



HAL
open science

The curse of isotropy: from principal components to principal subspaces

Tom Szwagier, Xavier Pennec

► **To cite this version:**

Tom Szwagier, Xavier Pennec. The curse of isotropy: from principal components to principal subspaces. 2024. hal-04171853v3

HAL Id: hal-04171853

<https://inria.hal.science/hal-04171853v3>

Preprint submitted on 16 Dec 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

The curse of isotropy: from principal components to principal subspaces

Tom Szwagier¹ Xavier Pennec¹

Abstract

This paper raises an important issue about the interpretation of principal component analysis. The *curse of isotropy* states that a covariance matrix with repeated eigenvalues yields rotation-invariant eigenvectors. In other words, principal components associated with equal eigenvalues show large intersample variability and are arbitrary combinations of potentially more interpretable components. However, empirical eigenvalues are never exactly equal in practice due to sampling errors. Therefore, most users overlook the problem. In this paper, we propose to identify datasets that are likely to suffer from the curse of isotropy by introducing a generative Gaussian model with repeated eigenvalues and comparing it to traditional models via the principle of parsimony. This yields an *explicit* criterion to detect the curse of isotropy in practice. We notably argue that in a dataset with 1000 samples, all the eigenvalue pairs with a relative eigengap lower than 21% should be assumed equal. This demonstrates that the curse of isotropy cannot be overlooked. In this context, we propose to transition from fuzzy principal components to much-more-interpretable principal subspaces. The final methodology—*principal subspace analysis*—is extremely simple and shows promising results on a variety of datasets from different fields.

1. Introduction

Principal component analysis (PCA) (Jolliffe, 2002) is a universal method in data analysis. It gives the main modes of variation in the data by diagonalizing the empirical covariance matrix. The eigenvectors associated with the largest eigenvalues are the *principal components*, and the subspace they span is used for dimension reduction and visualization. Additionally, principal components can be used for exploratory data analysis and interpretability purposes. It has been extensively used on structured anatomical data

¹Inria – Université Côte D’Azur, Sophia-Antipolis, France. Correspondence to: Tom Szwagier <tom.szwagier@inria.fr>.

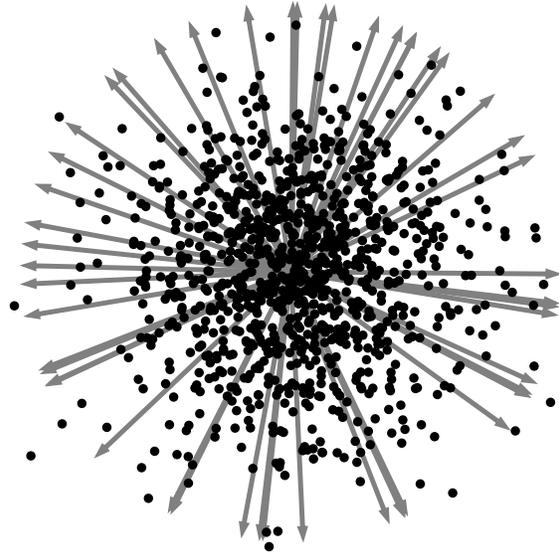


Figure 1. Covariance eigenvectors of a dataset sampled from a two-dimensional isotropic Gaussian, repeated independently 25 times. Principal components have an isotropic intersample variability.

(with components related to morphological features), in atmospheric sciences (with components related to climate patterns), computer vision (with so-called *eigenfaces* (Sirovich & Kirby, 1987)) and many other fields. We refer to the chapters 4 and 11 of Jolliffe (2002) for detailed examples of principal component interpretation.

Let us assume that a dataset has been sampled from a multivariate Gaussian distribution. If all the population covariance eigenvalues are *simple* (i.e. distinct), then we can associate to each eigenvalue a unique eigenvector (up to sign). Now, if some eigenvalues are *multiple*, then those are associated with multidimensional eigenspaces, i.e. an infinite number of eigenvectors. This implies that the principal components associated with those multiple eigenvalues exhibit a large *intersample variability*. More specifically, for any dataset size n , each independent n -sample from the distribution can yield totally different principal components, with a full rotational uncertainty within the eigenspaces. Therefore, under this multiple-eigenvalue assumption, principal components are unstable—regardless of n —which is fatal to data interpretability. We call this issue the *curse of isotropy*.

In real datasets, empirical covariance eigenvalues are never exactly equal (they are almost surely different from a measure-theoretical point of view, cf. proof of Thm. B.1), but some may be relatively close. In this case, it might be wiser to assume that *close* eigenvalues are actually *equal*—especially for small n —in order to avoid overfitting some spurious patterns caused by *sampling errors* (North et al., 1982). Under this assumption, the dataset suffers from the curse of isotropy and one must be careful about interpreting the associated principal components. Therefore, identifying the curse of isotropy in practice boils down to answering the following question: *when should we assume that the dataset has been sampled from a multivariate Gaussian distribution with repeated covariance eigenvalues?*

In this paper, we answer the question with an *explicit* guideline, derived from two key concepts: *parsimonious Gaussian modeling* and *flags of subspaces*. More specifically, we introduce a latent variable generative model called *principal subspace analysis* (PSA). This model assumes a Gaussian density with repeated eigenvalues, where the sequence of eigenvalue multiplicities is specified by the so-called *type* of the model. We show that PSA generalizes the celebrated Probabilistic PCA (PPCA) of Tipping and Bishop (Tipping & Bishop, 1999b) and unifies it with Isotropic PPCA (IPPCA) (Bouveyron et al., 2007; 2011)—a parsimonious version of PPCA suited to high dimensions. PSA models have a rich geometry relying on flag manifolds and stratify the space of covariance matrices. This enables us to assess the drop of model complexity caused by equalizing some eigenvalues and to perform efficient model selection based on parsimony-inducing criteria such as the Bayesian information criterion (BIC)—other criteria are investigated in Sec. C with similar conclusions. We show that two adjacent sample eigenvalues should be assumed equal when their relative eigengap is lower than a given threshold. This threshold depends on n but is independent of the dimension p .

The results are striking: in almost all the datasets that we analyze, the curse of isotropy arises. This questions the numerous scientific works relying on the *interpretation* of principal components. While this could sound fatal to exploratory data analysis, we show that the curse of isotropy can actually be leveraged to improve data interpretability. Indeed, in such a situation, we suggest to give up *principal components* and transition to more-interpretable *principal subspaces*. Taking advantage of our generative model and factor rotation methods, we propose several qualitative and quantitative methods to increase the interpretability of principal components. We test the resulting PSA methodology on synthetic and real datasets and get promising results. More precisely, while principal components associated with close eigenvalues may be fuzzy—as arbitrary linear combinations of more interpretable factors—principal subspaces show the emergence of *multidimensional attributes*.

2. The curse of isotropy

Let us consider a dataset sampled independently from a two-dimensional *isotropic* Gaussian distribution. This implies that the eigenvalues of the population covariance matrix are equal. The sample covariance matrix, however, is an approximation of the population covariance matrix, whose accuracy improves with the number of observed samples (Tyler, 1981). Notably, the empirical eigenvalues are almost surely distinct (cf. Thm B.1). Therefore, PCA outputs the unique eigenvectors (up to sign) associated with each eigenvalue. If we repeat this experiment several times independently and plot the principal components, we get Fig. 1. As we can see, the principal components are evenly spread in all directions—i.e. isotropically. We call this phenomenon *the curse of isotropy*. It is a curse since it yields principal components with high intersample variability and without any preferred direction. The observed components could therefore be random combinations of *actually* interpretable components.

A legitimate question might then be: *why (and when) should we assume that a given dataset has been sampled from a Gaussian distribution with repeated eigenvalues?* The Gaussian assumption is notably justified by the central limit theorem, the entropy maximization and the attractive computational properties that make Gaussian distributions the cornerstone of machine learning generative models (Bishop, 2006). Now, regarding the multiple-eigenvalue assumption, we have to go back to one of the founding principles of modeling that is the *law of parsimony*, also known as *Occam’s razor*: “The simplest explanation is usually the best one”. This principle is particularly applied in statistical modeling, where the limited number of observed samples makes overparameterized models overfitting (Myung et al., 2000). Notably, covariance matrices (which have $\mathcal{O}(p^2)$ parameters) can almost never be correctly estimated in practice, especially in high dimensions. Therefore, more parsimonious models have to be considered, like isotropic Gaussians (which have 1 parameter—the variance), where all the covariance eigenvalues are equal. In the following, we show that a Gaussian model with *repeated eigenvalues*, i.e. isotropic in some multidimensional eigenspaces, has less parameters than one with *distinct eigenvalues* and therefore provides a simpler explanation of the data. Then, using parsimonious model selection criteria such as the BIC, we are able to decide which eigenvalues should be assumed equal.

3. Identifying the curse of isotropy

In order to spot the curse of isotropy, we go through the lens of statistical modeling and introduce the PSA generative model. This model assumes a Gaussian distribution with repeated covariance eigenvalues. It enjoys an explicit

maximum likelihood estimate with a rich geometry enabling effective model selection.

3.1. PSA model

Let $\gamma := (\gamma_1, \dots, \gamma_d)$ be a *composition* of a positive integer p —i.e. a sequence of positive integers that sums up to p . We define the PSA model of *type* γ as the family of Gaussian distributions $p(x|\mu, \Sigma) := \mathcal{N}(x|\mu, \Sigma)$, where $\mu \in \mathbb{R}^p$ is a mean vector and $\Sigma = \sum_{k=1}^d \lambda_k Q_k Q_k^\top \in S_p^{++}$ is a covariance matrix with repeated eigenvalues $\lambda_1 > \dots > \lambda_d > 0$ of respective multiplicity $\gamma_1, \dots, \gamma_d$ and associated eigenspaces $\text{Im}(Q_1), \dots, \text{Im}(Q_d)$. These distributions can be rewritten as a (linear-Gaussian) latent variable generative model

$$x = \sum_{k=1}^{d-1} \sigma_k Q_k z_k + \mu + \epsilon, \quad (1)$$

where $\sigma_1 > \dots > \sigma_{d-1} > 0$ are decreasing scaling factors, $Q_k \in \mathbb{R}^{p \times \gamma_k}$ are mutually-orthogonal orthonormal γ_k -frames, $z_k \sim \mathcal{N}(0, I_{\gamma_k})$ are independent latent variables and $\epsilon \sim \mathcal{N}(0, \sigma^2 I_p)$ is an isotropic Gaussian noise. An illustration of the generative model is provided in Fig. 2. PPCA and IPPCA models can then be reinterpreted as PSA models, of respective types $\gamma = (1, \dots, 1, p - q)$ and $\gamma = (q, p - q)$, where $q < p$ is the intrinsic dimension (cf. Sec. B).

3.2. Geometry and inference

From a geometric point of view, the fitted density is isotropic on a sequence of mutually-orthogonal subspaces $\text{Im}(Q_1) \perp \dots \perp \text{Im}(Q_d)$ of respective dimensions $\gamma_1, \dots, \gamma_d$. Such a sequence is called a *flag* of linear subspaces of *type* γ . Therefore, flags of type γ —which are diffeomorphic to $\mathcal{O}(p)/(\mathcal{O}(\gamma_1) \times \dots \times \mathcal{O}(\gamma_d))$ (Arnold, 1972; Ye et al., 2022)—naturally parameterize PSA models. Consequently, Stiefel manifolds and Grassmannians—which are particular cases of flag manifolds—respectively parameterize PPCA and IPPCA models (cf. Sec. B). The remaining model parameters are the subspace variances $(\lambda_1, \dots, \lambda_d) \in \mathbb{R}^d$ and the mean $\mu \in \mathbb{R}^p$. Thus, the *complexity* (dimension of the parameter space) of the PSA model of type γ is

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

We can notably see that the decrease in model complexity is quadratic in the number of equalized eigenvalues.

One of the strength of the PSA models is that their maximum likelihood estimate is *explicit*, similarly to PPCA and IPPCA. In a nutshell, we show in Thm. B.1 that the most likely mean vector μ is the *empirical mean*, the most likely variances $\lambda_1, \dots, \lambda_d$ are the *block-averaged sample*

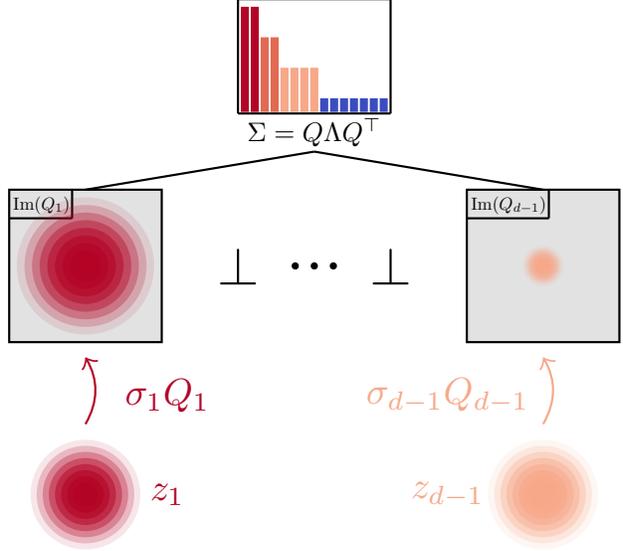


Figure 2. PSA generative model, assuming that the observed data was first sampled from a sequence of independent low-dimensional normal latent variables, then linearly mapped to mutually-orthogonal subspaces and finally shifted and added an isotropic Gaussian noise (1). The resulting density is a multivariate Gaussian with repeated eigenvalues, of respective multiplicities $\gamma = (2, 2, 4, 7)$.

eigenvalues according to the type γ , and the most likely flag $(\text{Im}(Q_1), \dots, \text{Im}(Q_d))$ is the sequence of mutually-orthogonal subspaces spanned by the associated eigenvectors. This yields the following expression for the maximum likelihood

$$\ln \hat{\mathcal{L}}(\gamma) = -\frac{n}{2} \left(p \ln(2\pi) + \sum_{k=1}^d \gamma_k \ln \bar{L}_k + p \right). \quad (3)$$

3.3. Identifying the curse of isotropy in practice

The Bayesian information criterion (Schwarz, 1978) is defined as

$$\text{BIC}(\gamma) := \kappa(\gamma) \ln n - 2 \ln \hat{\mathcal{L}}(\gamma). \quad (4)$$

It is a widely-used model selection criterion, making a trade-off between model complexity and goodness-of-fit, to prevent from overfitting given the number of observed samples. The formula results from an asymptotic approximation of the Bayesian model evidence. Given a dataset, one can compare the BIC of a PSA model with repeated eigenvalues to the BIC of a PSA model with distinct eigenvalues. The model with the lowest BIC is selected over the other one.

As discussed previously, two adjacent sample eigenvalues with a relatively small gap may be prone to isotropic

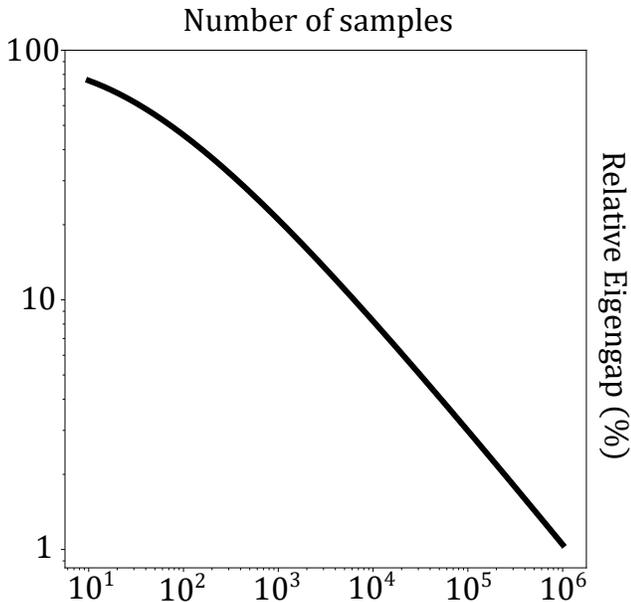


Figure 3. Plot of the relative eigengap threshold (5)—under which two adjacent sample eigenvalues should be assumed equal—as a function of n .

PC variability. To identify such situations where the curse of isotropy may arise, we compare a *full* covariance model $\gamma = (1, \dots, 1)$ with an *equalized* covariance model $\gamma' = (1, \dots, 1, 2, 1, \dots, 1)$ where eigenvalues j and $j + 1$ are assumed equal. Denoting $\delta_j := \frac{\ell_j - \ell_{j+1}}{\ell_j}$ the *relative eigengap* between the two sample eigenvalues, we show in Sec. C that

$$\text{BIC}(\gamma') < \text{BIC}(\gamma) \iff \frac{\delta_j}{2} < 1 - n^{\frac{2}{n}} + n^{\frac{1}{n}} \sqrt{n^{\frac{2}{n}} - 1}. \quad (5)$$

This condition—independent of p —is illustrated in Fig. 3. We notably deduce by substitution that for $n = 1000$ samples, all the adjacent sample eigenvalues with a relative eigengap lower than $\delta = 21\%$ should be assumed equal. In other words, given two sample eigenvalues of respective magnitude 1 and 0.8, one needs *at least* 1000 samples to overcome the curse of isotropy. *This is rarely the case in practice.* To illustrate this, we test the condition (5) on many classical datasets from the UCI Machine Learning Repository (cf. Sec. E), with n/p ratios ranging from 10 to 10^4 . For each dataset, we report the pairs of adjacent eigenvalues that are below the relative eigengap threshold in Fig. 4. The outcomes are striking: all datasets but one have some eigenvalue pairs below the threshold. This does not only concern the smallest eigenvalues—which are usually tossed away because considered as noise—but also the highest ones—which are usually interpreted by applied scientists. This shows that the curse of isotropy is not a negligible phenomenon at all and that particular care should be taken

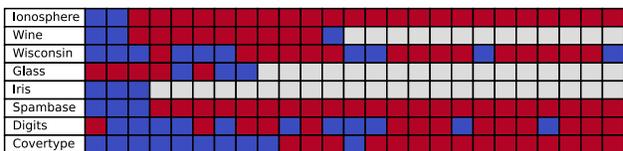


Figure 4. Practical identification of the curse of isotropy on several classical datasets. A red case in column j indicates that eigenvalues j and $j + 1$ have a relative eigengap below the threshold (5) and should be equalized. Blue is above and gray is undefined (we only plot the 25 leading eigenvalue pairs). We can clearly see that the curse of isotropy is not a negligible phenomenon in practice.

before interpreting the principal components. Note that (5) involves the *relative* eigengap between adjacent eigenvalues and not the *absolute* one, meaning that an exponentially-decreasing sample eigenvalue profile can actually highly suffer from the curse of isotropy. In other words, PSA models are not just suited to piecewise-constant-like sample covariance profiles.

The BIC is known for its tendency to select underparameterized models (Bishop, 2006). Therefore, we also investigate in Sec. C the eigenvalue-equalization guideline under other model selection criteria like the Akaike information criterion (AIC) (Akaike, 1974) and under empirical models (North et al., 1982). We get relative eigengaps around 10 – 20% for $n = 1000$, and experimental results substantiating the curse of isotropy’s importance.

3.4. Stratification and efficient model selection

We now explicit the stratified structure of PSA models and show how it enables to design efficient model selection strategies to choose which groups of eigenvalues to equalize. More details are given in Sec. C.

The space of symmetric matrices can be stratified according to the sequence of eigenvalue multiplicities (Arnold, 1972; Groisser et al., 2017; Breiding et al., 2018). This implies that the PSA models in dimension p form a stratified exponential family (Geiger et al., 2001) of cardinal 2^{p-1} , partially-ordered (Taeb et al., 2024) by the stratum-inclusion relation. We illustrate the family of 5-dimensional PSA models in Fig. 5.

In order to prevent from greedily exploring the whole family for model selection, we propose a simple yet efficient model selection technique based on the stratified structure of this family. The *hierarchical clustering strategy* consists in performing a hierarchical clustering of the sample eigenvalues, based on chosen *pairwise distance* (e.g. the relative eigengap $\delta_j = \frac{\ell_j - \ell_{j+1}}{\ell_j}$) and *cluster-linkage criterion* (e.g. single-linkage). This strategy yields a hierarchical subfamily of p models with decreasing complexity, from which we

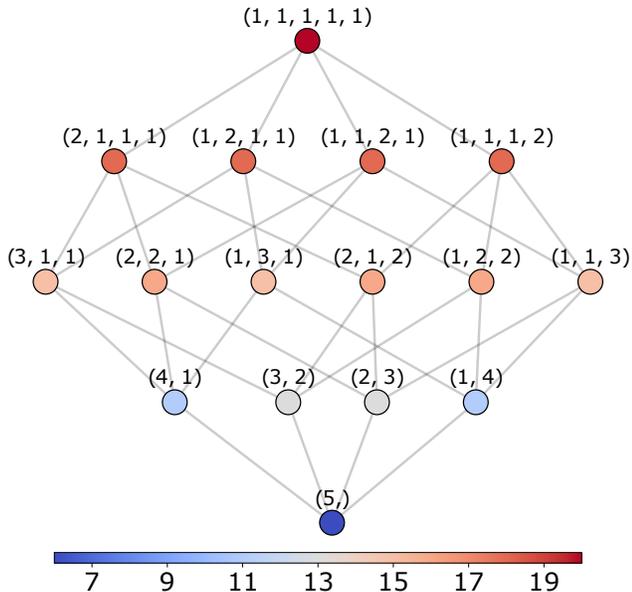


Figure 5. Hasse diagram of 5-dimensional PSA models. Each node represents a model. The associated label and color represent respectively the model type and its number of free parameters. The family contains 16 models: the isotropic Gaussian is the bottom node, the full covariance model is the top node,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

can more efficiently select the model minimizing the BIC. We prove the *asymptotic consistency* of the hierarchical clustering strategy in Prop C.6, as well as introduce other strategies.

4. From principal components to principal subspaces

To summarize the previous section, parsimonious considerations invite us to block-average eigenvalues whose relative gaps are close—given the number of observed samples. The associated PSA model is now parameterized with *eigenspaces* instead of individual *eigenvectors* and we are therefore facing the curse of isotropy. In this section, we propose to actually take advantage of the curse of isotropy and improve data interpretability by transitioning from *principal components* to *principal subspaces*.

A first idea, rather *qualitative*, is to work at the subspace level and generate samples from the multidimensional principal subspaces via Eq. (1) (cf. Fig. 2). Those samples might have common characteristics like low-frequencies or invariances for images (Hyvärinen & Hoyer, 2000). Since there is an isotropic Gaussian variability, one can also uniformly discretize the unit sphere included in the principal subspace—especially in 2D and 3D—to help visualization. Other ideas

involve orthogonal projections of explainable variability modes and correlation between subspace-projected dataset and co-variables.

A second idea, rather *quantitative*, is to look for rotations of principal components inside their principal subspace in order to increase interpretability. Indeed, as explained previously, the curse of isotropy might cause principal components to be rotated versions of more interpretable components. *Varimax* rotation (Kaiser, 1958) (on the components and not the projected data as sometimes done (Rohe & Zeng, 2023)) enables for instance to get rotated components with sparse loadings. Many other criteria can be considered depending on the data type, notably entropy, structured sparsity (Jenatton et al., 2010) or total variation for images.

5. Experiments

Eventually, PSA is grounded in a generative model with a rich geometry, yet the methodology is very simple and can be summarized in the three following steps: *eigendecomposition* of the sample covariance matrix, *block-averaging* of the eigenvalues with small relative eigengaps (or more formally, PSA model selection), *interpretation* of the resulting principal subspaces via factor rotation or latent subspace sampling. In this section we apply the PSA methodology to several synthetic and real datasets in a variety of fields. The experiments show that principal components associated with relatively-close eigenvalues are generally fuzzy due to the curse of isotropy. Therefore, equalizing the problematic eigenvalues and lifting the analysis to principal subspaces dramatically enhances exploratory data analysis.

5.1. Natural image patches

In this experiment, we consider patches extracted from natural images, as done in many seminal works investigating biological vision via unsupervised machine learning methods (Field, 1987; Olshausen & Field, 1996; Hyvärinen & Hoyer, 2000). We consider 10 flower images from the Natural Images database (cf. Sec. E) and randomly extract $n = 500$ (8, 8)-pixel patches from those. After removing the DC component (i.e. the mean value) to each patch, like usually done in such studies (Hyvärinen & Hoyer, 2000) and looking at the sample eigenvalue profile, we decide to fit a PSA model of type $\gamma' = (2, 3, 59)$ and compare it to the associated PPCA model $\gamma = (1, 1, 1, 1, 1, 59)$. We get a lower BIC with the PSA model. Then, we uniformly sample from the unit spheres included in the first and second feature subspaces, which are respectively 2D and 3D. We report the results in Fig. 6. While principal components do not look particularly interpretable individually, grouping them into principal subspaces with isotropic variability brings out low-frequency subspaces with rotational invariance. From a curse-of-isotropy point of view, the observed principal

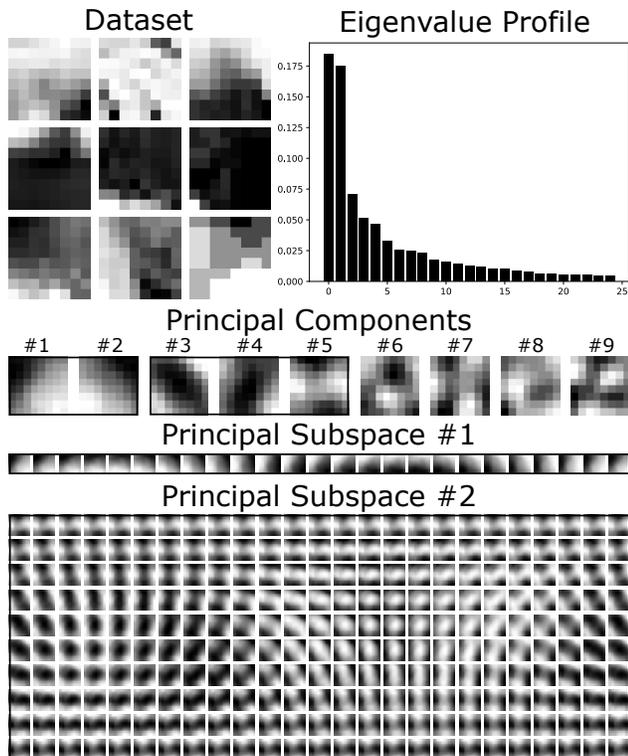


Figure 6. PCA vs PSA on the natural image patch dataset. Top: Dataset and covariance eigenvalue scree plot. Middle: Principal components. Bottom: Principal subspaces resulting from a PSA model of type $\gamma' = (2, 3, 59)$. We sample from those 2D and 3D subspaces and notice the emergence of low-frequency feature subspaces with rotational invariance. We insist on the fact that without principal subspace analysis, we would not-necessarily have been able to detect those multidimensional patterns.

components are random samples from the illustrated multidimensional subspaces.

5.2. Laplacian eigenfunctions

In this experiment, we generate a synthetic dataset consisting in linear combinations of Laplacian eigenfunctions—also known as *quasimodes* (Arnold, 1972)—with variance being a decreasing function of the Laplacian eigenvalue. This kind of generative model has been extensively used in many different areas, notably climate sciences (North et al., 1982) (for modeling atmospheric fields on the earth) and computer vision (for modeling shadows on faces under varying illumination conditions (Basri & Jacobs, 2003) or low-frequency patches in natural images (Field, 1987)). The global idea behind those models is that natural symmetries are present in the shapes under study (face, earth, square domain etc.) and lead to multiple eigenvalues in their Laplacian matrix, and therefore to multiple eigenvalues in the covariance matrix of homogeneous stochastic processes on

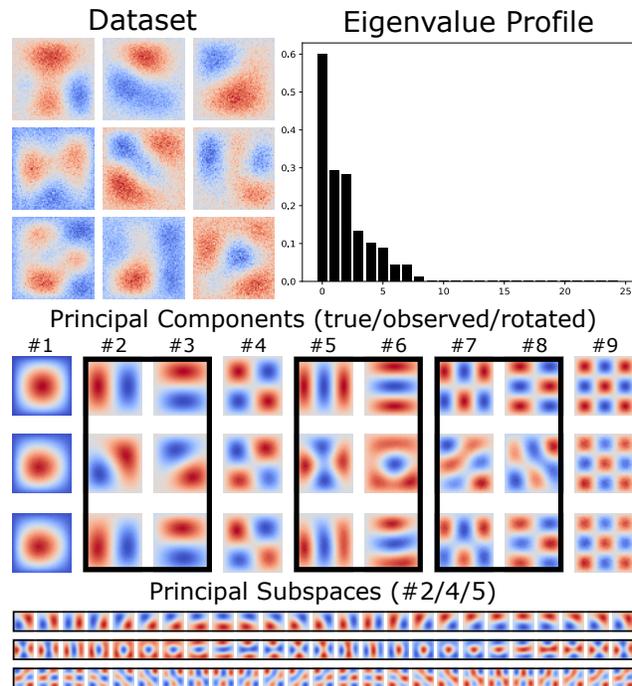


Figure 7. PCA vs PSA on the Laplacian eigenfunction dataset. Top: Dataset and covariance eigenvalue scree plot. Middle: Principal components (true, observed and rotated). The observed principal components are linear combinations of the true eigenmodes—i.e. quasimodes—especially (5, 6) and (7, 8). After factor rotation, one recovers the original eigenmodes. Bottom: Principal subspaces resulting from a PSA model of type $\gamma' = (1, 2, 1, 2, 2, 1, 4087)$. We sample from those 2D subspaces and obtain equal-frequency quasimodes.

those shapes.

We generate $n = 500$ points on a square grid with 64 pixels on each side. We take a combination of the $q = 9$ leading eigenmodes with variance scaling like $\exp(-2\lambda)$ (where λ is the Laplacian eigenvalue) and add an isotropic noise. We fit a PSA model of type $\gamma' = (1, 2, 1, 2, 2, 1, 4087)$ (corresponding to the expected Laplacian eigenvalue multiplicities on a square domain) and compare it to the associated PPCA model $\gamma = (1, \dots, 1, 4087)$. We get a lower BIC for the PSA model, and then perform factor rotation in the multidimensional eigenspaces. The results are shown in Fig. 7. We can see that the principal components are linear combinations of the original eigenmodes, i.e. quasimodes. With factor rotation in the associated subspaces, we better recover the original modes.

5.3. Eigenfaces (Sirovich & Kirby, 1987)

In this experiment, we consider the Carnegie Mellon University (CMU) Face Image database (cf. Sec. E). It consists

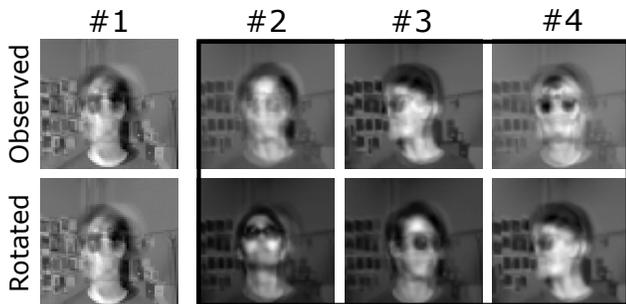


Figure 8. PCA vs PSA on the CMU Face Image dataset. Top: Eigenfaces. Bottom: Eigenfaces after entropy-related rotation within the second principal subspace (spanned by components 2, 3 and 4). This experiment shows that the curse of isotropy can yield blurry eigenfaces, corresponding to linear combinations of much more interpretable components that we can recover with PSA.

in 640 grayscale pictures of people from varying pose, expression, and eye conditions. We extract $n = 31$ (60, 64)-pixel images of the subject *Choon*. Inspired by the seminal paper (Basri & Jacobs, 2003) (which establishes a link between face shadowing under varying illumination conditions and spherical harmonics), we fit a PSA model of type $\gamma' = (1, 3, 5, 3831)$ and compare it to the associated PPCA model $\gamma = (1, \dots, 1, 3831)$. We get a lower BIC and perform entropy-guided subspace sampling in the second feature subspace, which is 3-dimensional. The results are illustrated in Fig. 8. While principal components 2, 3, and 4 are fuzzy and uninterpretable, we can see that they actually correspond to linear combinations of three *much more interpretable* factors.

5.4. Structured data

In this experiment, we consider a structured dataset taken from the UCI ML repository. The *Glass identification* dataset (cf. Sec. E) from the USA Forensic Science Service contains chemical features about different types of glasses, with applications in criminology. We fit a PSA model of type $\gamma' = (5, 4)$ and compare it to the associated PPCA model $\gamma = (1, \dots, 1, 4)$. We get a lower BIC and perform varimax rotation in the first feature subspace, of dimension 5. We report the loadings of the sample eigenvectors and compare them to the PSA factors after rotation in Fig. 9. We see that the PSA factors are more interpretable than the principal components, in the sense that they express as sparser combinations of the original variables. Moreover, contrary to classical factor rotation methods done after PCA (cf. Chapter 11 of Jolliffe (2002)), we here do not lose any hierarchy in the principal components in terms of explained variance, since under the PSA model of type $\gamma' = (5, 4)$, the five components have equal variance.

| | PC1 | PC2 | PC3 | PC4 | PC5 | RPC1 | RPC2 | RPC3 | RPC4 | RPC5 |
|----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| RI | -0.55 | -0.29 | -0.09 | -0.15 | 0.07 | -0.52 | 0.06 | 0.09 | -0.36 | -0.01 |
| Na | 0.26 | -0.27 | 0.38 | -0.49 | -0.15 | 0.29 | -0.18 | 0.61 | -0.2 | 0.17 |
| Mg | -0.11 | 0.59 | -0.01 | -0.38 | -0.12 | 0.33 | 0.6 | -0.01 | -0.21 | -0.08 |
| Al | 0.43 | -0.3 | -0.33 | 0.14 | -0.01 | 0.17 | -0.56 | -0.24 | -0.01 | -0.04 |
| Si | 0.23 | 0.16 | 0.46 | 0.65 | -0.01 | 0.02 | -0.02 | 0.05 | 0.84 | 0.0 |
| K | 0.22 | 0.15 | -0.66 | 0.04 | 0.31 | 0.16 | -0.12 | -0.71 | -0.21 | 0.11 |
| Ca | -0.49 | -0.35 | 0.0 | 0.28 | 0.19 | -0.68 | -0.07 | -0.02 | 0.03 | 0.03 |
| Ba | 0.25 | -0.48 | -0.07 | -0.13 | -0.25 | 0.09 | -0.52 | 0.22 | -0.19 | -0.13 |
| Fe | -0.19 | 0.06 | -0.28 | 0.23 | -0.87 | 0.01 | -0.01 | 0.0 | -0.0 | -0.97 |

Figure 9. PCA vs PSA on the Glass identification dataset. We fit a PSA model of type $\gamma' = (5, 4)$ and perform varimax rotation in the first principal subspace. We can see that the rotated PCs (right) are much sparser in the original variables than the PCs (left), while having the same estimated variance (under the PSA model). The colors are stratified similarly as in Section 4.1 of Jolliffe (2002) to help interpretability. For each eigenvector, the cases in dark red and dark blue correspond to coefficients whose absolute value is greater than half of the maximal absolute coefficient, the ones with light red and light blue to coefficients whose absolute value is between one quarter and one half, and the ones in gray are below, considered as negligible.

6. Related works

In the climate research community, a celebrated work (North et al., 1982)—often cited as *North’s rule-of-thumb*—warns scientists against close eigenvalues in the Karhunen-Loève expansion of a meteorological field. Indeed, the associated principal components—referred to as *empirical orthogonal functions* (EOF)—suffer from large sampling errors, which is very problematic due to the key role EOF’s play in this field for exploratory data analysis. The authors provide a perturbation-theoretical rule-of-thumb to decide which eigenvalues form *degenerate multiplets*. The rule as stated in the paper is quite vague, however we are able (cf. Sec. C) to reformulate its practical software implementation as a relative eigengap threshold and to compare it to our criterion (5). We show that this threshold is much lower than ours (e.g. 8.6% instead of 21% for 1000 samples), therefore our result has a much larger impact on the practical methodology of PCA.

More broadly, several works have mentioned close-eigenvalues in PCA or in general symmetric matrices. A paper from Jolliffe (Jolliffe, 1989) shows the advantages of factor rotation inside subspaces spanned by principal components with close eigenvalues for structured data. Eigenvalue equality has also been studied formally in the context of oscillatory systems (Arnold, 1972; Lazutkin, 1993; Gershkovich & Haritos, 2004) diffusion tensor imaging (Groisser et al., 2017), spectral geometry (Besson, 1988), spectral shape analysis (Lombaert et al., 2013), statistical tests (Anderson, 1963; Tyler, 1981; Rabenoro & Penneç, 2024) etc.

Finally, the use of flags for statistical analysis has been par-

ticularly well illustrated with the example of *independent subspace analysis* (Hyvärinen & Hoyer, 2000), from which the name of our model is drawn. The authors notice the emergence of phase and shift-invariant features by maximizing the independence between the norms of projections of samples into so-called *independent feature subspaces*. The learning algorithm is later recast as an optimization problem on flag manifolds (Nishimori et al., 2006). Flags also implicitly arise in general subspace methods under the name *mutually orthogonal subspaces*, like in the mutually-orthogonal class-subspaces of Watanabe and Pakvasa (Watanabe & Pakvasa, 1973) and the adaptive-subspace self-organizing maps of Kohonen (Kohonen, 1996). More recently, PCA was also reformulated as an optimization problem on flag manifolds (Pennec, 2018), raising perspectives for multilevel data analysis on manifolds.

7. Discussion

We raised an important issue—the *curse of isotropy*—about the isotropic variability of principal components under Gaussian models with repeated covariance eigenvalues, and showed that these models should often be assumed in practice according to the principle of parsimony. We developed a simple methodology—*principal subspace analysis*—based on generative modeling and flags of subspaces to spot this curse in practice and transition from fuzzy principal components to much-more-interpretable principal subspaces.¹

Principal subspace analysis paves the way to numerous extensions. First, one could deal with non-Gaussian data (exponential families, Gaussians on manifolds (Pennec, 2006), deep generative models etc.). In that case, the maximum likelihood estimate might not be explicit and one might require tools from optimization on Riemannian manifolds (Absil et al., 2009) (flag manifolds (Ye et al., 2022; Zhu & Shen, 2024) or symmetric positive-definite (SPD) matrices (Groisser et al., 2017)). Second, one should investigate more advanced model selection strategies from Bayesian statistics, like an extension of Minka’s criterion (Minka, 2000). If in addition the maximum likelihood estimate is not explicit, then model selection could be embedded in a shrinkage method (Tibshirani, 1996) or stratified optimization (Leygonie et al., 2023; Olikier et al., 2023) framework. Third, since PSA models are nothing but parsimonious Gaussian models, one could simply extend them into parsimonious Gaussian mixture models (Tipping & Bishop, 1999a). The eigenvalue-equalization principle could actually be applied to any problem relying on symmetric matrices, like variational Bayesian methods (Lambert et al., 2023) or spectral geometry problems. Notably, we think that the PSA methodology could extend to spectral graph theory and ap-

plications (Ng et al., 2001; Belkin & Niyogi, 2003; Lefevre et al., 2023), where relatively-close Laplacian eigenvalues are common (related to shape symmetries) and might be especially problematic for spectral embedding and spectral matching (Lombaert et al., 2013). Fourth, any method relying on flags of subspaces (Nishimori et al., 2006; Ma et al., 2021; Szwagier & Pennec, 2023; Mankovich & Birdal, 2023; Mankovich et al., 2024) could benefit from our framework to select an adapted flag type, whose choice has been canonical or heuristic up to now.

Materials and methods

The following self-contained appendix formally introduces the principal subspace analysis model, derives all the results from scratch with proofs, details the associated methodology, and provides additional evidence that the curse of isotropy is an important phenomenon that should not be overlooked in practice.

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.

¹The code is available on GitHub: <https://github.com/tomszwagier/principal-subspace-analysis>.

Abstract This appendix investigates a general family of covariance models with repeated eigenvalues extending probabilistic principal component analysis (PPCA). A geometric interpretation shows that these models are parameterized 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.

A. Reminders on probabilistic principal component analysis

Principal component analysis (PCA) is a ubiquitous tool in statistics, which however lacks a probabilistic formulation. Such a framework can indeed be useful in a variety of contexts like decision-making, generative modeling, missing data and model selection. The Probabilistic PCA model of [Tipping & Bishop \(1999b\)](#) circumvents this issue, and we describe it in this section.

A.1. Model

Let $(x_i)_{i=1}^n$ 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

$$x = Wz + \mu + \epsilon, \quad (6)$$

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

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

$$x \sim \mathcal{N}(\mu, WW^\top + \sigma^2 I_p). \quad (7)$$

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.

A.2. Maximum likelihood

The PPCA model parameters are the shift μ , the linear map W and the noise factor σ^2 . Let some observed dataset $(x_i)_{i=1}^n$, $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$ its mean and $S := \sum_{j=1}^p \ell_j v_j v_j^\top$ its sample covariance matrix, with its eigenvalues $\ell_1 \geq \dots \geq \ell_p \geq 0$ and associated eigenvectors $v_1 \perp \dots \perp v_p$. One can explicitly infer the parameters that are the most likely to have generated these data using maximum likelihood estimation. It is shown in the original PPCA paper 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 $L_q := \text{diag}(\ell_1, \dots, \ell_q)$ (up to the noise) and an orthogonal transformation by the associated q eigenvectors $V_q := [v_1 \dots v_q]$, and finally the most likely noise factor is the average of the $p - q$ discarded eigenvalues

$$\hat{\mu} = \bar{x}, \quad \hat{W} = V_q (L_q - \hat{\sigma}^2 I_q)^{\frac{1}{2}}, \quad \hat{\sigma}^2 = \frac{1}{p - q} \sum_{j=q+1}^p \ell_j. \quad (8)$$

One can then easily express the maximum log-likelihood

$$\ln \hat{\mathcal{L}}(q) := -\frac{n}{2} \left(p \ln(2\pi) + \sum_{j=1}^q \ln \ell_j + (p - q) \ln \left(\frac{1}{p - q} \sum_{j=q+1}^p \ell_j \right) + p \right). \quad (9)$$

A.3. Parsimony and model selection

The previously described PPCA is already a 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 the original PPCA paper

to have $\kappa(q) := pq - \frac{q(q-1)}{2} + 1$ free parameters—the removal of $\frac{q(q-1)}{2}$ parameters being due to the rotational-invariance of the latent variable $z \in \mathbb{R}^q$. 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 discussed in the next section, we can give an insightful geometric interpretation to the number of free parameters in the PPCA model using Stiefel manifolds.

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 (BIC) or more PPCA-oriented ones like Bayesian PCA (Bishop, 1998) or Minka’s criterion (Minka, 2000). They often rely on a tradeoff between goodness-of-fit (via maximum likelihood) and complexity (via the number of parameters), weighted by the number of samples.

A.4. Isotropic PPCA

Isotropic PPCA (IPPCA) 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

$$\Sigma := (a - b)UU^\top + bI_p. \quad (10)$$

Such a parsimonious model is shown to be efficient in high-dimensional classification problems (Bouveyron et al., 2007). 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.

B. Principal subspace 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 PSA, a covariance generative model with a general constraint on the sequence of eigenvalue multiplicities. PSA generalizes PPCA and IPPCA and unifies them in a new family of models parameterized by flag manifolds. Flag manifolds are themselves generalizations of Stiefel manifolds and Grassmannians, hence the link between PPCA, IPPCA and PSA that is detailed in this section.

B.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 $\gamma := (\gamma_1, \gamma_2, \dots, \gamma_d) \in \mathcal{C}(p)$ be a composition of a positive integer p . We define the PSA model of *type* γ as

$$x = \sum_{k=1}^{d-1} \sigma_k Q_k z_k + \mu + \epsilon. \quad (11)$$

In this formula, $\sigma_1 > \dots > \sigma_{d-1} > 0$ are decreasing scaling factors, $Q_k \in \mathbb{R}^{p \times \gamma_k}$ are mutually-orthogonal frames (i.e. they verify $Q_k^\top Q_{k'} = \delta_{kk'} I$ in Kronecker notation) and $z_k \sim \mathcal{N}(0, I_{\gamma_k})$ are independent latent variables. $\mu \in \mathbb{R}^p$, $\sigma^2 > 0$ and $\epsilon \sim \mathcal{N}(0, \sigma^2 I_p)$ are the classical shift, variance and isotropic noise present in PPCA.

Similarly as in PPCA, we can compute the population density

$$x \sim \mathcal{N} \left(\mu, \sum_{k=1}^{d-1} \sigma_k^2 Q_k Q_k^\top + \sigma^2 I_p \right). \quad (12)$$

The expression of the covariance matrix $\Sigma := \sum_k \sigma_k^2 Q_k Q_k^\top + \sigma^2 I_p \in \mathbb{R}^{p \times p}$ can be simplified by gathering all the orthonormal frames into one orthogonal matrix $Q := [Q_1 | \dots | Q_{d-1} | Q_d] \in \mathcal{O}(p)$ where $Q_d \in \mathbb{R}^{p \times \gamma_d}$ is an orthogonal completion of the previous frames. Writing $\Lambda := \text{diag}(\lambda_1 I_{\gamma_1}, \dots, \lambda_d I_{\gamma_d})$, with $\lambda_k := \sigma_k^2 + \sigma^2$ for $k \in [1, d-1]$ and $\lambda_d := \sigma^2$, one gets

$$\Sigma = Q \Lambda Q^\top. \quad (13)$$

Hence, the fitted density of PSA is a multivariate Gaussian with repeated eigenvalues $\lambda_1 > \dots > \lambda_d > 0$ of respective multiplicity $\gamma_1, \dots, \gamma_d$. An illustration of the generative model is provided in Fig. 2. Therefore, PPCA and IPPCA can be seen as PSA models, with respective types $\gamma = (1, \dots, 1, p - q)$ and $\gamma = (q, p - q)$. 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 $\gamma_1, \dots, \gamma_d$, whose direct sum generates the data space. Such a sequence is called a *flag* of linear subspaces of *type* γ . Hence flags are natural objects to geometrically interpret PSA, and so a fortiori PPCA and IPPCA. We detail this point in the next section.

B.2. Type

Just like the latent variable dimension $q \in [0, p - 1]$ is a central notion in PPCA, the type $\gamma \in \mathcal{C}(p)$ is a central notion in PSA. In this subsection, we introduce the concepts of *refinement* and γ -*composition* to make its analysis more convenient.

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

Let $\gamma := (\gamma_1, \gamma_2, \dots, \gamma_d) \in \mathcal{C}(p)$. 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* $\phi_\gamma: [1, p] \rightarrow [1, d]$ to be this surjective map, such that $\phi_\gamma(j)$ 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 γ -*composition* of an object. We give two examples. Let $Q \in \mathcal{O}(p)$. The γ -composition of Q is the sequence $Q^\gamma := (Q_1, \dots, Q_d)$ such that $Q_k \in \mathbb{R}^{p \times \gamma_k}, \forall k \in [1, d]$ and $Q = [Q_1 | \dots | Q_d]$. Let $L := (\ell_1, \dots, \ell_p)$ be a sequence of decreasing eigenvalues. The γ -composition of L is the sequence $L^\gamma := (L_1, \dots, L_d)$ such that $L_k \in \mathbb{R}^{\gamma_k}, \forall k \in [1, d]$ and $L = [L_1 | \dots | L_d]$. We call γ -*averaging* of L the sequence $\overline{L}^\gamma := (\overline{L}_1, \dots, \overline{L}_d) \in \mathbb{R}^d$ of average eigenvalues in L^γ .

B.3. Maximum likelihood

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

$$\ln \mathcal{L}(\mu, \Sigma) = -\frac{n}{2} (p \ln(2\pi) + \ln |\Sigma| + \text{tr}(\Sigma^{-1}C)), \quad (14)$$

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

Theorem B.1. *Let $(x_i)_{i=1}^n$ be a p -dimensional dataset, $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$ its mean and $S := \sum_{j=1}^p \ell_j v_j v_j^\top$ its sample covariance matrix, with $\ell_1 \geq \dots \geq \ell_p \geq 0$ its eigenvalues and $[v_1 | \dots | v_p] := V \in \mathcal{O}(p)$ some associated eigenvectors.*

The maximum likelihood parameters of PSA are

$$\hat{\mu} = \bar{x}, \quad \hat{Q} = V, \quad (\hat{\lambda}_1, \dots, \hat{\lambda}_d) = \overline{(\ell_1, \dots, \ell_p)}^\gamma. \quad (15)$$

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

Proof. Original results about the maximum likelihood estimation of covariance eigenvalues and eigenvectors from multivariate Gaussian distributions with repeated covariance eigenvalues date back from the celebrated paper of (Anderson, 1963). We provide an independent proof for completeness with a particular emphasis on geometry, flags of linear subspaces, and uniqueness. We successively find the optimal $\hat{\mu} \in \mathbb{R}^p$, $\hat{Q} \in \mathcal{O}(p)$ and $\hat{\lambda}_k \in \mathbb{R}$.

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

$$\ln \mathcal{L}(\mu) = -\frac{n}{2} \text{tr}(\Sigma^{-1}C) + \text{constant}, \quad (16)$$

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

$$\hat{\mu} = \arg \min_{\mu \in \mathbb{R}^p} \sum_{i=1}^n (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu) := f(\mu). \quad (17)$$

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

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

$$\ln \mathcal{L}(Q) = -\frac{n}{2} (\ln |\Sigma| + \text{tr}(\Sigma^{-1}S)) + \text{constant}, \quad (18)$$

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

$$\hat{Q} = \arg \min_{Q \in \mathcal{O}(p)} \text{tr}(\Sigma^{-1}S) = \text{tr}(Q\Lambda^{-1}Q^\top S) := g(Q). \quad (19)$$

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

$$dg_{\hat{Q}}(A\hat{Q}) = \text{tr}\left(\left((A\hat{Q})\Lambda^{-1}\hat{Q}^\top + \hat{Q}\Lambda^{-1}(A\hat{Q})^\top\right)S\right) = \text{tr}(A(\Sigma^{-1}S - S\Sigma^{-1})) = 0. \quad (20)$$

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

$$g(Q) = \sum_{k=1}^d \lambda_k^{-1} \left(\sum_{j \in \phi_\gamma^{-1}(\{k\})} \ell_{\psi(j)} \right), \quad (21)$$

where $\psi \in S_p$ is a permutation and $\phi_\gamma^{-1}(\{k\})$ is the set of indexes in the k -th part of the composition γ . We now need to find the permutation $\hat{\psi} \in S_p$ that minimizes g . First, since $\lambda_1 > \dots > \lambda_d > 0$ by assumption, then $(\lambda_1^{-1}, \dots, \lambda_d^{-1})$ is an increasing sequence. Therefore, $(\ell_{\hat{\psi}(\phi_\gamma^{-1}\{1\})}, \dots, \ell_{\hat{\psi}(\phi_\gamma^{-1}\{d\})})$ must be a non-increasing sequence, in that for $k_1 < k_2$, the eigenvalues in the k_1 -th part of γ must be greater than or equal to the eigenvalues in the k_2 -th part. Indeed, for $\lambda < \lambda'$, if $\ell < \ell'$, then $\lambda\ell' + \lambda'\ell < \lambda\ell + \lambda'\ell'$. 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 γ . If so, the minimizing $\hat{\psi}$ is unique up to permutations within each part of γ . Therefore, it is not \hat{Q} itself but the sequence of eigenspaces of \hat{Q} generated by its γ -composition that is unique, and we have $(\text{Im}(\hat{Q}_1), \dots, \text{Im}(\hat{Q}_d)) = (\text{Im}(V_1), \dots, \text{Im}(V_d))$. 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 when S is full-rank. 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 (as defined in Breiding et al. (2018))). Therefore, since sample covariance matrices are measurable functions with absolutely continuous (Gaussian) densities with respect to Lebesgue measure, a randomly drawn matrix S almost surely has distinct eigenvalues. Consequently, its type is $(1, \dots, 1)$, which is a refinement of any possible type in $\mathcal{C}(p)$. Note that the full-rank assumption avoids having multiple null eigenvalues with nonzero measure.

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

$$\ln \mathcal{L}(\Lambda) = -\frac{n}{2} (\ln |\Sigma| + \text{tr}(\Sigma^{-1}S)) + \text{constant}, \quad (22)$$

with $\Sigma = \hat{Q}\Lambda\hat{Q}^\top$. First, one has $\ln |\Sigma| = \sum_{k=1}^d \gamma_k \ln \lambda_k$. Second, according to the previous results, one has $\text{tr}(\Sigma^{-1}S) = \sum_{k=1}^d \lambda_k^{-1} \left(\sum_{j \in \phi_\gamma^{-1}\{k\}} \ell_j \right)$. The optimal eigenvalues $(\hat{\lambda}_1, \dots, \hat{\lambda}_d)$ are thus

$$(\hat{\lambda}_1, \dots, \hat{\lambda}_d) = \arg \min_{\lambda_1, \dots, \lambda_d \in \mathbb{R}} \sum_{k=1}^d \gamma_k \ln \lambda_k + \lambda_k^{-1} \left(\sum_{j \in \phi_\gamma^{-1}\{k\}} \ell_j \right) := h(\lambda_1, \dots, \lambda_d). \quad (23)$$

As $\frac{\partial h}{\partial \lambda_k} = \frac{\gamma_k}{\lambda_k} - \lambda_k^{-2} \left(\sum_{j \in \phi_{\gamma^{-1}}\{k\}} \ell_j \right)$, we get that $\hat{\lambda}_k = \frac{1}{\gamma_k} \left(\sum_{j \in \phi_{\gamma^{-1}}\{k\}} \ell_j \right) = \overline{L}_k$. \square

One can then easily express the maximum log-likelihood of PSA

$$\ln \hat{\mathcal{L}}(\gamma) = -\frac{n}{2} \left(p \ln(2\pi) + \sum_{k=1}^d \gamma_k \ln \overline{L}_k + p \right). \quad (24)$$

B.4. Geometric interpretation with flag manifolds

As discussed in the previous subsections, the appropriate parameter space for Q in PSA is the space of flags of type γ , noted $\text{Flag}(\gamma)$. The geometry of such a space is well known (Ye et al., 2022). In a few words, each subspace \mathcal{V}_k of dimension γ_k can be endowed with an orthonormal basis $Q_k := [q_k^1 | \dots | q_k^{\gamma_k}] \in \mathbb{R}^{p \times \gamma_k}$. This basis is invariant to rotations within the subspace—i.e. for $R_k \in \mathcal{O}(\gamma_k)$, $Q'_k := Q_k R_k$ is still an orthonormal basis of \mathcal{V}_k . Concatenating such orthonormal frames for all the mutually-orthogonal subspaces of a flag creates an orthogonal matrix $Q := [Q_1 | \dots | Q_d] \in \mathcal{O}(p)$. Eventually, $\text{Flag}(\gamma)$ is a smooth quotient manifold, consisting in equivalence classes of orthogonal matrices:

$$\text{Flag}(\gamma) \cong \mathcal{O}(p) / (\mathcal{O}(\gamma_1) \times \dots \times \mathcal{O}(\gamma_d)). \quad (25)$$

This result enables the accurate computation of the number of parameters in PSA. Before that, let us note that the other parameters are $\mu \in \mathbb{R}^p$ and $\Lambda \in \text{D}(\gamma) := \{\text{diag}(\lambda_1 I_{\gamma_1}, \dots, \lambda_d I_{\gamma_d}) \in \mathbb{R}^{p \times p} : \lambda_1 > \dots > \lambda_d > 0\}$, which can be seen as a convex cone of \mathbb{R}^d .

Proposition B.2. *The number of free parameters in PSA is*

$$\kappa(\gamma) := p + d + \frac{p(p-1)}{2} - \sum_{k=1}^d \frac{\gamma_k(\gamma_k-1)}{2}. \quad (26)$$

This geometric interpretation sheds light on PPCA, which—we remind—is a special case of PSA with $\gamma = (1, \dots, 1, p-q)$. First, as flags of type $(1, \dots, 1, p-q)$ belong to Stiefel manifolds (up to changes of signs), we can naturally parameterize 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. PSA 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 PSA with $\gamma = (q, p-q)$. Hence we can parameterize it with flags of type $(q, p-q)$, which belong to Grassmannians. With that in mind, we notice that our formula (26) differs from the one given in Bouveyron et al. (2011). We think that this paper overestimates 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.

C. Model selection

As discussed previously, 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 PSA model selection. First, we study the inference of two adjacent eigenvalues and their associated eigenvectors. We show that when the relative eigengap is small and the number of samples is limited, one should prefer a PSA model with repeated eigenvalues—i.e. block-average 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 PSA models.

C.1. Bayesian information criterion

The Bayesian information criterion (BIC) is defined as

$$\text{BIC}(\gamma) = \kappa(\gamma) \ln n - 2 \ln \hat{\mathcal{L}}(\gamma), \quad (27)$$

where κ is the number of free parameters (26) and $\ln \hat{\mathcal{L}}$ is the maximum log-likelihood (24). It is a widely-used model selection criterion, making a tradeoff between model complexity κ and goodness-of-fit $\hat{\mathcal{L}}$. The formula results from an asymptotic approximation of the model evidence. In this section, we use the BIC for PSA model selection. The model with lowest BIC is considered as the best model. In the two-eigenvalue case, we get an explicit criterion based on eigenvalue gaps to decide if we must assume that they are equal, and in the more general case, we propose efficient model comparison strategies. We also investigate other model selection criteria than the BIC for completeness in this section, and get similar conclusions.

C.2. The two-eigenvalue case

In order to better understand the dynamics of PSA model selection, we lead the experiment of quantifying the BIC variation induced by the equalization of two adjacent eigenvalues. More precisely and without loss of generality, we compare the BIC of a *full covariance model* $\gamma = (1, \dots, 1)$ to the one of an *equalized covariance model* $\gamma' = (1 \dots 1, 2, 1 \dots 1)$, where the eigenvalue λ_j has multiplicity 2.

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

$$\delta_j < 2 \left(1 - n^{\frac{2}{n}} + n^{\frac{1}{n}} \sqrt{n^{\frac{2}{n}} - 1} \right), \quad (28)$$

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

Proof. Since n and p are constant within model selection, the BIC can be rewritten (up to constant terms and factors) as

$$\text{BIC}(\gamma) := \left(d - \sum_{k=1}^d \frac{\gamma_k(\gamma_k - 1)}{2} \right) \frac{\ln n}{n} + \sum_{k=1}^d \gamma_k \ln \bar{L}_k. \quad (29)$$

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

$$\Delta \text{BIC} = p \frac{\ln n}{n} + \sum_{k=1}^p \ln \ell_k - (p-2) \frac{\ln n}{n} - \sum_{k \notin \{j, j+1\}} \ln \ell_k - 2 \ln \left(\frac{\ell_j + \ell_{j+1}}{2} \right), \quad (30)$$

$$= 2 \frac{\ln n}{n} + \ln \ell_j + \ln \ell_{j+1} - 2 \ln \left(\frac{\ell_j + \ell_{j+1}}{2} \right), \quad (31)$$

$$= 2 \frac{\ln n}{n} + \ln \ell_j + \ln (\ell_j (1 - \delta_j)) - 2 \ln \left(\frac{\ell_j (2 - \delta_j)}{2} \right), \quad (32)$$

$$= 2 \frac{\ln n}{n} + \ln (1 - \delta_j) - 2 \ln \left(1 - \frac{\delta_j}{2} \right), \quad (33)$$

$$= 2 \frac{\ln n}{n} - \ln \left(\frac{\left(1 - \frac{\delta_j}{2} \right)^2}{1 - \delta_j} \right). \quad (34)$$

Hence, one has

$$\Delta \text{BIC} = 0 \iff \exp \left(2 \frac{\ln n}{n} \right) = \frac{\left(1 - \frac{\delta_j}{2} \right)^2}{1 - \delta_j} \iff \frac{\delta_j^2}{4} - \left(1 - \exp \left(2 \frac{\ln n}{n} \right) \right) \delta_j + 1 - \exp \left(2 \frac{\ln n}{n} \right) = 0. \quad (35)$$

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

$$\delta(n) = 2 - 2 \exp \left(2 \frac{\ln n}{n} \right) + 2 \sqrt{\exp \left(4 \frac{\ln n}{n} \right) - \exp \left(2 \frac{\ln n}{n} \right)}. \quad (36)$$

□

C.3. Comparison with North’s rule-of-thumb

A rule-of-thumb for determining which sample eigenvalue pairs might lead to large PC sampling error is proposed in North et al. (1982). The authors show that the asymptotic sampling error of a population eigenvalue λ is $\Delta\lambda := \lambda(\frac{2}{n})^{\frac{1}{2}}$ in the Gaussian setting. North’s rule-of-thumb (NRT) states that when one population eigenvalue’s sampling error is comparable to or larger than its distance to an adjacent eigenvalue, then the PC’s sampling error is comparable to the associated adjacent PC. Note that this is not an explicit rule (compared to our relative eigengap threshold (28)) since one has to choose the level of uncertainty, and—most of all—it is based on the *true* eigenvalues (on which the confidence intervals are based) which are unknown. However, this rule has been applied in many contexts and it is commonly implemented in the following way (Taylor, 2022). For each sample eigenvalue pair $\ell_j \geq \ell_{j+1}$, compute the 1 sigma error intervals $I_j = [\ell_j - \ell_j\sqrt{\frac{2}{n}}, \ell_j + \ell_j\sqrt{\frac{2}{n}}]$ and $I_{j+1} = [\ell_{j+1} - \ell_{j+1}\sqrt{\frac{2}{n}}, \ell_{j+1} + \ell_{j+1}\sqrt{\frac{2}{n}}]$. If $I_j \cap I_{j+1} \neq \emptyset$, then the associated principal components suffer from large sampling errors and might be random mixtures of the true eigenvectors. We reformulate it as a relative eigengap threshold.

Proposition C.2. *North’s rule-of-thumb (as implemented in practice) boils down to the relative eigengap threshold*

$$\delta_j \leq \frac{2\sqrt{\frac{2}{n}}}{1 + \sqrt{\frac{2}{n}}}. \quad (37)$$

Proof. The sampling error interval overlap condition writes as

$$\ell_j - \sqrt{\frac{2}{n}}\ell_j \leq \ell_{j+1} + \sqrt{\frac{2}{n}}\ell_{j+1} \iff \frac{\ell_j - \ell_{j+1}}{\ell_j} \leq \sqrt{\frac{2}{n}} \left(1 + \frac{\ell_{j+1}}{\ell_j}\right), \quad (38)$$

$$\iff \frac{\ell_j - \ell_{j+1}}{\ell_j} \leq \sqrt{\frac{2}{n}} \left(2 - \frac{\ell_j - \ell_{j+1}}{\ell_j}\right), \quad (39)$$

$$\iff \frac{\ell_j - \ell_{j+1}}{\ell_j} \leq \frac{2\sqrt{\frac{2}{n}}}{1 + \sqrt{\frac{2}{n}}}. \quad (40)$$

□

This threshold is reported in Fig. 10, under the name NRT-1 (for 1 sigma sampling errors). We also report North’s rule-of-thumb for 2 sigma sampling errors (NRT-2), yielding a relative eigengap threshold of $\frac{4\sqrt{\frac{2}{n}}}{1+2\sqrt{\frac{2}{n}}}$. We see that the relative eigengap NRT-1 is much smaller than ours (e.g. 8.6% instead of 21% for 1000 samples). Therefore, although warning scientists about close sample eigenvalues in principal component analysis, North’s rule-of-thumb largely overlooks the curse of isotropy compared to our method. To see the practical effect of this lower threshold, we test this condition on the same real datasets as in Fig. 4. The results are in Fig. 11. We can see that the curse of isotropy remains a nonnegligible phenomenon with North’s rule, even though it is less marked than with the BIC. We think that North’s rule (as implemented in practice) underestimates the phenomenon, notably because it uses 1 sigma uncertainties and since it is based on sample eigenvalues instead of true eigenvalues in the implementations. We recall that 1 sigma uncertainties (NRT-1) correspond to 68% error bars while 2 sigma uncertainties (NRT-2) correspond to 95% error bars and yield a relative eigengap threshold of 16%, which is much closer to our results with the BIC. An interesting perspective would be to consider our guideline instead of the less-impactful North’s rule in seminal climate science papers which made some conclusions out of possibly degenerate principal components.

C.4. Comparison with other model selection criteria

Although being widely used in model selection, the BIC is well-known for its heavy complexity penalization, tending to select over-parsimonious models (Burnham & Anderson, 2004). Another widely-used criterion is the Akaike information criterion (Akaike, 1974). It is defined as

$$\text{AIC}(\gamma) = 2\kappa(\gamma) - 2\ln \hat{\mathcal{L}}(\gamma) \quad (41)$$

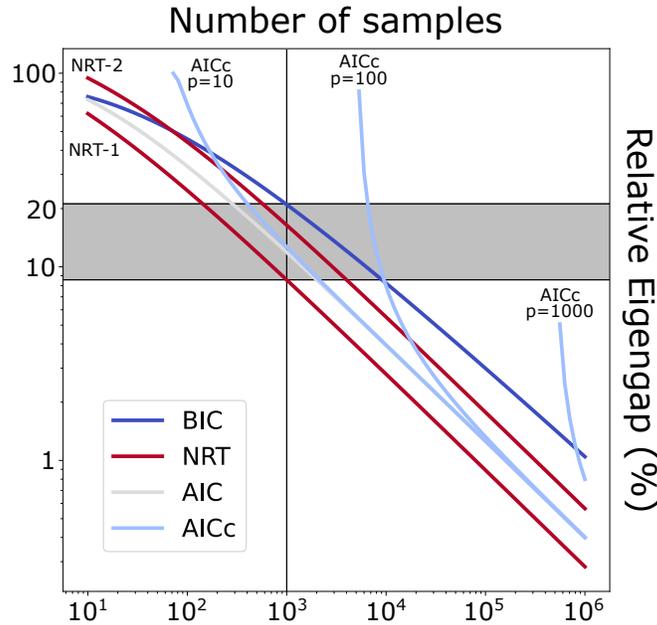


Figure 10. Theoretical relative eigengap curves—giving the relative distance under which two adjacent eigenvalues should be considered as equal—as a function of n , for the BIC (Thm. C.1) in dark blue, North’s rule-of-thumb (Prop. C.2) in dark red with 1 sigma (NRT-1) and 2 sigma (NRT-2) uncertainties, AIC (Prop. C.3) in gray, and AICc (Prop. C.4) in light blue with $p \in \{10, 100, 1000\}$. We can see that NRT-1 (which is the one implemented in practice) has globally the lowest relative eigengaps, followed by AIC, AICc, NRT-2 and BIC. Therefore, the BIC more frequently equalizes close-eigenvalues than other methods and has the biggest impact on the PCA methodology. Anyhow, all the methods substantiate the importance of the curse of isotropy. For instance for $n = 1000$, all the methods have a relative eigengap threshold roughly between 10% and 20%.

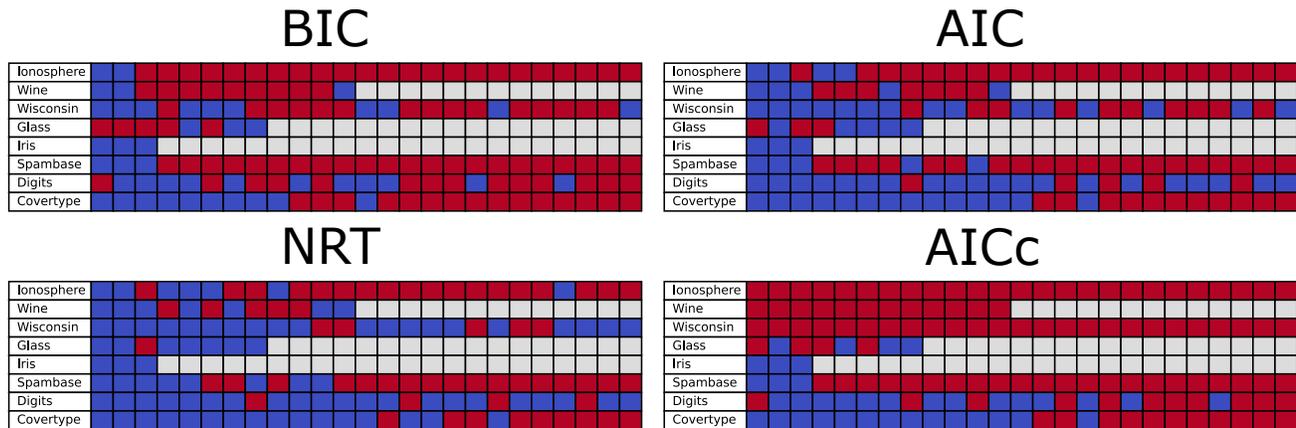


Figure 11. Practical effect of the different relative eigengap conditions using the BIC (Thm. C.1), NRT (Prop. C.2), AIC (Prop. C.3) and AICc (Prop. C.4) for several classical datasets from the UCI Machine Learning repository. A red case in column j indicates that the $(j, j + 1)$ eigenvalue pair is below the relative eigengap threshold and should be equalized. Blue indicates above, and gray that the pair does not exist. We only plot the 25 leading eigenvalue pairs. We can see that all the methods suggest to consider principal subspaces of dimension greater than 1, including North’s rule-of-thumb which has the lowest relative eigengap thresholds (cf. Fig. 10). Moreover, the number of eigenvalues to equalize seems to increase from NRT to AIC to BIC, but the low-sample correction of AIC seems to equalize even more eigenvalues than the BIC. This stresses that accounting for low-sample sizes is an important issue in the curse of isotropy.

where κ is the number of free parameters (26) and $\ln \hat{\mathcal{L}}$ is the maximum log-likelihood (24). Comparing an equalized covariance model to one with distinct eigenvalues like in Thm. C.1 but this time using the AIC yields another relative eigengap condition.

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

$$\delta_j < 2 \left(1 - e^{\frac{4}{n}} + e^{\frac{2}{n}} \sqrt{e^{\frac{4}{n}} - 1} \right) \quad (42)$$

then the equalized covariance model has a lower AIC than the full one.

Proof. The proof is essentially the same as the one of Thm. C.1. Since n and p are constant within model selection, the AIC can be rewritten (up to constant terms and factors) as

$$\text{AIC}(\gamma) := \left(d - \sum_{k=1}^d \frac{\gamma_k(\gamma_k - 1)}{2} \right) \frac{2}{n} + \sum_{k=1}^d \gamma_k \ln \bar{L}_k \quad (43)$$

Replacing $\frac{\ln n}{n}$ with $\frac{2}{n}$ in the proof of Thm. C.1, we finally get the result that $\delta(n) = 2 - 2 \exp\left(\frac{4}{n}\right) + 2 \sqrt{\exp\left(\frac{8}{n}\right) - \exp\left(\frac{4}{n}\right)}$. \square

This threshold is reported in Fig. 10. We see that this relative eigengap is smaller than ours (28) (e.g. 12% instead of 21% for 1000 samples), but higher than North's rule (37). This result is interesting since AIC is known for tending to select overparameterized models, especially for small sample sizes (Burnham & Anderson, 2004) (cf. next paragraph). Despite this, the relative eigengap condition with AIC is more impactful than North's rule. To see the practical effect of the AIC threshold of (42), we also report the relative eigengap condition on real datasets in Fig. 11. We see that many eigenvalue pairs should be assumed equal—slightly less than with BIC. Therefore, even with another model selection criterion, the curse of isotropy is still a nonnegligible phenomenon in real datasets, and the principal subspace analysis methodology enables to leverage it to improve interpretability.

Additionally, we provide a relative eigengap condition for the AICc (Hurvich & Tsai, 1989), which is a small-sample correction to the AIC. In practice, the AICc is advised over the AIC for $n/\kappa < 40$ (Burnham & Anderson, 2004). The AICc is defined as

$$\text{AICc}(\gamma) = 2\kappa(\gamma) \frac{n}{n - \kappa(\gamma) - 1} - 2 \ln \hat{\mathcal{L}}(\gamma) \quad (44)$$

where κ is the number of free parameters (26) and $\ln \hat{\mathcal{L}}$ is the maximum log-likelihood (24). One can see that this corrected criterion converges asymptotically to the AIC. Comparing an equalized covariance model to one with distinct eigenvalues like in Thm. C.1 but this time with the AICc yields the following relative eigengap condition.

Proposition C.4. *Let $(x_i)_{i=1}^n$ be a p -dimensional dataset with $n > \frac{p(p+3)}{2} + 1$ samples, $\ell_j \geq \ell_{j+1}$ two adjacent sample eigenvalues, $\delta_j = \frac{\ell_j - \ell_{j+1}}{\ell_j}$ their relative eigengap and $\varphi = \frac{4n-4}{(n - \frac{p(p+3)}{2})^2 - 1}$. If*

$$\delta_j < 2 \left(1 - e^\varphi + e^{\frac{\varphi}{2}} \sqrt{e^\varphi - 1} \right) \quad (45)$$

then the equalized covariance model has a lower AICc than the full one.

Proof. The proof is essentially the same as in Thm C.1 and Prop. C.3. Since n and p are constant within model selection, the AICc can be rewritten (up to constant terms and factors) as

$$\text{AICc}(\gamma) := \frac{2\kappa(\gamma)}{n - \kappa(\gamma) - 1} + \sum_{k=1}^d \gamma_k \ln \bar{L}_k \quad (46)$$

We compare the AICc of the full covariance model $\gamma = (1, \dots, 1)$ to the one of the equalized covariance model $\gamma' = (1, \dots, 1, 2, 1, \dots, 1)$ where the j -th eigenvalue has been equalized with the $j + 1$ -th. This boils down to studying the sign of the function $\Delta \text{AICc} = \text{AICc}(\gamma) - \text{AICc}(\gamma')$. One gets

$$\Delta \text{AICc} = \frac{p(p+3)}{n - \frac{p(p+3)}{2} - 1} - \frac{p(p+3) - 4}{n - \left(\frac{p(p+3)}{2} - 2\right) - 1} + \ln \ell_j + \ln \ell_{j+1} - 2 \ln \left(\frac{\ell_j + \ell_{j+1}}{2}\right) \quad (47)$$

$$= \frac{4n - 4}{\left(n - \frac{p(p+3)}{2}\right)^2 - 1} + \ln \ell_j + \ln \ell_{j+1} - 2 \ln \left(\frac{\ell_j + \ell_{j+1}}{2}\right) \quad (48)$$

Replacing $2 \frac{\ln n}{n}$ with $\varphi = \frac{4n-4}{\left(n - \frac{p(p+3)}{2}\right)^2 - 1}$ in the proof of Thm. C.1, we finally get the result that $\delta(n) = 2 - 2 \exp(\varphi) + 2\sqrt{\exp(2\varphi) - \exp(\varphi)}$. \square

Contrary to the other criteria (Thm C.1, Prop. C.2 and Prop. C.3), this threshold depends on the dimension p . Therefore, we plot it for several p in Fig. 10. We can see that this relative eigengap converges to the AIC for large n , but is higher than the one with the BIC (28) when the number of samples is close to the number of model parameters. We also test this condition on the same real datasets as in Fig. 4 and report the results in Fig. 11. We see that many eigenvalue pairs are ill-defined, especially in high-dimensional datasets where those are even more numerous than with the BIC.

C.5. Efficient model selection

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 .

The characterization of the PSA family structure is a bit more technical, as it requires to study the hierarchy of types, involving the concept of integer composition. Fortunately, this analysis can be lifted to the stratification of symmetric matrices according to the multiplicities of the eigenvalues, which is already well-known (Arnold, 1972; Groisser et al., 2017; Breiding et al., 2018). Therefore, without proof, we can state the following result.

Proposition C.5. *The family of p -dimensional PSA models induces a stratification of the space of symmetric positive-definite (SPD) matrices S_p^{++} according to the type γ . The refinement relation \preceq makes it a partially ordered set of cardinal 2^{p-1} .*

Hence the set of PSA models at a given data dimension can be represented using a Hasse diagram, as done in Fig. 5. We see that PSA contains PPCA, IPPCA, and many new models. PSA 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, therefore we need to define strategies for selecting only a few number of models to compare. The previously derived partial order \preceq on the set of PSA models allows simple efficient strategies for model selection. In the following subsections, we detail those strategies and prove additional properties.

C.5.1. RELATIVE EIGENGAP THRESHOLD CLUSTERING OF EIGENVALUES

The *relative eigengap threshold strategy* consists in clustering the eigenvalues whose relative eigengap $\delta_j := \frac{\ell_j - \ell_{j+1}}{\ell_j}$ is below a given threshold, e.g. the one of Thm. C.1. This clustering uniquely determines a PSA type γ , from which we apply maximum likelihood estimation, i.e. we block-average the corresponding eigenvalue clusters. This rule is extremely simple but it may select overly parsimonious models, since distant eigenvalues may end up in the same cluster by propagation. Therefore, we provide a more-advanced strategy in the following subsection.

C.5.2. HIERARCHICAL CLUSTERING OF EIGENVALUES

In this strategy, the subset of candidate models is generated by the *hierarchical clustering* 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. More precisely, here we choose a *continuous* pairwise distance δ between adjacent eigenvalues (such as the relative eigengap defined in Thm. C.1), and a linkage criterion Δ between eigenvalue clusters, making sense with respect to our model selection problem (such as the single-linkage

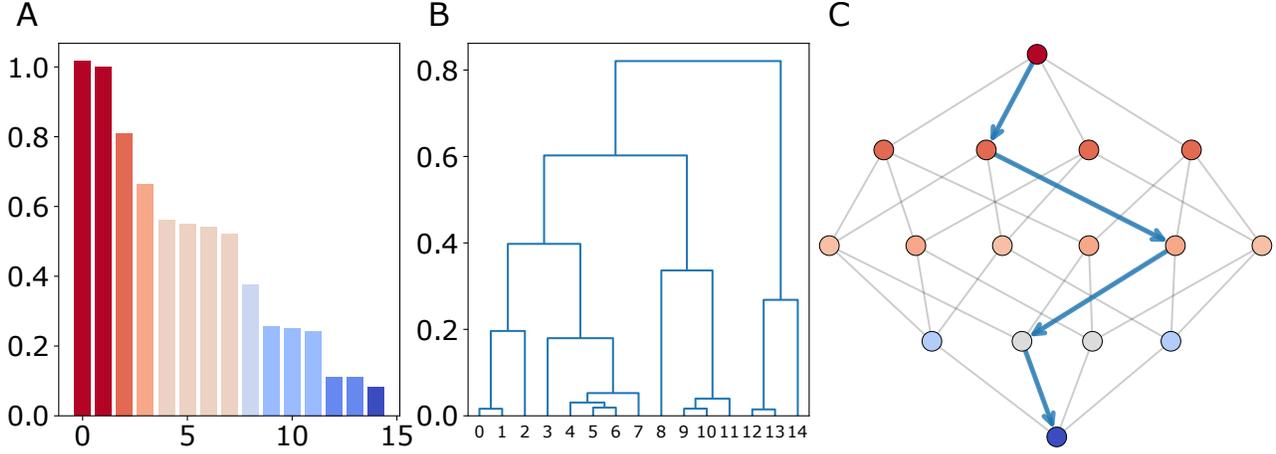


Figure 12. Hierarchical clustering of sample eigenvalues, using the relative eigengap distance for δ and the centroid-linkage criterion for Δ . (A) Sample eigenvalues, whose colors correspond to a given step $t = 8$ of the hierarchical clustering, with $\gamma^t = (2, 1, 1, 4, 1, 3, 2, 1)$. (B) Hierarchical clustering dendrogram. (C) Conceptual illustration of the hierarchical clustering strategy. This heuristic generates a sequence of PSA models $(\gamma^t)_{t=1}^p$ of decreasing complexity, starting from the full covariance model and ending at the isotropic covariance model. This can be visualized as a trajectory in the Hasse diagram of PSA models (cf. Fig. 5).

criterion $\Delta(\Lambda_1, \Lambda_2) = \min_{\ell_1, \ell_2 \in \Lambda_1 \times \Lambda_2} \delta(\ell_1, \ell_2)$ or the centroid-linkage criterion $\Delta(\Lambda_1, \Lambda_2) = \delta(\overline{\Lambda_1}, \overline{\Lambda_2})$. The method is detailed in Algorithm 1 and illustrated in Figure 12.

Algorithm 1 Hierarchical clustering strategy for PSA model selection

Input: $\ell_1 \geq \dots \geq \ell_p, \Delta$

Output: $(\gamma^t)_{t=1}^p$

$\gamma^1 \leftarrow (1, \dots, 1)$, $\Lambda^1 \leftarrow (\{\ell_1\}, \dots, \{\ell_p\})$

for $t = 1 \dots p - 1$ **do**

$\Delta^t \leftarrow (\Delta(\Lambda_1^t, \Lambda_2^t), \dots, \Delta(\Lambda_{p-t}^t, \Lambda_{p-t+1}^t))$

$k^t \leftarrow \arg \min \Delta^t$

$\Lambda^{t+1} \leftarrow (\Lambda_1^t, \dots, \Lambda_{k^t-1}^t, \Lambda_{k^t}^t \cup \Lambda_{k^t+1}^t, \Lambda_{k^t+2}^t, \dots, \Lambda_d^t)$

$\gamma^{t+1} \leftarrow (\gamma_1^t, \dots, \gamma_{k^t-1}^t, \gamma_{k^t}^t + \gamma_{k^t+1}^t, \gamma_{k^t+2}^t, \dots, \gamma_d^t)$

end for

sample eigenvalues and distance
hierarchical subfamily of PSA models
initialize with full covariance model

compute distances between adjacent clusters
find clusters with minimal distance
merge the two clusters of eigenvalues
update the model type

The hierarchical clustering strategy creates a *trajectory* $(\gamma^t)_{t=1}^p$ in the Hasse diagram of PSA models (cf. Fig. 5). The sequence starts from $\gamma^1 = (1, \dots, 1)$, 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, $\gamma^p = (p)$, 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 strategy.

The hierarchical clustering strategy 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 strategy, we show the following consistency result.

Proposition C.6 (Asymptotic consistency of the hierarchical clustering strategy). *The hierarchical clustering strategy generates a subfamily of PSA models that almost surely contains the true PSA model for n large enough.*

Proof. Let us assume that the true generative model is stratified with type $\gamma \in \mathcal{C}(p)$. We can then write the population covariance matrix as $\Sigma = \sum_{k=1}^d \lambda_k Q_k Q_k^\top$ with $\lambda_1 > \dots > \lambda_d > 0$ and $Q := [Q_1 | \dots | Q_d] \in \mathcal{O}(p)$. Let n be the number of independent samples and $S_n := \sum_{j=1}^p \ell_j(S_n) v_j(S_n) v_j(S_n)^\top$ with $\ell_1 \geq \dots \geq \ell_p$ and $V := [v_1 | \dots | v_p] \in \mathcal{O}(p)$. According to Tyler (1981), Lemma 2.1 (i), one then has almost surely, as n goes to infinity, $\ell_j(S_n) \rightarrow \lambda_{\phi_\gamma(j)}$, where ϕ_γ

is the γ -composition function. Hence for n large enough, by continuity of the distance function Δ , the gaps between eigenvalues in the same part of the γ -composition will be arbitrarily close to 0, while the other will be arbitrarily close to the true values $\{\Delta(\lambda_k, \lambda_{k+1}), k \in [1, d-1]\}$, which are all positive. Hence the hierarchical clustering method will first agglomerate the eigenvalues that are in the same part of γ , and second the distinct blocks, by increasing order of pairwise distance. The last model of the first phase will be exactly the true model. \square

Hence, the hierarchical clustering strategy generates a hierarchical subfamily of models of decreasing complexities, including the true PSA model for n large enough. The true model can be then recovered using asymptotically consistent model selection criteria on the subfamily. We now propose a second strategy 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.

C.5.3. PRIOR ON THE NUMBER OF DISTINCT EIGENVALUES

In this strategy, we perform model selection at a given level of the Hasse diagram (cf. Fig. 5). 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 strategy reduces the search space like the previous strategy, this time to $\binom{p-1}{d-1}$ models. In contrast to the hierarchical clustering strategy 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 strategy, we could use the BIC to choose the best model among this reduced family. For completeness, we provide an additional criterion that is nothing but the maximum likelihood itself. We indeed manage to extend to PSA the surprising result from [Bouveyron et al. \(2011\)](#) stating 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 PSA as well.

Proposition C.7 (Asymptotic consistency of the maximum likelihood for fixed d). *If the true PSA model has d distinct eigenvalues, then maximum likelihood model selection within the subfamily of PSA models of type-length d almost surely recovers the true model for n large enough.*

Proof. Let us assume that the true generative model is stratified with type $\gamma^* := (\gamma_1^*, \dots, \gamma_d^*)$, of length d , and let $\lambda_1 > \dots > \lambda_d > 0$ 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 PSA model of type $\gamma := (\gamma_1, \dots, \gamma_d)$, the maximum likelihood writes

$$\ln \hat{\mathcal{L}} \sim -\frac{n}{2} \left(p \ln 2\pi + \sum_{k=1}^d \gamma_k \ln \left(\frac{1}{\gamma_k} \sum_{j \in \phi_{\gamma^{-1}}\{k\}} \lambda_{\phi_{\gamma^*}(j)} \right) \right). \quad (49)$$

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

$$\arg \min_{\substack{\gamma \in \mathcal{C}(p) \\ \#\gamma=d}} \sum_{k=1}^d \gamma_k \ln \left(\frac{1}{\gamma_k} \sum_{j \in \phi_{\gamma^{-1}}\{k\}} \lambda_{\phi_{\gamma^*}(j)} \right) := f(\gamma). \quad (50)$$

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

$$f(\gamma) \geq \sum_{k=1}^d \gamma_k \left(\sum_{k'=1}^d \frac{c_{kk'}}{\gamma_k} \ln \lambda_{k'} \right) = \sum_{k,k'=1}^d c_{kk'} \ln \lambda_{k'} = \sum_{k'=1}^d \gamma_{k'}^* \ln \lambda_{k'} = f(\gamma^*). \quad (51)$$

To conclude, asymptotically, γ^* -PSA is the most likely model. Hence, the maximum likelihood criterion alone finds the true model among the family of PSA models with the same type length. \square

Hence we derived three simple strategies for model selection, taking into account the structure of the PSA models family.

Remark C.8. Many variants can be adopted depending on the problem at hand. For instance if the noise is known, or assumed with some explained variance ratio rules, one can first search for the associated intrinsic dimension q like in classical PCA, and then try to equalize some of the q first eigenvalues by optimizing the model selection criterion over the subfamily of models whose $p - q$ last eigenvalues are all equal.

Remark C.9. In high dimensions, some eigenvalues might be very small or even null. The case of small positive eigenvalues may yield high relative eigengaps in the last eigenvalue pairs—therefore PSA model selection tends to separate those eigenvalues—whereas those are traditionally considered as noise. The case of null eigenvalues yields undefined PSA models. To circumvent those two issues, a classical trick is the one of *covariance regularization*, consisting in adding a small constant to all the covariance eigenvalues. This somewhat boils down to adding an isotropic Gaussian noise to the data. This has notably the effect of diminishing the relative eigengaps, especially for the small positive or null eigenvalues. Another possibility would be to constrain the model types to have at least the last $p - q$ eigenvalues equal, where q is chosen sufficiently small such that the first q eigenvalues are large enough. This is left for future research.

D. Additional experiment — Dynamics of model selection

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

In order to better understand how this result applies in practice, we make the following PSA 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 PSA 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 in Figure 13, where we plot only a few models among the 16 for readability. First, on the BIC plots, we can see that for $n \leq 6000$, PSA discloses a whole family of models that better explain the 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 PSA 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, PSA can more confidently distinguish the sample eigenvalues. Third, on the Hasse diagram, we can see that PSA follows a trajectory as the number of available samples increases, which recalls the kind of subfamily generated by the hierarchical clustering strategy (cf. Figure 12). To conclude, we see on this synthetic example that PSA achieves a better complexity/goodness-of-fit tradeoff than PPCA in a wide range of sample sizes by equalizing the highest eigenvalues.

E. Information about datasets

In this section, we give a few more details about the data used for the experiments.

E.1. Natural image patches

In this experiment, we consider 10 flower images from the ImageNet database (Deng et al., 2009). Those were downloaded from Kaggle (<https://www.kaggle.com/datasets/prasunroy/natural-images>) and extracted from the folder `natural_images/flower/` from `flower_0000.jpg` up to `flower_0009.jpg`.

E.2. Eigenfaces

In this experiment, we consider 31 digital images from the CMU Face Images database (Mitchell, 1997). Those were downloaded from Kaggle (<https://www.kaggle.com/datasets/raviprakash22/cmu-face-images>) and extracted from the folder `faces/faces/choon`. We only extracted the $(60, 64)$ images, corresponding to all the files ending with `_2.pgm`.

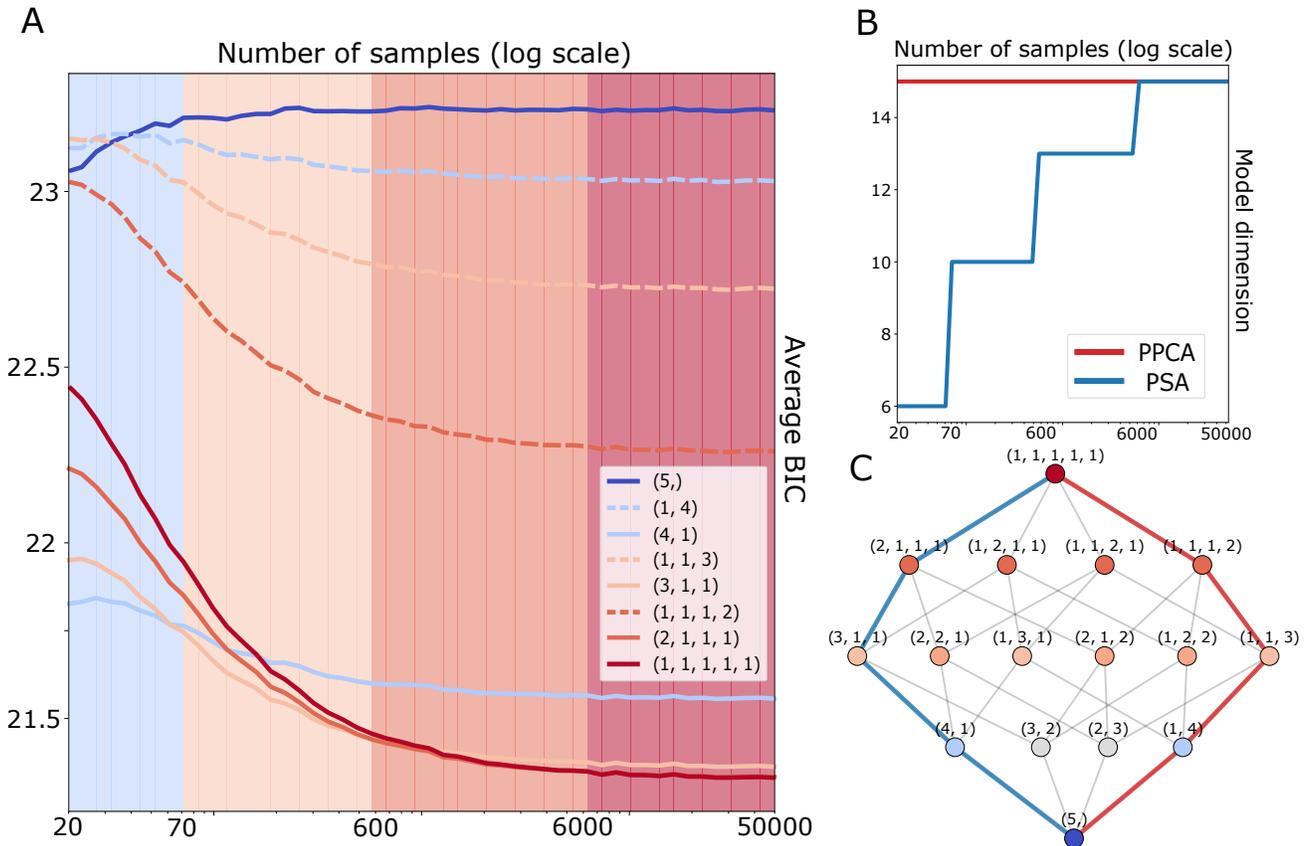


Figure 13. PSA model selection using the BIC for an increasing number of available samples. (A) Each curve represents the average BIC of a given PSA model over several independent experiments. The lowest curve at a given n (horizontal coordinate) therefore corresponds to the most selected model. The 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 $\gamma = (4, 1)$. For $n \in [70, 600]$ (light orange), it is $\gamma = (3, 1, 1)$. For $n \in [600, 6000]$ (orange), it is $\gamma = (2, 1, 1, 1)$. And for $n \in [6000, 50000]$ (red), it is $\gamma = (1, 1, 1, 1, 1)$. (B) Comparison of the complexities of the mostly selected models within the whole PSA family (blue) and within the PPCA family only (red). (C) PSA Hasse diagram. The blue curve corresponds to the trajectory followed by the optimal PSA 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.

E.3. Structured data

For the structured data experiment (cf. Fig. 9) and the relative eigengap tables (cf. Fig. 4 and Fig. 11), we consider data from the UCI Machine Learning Repository (<https://archive.ics.uci.edu/>): Ionosphere (Sigillito et al., 1989), Wine (Aeberhard & Forina, 1991), Wisconsin (Wolberg et al., 1995), Glass (German, 1987), Iris (Fisher, 1936), Spambase (Hopkins et al., 1999), Digits (Alpaydin & Kaynak, 1998), Coverttype (Blackard, 1998).

References

- Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009. ISBN 978-1-4008-3024-4.
- Aeberhard, S. and Forina, M. Wine, 1991. *UCI Machine Learning Repository*. doi: 10.24432/C5PC7J.
- Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. ISSN 1558-2523. doi: 10.1109/TAC.1974.1100705.

- Alpaydin, E. and Kaynak, C. Optical Recognition of Handwritten Digits, 1998. *UCI Machine Learning Repository*. doi: 10.24432/C50P49.
- Anderson, T. W. Asymptotic Theory for Principal Component Analysis. *The Annals of Mathematical Statistics*, 34(1): 122–148, 1963. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177704248.
- Arnold, V. I. Modes and quasimodes. *Functional Analysis and Its Applications*, 6(2):94–101, 1972. ISSN 1573-8485. doi: 10.1007/BF01077511.
- Basri, R. and Jacobs, D. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003. ISSN 1939-3539. doi: 10.1109/TPAMI.2003.1177153.
- Belkin, M. and Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373–1396, 2003. ISSN 0899-7667. doi: 10.1162/089976603321780317.
- Besson, G. On the multiplicity of the eigenvalues of the Laplacian. In *Geometry and Analysis on Manifolds*, pp. 32–53. Springer, 1988. ISBN 978-3-540-45930-9. doi: 10.1007/BFb0083045.
- Bishop, C. Bayesian PCA. In *NeurIPS*, volume 11, 1998.
- Bishop, C. M. *Pattern recognition and machine learning*. Springer, 2006. ISBN 978-0-387-31073-2.
- Blackard, J. Covertypes, 1998. *UCI Machine Learning Repository*. doi: 10.24432/C50K5N.
- Bouveyron, C., Girard, S., and Schmid, C. High-dimensional data clustering. *Computational Statistics & Data Analysis*, 52(1):502–519, 2007. ISSN 0167-9473. doi: 10.1016/j.csda.2007.02.009.
- Bouveyron, C., Celeux, G., and Girard, S. Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic PCA. *Pattern Recognition Letters*, 32(14):1706–1713, 2011. doi: 10.1016/j.patrec.2011.07.017.
- Breiding, P., Kozhasov, K., and Lerario, A. On the Geometry of the Set of Symmetric Matrices with Repeated Eigenvalues. *Arnold Mathematical Journal*, 4(3):423–443, 2018. doi: 10.1007/s40598-018-0095-0.
- Burnham, K. P. and Anderson, D. R. *Model Selection and Multimodel Inference*. Springer, 2004. ISBN 978-0-387-95364-9. doi: 10.1007/b97636.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
- Field, D. J. Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A*, 4(12): 2379–2394, 1987. ISSN 1520-8532. doi: 10.1364/JOSAA.4.002379.
- Fisher, R. A. Iris, 1936. *UCI Machine Learning Repository*. doi: 10.24432/C56C76.
- Geiger, D., Heckerman, D., King, H., and Meek, C. Stratified exponential families: Graphical models and model selection. *The Annals of Statistics*, 29(2):505–529, 2001. doi: 10.1214/aos/1009210550.
- German, B. Glass Identification, 1987. *UCI Machine Learning Repository*. doi: 10.24432/C5WW2P.
- Gershkovich, V. and Haritos, N. Problem of close eigenvalues in the vibration testing of structures. *ANZIAM Journal*, 46: C658–C671, 2004. ISSN 1445-8810. doi: 10.21914/anziamj.v46i0.982.
- Groisser, D., Jung, S., and Schwartzman, A. 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): 1092–1159, 2017. ISSN 1935-7524, 1935-7524. doi: 10.1214/17-EJS1250.
- Hopkins, M., Reeber, E., Forman, G., and Suermondt, J. Spambase, 1999. *UCI Machine Learning Repository*. doi: 10.24432/C53G6X.
- Hurvich, C. M. and Tsai, C.-L. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989. doi: 10.1093/biomet/76.2.297.

- Hyvärinen, A. and Hoyer, P. Emergence of Phase- and Shift-Invariant Features by Decomposition of Natural Images into Independent Feature Subspaces. *Neural Computation*, 12(7):1705–1720, 2000. doi: 10.1162/089976600300015312.
- Jenatton, R., Obozinski, G., and Bach, F. Structured Sparse Principal Component Analysis. In *AISTATS*, pp. 366–373. PMLR, 2010.
- Jolliffe, I. T. Rotation of Ill-Defined Principal Components. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 38(1):139–147, 1989. ISSN 0035-9254. doi: 10.2307/2347688.
- Jolliffe, I. T. *Principal Component Analysis*. Springer, 2002.
- Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958. ISSN 1860-0980. doi: 10.1007/BF02289233.
- Kohonen, T. Emergence of invariant-feature detectors in the adaptive-subspace self-organizing map. *Biological Cybernetics*, 75(4):281–291, 1996. ISSN 1432-0770. doi: 10.1007/s004220050295.
- Lambert, M., Bonnabel, S., and Bach, F. The limited-memory recursive variational Gaussian approximation (L-RVGA). *Statistics and Computing*, 33(70), 2023. doi: 10.1007/s11222-023-10239-x.
- Lazutkin, V. F. *KAM Theory and Semiclassical Approximations to Eigenfunctions*. Springer, 1993. ISBN 978-3-642-76249-9 978-3-642-76247-5. doi: 10.1007/978-3-642-76247-5.
- Lefevre, J., Fraize, J., and Germaud, D. Perturbation of Fiedler Vector: Interest for Graph Measures and Shape Analysis. In *GSI*, LNCS, pp. 593–601. Springer, 2023. doi: 10.1007/978-3-031-38299-4_61.
- Leygonie, J., Carrière, M., Lacombe, T., and Oudot, S. A gradient sampling algorithm for stratified maps with applications to topological data analysis. *Mathematical Programming*, 202(1):199–239, 2023. ISSN 1436-4646. doi: 10.1007/s10107-023-01931-x.
- Lombaert, H., Sporning, J., and Siddiqi, K. Diffeomorphic Spectral Matching of Cortical Surfaces. In *IPMI*, volume 7917 of LNCS, pp. 376–389. Springer, 2013. ISBN 978-3-642-38868-2. doi: 10.1007/978-3-642-38868-2_32.
- Ma, X., Kirby, M., and Peterson, C. The Flag Manifold as a Tool for Analyzing and Comparing Sets of Data Sets. In *ICCV Workshops*, pp. 4168–4177, 2021. doi: 10.1109/ICCVW54120.2021.00465.
- Mankovich, N. and Birdal, T. Chordal Averaging on Flag Manifolds and Its Applications. In *ICCV*, pp. 3881–3890, 2023.
- Mankovich, N., Camps-Valls, G., and Birdal, T. Fun with Flags: Robust Principal Directions via Flag Manifolds. In *CVPR*, pp. 330–340, 2024.
- Minka, T. Automatic Choice of Dimensionality for PCA. In *NeurIPS*, volume 13. MIT Press, 2000.
- Mitchell, T. CMU Face Images, 1997. *UCI Machine Learning Repository*. doi: 10.24432/C5JC79.
- Myung, I. J., Balasubramanian, V., and Pitt, M. A. Counting probability distributions: Differential geometry and model selection. *PNAS*, 97(21):11170–11175, 2000. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.170283897.
- Ng, A., Jordan, M., and Weiss, Y. On Spectral Clustering: Analysis and an algorithm. In *NeurIPS*, volume 14. MIT Press, 2001.
- Nishimori, Y., Akaho, S., and Plumbley, M. D. Riemannian Optimization Method on the Flag Manifold for Independent Subspace Analysis. In *ICA*, LNCS, pp. 295–302. Springer, 2006. doi: 10.1007/11679363_37.
- North, G. R., Bell, T. L., Cahalan, R. F., and Moeng, F. J. Sampling Errors in the Estimation of Empirical Orthogonal Functions. *Monthly Weather Review*, 110(7):699–706, 1982. ISSN 1520-0493, 0027-0644. doi: 10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2.
- Olikier, G., Gallivan, K. A., and Absil, P.-A. First-order optimization on stratified sets, 2023. doi: 10.48550/arXiv.2303.16040.

- Olshausen, B. A. and Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. ISSN 1476-4687. doi: 10.1038/381607a0.
- Penneç, X. Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006. ISSN 1573-7683. doi: 10.1007/s10851-006-6228-4.
- Penneç, X. Barycentric Subspace Analysis on Manifolds. *The Annals of Statistics*, 46(6A):2711–2746, 2018. doi: 10.1214/17-AOS1636.
- Rabenoro, D. and Penneç, X. A geometric framework for asymptotic inference of principal subspaces in PCA, 2024. doi: 10.48550/arXiv.2209.02025.
- Rohe, K. and Zeng, M. Vintage factor analysis with Varimax performs statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1037–1060, 2023. ISSN 1369-7412. doi: 10.1093/jrssi/bqkad029.
- Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.
- Sigillito, V. G., Wing, S. P., Hutton, L. V., and Baker, K. B. Ionosphere, 1989. *UCI Machine Learning Repository*. doi: 10.24432/C5W01B.
- Sirovich, L. and Kirby, M. Low-dimensional procedure for the characterization of human faces. *JOSA A*, 4(3):519–524, 1987. ISSN 1520-8532. doi: 10.1364/JOSAA.4.000519.
- Szwagier, T. and Penneç, X. Rethinking the Riemannian Logarithm on Flag Manifolds as an Orthogonal Alignment Problem. In *GSI, LNCS*, pp. 375–383. Springer, 2023. doi: 10.1007/978-3-031-38271-0_37.
- Taeb, A., Bühlmann, P., and Chandrasekaran, V. Model selection over partially ordered sets. *PNAS*, 121(8):e2314228121, 2024. doi: 10.1073/pnas.2314228121.
- Taylor, M. *sinkr: Collection of functions with emphasis in multivariate data analysis*, 2022. R package version 0.7.
- Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 0035-9246.
- Tipping, M. E. and Bishop, C. M. Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation*, 11(2): 443–482, 1999a. ISSN 0899-7667. doi: 10.1162/089976699300016728.
- Tipping, M. E. and Bishop, C. M. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 1999b. doi: 10.1111/1467-9868.00196.
- Tyler, D. E. Asymptotic Inference for Eigenvectors. *The Annals of Statistics*, 9(4):725–736, 1981. doi: 10.1214/aos/1176345514.
- Watanabe, S. and Pakvasa, N. Subspace method of pattern recognition. In *Proc. 1st. IJCP*, pp. 25–32, 1973.
- Wolberg, W. H., Mangasarian, O. L., and Street, W. N. Breast Cancer Wisconsin (Diagnostic), 1995. *UCI Machine Learning Repository*. doi: 10.24432/C5DW2B.
- Ye, K., Wong, K. S.-W., and Lim, L.-H. Optimization on flag manifolds. *Mathematical Programming*, 194(1):621–660, 2022. ISSN 1436-4646. doi: 10.1007/s10107-021-01640-3.
- Zhu, X. and Shen, C. Practical gradient and conjugate gradient methods on flag manifolds. *Computational Optimization and Applications*, 88(2):491–524, 2024. ISSN 1573-2894. doi: 10.1007/s10589-024-00568-6.