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 models that stratifies the space of covariance matrices by eigenvalue multiplicity. This family, coined Stratified Principal Component Analysis (SPCA), includes in particular Probabilistic PCA (PPCA) models, where the noise component is assumed to be isotropic. We provide an explicit maximum likelihood and a geometric characterization relying on flag manifolds. A key outcome of this analysis is that PPCA's parsimony—with respect to the full covariance model—is due to the eigenvalue-equality constraint in the noise space and the subsequent inference of a multidimensional eigenspace. The sequential nature of flag manifolds enables to extend this constraint to the signal space and bring more parsimonious models. Moreover, the stratification and the induced partial order on SPCA yield efficient model selection heuristics. Experiments on simulated and real datasets substantiate the interest of equalising adjacent sample eigenvalues when the gaps are small and the number of samples is limited. They notably demonstrate that SPCA models achieve a better complexity/goodness-of-fit tradeoff than PPCA.

Key words: Probabilistic principal component analysis; Parsimony; Eigenvalue multiplicity; Flag manifolds; Stratified space

1. Introduction

Principal Component Analysis (PCA) (Pearson, 1901) 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 (Thorndike, 1953) or the percentage of explained variance. However, those heuristics lack statistical foundation and do not depend on the size of the dataset. Some more statistically grounded dimension selection methods rely on a generative modelling formulation of PCA, called *Probabilistic PCA (PPCA)* (Tipping and Bishop, 1999). PPCA can be seen as a covariance model where the lowest eigenvalues, representing the noise, are all equal. The choice of the cutoff dimension is then rather based on the *principle of parsimony*: 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 favors *low-complexity* over *goodness-of-fit* when the number of available samples is limited (*small-data* regime). As the eigenvalues of full-rank sample covariance matrices are almost surely all distinct (see discussion in Appendix A.2), PPCA makes an error when modelling the sample covariance matrix with equal noise eigenvalues, but this error is balanced by the complexity drop. One may wonder however if such a complexity drop is enough, especially in the small-data regime. The eigenvalue-equalisation principle 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 that is stratified by eigenvalue multiplicity (Arnold, 1995). Those models, coined *Stratified PCA (SPCA)*, enjoy an explicit maximum likelihood estimate and a unifying geometric characterization relying on flag manifolds. SPCA enables to answer a first key question on the identifiability of two adjacent sample eigenvalues. 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, which is rarely satisfied by real datasets. To extend this result to more than two eigenvalues, we must perform model selection among the whole family of SPCA models, which contains PPCA. As the number of candidate models grows exponentially with the data dimension, we are encouraged to design non-greedy model selection heuristics. Fortunately, the stratification of the family of SPCA models and the induced partial order on the sequence of eigenvalue multiplicities enables 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 justified 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 organized 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 PPCA (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 groups of adjacent eigenvalues. We show that SPCA extends PPCA and IPPCA and comes with an insightful geometric interpretation 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. This one first enables to answer a key question on the distinguishability of two adjacent sample eigenvalues and second to go beyond two eigenvalues using heuristics based on the structure of the SPCA family. 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

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 *Probabilistic PCA (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 $\boldsymbol{z} \sim \mathcal{N}(0, I_q)$, $W \in \mathbb{R}^{p\times q}$, $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_p)$. An illustration of the generative model is provided in Figure 1.

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 multivariate 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 that is an equality constraint on the lowest $p-q$ eigenvalues.

2.2. Maximum Likelihood

The PPCA model parameters are the shift $\boldsymbol{\mu}$, the linear map W and the noise factor σ^2. Given some observed data $(\boldsymbol{x}_i)_{i=1}^n$, $\overline{\boldsymbol{x}} := \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ 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, we can explicitly infer the parameters that are the most likely to have generated these data using maximum likelihood estimation. The most likely shift is the empirical mean; the

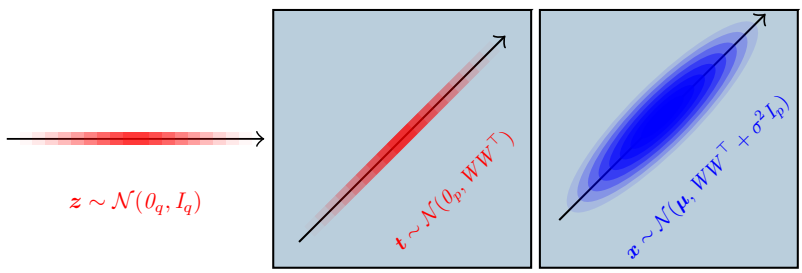


Fig. 1: PPCA generative model (1), assuming that the observed data was first sampled from a lower dimensional normal latent variable, then affinely mapped to the ambient space and finally added an isotropic Gaussian noise. *Left:* Latent variable. *Middle:* Linear map to data space. *Right:* Additive shift and noise.

most likely linear map is the composition of a scaling by the q highest eigenvalues $\Lambda_q := \operatorname{diag}(\lambda_1, \dots, \lambda_q)$ (up to the noise) and an orthogonal transformation by the associated q eigenvectors $V_q := [\boldsymbol{v}_1 | \dots | \boldsymbol{v}_q]$; 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-maximizing 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 parameterized 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 a more 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 optimizes a criterion, like the Bayesian Information Criterion 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), modulated by the number of samples.

2.4. Isotropic Probabilistic Principal Component Analysis

Isotropic PPCA (IPPCA) (Bouveyron et al., 2011) is an even more constrained covariance model with only two distinct eigenvalues. For $a > b$ and $U \in \operatorname{St}(p, q)$—the Stiefel manifold of $p \times q$ orthonormal frames—one defines it as

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

Such a 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 *Stratified PCA (SPCA)*, a covariance model with a general constraint on the sequence of eigenvalue multiplicities. SPCA generalizes PPCA and IPPCA and unifies them in a new family of models parameterized with flag manifolds (Monk, 1959). Flag manifolds are themselves generalizations 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 *Stratified PCA* 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, $U_k \in \mathrm{St}\,(p, \gamma_k)$ are mutually orthogonal frames (belonging to Stiefel manifolds) and $\boldsymbol{z}_k \sim \mathcal{N}\,(0_{\gamma_k}, I_{\gamma_k})$ are independent latent variables. $\boldsymbol{\mu} \in \mathbb{R}^p$, $\sigma^2 > 0$ and $\boldsymbol{\epsilon} \sim \mathcal{N}\,(0_p, \sigma^2 I_p)$ are the classical shift and isotropic noise present in PPCA. An illustration is provided in Figure 2.

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 $U_d \in \mathrm{St}\,(p, \gamma_d)$ is an orthogonal completion of the previous frames. Writing $L := \mathrm{diag}\,(\ell_1 I_{\gamma_1}, \ldots, \ell_d I_{\gamma_d})$, with $\ell_k := \sigma_k^2 + \sigma^2$ for $k \in [1 \,..\, d-1]$ and $\ell_d := \sigma^2$, one gets

$$\Sigma = QLQ^\top. \tag{8}$$

Hence, the fitted density of $\boldsymbol{\gamma}$-SPCA is a multivariate Gaussian with repeated eigenvalues $\ell_1 > \cdots > \ell_d > 0$ of respective multiplicity $\gamma_1, \ldots, \gamma_d$. 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 interprete SPCA, and so a fortiori PPCA and IPPCA. We detail this point later in Subsection 3.4.

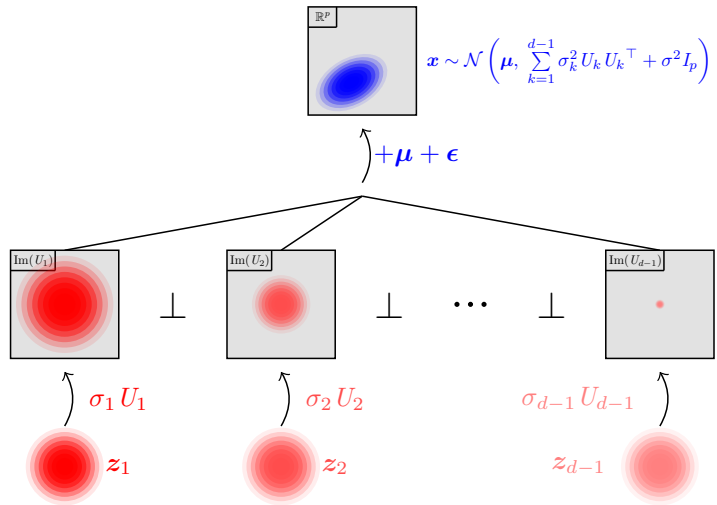


Fig. 2: $\boldsymbol{\gamma}$-SPCA generative model (6), assuming that the observed data was first sampled from a sequence of independent lower dimensional normal latent variables, then linearly mapped to mutually orthogonal subspaces and finally shifted and added an isotropic Gaussian noise. The resulting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

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 $(2,3) \preceq (1,1,2,1)$.

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 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 will 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 $Q^{\boldsymbol{\gamma}} := (Q_1, \ldots, Q_d)$ such that $Q_k \in \mathbb{R}^{p \times \gamma_k}, \forall k \in [1\,..\,d]$ and $Q = [Q_1 | \ldots | Q_d]$. Let $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_p)$ be a sequence of decreasing eigenvalues. The $\boldsymbol{\gamma}$-composition of $\boldsymbol{\lambda}$ is the sequence $\boldsymbol{\lambda}^{\boldsymbol{\gamma}} := \left(\boldsymbol{\lambda}^1, \ldots, \boldsymbol{\lambda}^d\right)$ such that $\boldsymbol{\lambda}^k \in \mathbb{R}^{\gamma_k}$, $\forall k \in [1\,..\,d]$ and $\boldsymbol{\lambda} = \left[\boldsymbol{\lambda}^1 | \ldots | \boldsymbol{\lambda}^d\right]$. We will call $\boldsymbol{\gamma}$-*averaging* of $\boldsymbol{\lambda}$ the sequence $\overline{\boldsymbol{\lambda}^{\boldsymbol{\gamma}}} := \left(\overline{\boldsymbol{\lambda}^1}, \ldots, \overline{\boldsymbol{\lambda}^d}\right) \in \mathbb{R}^d$ 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 will 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 geometric 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 optimization 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 $\operatorname{Flag}(\boldsymbol{\gamma})$. The geometry of such a set is well known (Monk, 1959). $\operatorname{Flag}(\boldsymbol{\gamma})$ is a smooth quotient manifold, consisting in equivalence classes of orthogonal matrices

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

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 $L \in \mathrm{D}(\boldsymbol{\gamma}) := \left\{\operatorname{diag}\left(\ell_1 I_{\gamma_1}, \ldots, \ell_d I_{\gamma_d}\right) \in \mathbb{R}^{p\times p} : \ell_1 > \cdots > \ell_d > 0\right\}$, which can be seen as a convex cone of $\mathbb{R}^d$.

Proposition 2 *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 $(1, \dots, 1, p-q)$ are nothing but Stiefel manifolds (up to changes of signs), we can naturally parameterize PPCA models with those spaces, which is already commonly done in the literature. 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 parameterize it with flags of type $(q, p-q)$, which are nothing but Grassmannians.

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 all the parameters—that are the eigenvectors and the eigenvalues—can be accurately identified, 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 to provide a possible answer, through SPCA model selection. First, we study the identifiability of two adjacent sample eigenvalues and deduce that one rarely has enough samples to distinguish them. We conclude 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

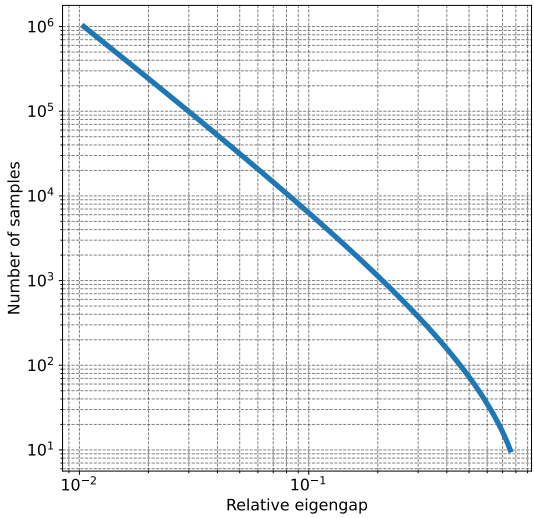
In this work, we focus on one simple model selection criterion to set up the ideas. The *Bayesian Information Criterion (BIC)* is defined as

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

where κ is the number of free parameters—computed in Proposition 2—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 3 *The SPCA model minimizing the BIC is*

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



#Samples	Eigengap
10	76%
100	46%
1000	21%
10000	8%

Fig. 3: Plot of the threshold function δ (Proposition 4), corresponding to the minimal number of samples needed to distinguish two adjacent eigenvalues, separated by a given relative eigengap. Real datasets that fulfill this condition for any pair of adjacent eigenvalues are pretty scarce.

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 variation of the BIC 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* $\boldsymbol{\gamma}' = (1, \dots, 1, 2, 1, \dots 1)$.

Proposition 4 *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*

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

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

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

A few values of the *threshold function* $\delta(n)$ are reported in Figure 3. The table can be read in the following way: if a pair of sample eigenvalues has a relative eigengap lower than 21%, then we need at least 1000 data points to statistically distinguish

them. This is an important result, as 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 4. 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 Stratified PCA 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 characterization 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 by eigenvalue multiplicity (Arnold, 1995; Groisser et al., 2017; Breiding et al., 2018). Hence, without proof, we can state the following result.

Proposition 5 *The family of p-dimensional SPCA models induces a stratification of the space of full-rank $p \times p$ covariance matrices by eigenvalue multiplicity. 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 4. 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 non-greedy 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.

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,

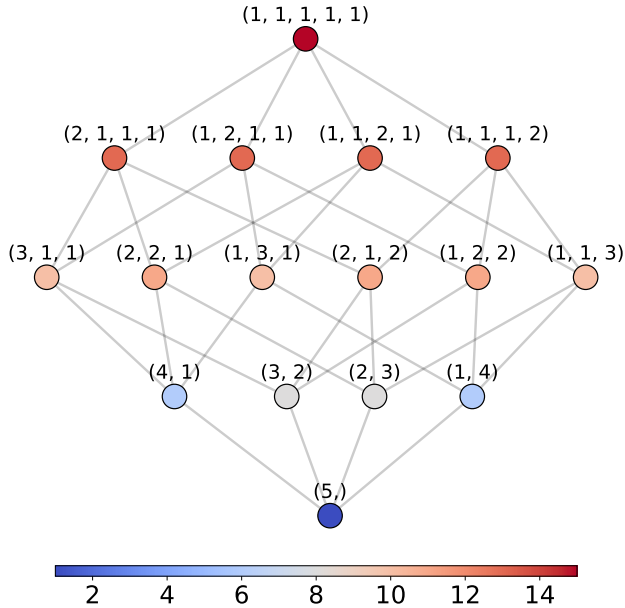


Fig. 4: Hasse diagram of 5-dimensional SPCA models. Each node represents a model. The associated label and color represent respectively the model type and its number of free parameters. The family contains 16 models: the isotropic Gaussian is the bottom node, the full covariance model is the top node, the five PPCA models are on the right part and the four IPPCA models are on the first floor.

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 4), 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 5. The hierarchical clustering heuristic creates a *trajectory* in the Hasse diagram of SPCA types $(\boldsymbol{\gamma}^t)_{t=1}^p$. 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 we reach the isotropic covariance model, $\boldsymbol{\gamma}^p = (p)$, in which all the eigenvalues are in the same cluster.

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.

Proposition 6 *If the true generative model belongs to SPCA, then the hierarchical clustering heuristic* (4.4.1) *will asymptotically consistently select it.*

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
  $\boldsymbol{\gamma}^1 \Leftarrow (1, \dots, 1)\,, \quad \boldsymbol{\lambda}^1 \Leftarrow \boldsymbol{\lambda} := (\lambda_1, \dots, \lambda_p)$ ▷ full covariance model
  for $t = 1 \,..\, p-1$ **do**
    $\Delta^t \Leftarrow \Delta\left(\lambda_k^t, \lambda_{k+1}^t\right)_{k=1}^{p-t}$ ▷ distance between adjacent clusters
    $k^t = \arg\min \Delta^t$ ▷ minimal distance
    $\boldsymbol{\gamma}^{t+1} = \left(\gamma_1^t, \dots, \gamma_{k^t-1}^t, \gamma_{k^t}^t + \gamma_{k^t+1}^t, \gamma_{k^t+2}^t, \dots, \gamma_d\right)$ ▷ type agglomeration
    $\boldsymbol{\lambda}^{t+1} = \overline{\boldsymbol{\lambda}^{\boldsymbol{\gamma}^{t+1}}}$ ▷ $\boldsymbol{\gamma}$-averaging
  end for

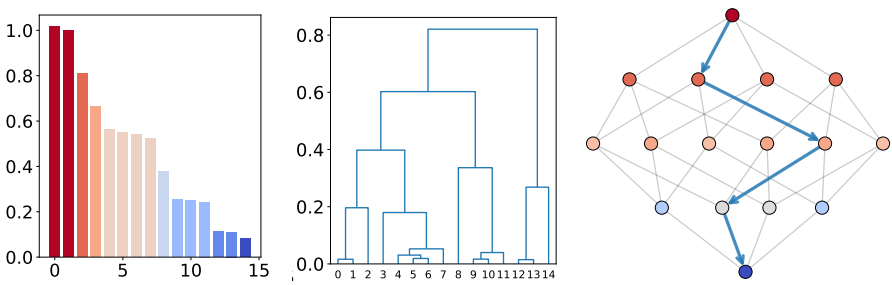


Fig. 5: Hierarchical clustering of sample eigenvalues, using the Euclidean distance. The successive steps in the hierarchical clustering generate a subfamily of SPCA models, of cardinal p. *Left:* sample eigenvalues at a given step of the hierarchical clustering, $\boldsymbol{\gamma}^t = (2, 1, 1, 4, 1, 3, 2, 1)$. The colors correspond to the parts of $\boldsymbol{\gamma}^t$. *Middle:* hierarchical clustering dendrogram. *Right:* conceptual view of the hierarchical clustering trajectory on the SPCA Hasse diagram.

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

Hence, the hierarchical clustering heuristic generates a hierarchical family of models of different complexities, and provided enough data, the true model will be included. Using asymptotic model selection criteria making a tradeoff between goodness-of-fit and complexity like the BIC will then allow to select 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 4). More precisely, we consider for selection only the models that have a given type length d, like done in IPPCA with $d = 2$. Similarly as for the hierarchical clustering heuristic, the type-length prior heuristic drastically reduces the search space, this time to $\binom{p-1}{d-1}$ models.

Similarly as in the hierarchical clustering heuristic (4.4.1), we could then use the BIC to choose the best model among this reduced family. 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. 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 7 *If the true generative model belongs to SPCA, then the maximum likelihood criterion alone will asymptotically consistently select it within the type-length prior heuristic* (4.4.2)*.*

Proof – The proof is given in Appendix B.3. We emphasize the use of Jensen's inequality, which elegantly generalizes 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 3, 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.

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 6, where we plot only a few models among the 16 for readability. First, on the BIC plots, we can see that for $n \leq 6000$, the best models do not belong to PPCA. This shows that even for a very large number of samples with respect to the dimension, distinguishing the first eigenvalues 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

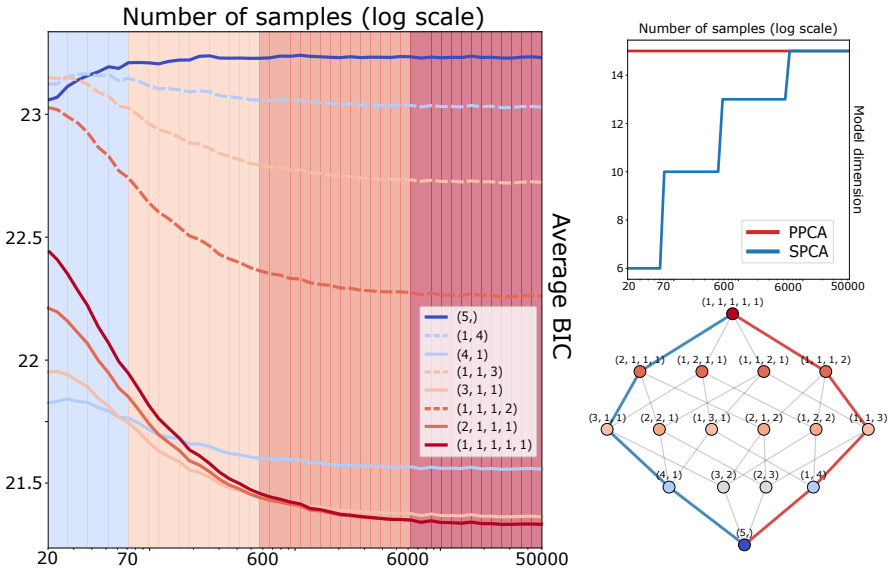


Fig. 6: SPCA model selection using the BIC for an increasing number of available samples. *Left:* Each curve represents the average BIC of a given SPCA model over several independent experiments. The lowest curve at a given n (horizontal coordinate) therefore corresponds to the most selected model. The 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 $\boldsymbol{\gamma} = (1, 1, 1, 1, 1)$. We conclude that for $n \leq 6000$, SPCA discloses a whole family of models that better explain the observed data than PPCA. *Top Right:* Comparison of the complexities of the mostly selected models within the whole SPCA family (blue) and within the PPCA family only (red). We can see a consistent increase of complexity of the fitted SPCA models with the number of samples, while PPCA always selects the full covariance model which is of higher complexity. *Bottom Right:* 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.

less complex models along the whole trajectory. Moreover, interestingly, we note the consistent increase of model complexity with the number of samples. 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, going up with the

Table 1. Comparison of PPCA and SPCA best models on several real datasets. We can see that for any dataset, SPCA finds new models that have a lower BIC. 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. 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	$\boldsymbol{\gamma}$	BIC	$\boldsymbol{\gamma}$	BIC
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
Wine	48	13	$(1^3, 10)$	$+36.35$	$(8, 5)$	$+35.57$
WDBC	357	30	(1^{30})	$+25.12$	$(2, 1, 2, 1, 2, 5, 1, 2, 1, 3^2, 4, 1^3)$	$+24.72$

number of available samples, which recalls the kind of subfamily generated by the hierarchical clustering heuristic (cf. Figure 5). 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*, *Ionosphere*, *Wine* and *Breast Cancer Wisconsin (WDBC)*. Due to the high dimensionality of some datasets, we cannot apply greedy SPCA model selection techniqes, therefore we use the hierarchical clustering heuristic introduced in Subsection 4.4. 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 see that SPCA achieves again a better complexity/goodness-of-fit tradeoff than PPCA by equalising some eigenvalues with small gaps. For conciseness, we do not report the sample eigenvalue profiles of those datasets, but we can check that none of them satisfies the relative eigengap condition of Proposition 4. Hence, the use of SPCA to model real datasets is justified.

In addition to the previous experiment, we also perform a floor-by-floor model comparison on the Glass dataset. More precisely, for each type length, we compare the unique associated PPCA model to the best SPCA one using the type-length prior heuristic introduced in Subsection 4.4. The results are reported on Table 2. We can see that the rich family of SPCA models with a prespecified number of distinct eigenvalues $d \in [1\,..\,p]$, which is of cardinal $\binom{p-1}{d-1}$, importantly increases the modelling power of PPCA, which only contains one model for each d.

Table 2. Comparison of PPCA and SPCA best models in the fixed-type-length setting (4.4.2) on the Glass dataset. For a given $d \in [1\,..\,p]$, PPCA contains only one model ($q = d - 1$), while SPCA contains $\binom{p-1}{d-1}$, which increases the modelling power. We can see for instance that the SPCA model of type $(8, 1)$ outperforms the PPCA model of type $(1, 8)$ on this dataset, therefore assuming a principal subspace of dimension 8 with isotropic variability is more likely than assuming a principal subspace of dimension 1.

PPCA		**SPCA**	
$\boldsymbol{\gamma}$	BIC	BIC	$\boldsymbol{\gamma}$
$(9,)$	$+4.20$	$+4.20$	$(9,)$
$(1, 8)$	-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	$(1, 2, 3, 1, 1, 1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(1, \ldots\ldots\ldots, 1)$	-16.77	-16.77	$(1, \ldots\ldots\ldots, 1)$

6. Discussion

We introduced in this paper a generative covariance model with repeated eigenvalues called *Stratified PCA (SPCA)*, which generalizes *Probabilistic PCA (PPCA)* (Tipping and Bishop, 1999) and *Isotropic PPCA (IPPCA)* (Bouveyron et al., 2011). The geometric interpretation of its parameter space shed light on the parsimony of PPCA and raised the natural question of extending its eigenvalue-equalisation principle to the signal space. We indeed argued that assuming all the eigenvalues and eigenvectors in the signal space to be identifiable is not justified in many settings. Hence, SPCA could circumvent this issue by equalising the adjacent eigenvalues with small gaps and gathering the associated eigenvectors into a multidimensional eigenspace. We confirmed our expectations on synthetic and real datasets, showing how SPCA models achieve a better complexity/goodness-of-fit tradeoff than PPCA.

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 favor under-parameterized 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) 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 optimization 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 optimization scheme on the space of all SPCA models. The latter being

stratified by eigenvalue multiplicities, we could benefit from recent works on stratified optimization (Leygonie et al., 2023; Olikier et al., 2023).

SPCA also comes with several exciting 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, the perfect model might lie in the SPCA family. Second, the multidimensional eigenspaces obtained by gathering eigenvectors associated to 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 (Nishimori et al., 2006) or principal (Ma et al., 2021) component analysis 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 minimizer of the *accumulated unexplained variance* criterion (Pennec, 2018), we enrich the previous works and enable for instance to perform flag-type selection.

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 $\underline{\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}}$.

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

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

As g is a smooth function on $\mathcal{O}(p)$ which is a compact manifold, $\hat{Q}$ exists and $dg_{\hat{Q}} : \mathcal{T}_{\hat{Q}}(\mathcal{O}(p)) \ni \delta \mapsto \operatorname{tr}\left(\left(\delta L^{-1}\hat{Q}^\top + \hat{Q}L^{-1}\delta^\top\right)S\right) \in \mathbb{R}$ 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 diagonalizable 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 minimizes 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, $\left(\lambda_{\hat{\psi}(\phi_{\boldsymbol{\gamma}}^{-1}\{k\})}\right)_{k=1}^d$ 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 minimizing $\hat{\psi}$ 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 $(\operatorname{Im}(\hat{Q}_1), \ldots, \operatorname{Im}(\hat{Q}_d)) = (\operatorname{Im}(V_1), \ldots, \operatorname{Im}(V_d))$. 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

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

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 $\operatorname{tr}\left(\Sigma^{-1}S\right) = \sum_{k=1}^d \ell_k^{-1}\left(\sum_{j\in\phi_{\boldsymbol{\gamma}}^{-1}\{k\}}\lambda_j\right)$. The optimal eigenvalues $\left(\hat{\ell}_1,\dots,\hat{\ell}_d\right)$ are thus

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

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

B. Other proofs

B.1. Proof of Proposition 4 (Eigenvalue equalisation)

We compare the BIC of the full covariance model $\boldsymbol{\gamma} = (1,\dots,1)$ to the one of the equalised covariance model $\boldsymbol{\gamma}' = (1,\dots,1,2,1,\dots 1)$ 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

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

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

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

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

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

Hence, one has

$$\Delta\,\mathrm{BIC} = 0 \iff e^{2\frac{\ln n}{n}} = \frac{\left(1-\frac{\delta_j}{2}\right)^2}{1-\delta_j} \iff \frac{\delta_j^2}{4} - \left(1-e^{2\frac{\ln n}{n}}\right)\delta_j + 1 - e^{2\frac{\ln n}{n}} = 0.$$

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

$$\delta(n) := 2 - 2e^{2\frac{\ln n}{n}} + 2\sqrt{e^{4\frac{\ln n}{n}} - e^{2\frac{\ln n}{n}}}. \tag{30}$$

B.2. Proof of Proposition 6 (Asymptotic consistency of the 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 $Q := [Q_1|\dots|Q_d] \in \mathcal{O}(p)$. 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 $V := [\boldsymbol{v}_1|\dots|\boldsymbol{v}_p] \in \mathcal{O}(p)$. 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, 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 $\{\Delta(\ell_k, \ell_{k+1}), k \in [1\,..\,d-1]\}$, 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. Asymptotic criteria like the BIC will thus consistently choose the true model among this reduced subfamily of cardinal p.

B.3. Proof of Proposition 7 (Asymptotic consistency of the type-length prior)

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 last 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

$$\ln \hat{\mathcal{L}} \sim -\frac{n}{2} \left(p \ln 2\pi + \sum_{k=1}^{d} \gamma_k \ln \left(\frac{1}{\gamma_k} \sum_{j \in \phi_{\boldsymbol{\gamma}}^{-1}\{k\}} \ell_{\phi_{\boldsymbol{\gamma}^*}(j)} \right) \right). \tag{31}$$

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

$$\underset{\substack{\boldsymbol{\gamma} \in \mathcal{C}(p) \\ \#\boldsymbol{\gamma} = d}}{\arg\min} \sum_{k=1}^{d} \gamma_k \ln \left(\frac{1}{\gamma_k} \sum_{j \in \phi_{\boldsymbol{\gamma}}^{-1}\{k\}} \ell_{\phi_{\boldsymbol{\gamma}^*}(j)} \right) := f(\boldsymbol{\gamma}). \tag{32}$$

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,

$$f(\boldsymbol{\gamma}) \geq \sum_{k=1}^{d} \gamma_k \left(\sum_{k'=1}^{d} \frac{c_{kk'}}{\gamma_k} \ln \ell_{k'} \right) = \sum_{k,k'=1}^{d} c_{kk'} \ln \ell_{k'} = \sum_{k'=1}^{d} \boldsymbol{\gamma}_{k'}^* \ln \ell_{k'} = f(\boldsymbol{\gamma}^*). \tag{33}$$

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

Arnold, V. I. (1995). Remarks on eigenvalues and eigenvectors of Hermitian matrices, berry phase, adiabatic connections and quantum Hall effect. *Selecta Mathematica*, 1(1):1–19.

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

Bishop, C. (1998). Bayesian PCA. In *Advances in Neural Information Processing Systems*, volume 11.

Bouveyron, C., Celeux, G., and Girard, S. (2011). Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic PCA. *Pattern Recognition Letters*, 32(14):1706–1713.

Bouveyron, C. and Girard, S. (2009). Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, 42(11):2649–2658.

Breiding, P., Kozhasov, K., and Lerario, A. (2018). On the Geometry of the Set of Symmetric Matrices with Repeated Eigenvalues. *Arnold Mathematical Journal*, 4(3):423–443.

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

Drton, M. and Plummer, M. (2017). A Bayesian Information Criterion for Singular Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(2):323–380.

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

Groisser, D., Jung, S., and Schwartzman, A. (2017). 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).

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

Leygonie, J., Carrière, M., Lacombe, T., and Oudot, S. (2023). A gradient sampling algorithm for stratified maps with applications to topological data analysis. *Mathematical Programming*.

Ma, X., Kirby, M., and Peterson, C. (2021). 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.

Minka, T. (2000). Automatic Choice of Dimensionality for PCA. In *Advances in Neural Information Processing Systems*, volume 13.

Monk, D. (1959). The Geometry of Flag Manifolds. *Proceedings of the London Mathematical Society*, s3-9(2):253–286.

Nishimori, Y., Akaho, S., and Plumbley, M. D. (2006). Riemannian Optimization Method on the Flag Manifold for Independent Subspace Analysis. In *Independent*

Component Analysis and Blind Signal Separation, pages 295–302.

Olikier, G., Gallivan, K. A., and Absil, P.-A. (2023). First-order optimization on stratified sets. Preprint.

Pearson, K. (1901). 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.

Pennec, X. (2018). Barycentric Subspace Analysis on Manifolds. *The Annals of Statistics*, 46(6A):2711–2746.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27(2):125–140.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, 18(4):267–276.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622.

Tyler, D. E. (1981). Asymptotic Inference for Eigenvectors. *The Annals of Statistics*, 9(4):725–736.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244.

Ye, K., Wong, K. S.-W., and Lim, L.-H. (2021). Optimization on flag manifolds. *Mathematical Programming*.