lambda}$ ▷ full covariance model
  for $t = 1 \,..\, p-1$ **do**
    $\Delta^t \leftarrow (\Delta(\lambda_k^t, \lambda_{k+1}^t))_{k=1}^{p-t}$ ▷ distance between adjacent clusters
    $k^t = \arg\min \Delta^t$ ▷ minimal distance
    $\boldsymbol{\gamma}^{t+1} = (\gamma_1^t, \dots, \gamma_{k^t-1}^t, \gamma_{k^t}^t + \gamma_{k^t+1}^t, \gamma_{k^t+2}^t, \dots, \gamma_d^t)$ ▷ type agglomeration
    $\boldsymbol{\lambda}^{t+1} = \overline{\boldsymbol{\lambda}^{\boldsymbol{\gamma}^{t+1}}}$ ▷ $\boldsymbol{\gamma}$-averaging
  end for

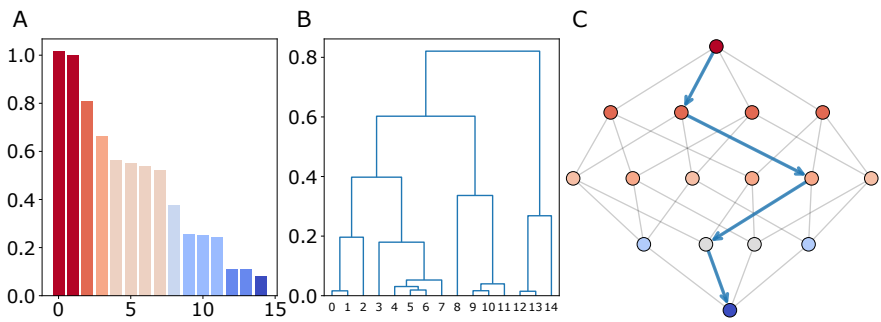

Fig. 4: Hierarchical clustering of sample eigenvalues, using the relative eigengap distance. (A) Sample eigenvalues, whose colors correspond to a given step $t = 8$ of the hierarchical clustering, with $\boldsymbol{\gamma}^t = (2, 1, 1, 4, 1, 3, 2, 1)$. (B) Hierarchical clustering dendrogram. (C) Conceptual view of the hierarchical clustering trajectory on the SPCA Hasse diagram.

Proposition 5 *The hierarchical clustering heuristic* (4.4.1) *is consistent, in the sense that almost surely, given enough samples, it generates a subfamily of SPCA models of size* p *that contains the true model.*

Proof – The proof is given in Appendix B.2. □

Hence, the hierarchical clustering heuristic generates a hierarchical subfamily of models of decreasing complexities, and provided enough data, the true model will be included. Thereafter, using consistent model selection criteria on this reduced subfamily, one can asymptotically recover the true model. We now propose a second heuristic that is not hierarchical but instead makes a prior assumption on the model complexity and then selects the one that has the maximum likelihood among all the candidates.

4.4.2. Prior on the length of the type

In this heuristic, we perform model selection at a given floor of the Hasse diagram (cf. Figure 3). More precisely, we consider as candidates only the models that have a given type length d, like done in IPPCA with $d = 2$. The type-length prior heuristic reduces the search space like the previous heuristic, this time to $\binom{p-1}{d-1}$ models. In contrast to the hierarchical clustering heuristic which creates a hierarchy of models with decreasing complexity, we here rather fix the complexity range of the candidate models, by working on one floor of the Hasse diagram, and then try to find the model of best fit.

Just like in the hierarchical clustering heuristic (4.4.1), we could use the BIC to choose the best model among this reduced family (as done in the second experiment of Subsection 5.2). For completeness, we provide an additional criterion that is nothing but the maximum likelihood itself. We indeed manage to extend to SPCA the surprising result from Bouveyron et al. [2011] that the maximum likelihood criterion alone asymptotically consistently finds the true intrinsic dimension within the IPPCA setting. Intuitively, this can be explained by the fact that we a priori fix the complexity of the candidate models and therefore we can focus on the other side of the weighing scale that is the goodness of fit. As this criterion empirically yields competitive results with respect to other classical model selection criteria in the large sample, low signal-to-noise ratio regime, we expect it to be of interest in SPCA as well.

Proposition 6 *The maximum likelihood is asymptotically consistent within the subfamily of SPCA models with a given type length d.*

Proof – The proof is given in Appendix B.3. We emphasize the use of Jensen's inequality, which elegantly generalises the proof of Bouveyron et al. [2011]. □

Hence we derived two simple heuristics for model selection, taking into account the structure of the SPCA models family. We now have all the tools needed for inference and model selection using SPCA.

5. Experiments

As seen in the previous sections, given a dataset and its sample covariance matrix, SPCA equalises the eigenvalues and gives rise to new multidimensional eigenspaces. This causes an additional drop of complexity with respect to PPCA which, according to Figure 2, seems justified when the eigenvalue gaps are small in view of the number of available samples. In this section, we confirm experimentally this hypothesis on some synthetic and real datasets.

5.1. Simpler models for all sample size

A key result in the previous section is that we rarely have enough available samples to confidently assert that two adjacent sample eigenvalues are distinct. Consequently, PPCA models could be made more parsimonious by equalising the adjacent sample eigenvalues with small gaps in the signal space as well.

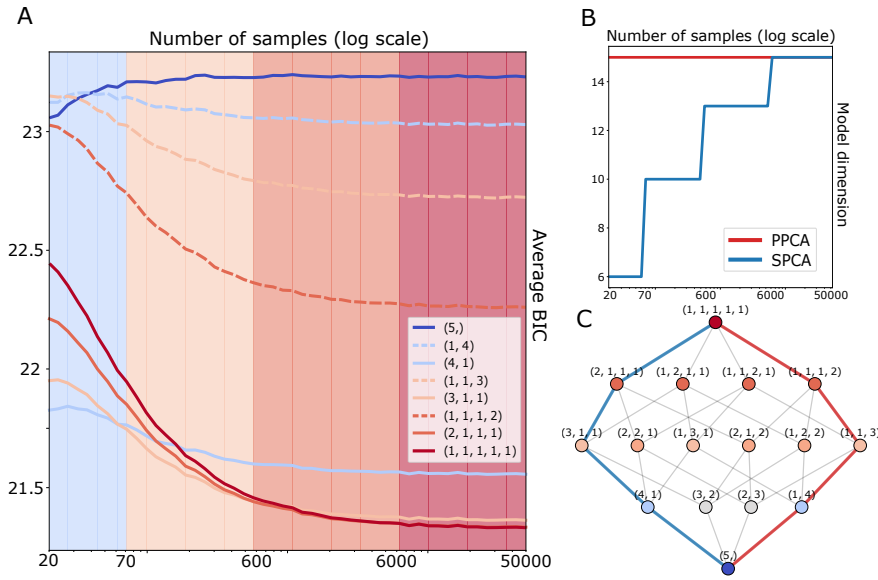

Fig. 5: SPCA model selection using the BIC for an increasing number of available samples. (A) Each curve represents the average BIC of a given SPCA model over several independent experiments. The lowest curve at a given n (horizontal coordinate) therefore corresponds to the most selected model. The curves corresponding to PPCA models are dashed. The curve color is related to the number of free parameters, from low (blue) to high (red). The background color then corresponds to the most selected model at a given sample size. For instance, we can see that for $n \in [20, 70]$ (light blue), the model that is the most selected is $\boldsymbol{\gamma} = (4, 1)$. For $n \in [70, 600]$ (light orange), it is $\boldsymbol{\gamma} = (3, 1, 1)$. For $n \in [600, 6000]$ (orange), it is $\boldsymbol{\gamma} = (2, 1, 1, 1)$. And for $n \in [6000, 50000]$ (red), it is $\boldsymbol{\gamma} = (1, 1, 1, 1, 1)$. (B) Comparison of the complexities of the mostly selected models within the whole SPCA family (blue) and within the PPCA family only (red). (C) SPCA Hasse diagram. The blue curve corresponds to the trajectory followed by the optimal SPCA selected model as the number of samples increases. We could expect that the PPCA models on the right follow the same kind of trajectory (in red), but it actually only stays on the top node as the other available models do not fit well the data distribution.

Willing to better understand how this result applies in practice, we make the following SPCA model selection experiment. We consider a given multivariate Gaussian population density, with covariance matrix eigenvalues $(10, 9, 7, 4, 0.5)$, and sample $n \in [20, 50000]$ data points from it. We fit all the SPCA models to this data distribution and select the one with the lowest BIC. The experiment is repeated several times independently for each n, and the results are reported on Figure 5, where we plot only a few models among the 16 for readability. First, on the BIC plots, we can see that for $n \leq 6000$, SPCA discloses a whole family of models that better explain the

Table 1. Comparison of PPCA and SPCA best models on several real datasets. To shrink long types, we use the power notation to indicate repetition of elements; for instance $(1, 1, 1, 2, 2, 3) := (1^3, 2^2, 3)$.

Dataset			PPCA		SPCA	
Name	n	p	γ	BIC	γ	BIC
Wine	48	13	$(1^3, 10)$	+36.35	$(8, 5)$	+35.57
Glass	17	9	(1^9)	−16.77	$(1, 2, 3, 1^3)$	−17.49
Ion	224	32	$(1^{30}, 2)$	−26.59	$(1^5, 2, 13, 6, 4, 2)$	−28.50
WDBC	357	30	(1^{30})	+25.12	$(2, 1, 2, 1, 2, 5, 1, 2, 1, 3^2, 4, 1^3)$	+24.72

observed data than PPCA. This shows that even for a very large number of samples with respect to the dimension, distinguishing the first eigenvalues and eigenvectors like PPCA does is not justified. Second, on the complexity plots, we can see that PPCA mostly selects the full covariance model for any sample size, while SPCA finds less complex models along the whole trajectory. Moreover, interestingly, we note the consistent increase of model complexity with the number of samples. We deduce that as the sample size increases, SPCA can more confidently distinguish the sample eigenvalues. Third, on the Hasse diagram, we can see that SPCA follows a trajectory as the number of available samples increases, which recalls the kind of subfamily generated by the hierarchical clustering heuristic (cf. Figure 4). To conclude, we see on this synthetic example that SPCA achieves a better complexity/goodness-of-fit tradeoff than PPCA in a wide range of sample sizes by equalising the highest eigenvalues.

5.2. Parsimony on real data

As the previous experiment was synthetic, we naturally wonder whether the same conclusions can be made out of real data. Indeed, as real datasets follow rather non-linear and multimodal distributions, the application of a simple linear-Gaussian model like SPCA to real datasets seems limited. However, PPCA has the same limits and remains quite used as a simple representation.

In this experiment, we compare PPCA to SPCA on several classical real datasets extracted from the open source UCI Machine Learning Repository: *Glass Identification* [German, 1987], *Ionosphere* [Sigillito et al., 1989], *Wine* [Aeberhard and Forina, 1991] and *Breast Cancer Wisconsin (WDBC)* [Wolberg et al., 1995]. Due to the high dimensionality of some datasets, we cannot perform an exhaustive comparison between all the SPCA models, therefore we use the hierarchical clustering heuristic introduced in Subsection 4.4 with the relative eigengap distance. As those datasets are made for classification problems, we keep only one class in order to make the data distribution more unimodal. For each dataset, we compare the best SPCA model to the best PPCA model (in terms of BIC). The results are reported in Table 1. We can see that for any dataset, SPCA achieves again a better complexity/goodness-of-fit tradeoff than PPCA. For instance, on the Wine dataset, PPCA finds a principal subspace of dimension 3 with distinct eigenvalues, while SPCA finds a principal subspace of dimension 8 with isotropic variability. For conciseness, we do not report the sample eigenvalue profiles of those datasets, but we can check that none of them

Table 2. Floor-by-floor comparison of PPCA and SPCA best models on the Glass dataset.

PPCA		SPCA	
γ	BIC	BIC	γ
$(9,)$	$+4.20$	$+4.20$	$(9,)$
$(1,8)$	-0.78	-8.21	$(8,1)$
$(1,1,7)$	-3.45	-15.92	$(3,5,1)$
$(1,1,1,6)$	-5.97	-16.93	$(3,3,2,1)$
$(1,1,1,1,5)$	-6.36	-17.38	$(1,2,3,2,1)$
$(1,1,1,1,1,4)$	-6.55	-17.49	$(1,2,3,1,1,1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(1,\dots\dots\dots,1)$	-16.77	-16.77	$(1,\dots\dots\dots,1)$

satisfies the relative eigengap condition of Proposition 3 given the number of available samples. Hence, the SPCA model is indeed justified for modelling real datasets.

In addition to the previous experiment, we also perform a floor-by-floor model comparison on the Glass dataset. More precisely, for a given type length $d \in [1\,..\,p]$, we compare the unique associated PPCA model ($q = d-1$) to the best SPCA model among the $\binom{p-1}{d-1}$. The results are reported on Table 2. We can see that the rich family of SPCA models with a prespecified number of distinct eigenvalues d importantly increases the modelling power of PPCA. For instance, the SPCA model of type $(8,1)$ has a lower BIC than the PPCA model of type $(1,8)$. This suggests that a principal subspace of dimension 8 with isotropic variability better models the data distribution than a principal subspace of dimension 1.

6. Discussion

We introduced in this paper a generative covariance model with repeated eigenvalues called SPCA, which generalises PPCA [Tipping and Bishop, 1999] and IPPCA [Bouveyron et al., 2011] under a unique geometric framework relying on flag manifolds. We noticed that the parsimony of PPCA comes from the low-rank model and the emergence of a multidimensional isotropic noise eigenspace. This raised the natural question of extending the isotropy constraint to the signal space. The SPCA model showed that assuming distinct eigenvalues in the signal space—as PPCA does—is not justified in practice. Hence, SPCA could circumvent this issue by equalising the adjacent eigenvalues with small gaps and gathering the associated eigenvectors into multidimensional eigenspaces. We confirmed our expectations on synthetic and real datasets, showing how SPCA models achieve a better complexity/goodness-of-fit tradeoff than PPCA. The code is available on GitHub[1].

SPCA is at an early stage of research and its development has been requiring several limiting choices that could be relaxed and improved in future works. A first limit is the choice of the BIC for model selection. Indeed, the BIC is known to

[1] https://github.com/tomszwagier/stratified-pca

favor under-parameterised models and not work very well in the small-data regime. However, this does not prevent it from being widely used due to its simplicity. Therefore, it provides an elementary way to highlight the interest of SPCA, similarly as Tipping and Bishop [1999] used a simple model selection criterion when introducing PPCA. One could later investigate extensions of Minka [2000] (which is relying on a geometric interpretation of PPCA with Stiefel manifolds) and Drton and Plummer [2017] to SPCA models. A second limit is the linear-Gaussian nature of SPCA which is not suited to real data. Some nonlinear and non-Gaussian extensions could therefore be considered in the future. The probable lack of analytic solution would involve optimisation on flag manifolds [Ye et al., 2021]. Due to the cost of inference for each model, we might need to replace discrete model selection with a global optimisation scheme on the space of all SPCA models. The latter being stratified by eigenvalue multiplicity, we could benefit from recent works on stratified optimisation [Leygonie et al., 2023, Olikier et al., 2023].

SPCA also comes with several interesting perspectives. First, it unleashes a whole new family of parsimonious linear-Gaussian models interpolating between the isotropic model and the full covariance one. Hence when a PPCA model overfits and the associated IPPCA model underfits, a better model might lie in the SPCA family. Second, the multidimensional eigenspaces obtained by gathering eigenvectors associated with distinct sample eigenvalues could provide robust, invariant and interpretable feature subspaces [Hyvärinen and Hoyer, 2000]. Indeed, just like the first eigenvectors can be interpreted as modes of variation [Castro et al., 1986], the eigenspaces inferred from SPCA could be interpreted as multidimensional attributes, and the norms of projection onto them as their level of expressiveness. Third, SPCA brings a statistical framework to the flag-based multiscale modeling of datasets. Indeed, several works use flags to represent datasets, be it in an independent component analysis [Nishimori et al., 2006] or principal component analysis [Ma et al., 2021] context, enriching the already well developed literature on Grassmannians and Stiefel manifolds for dimension reduction [Edelman et al., 1998]. In this paper, by introducing a generative model whose maximum likelihood estimate coincides with the minimiser of the *accumulated unexplained variance* criterion [Pennec, 2018], we enrich the previous works and enable for instance to perform flag-type selection. Fourth, beyond statistical modelling, SPCA provides a low-dimensional approximation of any symmetric matrix. Applications could therefore be investigated in spectral clustering [Ng et al., 2001] and shape analysis [Lefevre et al., 2023], where repeated eigenvalues in the graph Laplacian are prone to occur, as well as in variational Bayesian methods, where parsimonious Gaussian models can be used to approximate posterior distributions of parameters [Lambert et al., 2023].

Acknowledgements

This work was supported by the ERC grant #786854 G-Statistics from the European Research Council under the European Union's Horizon 2020 research and innovation program and by the French government through the 3IA Côte d'Azur Investments ANR-19-P3IA-0002 managed by the National Research Agency.

A. Proof of Theorem 1 (Maximum likelihood of SPCA)

We successively find the optimal $\hat{\boldsymbol{\mu}} \in \mathbb{R}^p$, $\hat{Q} \in \mathcal{O}(p)$ and $\hat{\ell}_k \in \mathbb{R}$.

A.1. Expression of $\hat{\boldsymbol{\mu}}$

The log-likelihood expresses as a function of $\boldsymbol{\mu} \in \mathbb{R}^p$ in the following way

$$\ln \mathcal{L}(\boldsymbol{\mu}) = -\frac{n}{2} \operatorname{tr}\left(\Sigma^{-1} C\right) + \text{constant} \tag{17}$$

with $C = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top$. The optimal shift $\hat{\boldsymbol{\mu}}$ is thus

$$\hat{\boldsymbol{\mu}} = \operatorname*{arg\,min}_{\boldsymbol{\mu} \in \mathbb{R}^p} \sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) := f(\boldsymbol{\mu}). \tag{18}$$

The gradient of $\boldsymbol{x} \mapsto (\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ is $\boldsymbol{x} \mapsto 2\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$. Hence, setting the gradient of f to 0 at $\hat{\boldsymbol{\mu}}$, one gets $\sum_i 2\Sigma^{-1}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}) = 0$, whose solution is $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$.

Hence $\hat{C}$ is actually the sample covariance matrix of the dataset, which will be denoted S (as in the theorem statement) from now on.

A.2. Expression of $\hat{Q}$

The log-likelihood expresses as a function of Q in the following way

$$\ln \mathcal{L}(Q) = -\frac{n}{2}\left(\ln|\Sigma| + \operatorname{tr}\left(\Sigma^{-1} S\right)\right) + \text{constant} \tag{19}$$

with $\Sigma = QLQ^\top$. Hence $|\Sigma|$ is independent of Q and the optimal orthogonal transformation $\hat{Q}$ is

$$\hat{Q} = \operatorname*{arg\,min}_{Q \in \mathcal{O}(p)} \operatorname{tr}\left(\Sigma^{-1} S\right) = \operatorname{tr}\left(QL^{-1}Q^\top S\right) := g(Q). \tag{20}$$

As g is a smooth function on $\mathcal{O}(p)$ which is a compact manifold, $\hat{Q}$ exists and $dg_{\hat{Q}} \colon \mathcal{T}_{\hat{Q}}(\mathcal{O}(p)) \ni \delta \mapsto \operatorname{tr}\left(\left(\delta L^{-1}\hat{Q}^\top + \hat{Q}L^{-1}\delta^\top\right) S\right) \in \mathbb{R}$ vanishes. It is known that $\mathcal{T}_{\hat{Q}}(\mathcal{O}(p)) = \operatorname{Skew}_p \hat{Q}$, therefore one has for all $A \in \operatorname{Skew}_p$

$$dg_{\hat{Q}}(A\hat{Q}) = \operatorname{tr}\left(\left((A\hat{Q})L^{-1}\hat{Q}^\top + \hat{Q}L^{-1}(A\hat{Q})^\top\right) S\right) = \operatorname{tr}\left(A(\Sigma^{-1}S - S\Sigma^{-1})\right) = 0. \tag{21}$$

Therefore $\Sigma^{-1}S - S\Sigma^{-1} = 0$. Hence, S and Σ^{-1} are two symmetric matrices that commute, so they must be simultaneously diagonalisable in an orthonormal basis. Since the trace is basis-invariant, g simply rewrites as a function of the eigenvalues

$$g(Q) = \sum_{k=1}^d \ell_k^{-1} \left(\sum_{j \in \phi_{\boldsymbol{\gamma}}^{-1}(\{k\})} \lambda_{\psi(j)} \right), \tag{22}$$

where $\psi \in S_p$ is a permutation and $\phi_{\boldsymbol{\gamma}}^{-1}(\{k\})$ is the set of indexes in the k-th part of the composition $\boldsymbol{\gamma}$ (cf. Subsection 3.2). We now need to find the permutation $\hat{\psi} \in S_p$ that minimises g. First, since $\ell_1 > \cdots > \ell_d > 0$ by assumption, then

$\left(\ell_k^{-1}\right)_{k=1}^d$ is an increasing sequence. Therefore, $\left(\lambda_{\hat{\psi}\left(\phi_\gamma^{-1}\{k\}\right)}\right)_{k=1}^d$ must be a non-increasing sequence, in that for $k_1 < k_2$, the eigenvalues in the k_1-th part of $\boldsymbol{\gamma}$ must be greater than or equal to the eigenvalues in the k_2-th part. Indeed, for $\ell < \ell'$, if $\lambda < \lambda'$, then $\ell\lambda' + \ell'\lambda < \ell\lambda + \ell'\lambda'$. Second, for such a $\hat{\psi}$ sorting the eigenvalues in non-increasing order in between parts, we can easily check that the inequality between eigenvalues of distinct parts is strict if and only if the type of Σ is a refinement of $\boldsymbol{\gamma}$. If so, the minimising $\hat{\psi}$ is unique up to permutations within each part of $\boldsymbol{\gamma}$. Therefore, it is not $\hat{Q}$ itself but the sequence of eigenspaces of $\hat{Q}$ generated by its $\boldsymbol{\gamma}$-composition (cf. Subsection 3.2) that is unique, and we have $(\mathrm{Im}(\hat{Q}_1), \ldots, \mathrm{Im}(\hat{Q}_d)) = (\mathrm{Im}(V_1), \ldots, \mathrm{Im}(V_d))$. Hence, the accurate space to describe the parameter $\hat{Q}$ is actually the space of flags of type γ.

An important remark is that the uniqueness condition will almost surely be met. Indeed, the set of $p \times p$ symmetric matrices with repeated eigenvalues has null Lebesgue measure (it is a consequence of Sard's theorem applied to the discriminant polynomial function [Breiding et al., 2018]). Therefore, for $n \geq p$ and any density with respect to Lebesgue measure on the set of sample covariance matrices, a randomly drawn matrix S almost surely has distinct eigenvalues. Consequently, its type is $(1, \ldots, 1)$, which is a refinement of any possible type in $\mathcal{C}(p)$.

A.3. Expression of $\hat{L}$

The log-likelihood expresses as a function of L in the following way

$$\ln \mathcal{L}(L) = -\frac{n}{2}\left(\ln|\Sigma| + \mathrm{tr}\left(\Sigma^{-1}S\right)\right) + \text{constant} \tag{23}$$

with $\Sigma = \hat{Q}L\hat{Q}^\top$. First, one has $\ln|\Sigma| = \sum_{k=1}^d \gamma_k \ln \ell_k$. Second, according to the previous results, one has $\mathrm{tr}\left(\Sigma^{-1}S\right) = \sum_{k=1}^d \ell_k^{-1}\left(\sum_{j\in\phi_\gamma^{-1}\{k\}} \lambda_j\right)$. The optimal eigenvalues $\left(\hat{\ell}_1, \ldots, \hat{\ell}_d\right)$ are thus

$$\left(\hat{\ell}_1, \ldots, \hat{\ell}_d\right) = \underset{\ell_1,\ldots,\ell_d\in\mathbb{R}}{\arg\min} \sum_{k=1}^d \gamma_k \ln \ell_k + \ell_k^{-1}\left(\sum_{j\in\phi_\gamma^{-1}\{k\}} \lambda_j\right) := h(\ell_1, \ldots, \ell_d). \tag{24}$$

As $\frac{\partial h}{\partial \ell_k} = \frac{\gamma_k}{\ell_k} - \ell_k^{-2}\left(\sum_{j\in\phi_\gamma^{-1}\{k\}} \lambda_j\right)$, we get that $\hat{\ell}_k = \frac{1}{\gamma_k}\left(\sum_{j\in\phi_\gamma^{-1}\{k\}} \lambda_j\right)$.

B. Other proofs

B.1. Proof of Proposition 3 (Eigenvalue equalisation)

We compare the BIC of the full covariance model $\boldsymbol{\gamma} = (1, \ldots, 1)$ to the one of the equalised covariance model $\boldsymbol{\gamma}' = (1, \ldots, 1, 2, 1, \ldots 1)$ where the j-th eigenvalue has been equalised with the $j+1$-th. This boils down to studying the sign of the function

$\Delta\,\mathrm{BIC} := \mathrm{BIC}(\boldsymbol{\gamma}) - \mathrm{BIC}(\boldsymbol{\gamma}')$. One gets

$$\Delta\,\mathrm{BIC} = p\frac{\ln n}{n} + \sum_{k=1}^{p} \ln\lambda_k - (p-2)\,\frac{\ln n}{n} - \sum_{k\notin\{j,j+1\}} \ln\lambda_k - 2\ln\left(\frac{\lambda_j+\lambda_{j+1}}{2}\right) \tag{25}$$

$$= 2\frac{\ln n}{n} + \ln\lambda_j + \ln\lambda_{j+1} - 2\ln\left(\frac{\lambda_j+\lambda_{j+1}}{2}\right) \tag{26}$$

$$= 2\frac{\ln n}{n} + \ln\lambda_j + \ln\left(\lambda_j\,(1-\delta_j)\right) - 2\ln\left(\frac{\lambda_j\,(2-\delta_j)}{2}\right) \tag{27}$$

$$= 2\frac{\ln n}{n} + \ln\left(1-\delta_j\right) - 2\ln\left(1-\frac{\delta_j}{2}\right) \tag{28}$$

$$= 2\frac{\ln n}{n} - \ln\left(\frac{\left(1-\frac{\delta_j}{2}\right)^2}{1-\delta_j}\right). \tag{29}$$

Hence, one has

$$\Delta\,\mathrm{BIC} = 0 \iff \exp\left(2\frac{\ln n}{n}\right) = \frac{\left(1-\frac{\delta_j}{2}\right)^2}{1-\delta_j} \tag{30}$$

$$\iff \frac{\delta_j^2}{4} - \left(1-\exp\left(2\frac{\ln n}{n}\right)\right)\delta_j + 1 - \exp\left(2\frac{\ln n}{n}\right) = 0. \tag{31}$$

It is a polynomial equation whose positive solution is unique when $n \geq 1$ and is

$$\delta(n) := 2 - 2\exp\left(2\frac{\ln n}{n}\right) + 2\sqrt{\exp\left(4\frac{\ln n}{n}\right) - \exp\left(2\frac{\ln n}{n}\right)}. \tag{32}$$

B.2. Proof of Proposition 5 (Consistency of hierarchical clustering)

Let us assume that the true generative model is stratified with type $\boldsymbol{\gamma} \in \mathcal{C}(p)$. We can then write the population covariance matrix as $\Sigma = \sum_{k=1}^{d} \ell_k\, Q_k\, Q_k{}^\top$ with $\ell_1 > \cdots > \ell_d > 0$ and $Q := [Q_1|\ldots|Q_d] \in \mathcal{O}(p)$. Let n be the number of independent samples and $S_n := \sum_{j=1}^{p} \lambda_j(S_n)\boldsymbol{v}_j(S_n)\boldsymbol{v}_j(S_n)^\top$ with $\lambda_1 \geq \cdots \geq \lambda_p$ and $V := [\boldsymbol{v}_1|\ldots|\boldsymbol{v}_p] \in \mathcal{O}(p)$. According to [Bouveyron et al., 2011, Proposition 1] and [Tyler, 1981, Lemma 2.1 (i)], one then has almost surely, as n goes to infinity, $\lambda_j(S_n) \to \ell_{\phi_{\boldsymbol{\gamma}}(j)}$, where $\phi_{\boldsymbol{\gamma}}$ is the $\boldsymbol{\gamma}$-composition function (cf. Subsection 3.2). Hence for n large enough, by continuity of the distance function Δ, the gaps between eigenvalues in the same part of the $\boldsymbol{\gamma}$-composition will be arbitrarily close to 0, while the other will be arbitrarily close to the true values $\{\Delta\left(\ell_k, \ell_{k+1}\right), k \in [1\,..\,d-1]\}$, which are all positive. Hence the hierarchical clustering method will first agglomerate the eigenvalues that are in the same part of $\boldsymbol{\gamma}$, and second the distinct blocks, by increasing order of pairwise distance. The last model of the first phase will be exactly the true model.

Note that one can thereafter perform model selection within the reduced subfamily of SPCA models obtained by the hierarchical clustering heuristic and asymptotically recover the true model using a consistent criterion.

B.3. Proof of Proposition 6 (Consistency of maximum likelihood)

Let us assume that the true generative model is stratified with type $\boldsymbol{\gamma}^* := (\gamma_1^*, \dots, \gamma_d^*)$, of length d, and let $\ell_1 > \cdots > \ell_d > 0$ be the eigenvalues of the associated population covariance matrix. Then, similarly as in the previous proof, almost surely, asymptotically, the sample covariance matrix eigenvalues are the ones of the population covariance matrix. Hence, for any SPCA model of type $\boldsymbol{\gamma} := (\gamma_1, \dots, \gamma_d)$, the maximum likelihood writes

$$\ln \hat{\mathcal{L}} \sim -\frac{n}{2} \left(p \ln 2\pi + \sum_{k=1}^{d} \gamma_k \ln \left(\frac{1}{\gamma_k} \sum_{j \in \phi_{\boldsymbol{\gamma}}^{-1}\{k\}} \ell_{\phi_{\boldsymbol{\gamma}^*}(j)} \right) \right). \tag{33}$$

As n and p are fixed when we compare the models, they do not intervene in the model selection. Hence, the search of the optimal model in terms of maximum likelihood boils down to the following problem

$$\underset{\substack{\boldsymbol{\gamma} \in \mathcal{C}(p) \\ \#\boldsymbol{\gamma} = d}}{\arg\min} \sum_{k=1}^{d} \gamma_k \ln \left(\frac{1}{\gamma_k} \sum_{j \in \phi_{\boldsymbol{\gamma}}^{-1}\{k\}} \ell_{\phi_{\boldsymbol{\gamma}^*}(j)} \right) := f(\boldsymbol{\gamma}). \tag{34}$$

One has $f(\boldsymbol{\gamma}) = \sum_{k=1}^{d} \gamma_k \ln \left(\frac{1}{\gamma_k} \sum_{k'=1}^{d} c_{kk'} \ell_{k'} \right)$, where $c_{kk'}$ is the cardinal of the intersection of the k-th part of $\boldsymbol{\gamma}$ with the k'-th part of $\boldsymbol{\gamma}^*$. Then, by definition, one has $\sum_{k'=1}^{d} c_{kk'} = \gamma_k$ and $\sum_{k=1}^{d} c_{kk'} = \boldsymbol{\gamma}_{k'}^*$. Hence, using Jensen's inequality,

$$f(\boldsymbol{\gamma}) \geq \sum_{k=1}^{d} \gamma_k \left(\sum_{k'=1}^{d} \frac{c_{kk'}}{\gamma_k} \ln \ell_{k'} \right) = \sum_{k,k'=1}^{d} c_{kk'} \ln \ell_{k'} = \sum_{k'=1}^{d} \boldsymbol{\gamma}_{k'}^* \ln \ell_{k'} = f(\boldsymbol{\gamma}^*). \tag{35}$$

To conclude, asymptotically, $\boldsymbol{\gamma}^*$-SPCA is the most likely model. Hence, the maximum likelihood criterion alone finds the true model among the family of SPCA models with the same type length.

References

S. Aeberhard and M. Forina. Wine. UCI Machine Learning Repository, 1991. URL https://doi.org/10.24432/C5PC7J.

V. I. Arnold. Remarks on eigenvalues and eigenvectors of Hermitian matrices, berry phase, adiabatic connections and quantum Hall effect. *Selecta Mathematica*, 1(1): 1–19, Mar. 1995. URL https://doi.org/10.1007/BF01614072.

F. Bergeron, M. Bousquet-Melou, and S. Dulucq. Standard Paths in the Composition Poset. *Annales des sciences mathématiques du Québec*, 19(2):139–151, 1995.

C. Bishop. Bayesian PCA. In *Advances in Neural Information Processing Systems*, volume 11, 1998. URL https://papers.nips.cc/paper/1998/hash/c88d8d0a6097754525e02c2246d8d27f-Abstract.html.

C. Bouveyron and S. Girard. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, 42(11):2649–2658, Nov. 2009. URL https://doi.org/10.1016/j.patcog.2009.03.027.

C. Bouveyron, G. Celeux, and S. Girard. Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic PCA. *Pattern Recognition Letters*, 32(14): 1706–1713, Oct. 2011. URL https://doi.org/10.1016/j.patrec.2011.07.017.

P. Breiding, K. Kozhasov, and A. Lerario. On the Geometry of the Set of Symmetric Matrices with Repeated Eigenvalues. *Arnold Mathematical Journal*, 4(3):423–443, Dec. 2018. URL https://doi.org/10.1007/s40598-018-0095-0.

P. E. Castro, W. H. Lawton, and E. A. Sylvestre. Principal Modes of Variation for Processes with Continuous Sample Curves. *Technometrics*, 28(4):329–337, 1986. URL https://www.jstor.org/stable/1268982.

M. Drton and M. Plummer. A Bayesian Information Criterion for Singular Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(2): 323–380, Mar. 2017. URL https://doi.org/10.1111/rssb.12187.

A. Edelman, T. A. Arias, and S. T. Smith. The Geometry of Algorithms with Orthogonality Constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, June 1998. URL https://doi.org/10.1137/S0895479895290954.

D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified exponential families: Graphical models and model selection. *The Annals of Statistics*, 29(2):505–529, Apr. 2001. URL https://doi.org/10.1214/aos/1009210550.

B. German. Glass Identification. UCI Machine Learning Repository, 1987. URL https://doi.org/10.24432/C5WW2P.

D. Groisser, S. Jung, and A. Schwartzman. Geometric foundations for scaling-rotation statistics on symmetric positive definite matrices: minimal smooth scaling-rotation curves in low dimensions. *Electronic Journal of Statistics*, 11(1), Jan. 2017. URL https://doi.org/10.1214/17-EJS1250.

A. Hyvärinen and P. Hoyer. Emergence of Phase- and Shift-Invariant Features by Decomposition of Natural Images into Independent Feature Subspaces. *Neural Computation*, 12(7):1705–1720, July 2000. URL https://doi.org/10.1162/089976600300015312.

I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2002. URL http://link.springer.com/10.1007/b98835.

M. Lambert, S. Bonnabel, and F. Bach. The limited-memory recursive variational Gaussian approximation (L-RVGA). *Statistics and Computing*, 33(3):70, Apr. 2023. URL https://doi.org/10.1007/s11222-023-10239-x.

J. Lefevre, J. Fraize, and D. Germanaud. Perturbation of Fiedler Vector: Interest for Graph Measures and Shape Analysis. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information*, Lecture Notes in Computer Science, pages 593–601, Cham, 2023. Springer Nature Switzerland. URL https://doi.org/10.1007/978-3-031-38299-4_61.

J. Leygonie, M. Carrière, T. Lacombe, and S. Oudot. A gradient sampling algorithm for stratified maps with applications to topological data analysis. *Mathematical Programming*, Mar. 2023. URL https://doi.org/10.1007/s10107-023-01931-x.

X. Ma, M. Kirby, and C. Peterson. The Flag Manifold as a Tool for Analyzing and Comparing Sets of Data Sets. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4168–4177, Oct. 2021. URL https://doi.org/10.1109/ICCVW54120.2021.00465.

T. Minka. Automatic Choice of Dimensionality for PCA. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper_files/paper/2000/file/7503cfacd12053d309b6bed5c89de212-Paper.pdf.

D. Monk. The Geometry of Flag Manifolds. *Proceedings of the London Mathematical Society*, s3-9(2):253–286, 1959. URL https://doi.org/10.1112/plms/s3-9.2.253.

A. Ng, M. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper/2001/hash/801272ee79cfde7fa5960571fee36b9b-Abstract.html.

Y. Nishimori, S. Akaho, and M. D. Plumbley. Riemannian Optimization Method on the Flag Manifold for Independent Subspace Analysis. In *Independent Component Analysis and Blind Signal Separation*, pages 295–302, 2006. URL https://doi.org/10.1007/11679363_37.

G. Olikier, K. A. Gallivan, and P.-A. Absil. First-order optimization on stratified sets, Mar. 2023. URL http://arxiv.org/abs/2303.16040. Preprint.

X. Pennec. Barycentric Subspace Analysis on Manifolds. *The Annals of Statistics*, 46(6A):2711–2746, 2018. URL https://www.jstor.org/stable/26542880.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2): 461–464, 1978. URL https://www.jstor.org/stable/2958889.

R. N. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27(2):125–140, June 1962. URL https://doi.org/10.1007/BF02289630.

V. G. Sigillito, S. P. Wing, L. V. Hutton, and K. B. Baker. Ionosphere. UCI Machine Learning Repository, 1989. URL https://doi.org/10.24432/C5W01B.

M. E. Tipping and C. M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3): 611–622, 1999. URL https://www.jstor.org/stable/2680726.

D. E. Tyler. Asymptotic Inference for Eigenvectors. *The Annals of Statistics*, 9(4): 725–736, July 1981. URL https://doi.org/10.1214/aos/1176345514.

J. H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, Mar. 1963. URL https://doi.org/10.1080/01621459.1963.10500845.

W. H. Wolberg, O. L. Mangasarian, and W. N. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1995. URL https://doi.org/10.24432/C5DW2B.

K. Ye, K. S.-W. Wong, and L.-H. Lim. Optimization on flag manifolds. *Mathematical Programming*, June 2021. URL https://doi.org/10.1007/s10107-021-01640-3.