

Exploring dimension learning via a penalized probabilistic principal component analysis

Wei Q. Deng & Radu V. Craiu

To cite this article: Wei Q. Deng & Radu V. Craiu (2023) Exploring dimension learning via a penalized probabilistic principal component analysis, Journal of Statistical Computation and Simulation, 93:2, 266-297, DOI: [10.1080/00949655.2022.2100890](https://doi.org/10.1080/00949655.2022.2100890)

To link to this article: <https://doi.org/10.1080/00949655.2022.2100890>



Published online: 02 Aug 2022.



Submit your article to this journal [↗](#)



Article views: 26





View related articles [↗](#)



View Crossmark data [↗](#)



Exploring dimension learning via a penalized probabilistic principal component analysis

Wei Q. Deng ^{a,b} and Radu V. Craiu ^c

^aDepartment of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Canada; ^bPeter Boris Centre for Addictions Research, St. Joseph's Healthcare Hamilton, Hamilton, Canada; ^cDepartment of Statistical Sciences, University of Toronto, Toronto, Canada

ABSTRACT

Establishing a low-dimensional representation of the data leads to efficient data learning strategies. In many cases, the reduced dimension needs to be explicitly stated and estimated from the data. We explore the estimation of dimension in finite samples as a constrained optimization problem, where the estimated dimension is a maximizer of a penalized profile likelihood criterion within the framework of a probabilistic principal components analysis. Unlike other penalized maximization problems that require an 'optimal' penalty tuning parameter, we propose a data-averaging procedure whereby the estimated dimension emerges as the most favourable choice over a range of plausible penalty parameters. The proposed heuristic is compared to a large number of alternative criteria in simulations and an application to gene expression data. Extensive simulation studies reveal that none of the methods uniformly dominate the other and highlight the importance of subject-specific knowledge in choosing statistical methods for dimension learning. Our application results also suggest that gene expression data have a higher intrinsic dimension than previously thought. Overall, our proposed heuristic strikes a good balance and is the method of choice when model assumptions deviated moderately.

ARTICLE HISTORY

Received 10 November 2021
Accepted 8 July 2022

KEYWORDS

Dimension estimation; model selection; penalization; principal component analysis; probabilistic principal component analysis; profile likelihood

1. Introduction


Consider a data matrix $X \in \mathbb{R}^{n \times m}$ that has been column and row centred such that

$$\sum_{i=1}^n x_{ij} = \sum_{j=1}^m x_{ij} = 0; \quad \text{for } i = 1, \dots, n; \text{ and } j = 1, \dots, m,$$

we are interested in a linear decomposition of X to a signal component, driven by variance in the top singular values, and a noise component of the form:

$$X = WL + F, \tag{1}$$

CONTACT Wei Q. Deng  dengwq@mcmaster.ca

 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/00949655.2022.2100890>

where $W \in \mathbb{R}^{n \times k}$ is a constant matrix with rank $k < n$, $L \in \mathbb{R}^{k \times m}$ is an arbitrary matrix with orthonormal columns, and F is a matrix whose rows are uncorrelated and have equal variance. The dimension of interest depends on W as it is the minimal rank k such that rows of $X - WL$ are uncorrelated and have isotropic covariance. Henceforth, we refer to k as the *effective rank* of the data because, intuitively, correlation structure in the rows of X reduces data dimension attributed to the signal component (WL) from $\min(n, m)$ to k .

Estimation of k has been studied in various contexts as the linear model (1) has many alternative forms and names, such as a principal component analysis (PCA; [1–3]), a truncated singular value decomposition (SVD), a factor analysis model [4], and a spiked population model [5], where the effective rank coincides with the definition of the number of spikes.

The approaches to determine k as the number of principal components (PCs), can be summarized under roughly three categories according to Jolliffe [6]. The first type is a variety of *ad-hoc* rules that have an empirical basis, such as the scree test [7] or Kaiser rule. To automate the decision, Zhu and Ghodsi proposed a profile likelihood criterion that detects a ‘gap’ in the sample eigenvalues [8]. A second class of methods rely on asymptotic tests, such as the likelihood ratio test for equality of eigenvalues [9–13], which differ according to asymptotic conditions on the data dimensions. Instead of an asymptotic test, Choi et al. [14] recently proposed an exact method for hypothesis testing of signals in a noisy matrix to estimate the number of PCs that showed promising results in simulations. Finally, for small datasets, computational methods such as bootstrap, permutation and cross-validation can be implemented in a timely manner. Among them, cross-validation is frequently used [15] with a general cross-validation (GCV) criterion [16] that also works well with large datasets.

Using a truncated SVD, Gavish and Donoho [17] proposed to remove the underlying noise in the singular values via a hard threshold-based approach. In this case, the stopping rule based on a single threshold could be useful for recovering the original data in the sense of asymptotic mean squared error, but does not directly inform the minimal rank of the noise reduced data. Similarly in isotropic factor analysis, Bai and Ng [18] proposed to estimate the number of factors by finding some threshold to separate large and small eigenvalues of the data covariance matrix that leverages various penalty functions, but the approach depends on the correct estimation of error variance. Using a different strategy, Passemier et al. [19] tackled the estimation of the noise variance, which led to a bias-corrected criterion for estimating k when $n \gg m$.

Here we focus on reviewing model-based methods where the solution arises from various model selection criteria. Probabilistic principal component analysis (PPCA), introduced in the seminal paper of Tipping and Bishop [20], allows the estimation of k as a likelihood optimization problem. An alternative Bayesian approach was proposed in [21], with the caveat that the full Bayesian estimation using Markov Chain Monte Carlo can be computationally prohibitive for large datasets [22] and approximations are needed. Indeed, Minka implemented Laplace’s method to approximate the posterior likelihood [23] and showed it to be often superior to cross-validation and variational inference [24,25] with the added benefit of fast computation. An exact marginal likelihood criterion based on a normal-gamma prior distribution has been developed that is competitive with both Bayesian and frequentist methods in low dimensional settings [26]. For high-dimensional data with a small number of observations, Hoyle [27] noted the unsatisfactory performance

of Laplace's approximation and proposed to modify the Bayesian model using a Gaussian parametrization that showed improved performance. Observing the symmetry in the data structure, Sobczyk et al. [28] approximated the Bayesian models for both X and X^T , and thus proposed two separate criteria that work well under divergence of either the number of observations (m) or samples (n), while the other one is constant.

Penalized maximum likelihood approaches are widely used to induce sparsity in the number of parameters used to characterize statistical models and have proven suitable for model selection. Here we explore using penalized probabilistic PCA models to estimate the effective rank and propose an accompanying data-driven heuristic to estimate the dimension. This heuristic has theoretical basis, was examined in extensive simulations and applied to a microarray gene expression dataset to inform the data dimension. We find that the penalized approach is competitive when compared to Bayesian and empirical alternatives in both simulated and application data, especially under departure from independence and normality assumptions. None of the methods uniformly dominate the others across the wide range of conditions, highlighting the importance of verifying the assumptions underlying each method.

This paper is structured as follows. We first revisit the probabilistic principal component model in Section 2. In Section 3, we explore using the penalized probabilistic PCA to model the data dimension as part of the optimization problem and present a data-driven algorithm for dimension learning. Results from an extensive simulation study comparing different classes of methods are presented in Section 4 and an application to gene expression data is presented in Section 5. In the last section, we conclude the paper with general remarks on the proposed penalized approach and our practical recommendation to dimension learning in data applications.

2. Probabilistic principal components analysis

Given data $X \in \mathbb{R}^{n \times m}$, we seek a low-dimensional representation in the columns of X , $X_j \in \mathbb{R}^n, j = 1, \dots, m$. Suppose there exists a fixed dimension $k \in \mathbb{Z}_+(1 \leq k \leq n - 1)$ such that:

$$X_j = \mu + Wl_j + f_j, \quad j = 1, \dots, m \quad (2)$$

where μ is the mean vector, $W \in \mathbb{R}^{n \times k}$ is a constant matrix, $l_j \in \mathbb{R}^k$ is a latent vector, and $f_j \in \mathbb{R}^n$ is noise in the data. In order to identify the data decomposition to signal (WL) and noise (F) components, we make the assumption that both the latent vector and the noise component are spherical Gaussian. This decomposition implies that the n -dimensional vector X_j is obtained as a linear transformation of a k -dimensional latent vector. Therefore, the spanned subspace of X_1, \dots, X_m has effective dimension k . The value of k is unknown in realistic examples and needs to be estimated from the data X . The usual PCA decomposition is obtained when the dimension is $k = n$, and in this case, F in Equation (1) reduces to $\mathbf{0}$.

In this paper, we assume $f_j \sim \mathcal{N}(0, \zeta^2 I_n)$ and $l_j \sim \mathcal{N}(0, I_k)$, which imply that for any $1 \leq j \leq m$, X_j follows the Gaussian distribution:

$$X_j \sim \mathcal{N}(\mu, WW^T + \zeta^2 I_n). \quad (3)$$

Denote the covariance matrix of X_j by $\Phi = WW^T + \zeta^2 I_n$ and under model (3) it has a maximum of $k + 1$ unique eigenvalues: $\lambda_1, \dots, \lambda_k$ and ζ^2 . This model forces the samples,

represented by the rows of X , to be conditionally independent given the random vectors, $\{l_1, \dots, l_k\}$, and thus the covariance matrix $WW^T + \zeta^2 I_n$ can take on a more parsimonious representation. In general, the latent vectors may not have a Gaussian distribution and can be used to specify non-Gaussian signal components, such as those in a linear noisy independent component analysis model.

The log-likelihood function with respect to the unknown parameters W and ζ^2 , given independent observations $X = (x_1, \dots, x_m)$, is denoted by

$$l(\mu, W, \zeta^2; X) = -\frac{m}{2} [n \log(2\pi) + \log |WW^T + \zeta^2 I| + \text{tr}\{(WW^T + \zeta^2 I)^{-1} \hat{\Phi}\}], \quad (4)$$

where $\hat{\Phi} = m^{-1} \sum_{j=1}^m (X_j - \hat{\mu})(X_j - \hat{\mu})^T$ is the sample covariance matrix. Assuming $m > n$ and $\hat{\Phi}$ is full rank, the maximum likelihood estimator (MLE) for μ is simply the sample mean $\hat{\mu} = \frac{1}{m} \sum_{j=1}^m X_j$. Without loss of generality, $\hat{\mu}$ can be replaced by zero provided that the data X had been row centred. For convenience, we also assume the data had also been row standardized such that the diagonal elements of $\hat{\Phi}$ equal to 1.

This intrinsic data dimension, $\text{rank}(W) = \text{rank}(W^T W) = k$, is only implicitly involved in the log-likelihood. It has been shown in [20] that for any integer $q \in \{1, \dots, n - 1\}$, (4) is maximized by:

$$\hat{W}_q = U_q \hat{D}_q B_q, \quad \text{and} \quad \hat{\zeta}_q^2 = \frac{\sum_{i=q+1}^n \hat{\lambda}_i}{n - q}, \quad (5)$$

where $\{\hat{\lambda}_i\}_i$'s are the sample eigenvalues of $\hat{\Phi}$, U_q is an $n \times q$ matrix with columns corresponding to the first q eigenvectors of $\hat{\Phi}$, $\hat{H}(q)$ is a diagonal matrix with the first q non-zero entries each given by $\hat{\eta}_i = \sqrt{\hat{\lambda}_i - \hat{\zeta}_q^2}$, and $B_q \in \mathbb{R}^{q \times q}$ is an arbitrary orthogonal matrix. The integer q needs not be specified, but the form of (5) suggests that the division between the first q and the last $n - q$ eigenvalues/eigenvector is the key to maximizing (4). In other words, for every value of q , we can identify the corresponding MLEs given in (5), but the different choices of q cannot be distinguished under the current likelihood model.

Let l_p denote the profile log-likelihood. If we considered the parameters W, ζ^2 to be nuisance parameters, a profile log-likelihood in q is obtained by substituting the solutions in (4):

$$l_p(q; \hat{\lambda}_i) = -\frac{m}{2} \left\{ n \log(2\pi) + \sum_{i=1}^q \log \hat{\lambda}_i + (n - q) \log \hat{\zeta}_q^2 + n \right\}. \quad (6)$$

The formulation (6) clarifies that data dimension is implicitly involved in defining the parameters of the model, and one might be tempted to find the maximizer (in q) of the profile log-likelihood as the estimate of effective rank. However, the following result suggests that the profile log-likelihood alone is not sufficient to identify the intrinsic data dimension.

Proposition 2.1: *Consider a sample $X \in \mathbb{R}^{n \times m}$ with each column following a multivariate Gaussian distribution $N(0, WW^T + \zeta^2 I)$. If the sample row covariance matrix of X is positive semi-definite and $k = \text{rank}(WW^T)$, then the profile log-likelihood $l_p(q)$ is non-decreasing in $q \in \mathbb{Z}_+(1 \leq q \leq n - 1)$.*

Proof is included in Supplementary Materials.

This result shows that the profile log-likelihood is monotonically non-decreasing in q , suggesting that it can not be used as a criterion to select k , the data dimension, in finite samples. The choice of k thus becomes a model selection-type problem, with decreasing values of q corresponding to more constraint models and $q = n$ corresponds to a fully non-parametric, conventional PCA.

Remark 2.1: Proposition 2.1 demonstrates that the saturated model with $q = n - 1$ is always preferred. If one permits $q = 0$, then $\hat{W} = \mathbf{0}$ and the likelihood is minimized. The same conclusion can be reached by observing the proportion of variance explained by the PPCA model with true rank k :

$$\text{tr}(WW^T) = \sum_{i=1}^k d_i^2 = n(1 - \zeta^2),$$

where $\{d_i\}_{i=1,\dots,k}$ are the singular values of W . When ζ^2 is equal to 0 (or $k = n$), the model corresponds to PCA with a full-rank loading matrix and is completely deterministic; and when ζ^2 is equal to 1 (or $k = 0$), the model reduces to an isotropic Gaussian distribution and $W = \mathbf{0}$. In order to avoid degenerate situations, in this paper we restrict the range of k to $\{1, 2, \dots, n - 1\}$.

Remark 2.2: The generative model (2) has a specific dimension k , which is embedded in the parameter W through the data generative process. At the same time, the data generated can support each possible q if we evaluate the model likelihood alone without any constraint on the error variance or model complexity.

3. Effective rank selection heuristics based on a penalized probabilistic principal components analysis

Penalized maximum likelihood approaches are widely used to induce sparsity in statistical models. The level of penalty imposed on the model is regularized via a tuning parameter, which controls the trade-off between goodness-of-fit and complexity [29–31]. In the problem considered here, the model complexity, defined by the number of free parameters $nk + 1 - k(k - 1)/2$, is directly related to the data dimension, while the fit corresponds to the amount of variance explained, i.e. $\text{tr}(\Phi) - n\zeta^2$. The natural guiding principle is to favour a parsimonious representation for the covariance by simultaneously penalizing small explained variance and large k .

The penalized log-likelihood has the form:

$$l(W, \zeta^2; \delta) = \frac{-m}{2} \{ \log |WW^T + \zeta^2 I| + \text{tr}[(WW^T + \zeta^2 I)^{-1} \hat{\Phi}] - \delta \text{pen}(W, \zeta^2) \},$$

where the tuning parameter $\delta > 0$ controls the amount of penalty due to a penalty function, $\text{pen}(W, \zeta^2)$. Notice that m is a scaling factor and does not directly affect the maximization other than through the convergence of $\hat{\Phi}$ to the true covariance Φ .

The penalty function should depend on (W, ζ^2) and thus be able to capture the model dimension embedded in W and the amount of error variance ζ^2 . At the same time, the two parameters combine in the case of standardized data because $\text{tr}(WW^T) + n\zeta^2 = n$.

By maximizing the penalized log-likelihood function, it will also be possible to express the penalized MLEs indexed by q and thus to motivate the penalized profile log-likelihood as a vehicle for intrinsic data dimension selection.

Unlike in other constrained optimization problems, the estimation of individual entries of W is not the primary objective. Rather, we are interested in penalty functions that diverge when the estimated eigenvalues (i.e. the sum of ζ^2 and each squared singular value of W) are close to 1, or alternatively, when ζ^2 is close to 0. Here we explore the following penalty functions that capture both the amount of variance explained and the complexity of the model:

$$\text{pen}_1(W, \zeta^2) = \text{rank}(W) \log \zeta^2 = k \log \zeta^2 \tag{7a}$$

$$\text{pen}_2(W, \zeta^2) = -\frac{\text{rank}(W)}{\zeta^2} = -\frac{k}{\zeta^2} \tag{7b}$$

$$\text{pen}_3(W, \zeta^2) = \beta \text{pen}_1(W, \zeta^2) + (1 - \beta) \text{pen}_2(W, \zeta^2), \quad \beta \in (0, 1). \tag{7c}$$

In our experience, the penalties lead to equivalent analyses since the tuning parameters will adjust to yield similar results. Ultimately, the choice to use (7a) over the others is driven by convenience because it leads to simpler analytical derivations and intuitive heuristics.

3.1. Penalized maximum likelihood estimators

The penalized log-likelihood using the proposed penalty function (7a) becomes:

$$l(W, \zeta^2; \delta) = \frac{-m}{2} \{ \log |WW^T + \zeta^2 I| + \text{tr}[(WW^T + \zeta^2 I)^{-1} \hat{\Phi}] - \delta \text{rank}(W) \log \zeta^2 \}.$$

Similarly to (5), the penalized MLEs, \tilde{W} and $\tilde{\zeta}^2$, are functions of q . Due to a non-zero δ -value, the penalized MLE of ζ^2 is expressed in terms of δ and $\hat{\zeta}_q^2$:

$$\tilde{\zeta}_q^2 = \frac{\sum_{i=q+1}^n \hat{\lambda}_i}{n - q - \delta q} = \frac{n - q}{n - q - \delta q} \hat{\zeta}_q^2. \tag{8}$$

Taking derivative with respect to W yields the same relationship between the squared singular values of W and ζ^2 :

$$\hat{\lambda}_i = \begin{cases} \tilde{\eta}_i^2(q) + \tilde{\zeta}_q^2, & \text{if } i \leq q; \\ \tilde{\zeta}_q^2, & \text{otherwise,} \end{cases}$$

where $\tilde{\eta}_i^2(q)$ denotes the i th estimated value when the estimated effective rank is q . For a fixed q , $\tilde{\zeta}_q^2$ is unbounded as $n - q - \delta q$ can be very close to 0 or even negative for large δ -values. This implies that the choice of q poses a restriction of the range of δ , and vice versa. Thus, the theoretical range of δ has an upper bound at $n/q - 1$ so that $\tilde{\zeta}_q^2$ is positive. Henceforth, we reparametrized the tuning parameter to $\tilde{\delta} = \delta/n \in [0, 1/q - 1/n]$.

Interestingly, the penalized MLEs of ζ^2 under (7a) and (7b) are closely related to those estimated under an approximated posterior likelihood assuming an inverse-gamma prior [23], with $\tilde{\delta}$ corresponding to linear functions of the hyperparameters, see Appendix 1 for more details.

Substituting the penalized MLEs given q , we obtain the penalized profile log-likelihood, denoted by $l_p(q; \tilde{\delta})$, as a function of q for a fixed $\tilde{\delta}$:

$$l_p(q; \tilde{\delta}) = l_p(q) - \frac{m}{2} \left[n \left(1 - \frac{q}{n} - \tilde{\delta}q \right) \log \frac{n - q}{n - q - nq\tilde{\delta}} - \tilde{\delta}nq(\log \hat{\zeta}_q^2 + 1) \right]. \quad (9)$$

The penalized profile log-likelihood criterion favours a more parsimonious model by penalizing large values of q as well as retaining as much explained variance as possible. Given suitable choices of $\tilde{\delta}$, the following results establish the conditions under which the penalized criterion will find the correct dimension:

Proposition 3.1: *Consider a sample $X \in \mathbb{R}^{n \times m}$ with each column following a multivariate Gaussian distribution $\mathcal{N}(0, WW^T + \zeta^2 I)$. If $\hat{\Phi}$, the sample covariance matrix of X^T , is positive semi-definite, then there exists $\tilde{\delta}_o \in (0, 1 - 1/n)$ such that $l_p(q; \tilde{\delta}_o)$ is maximized at k ($1 < k < n$), the rank of W or the effective rank of X .*

Proof is included in Supplementary Materials.

3.2. A data-driven voting strategy to estimate the effective rank

The introduction of penalty changes the monotonicity property of the profile likelihood (6), and thus makes it possible to select the correct dimension k for appropriate choices of $\tilde{\delta}$ -value. The selection of appropriate tuning parameter values in other well-known problems, such as the selection of shrinkage tuning parameter in lasso [29,32], uses either a model selection criterion, e.g. Akaike or Bayesian information criterion, or cross-validation. However, the use of a cross-validation approach is based on optimizing a certain objective function that can be analytically expressed, a task that is difficult when of interest is determining the dimension. Our attempts at using an off-the-shelf information criterion produced modest results in simulations under the correct model specification, but failed to identify a sensible estimate when the data generative model deviated from assumptions.

So far, a data-driven heuristic gave the best balance in performance. It entails a voting strategy in which each value of $\tilde{\delta}$ over a plausible range, determined from the data, will lead to a vote for a particular value of q as the estimate. Since the same estimate of k can result from multiple $\tilde{\delta}$ -values, ultimately the estimated dimension that has been obtained most often is selected.

The search for the intrinsic dimension implies a grid search for $\tilde{\delta}$ whose values $\{\tilde{\delta}_1, \dots, \tilde{\delta}_T\}$ are selected using a sequence of T equidistant points on log scale. The user-specified integer T needs to be large enough to identify a mode, and in simulations we used $T = 5,000$ or roughly $50n$, with values of the same order of magnitude leading to the same results. Each $\tilde{\delta}_t$ will result in (9) supporting a possible value for k ($1 \leq k \leq n - 1$), which is the maximizer of $l_p(q; \tilde{\delta}_t)$ in q . Then, the number of times that a value of k maximizes the penalized profile log-likelihood is counted and the one with the highest vote count is selected. Define $|A(j)| = \#\{t : \arg \max_q l_p(q, \tilde{\delta}_t) = j\}$ and the estimate is denoted by $\tilde{k} = \arg \max_j |A(j)|$. The data-driven procedure is described in Algorithm 1.

The penalized approach requires a proper calibration of $\tilde{\delta}$ so that the true dimension, k , identifies as the global maximizer of $l_p(q; \tilde{\delta})$ most often. In theory, $\tilde{\delta}$ could take any value

Algorithm 1 A data-driven voting strategy to estimate the effective rank

Require: integer T , $\{\hat{\lambda}_i\}_{i=1,\dots,n-1}$, $\kappa = 0.001$
 initialization; setting $n_{\max} = \min\{i : \hat{\lambda}_i < \kappa\} - 1$
if $n_{\max} > 1$ **then**
 while $q = 1$ **do**
 find $\tilde{\delta}_T = \arg \min_{\tilde{\delta}} \{u : l_p(q = 1; u) > l_p(q = 2; u)\}$;
 end while
 while $q = n - 2$ **do**
 find $\tilde{\delta}_1 = \arg \max_{\tilde{\delta}} \{u : l_p(q = n - 2; u) > l_p(q = n - 1; u)\}$;
 end while
 construct $\{\tilde{\delta}_1, \dots, \tilde{\delta}_T\}$;
 while $j \leq n - 1$ **do**
 $|A(j)| = \#\{t : \arg \max_q l_p(q, \tilde{\delta}_t) = j\}$
 end while
 $\tilde{k} = \arg \max_j |A(j)|$
else
 $\tilde{k} = 1$
end if

in $[0, \infty)$, but for practical considerations, it has a finite range depending on the maximum and minimum q to avoid degenerate cases. The connection between q and $\tilde{\delta}$, given by $\tilde{\delta} \in (0, (1/q - 1/n)[1 - \hat{\zeta}_q^2])$, is derived in Appendix 2. A theoretical justification of the voting method based on the log-scale is provided in Lemma A.5 in Appendix 2. A detailed illustration of the method on simulated data can be found in Supplementary Materials.

To make the methods accessible, we implemented the voting procedure in a statistical software R package, available at <https://github.com/WeiAkaneDeng/SPAC2>.

4. Simulation studies

4.1. Data simulation

Given the true dimension k , error variance ζ^2 , and observed dimensions (n, m) , we can generate the data by specifying either (1) the signal components of the first k true eigenvalues $(\eta_1^2, \dots, \eta_k^2)$ directly or, (2) a trend in the first k signal components. The residual noise was assumed to have a multivariate distribution with mean vector zero and covariance $\zeta^2 I_n$. The maximum dimension (n , when $n < m$) is often directly associated with the difficulty of recovering the true dimension and was kept fixed at $n = 100$.

We explored four data generation scenarios: the first scenario, denoted by S0, is a baseline case where each observation is independent and identically distributed (i.i.d) following a standard normal distribution; the second scenario encompassed the spiked covariance model with either the first k true eigenvalues being equal, a homogeneous setting (scenario S1.1), or decaying with a linear or an exponential trend, the heterogeneous settings (scenario S1.2); the third scenario, S2, explored varying data dimensions whereby the row covariance matrix could also be rank-deficient; and finally, scenario S3, examined the

Table 1. Simulation scenarios.

| Scenario | Description of scenarios | Dimensions | Error distribution | | |
|----------|------------------------------------|------------|--|---------|---|
| S0 | i.i.d. | $n < m$ | $\mathcal{N}(0, 1)$ | | |
| S1 | Homogeneity | $n < m$ | $\mathcal{N}(0, \zeta^2)$ | | |
| | Heterogeneity (linear/exponential) | $n < m$ | $\mathcal{N}(0, \zeta^2)$ | | |
| S2 | Heterogeneity (exponential) | $n > m$ | $\mathcal{N}(0, \zeta^2)$ | | |
| | Non-normality | $n < m$ | MV – lognormal($\mu = 2, \sigma = 1, \Sigma = I$) MV – exponential(rate = 1, $\Sigma = I$) $t_5(\Sigma = I)$ | | |
| S3 | Correlated observations | $n < m$ | $\mathcal{N}(0, \Sigma)$ with Σ specified by AR1(ρ), $\rho = \{0.4, 0.7\}$ AR2 $\phi_1 = 0.4, \phi_2 = 0.4^2$, Polynomial kernel with offset $\{0, 1\}$ Gaussian kernel (scale parameter of 0.1) Laplacian kernel (scale parameter of 0.1) | | |
| | | | Both | $n < m$ | MV – lognormal($\mu = 2, \sigma = 1, \Sigma$) MV – exponential(rate = 1, Σ) $t_5(\Sigma)$ |

impact of model violations, such as heavy tails and correlated observations. These scenarios are summarized in Table 1. For each condition, the simulation was repeated 100 times and the number of observations was fixed at $m = 5000$ except in scenario S2. Though there is no explicit assumption requiring $m > n$, the choice for a larger m is to ensure some consistency in the sample eigenvalues, which is essential to the majority of the methods.

We applied double standardization to each simulated dataset and then calculated the sample eigenvalues. For data generated under S3, the sum of the sample eigenvalues could potential exceed n as data deviated from normality, thus the sample eigenvalues were scaled to sum to n prior to analysis. Meanwhile, when the row covariance is rank-deficient, the trailing sample eigenvalues could be negative; in this case, we adjusted the search space to $\{1, 2, \dots, n_{\max}\}$, where $n_{\max} = \max_i(\hat{\lambda}_i > 0.001)$.

4.2. Alternative methods

The performance of the proposed approach, denoted by $pPPCA$ for penalty (7a), is compared with a list of alternative methods (mathematical constructions in Appendix 3). For completeness, we also included $pPPCA2$ for penalty (7b), and $pPPCA3$ for penalty (7c) with $\beta = 1/2$. Briefly, we focused on the class of model selection criteria, including Akaike information criterion (AIC); a simplification to the Laplace’s method using BIC approximation [33], denoted by BIC ; an approximation to the posterior likelihood using Laplace’s method proposed in [23], denoted by $Laplace$; the best performer from a class of Bayesian criteria under different diverging assumptions, Penalized Semi-integrated Likelihood (PESEL; [28]). There is another class of methods that focused on the estimation, including a bias-corrected criterion for estimating k by Passemier et al. [19], denoted by $Passemier$, and a list of Bai and Ng’s criteria [18], denoted by BN . A hypothesis testing criterion for the equality of the last $n - k$ eigenvalues [10] was also selected, denoted by $Lawley$.

The hard threshold-based approach [17] removes the underlying noise in the singular values, and is denoted by *Donoho*. Finally, the last class of methods attempt to detect an ‘elbow’ in the scree plot produced by the sample eigenvalues: a list of empirical approaches, as well as a simple profile likelihood-based criterion (*ProfileL*) by Zhu and Ghodsi [8] were included in the comparison.

Berthet and Rigollet [34] considered the minimal value of $\theta (> 0)$ in a more restrictive spiked covariance model $I + \theta vv^T$ that can be theoretically distinguished from I , where $v = (v_1, \dots, v_k)$ is a set of n -dimensional unit vectors. This is equivalent to our problem when the top k eigenvalues are equal. For each true k , a corresponding critical value is given and shown to be of order $k\sqrt{\log(n/k)/m}$ [34], implying that as the true k increases, the signal needs to increase relatively for detection. Results from this study, though not directly applicable for method comparison, provide insight for the simulation study that follows.

Some of the methods we do not consider in the comparison are automatic relevance determination [21] and related methods that followed it [35,36] as they have been shown to be outperformed by methods based on the Laplace approximation [23]. Variational approximation methods [24,25,37] are also excluded, as [37] does not directly estimate the number of PCs, while [25] has been shown to be suboptimal to [27]. We have also excluded Bayesian methods that rely on MCMC sampling [22], as they become computationally prohibitive when either n or m is large (> 1000). The large number of observations is why cross-validation is difficult to implement beyond the heavy computational burden as data splitting can sometimes create biased signal in the data depending on how the held-out datasets are obtained, i.e. when the covariance structure is local to a subset of the observations. For this reason, we excluded cross-validation, but included the general cross-validation (GCV) criterion of [16] that has better scalability properties.

4.3. Scenario 0: independent identically distributed

As a baseline scenario, we compared methods when the data were drawn from a multivariate normal distribution with zero mean and an identity covariance. Depending on what is considered independent signal and noise, the effective rank could be 0 or a value close to the maximum possible rank $n-1$ (due to the standardization). Unsurprisingly, most methods estimated either 1 or $n-1$ majority of the time (Figure 1), with *pPPCA* preferring $n-1$ and most other model selection methods choosing 1. In this case, *Lawley*, *profileL*, and some ‘elbow’-based empirical approaches do not work very well, giving estimates ranging between 60–80, capturing the fluctuation in sampling distribution of the bottom eigenvalues.

4.4. Scenario 1.1: homogeneous eigenvalues

The first experiment consisted of k equal squared singular values, where we used $\zeta_k^2 = \{0.8, 0.81, \dots, 0.99\}$ and $k = \{5, 10, 20\}$ to capture a range of signal to noise (SNR) values, defined by the ratio of η_k^2 and ζ_k^2 rather than $(1 - \zeta_k^2)\zeta_k^{-2}$. The theoretical lower bounds of $k\sqrt{\log(n/k)/m}$ roughly correspond to $\zeta_k^2 = 0.98$ for $k = 10$ and $\zeta_k^2 = 0.93$ for $k = 20$.

The best performer from each class of methods is presented in Figure 2. The results of all methods can be found in Supplementary Figure 1. Most methods exhibited a decreasing

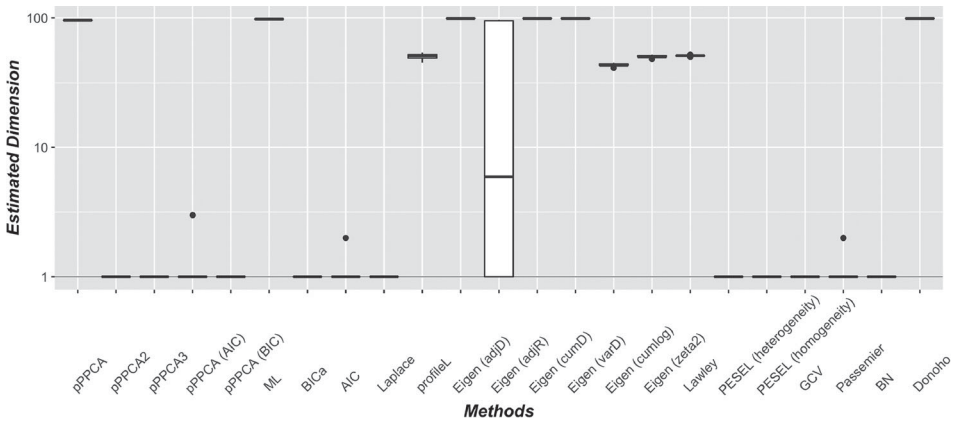


Figure 1. Distribution of the estimated k over 100 replicates when data are i.i.d.

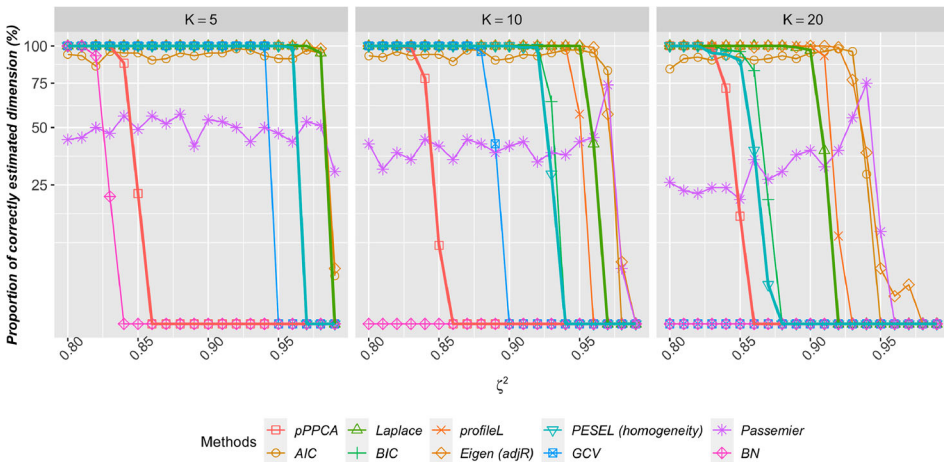


Figure 2. Proportion of correctly estimated k over 100 replicates as a function of ζ^2 assuming the first k squared singular values are equal. The coloured line corresponds to each method among the subset with reasonably good performance.

relationship between correctly estimated dimension as a function of increasing SNR (i.e. small ζ^2 and small k), with the exception of *AIC* and *Passemier*, where both methods have good performance most of the time. Interestingly, though *pPPCA* showed a decreasing trend as SNR increased for each true k , its performance did not deteriorate terribly across the different true k . The other methods were clearly more sensitive to the SNR as they approached the theoretical lower bounds for detection, with *AIC*, *Laplace* having the best performance among model selection approaches and *profileL* and *Eigen (adjR)* having the best performance among empirical approaches. The two *PESEL* criteria were similar to *BIC* and both had better performance than *pPPCA*.

We expected methods that take advantage of the homogeneity in the true eigenvalues to have reasonably good performance, such as *profileL* and *PESEL (homogeneity)*. But in fact,

profileL was better than *PESEL (homogeneity)* as SNR approached the theoretical minimum at $\zeta_k^2 = 0.98$ for $k = 10$, and even better than *Laplace* when $k = 20$.

At this point, we can eliminate both *Donoho* and *ML* from the list of methods as they were not designed to detect the effective rank, as the former aims to detect a theoretical minimum in terms of mean squared error loss, while the latter is a flawed information measure for model selection.

4.5. Scenario 1.2: heterogeneous eigenvalues

A more interesting and realistic scenario is when the true eigenvalues decrease according to a linear or exponential trend. In this case, the singular values can be determined by varying the two parameters ζ_k^2 and η_k^2 for a given k . We chose $\zeta_k^2 = \{0.1, \dots, 0.8\}$, $\eta_k^2 = \{0.1, 0.3\}$, and $k = \{5, 10\}$. The performance of methods could possibly be impacted by the following factors, including (1) the trend in $\{\eta_i^2\}_{i=1, \dots, k}$, the signal components, (2) true dimension k , and (3) the error variance ζ^2 .

Laplace had the best performance across the conditions, followed by the proposed *pPPCA*, *PESEL (heterogeneity)*, where both method would underestimate by 1. For most methods, we observed little impact on the performance of methods due to the choice of a linear and an exponential trends (Figures 3 and 4). However, performance of *pPPCA* was superior for a linear trend when true $k = 5$ (Figures 3) or an exponential trend (Figure 4) for a larger $k = 10$, possibly related to the fact that the empirical range of the penalty parameter influenced the sampling distribution of the first k sample eigenvalues.

Contrary to the homogeneous case, the decreasing trend in the signal component posed difficulty most noticeably for methods that assumed homogeneity. For example, both *profileL* and *PESEL (homogeneity)* completely failed to recover the correct dimension and underestimated. Again, we observed *PESEL (heterogeneity)* to be near identical to *BIC* and that *AIC* and *Passemier* would estimated correctly most of the time, but both are inconsistent.

4.6. Scenario 2: data dimensions

One of the data attributes encountered in real world applications is the varying ratios of m , the number of observations, and n , the maximum dimension. To evaluate the performance with respect to different ratios, we assumed the first k ($= \{5, 10\}$) squared singular values were equal (i.e. homogeneous) or decayed linearly or at an exponential rate with their values determined by fixing $\eta_k^2 = 0.3$, $\zeta_k^2 = 0.5$. The choice of m was set to be 50, 500, 1000, 5000, 10,000, and 20,000.

Informed by results in Section 4.5, we compared only methods that correctly estimated at least 5% for this slightly challenging scenario, including *AIC*, *BIC*, *Eigen* (ζ^2), *Laplace*, *Passemier*, *PESEL (heterogeneity)*, and *pPPCA*.

As m was increased, estimates from *BIC*, *Laplace*, *PESEL (heterogeneity)*, and *pPPCA* all approached the correct dimension 100% (Figure 5). Across different m/n ratios, *pPPCA* had the best performance when the signal was homogeneous; while there was no dominant method when the signals were heterogeneous, *Passemier*, *AIC* or *Laplace* were competitive depending on values of m/n . Among methods that are empirically consistent, *Laplace* had superior performance than both *PESEL (heterogeneity)* and *pPPCA*. Between these two,

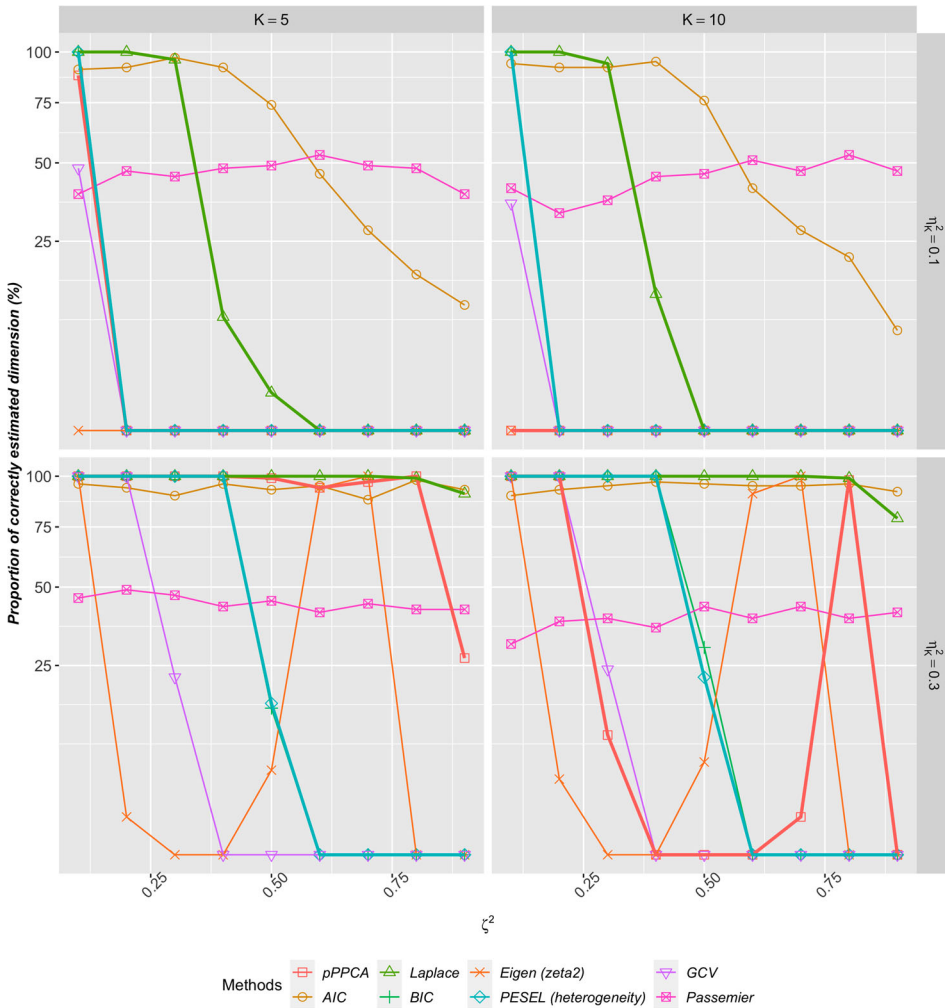


Figure 3. Proportion of correctly estimated k over 100 replicates as a function of ζ^2 assuming a linear decay in the first k squared singular values.

there was no universally better method across the combinations of k and linear/exponential trends. Again, we observed that the type of trend has a bigger impact on the performance of the proposed $pPPCA$ than other methods, preferring a linear trend when $k = 5$ or an exponential trend for a larger $k = 10$.

4.7. Scenario 3: departure from model assumptions

In many applications, noise in the data might not be independently nor normally distributed. We investigated cases where the observed error was drawn from a multivariate lognormal, exponential or, student’s t -distribution and with covariance matrix Σ specified by an identity matrix (not necessary independent), a first-order (AR1) or a second-order (AR2) autoregressive structure, polynomial kernels with no offset or an offset of 1, Gaussian

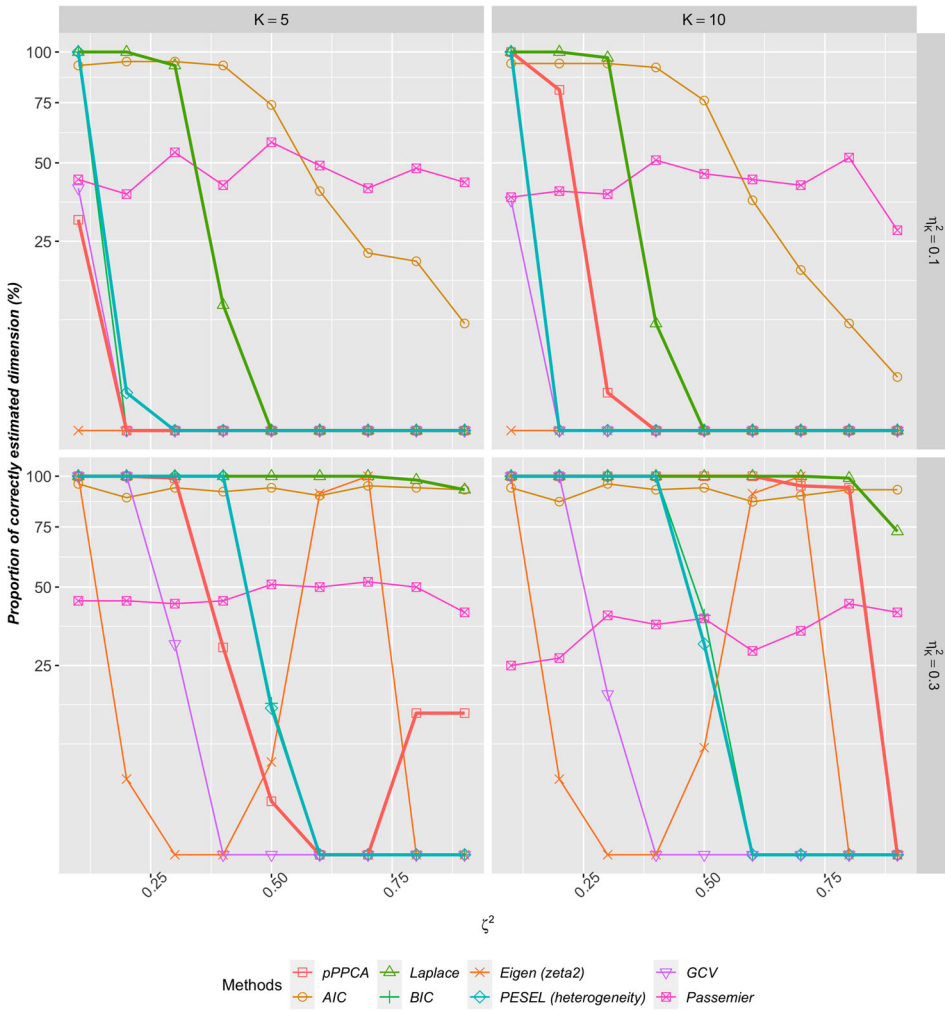


Figure 4. Proportion of correctly estimated k over 100 replicates as a function of ζ^2 assuming an exponential decay in the first k squared singular values.

radial basis function (RBF) kernel with scale parameter equals to 0.1, and a Laplacian kernel with scale parameter equals to 0.1. Here we considered two AR1 models with parameter value equals to 0.4 or 0.7 and one AR2 model with parameter values equal to 0.4 and 0.4^2 . The true eigenvalues of the error covariance are shown in Supplementary Figure 2, with the Gaussian RBF and Laplacian kernel being the most aggressive in terms of elevating the leading eigenvalues and shrinking the trailing eigenvalues towards zero. For simplicity, we set the remaining parameter values to $\zeta^2 = \{0.1, 0.2, \dots, 0.8\}$, $\eta_k^2 = 0.3$, and specified an exponential decay for the signal component of the first k eigenvalues. The true dimension was $k = 10$.

Both non-normal error distribution and correlated features are expected to induce a change in the spectrum of the observed eigenvalues while the total amount of variance in X (i.e. the sums of squared singular values) remains constant after standardization

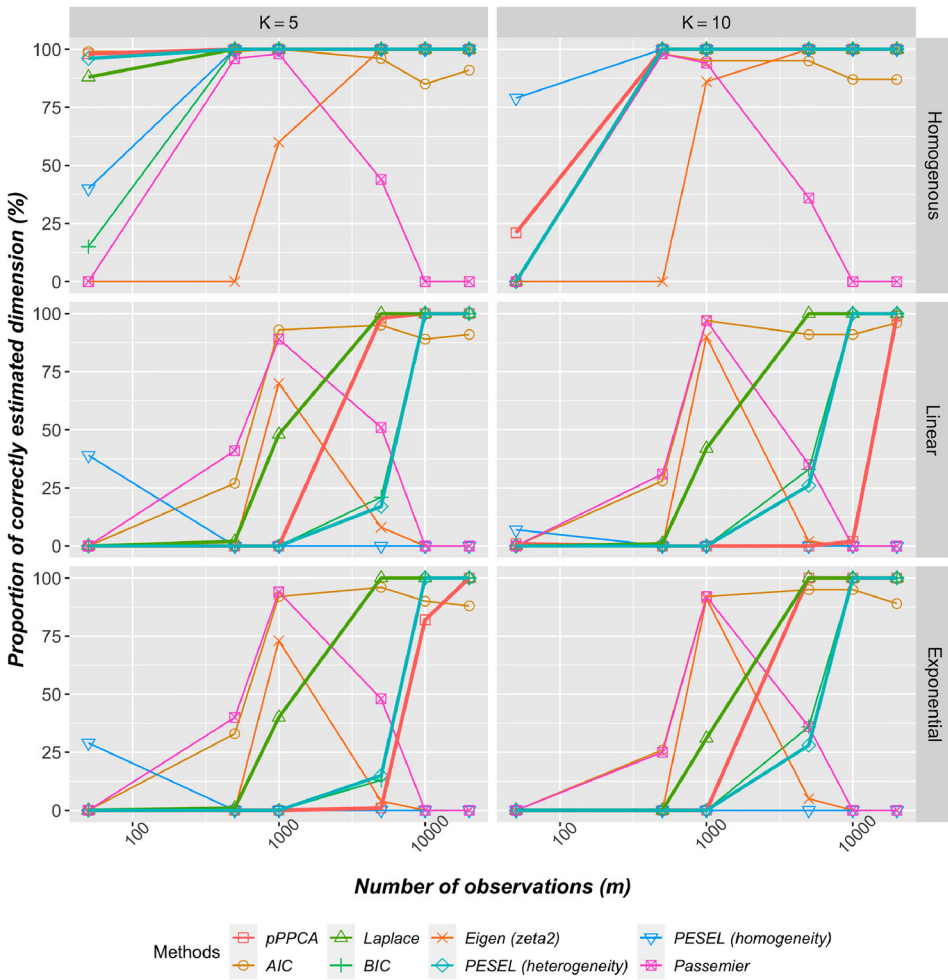


Figure 5. Proportion of correctly estimated dimension over 100 replicates as a function of m assuming homogeneity, a linear or an exponential decay in the first k squared singular values.

($= n(m - 1)$). Results from the multivariate normal error condition suggested that moderately correlated observations had little to no impact on *pPPCA* nor *Laplace*, as long as the decay was relatively smooth around the true dimension, as is the case for AR1 with correlation coefficient of 0.4, AR2 with parameter values 0.4 and 0.4^2 , and polynomial kernels (Supplementary Figure 3). We observed similar results for multivariate lognormal or exponential error terms when the covariance was an identity matrix, but slightly deteriorated performance from *pPPCA* when the covariance deviated from identity (Supplementary Figures 4 and 5). The most noticeable error distribution that created a shift in the eigenvalue spectrum was the multivariate *t*-distributed error (Supplementary Figure 6), modifying the true SNR and thus making the estimation of effective rank more difficult across all covariance structures for all methods (Figure 6).

All methods except *GCV*, *PESEL (heterogeneity)*, *Eigen (zeta2)* and *pPPCA*, failed completely at identifying the true k for AR1 with stronger correlation coefficient (0.7), Gaussian

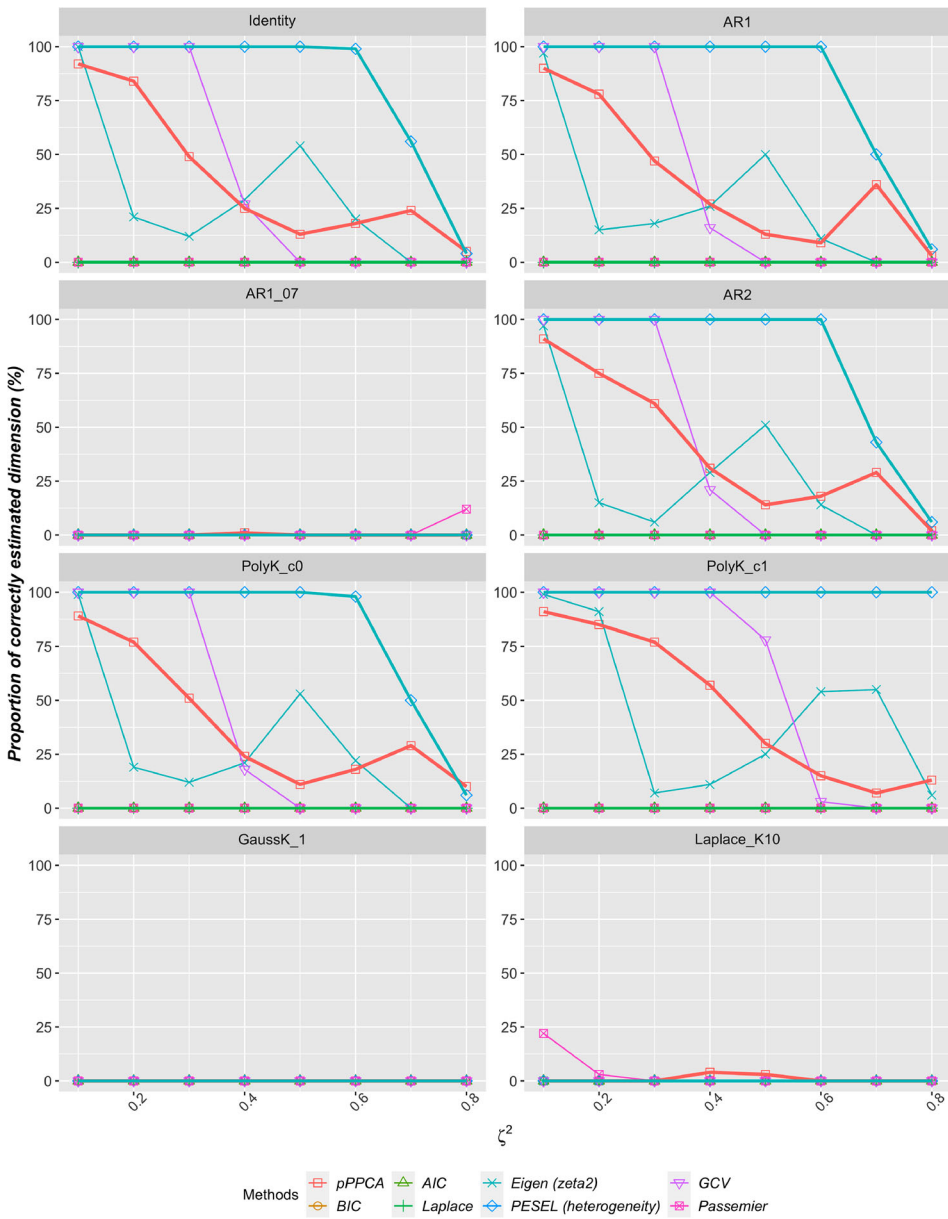


Figure 6. Proportion of correctly estimated dimension over 100 replicates as a function of ζ^2 assuming an exponential decay in the first k squared singular values under non-normality.

and Laplacian kernels (Figure 6). *PESEL (heterogeneity)* is the most competitive when the errors were sampled from a multivariate t -distribution while the proposed *pPPCA* is more robust to correlation than distribution with fat tails for a range of SNRs, suggesting tolerance for moderated correlated data (Supplementary Figure 4 and 5). Though *GCV* was not as strong compared to *pPPCA* nor *PESEL (heterogeneity)*, but its performance was consistent across most S3 conditions (Supplementary Figures 4, 5 and 6). Indeed, *GCV*

approximates a cross-validation criterion and the underlying error distribution had less influence on its performance.

The fat tails and correlation in the error distribution present a challenge to *Laplace* as the criterion were derived based under the normal asymptotic conditions (Figure 6). Naturally, a poor estimation of the residual variance ζ^2 leads to an incorrect estimation of k , which affects all methods under comparison. Indeed, the residual variance would impact the estimated dimension through its relative size to the explained variance. A biased ζ^2 estimate has a direct impact on the estimated dimension provided that the signal remains the same: a smaller \hat{k} is expected for an upward biased ζ^2 estimates, while a larger \hat{k} is expected for a downward biased ζ^2 estimate. In the setting considered here, $n < m$, a data rich case as our interest is in the samples rather than features, the MLE estimator of ζ^2 is consistent and approximately unbiased. However, when $n > m$, there will be a downward bias that requires the use of a biased corrected estimator [19,38].

4.8. Computational considerations

The computational complexity of each method largely depends on the computation of covariance and its associated eigenvalues, which can be computed once for all methods. Specific to pPPCA, given the grid size, the computational cost is linear in the true dimension (Supplementary Figure 7), as the algorithm first examines the possible penalty values for each searchable dimension before aggregating them to support an estimated dimension.

5. Application to microarray gene expression data

Large-scale gene expression data over multiple tissues have made it possible for scientists to study the global structure of expression profiles [39] and extract biologically relevant information. It has been reported that linear projections of expression data have intrinsically low dimensions, but higher than previously thought [40–42]. Here we apply the proposed method to a heterogeneous gene expression dataset to inform the effective rank.

5.1. NCI60 data

This data contained gene expression measured across 9 types of human cancer cell lines [43], and has been recently profiled using microarray technology at $m = 41,000$ gene probes [44]. The pre-processed data were obtained from the European Bioinformatics Institute database and a total of $n = 60$ samples were analysed after removing 65 duplicated cell line samples (Table 2).

As only 30–40% of genes are expected to expressed in each tissue [45], a standard variance filter was applied to remove gene probes with variance lower than their 10% percentile value. In many cases, the excessively large variance corresponds to expression with bimodal or even multi-modal distribution, and thus we removed gene probes with variance above 95% percentile. The sizes of variance filters roughly correspond to 0.2 and 5.8 on the \log_2 scale, which reduced the number of gene probes from $m = 41,000$ to $m = 34,850$. See Supplementary Figure 4 for a summary of the sample and gene variance, as well as gene-based skewness and kurtosis prior to filtering. For each gene probe, the expression

values were further standardized across samples to have a sample mean of zero and variance of 1. The sample eigenvalues were calculated based on the singular values of X after standardization to be $\hat{\lambda}_i = \frac{\hat{d}_i^2}{m}$.

5.2. Data analysis

Since correlation in both rows and columns is expected of gene expression data, we assessed the burden of such correlation using the averaged squared Pearson’s correlation coefficient for each gene or sample (Figure 7). In addition, gene expression distribution can be notoriously non-normal, with more than 50% of gene features exhibiting heavy tails, skewness, and even multiple modes [46,47]. For a given dataset, we compared results on both the standardized data and those undergoing a rank-based inverse normal transformation for each gene feature. For alternative methods, only the most sensible estimate from a class of methods was reported, i.e. the value closest to the reported number of cell lines. Note that the reported results are exploratory in nature and had not been rigorously validated in terms of their biological interpretation nor clinical relevance.

As a follow-up analysis, we first estimated the dimension for the melanoma cell line alone since it had the highest number of samples (Table 2), and then increased the number

Table 2. NCI 60 cell line classes.

| Tissue of origin | Number of samples |
|------------------------|-------------------|
| Breast | 6 |
| Central nervous system | 6 |
| Colon | 10 |
| Leukaemia | 7 |
| Melanoma | 11 |
| Non-small cell lung | 8 |
| Ovarian | 7 |
| Prostate | 2 |
| Renal | 9 |

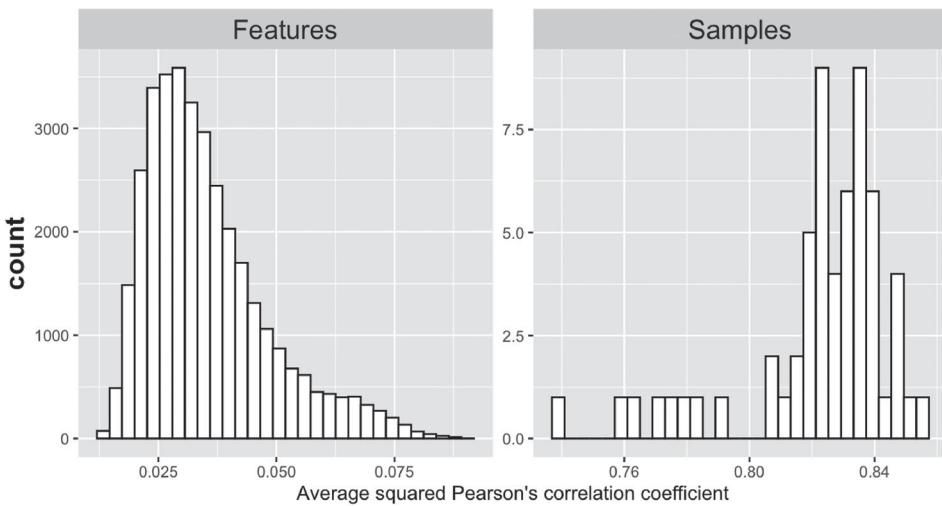


Figure 7. Averaged squared Pearson’s correlation coefficient for each sample or feature.

of samples by introducing additional cell line groups one by one, in the order of decreasing sample sizes per cell line (colon, renal, etc.). We hope the trend in estimated dimension as the data dimension increased can shed light on the structure of microarray data as they become increasingly more heterogeneous.

The estimated data dimension can also be used for supervising learning. Thus, we evaluated the clustering quality of samples using testing data consist of 17,425 gene probes while the number of clusters was determined from the estimated dimension on training data consist of the remaining gene probes. We applied the hierarchical clustering algorithm using Ward’s method [48] and the clustering quality was assessed using the adjusted Rand index [49], which compares the similarity between the assigned cluster and true cluster. The adjusted Rand index takes into account the number of clusters and larger values indicate better agreement. Each random data-splitting was repeated 1000 times.

5.3. Results

There was no visible difference in the sample eigenvalues for data irrespective of a rank-based inverse normal transformation: in both cases we observed a smooth decay with no clear elbow (Figure 8). The penalized approach estimated $\tilde{k} = 10$ for both the standardized and the transformed data, suggesting robustness to non-normal features of the data. By design, empirical methods that are sensitive to the presence of a gap also gave similar estimates, for example, *profileL*, *GCV*, *Lawley*, and elbow based approaches. Notably, *GCV*, *Lawley*, and the best of the elbow approach are in agreement with our penalized approach (Table 3), giving estimates roughly in line with the number of cancer cell lines ($k = 9$). On the other hand, model-based methods, such as *AIC*, *BIC*, and *Laplace* were unable to give sensible estimates. In particular, many overwhelmingly identified the boundary points at around $k = n - 1$ or $k = 1$. This observation agrees with Minka’s comments in [23] that Bayesian methods do not perform well when data deviated from a reasonable level of normality and when the last $n - k$ sample eigenvalues decay faster than expected under the model, a result of either severe non-normality or the true eigenvalues of the last

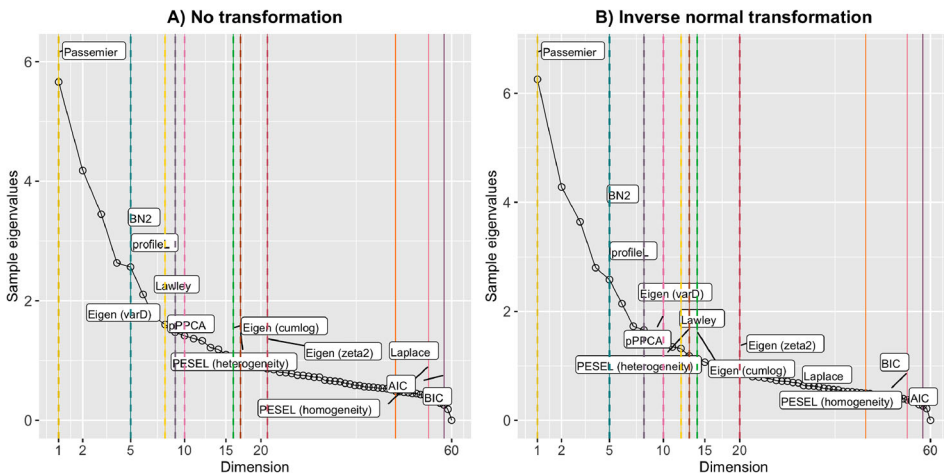


Figure 8. Estimated effective rank by each method with respect to the sample eigenvalue scree plot.

Table 3. Estimated dimension of the NCI60 dataset.

| Methods | No transformation | Inverse normal transformation |
|-----------------------|-------------------|-------------------------------|
| pPPCA | 10 | 10 |
| AIC | 58 | 58 |
| BN | 5 | 5 |
| BIC | 58 | 58 |
| Best elbow approach | 9 | 8 |
| GCV | 10 | 12 |
| Laplace | 46 | 44 |
| Lawley | 8 | 12 |
| PESEL (heterogeneity) | 17 | 13 |
| PESEL (homogeneity) | 54 | 54 |
| Passemier | 58 | 58 |
| ProfileL | 5 | 5 |

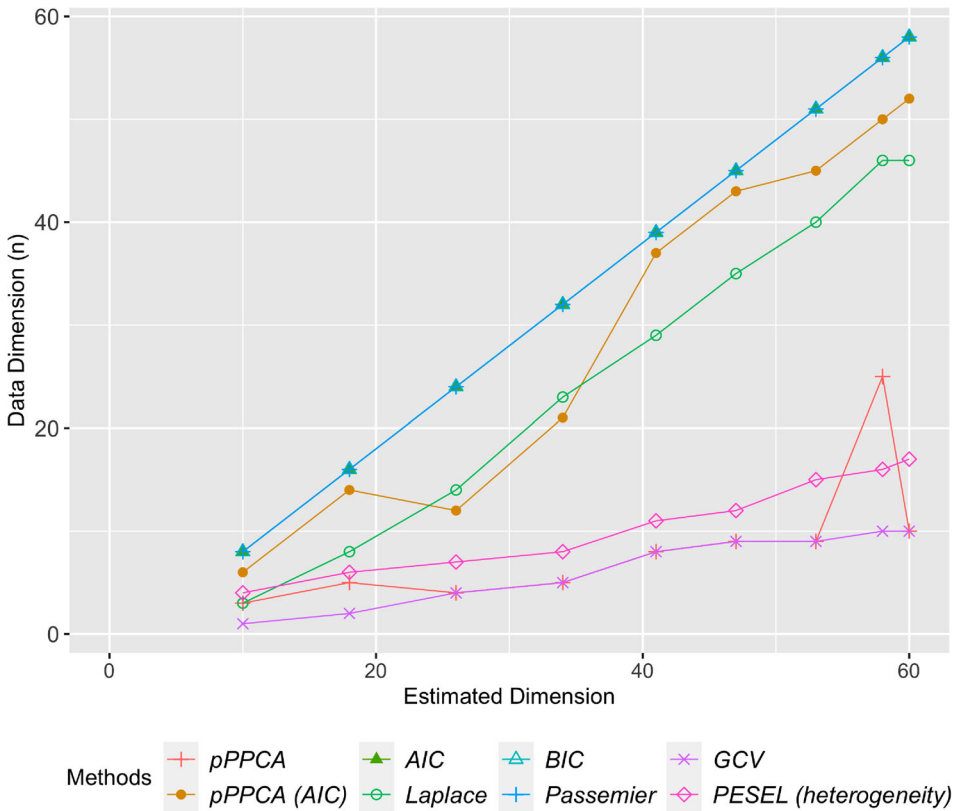


Figure 9. Estimated effective rank with respect to increasing numbers of cell lines by each method.

$n-k$ principal directions not being constant. The performance of *PESEL (homogeneity)* seemed to suggest the later is more likely as it had shown fairly good performance under non-normality in simulations.

Since the expression data are heterogeneous coming from multiple cell lines, we sought to examine the data dimension as a function of increasing data complexity. Figure 9 reveals that the dimension increased with with additional cell line being included in the data.

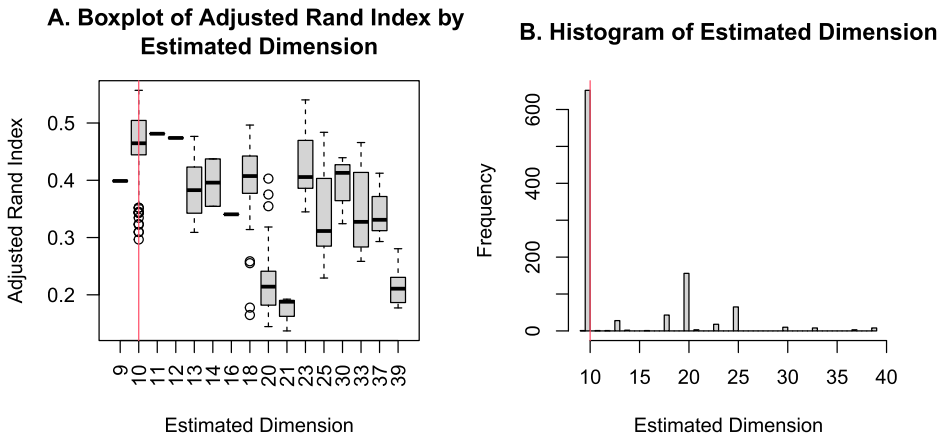


Figure 10. The figure shows a boxplot of the adjusted Rand Index in testing data as a function of estimated dimension in training data, and a histogram of estimated dimension from 1000 replications. The solid line corresponds to the estimated dimension at $\tilde{k} = 10$.

The dimension estimated from the overall dataset using pPPCA ($\tilde{k} = 10$) was chosen most frequently in the 1000 replications using the testing expression data ($n = 60$, $m = 17, 425$). The clustering quality of samples to cancer cell line groups using the estimated dimension of 10 gave the best performance as evaluated by the adjusted Rand index (Figure 10).

6. Concluding remarks

Both Bayesian methods and penalized approaches are often linked to improved prediction performance as a result of internally choosing the more parsimonious model. Here we compared their performance on the non-supervised learning of data dimension. Rather than an out-of-sample criterion, the estimation of dimension is very much ‘in sample’ as we are primarily interested in the representation of this particular dataset and do not expect it to generalize beyond very homogeneous populations.

The comprehensive simulation design covered a wide range of theoretical and realistic data scenarios, focusing on the impact of SNR, patterns of eigenvalue spectrum, relative sizes of m and n , and correlated and non-normal error. The proposed pPPCA strikes a balance between capturing the ‘gap’ in the top sample eigenvalues via the voting strategy as well as modelling the error variance via a likelihood penalization. Thus, its complementary performance to the approximated Bayesian posterior likelihood and ‘elbow’ based approaches is unsurprising. This also explains its good performance when data deviated from the independence assumption, an advantage in applications where one might be uncertain of the characteristics of the data generating process.

Even though the proposed method was not the ‘best’ in every scenario, its overall performance was competitive. Irrespective of other simulation conditions, it has good performance for large k as the penalty on the estimated dimension is mostly driven by $\log(\zeta^2)$, which favours a model that is more flexible than preferred by Bayesian model selection.

Supported by the application results, we recommend applying $pPPCA$ to explore the dimension of gene expression data when there is a good separation between signal and noise, and proper data transformation applied. As a possible follow-up analysis, the data could be better modelled assuming \tilde{k} distinct error variance parameters using a generalized factor analysis model. Though in an exploratory analysis, the assumption of isotropic error covariance should suffice as a first step to identify the hidden dimension.

Acknowledgements

The authors would like to thank the anonymous reviewers and associate editor for their careful reading of our manuscript and comments. The authors would also like to acknowledge Professor Andrey Feuerverger for helpful suggestions, Professors Dehan Kong, Lei Sun, Qiang Sun, Stanislav Volgushev, and Fang Yao, for a critical reading of the original version of the paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by Alexander Graham Bell Canada Graduate Scholarships – Doctoral Program from the Natural Sciences and Engineering Research Council of Canada [CGSD-459873-2014].

ORCID

Wei Q. Deng  <http://orcid.org/0000-0003-4212-2607>

Radu V. Craiu  <http://orcid.org/0000-0002-1348-8063>

References

- [1] Pearson K. Liii. On lines and planes of closest fit to systems of points in space. Lond Edinb Dublin Philos Mag J Sci. 1901;2(11):559–572.
- [2] Hotelling H. Relations between two sets of variates. In: Kotz S, Johnson NL, editors. Break-throughs in Statistics. New York (NY): Springer; 1992. p. 162–190.
- [3] Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 1933;24(6):417.
- [4] Bartholomew DJ. Latent variable models and factors analysis. New York (NY): Oxford University Press, Inc.; 1987.
- [5] Johnstone IM. On the distribution of the largest eigenvalue in principal components analysis. Ann Stat. 2001;29(2):295–327.
- [6] Jolliffe IT. Principal component analysis. 2nd. New York (NY): Springer; 2002. p. 111–149.
- [7] Cattell RB. The scree test for the number of factors. Multivariate Behav Res. 1966;1(2):245–276.
- [8] Zhu M, Ghodsi A. Automatic dimensionality selection from the scree plot via the use of profile likelihood. Comput Stat Data Anal. 2006;51(2):918–930.
- [9] Bartlett MS. A note on the multiplying factors for various χ_2 approximations. J R Stat Soc. 1954;16(2):296–298.
- [10] Lawley D. Tests of significance for the latent roots of covariance and correlation matrices. Biometrika. 1956;43(1/2):128–136.
- [11] Ledoit O, Wolf M. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. Ann Stat. 2002 08;30(4):1081–1102 <http://dx.doi.org/10.1214/aos/1031689018> .

- [12] Schott JR. A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix. *J Multivariate Anal.* 2006;97(4):827–843.
- [13] Forzani L, Gieco A, Tolmasky C. Likelihood ratio test for partial sphericity in high and ultra-high dimensions. *J Multivariate Anal.* 2017;159(Supplement C):18–38.
- [14] Choi Y, Taylor J, Tibshirani R, et al. Selecting the number of principal components: estimation of the true rank of a noisy matrix. *Ann Stat.* 2017;45(6):2590–2617.
- [15] Mardia K, Kent J, Bibby J. *Multivariate analysis.* London; New York: Academic Press; 1979.
- [16] Josse J, Husson F. Selecting the number of components in principal component analysis using cross-validation approximations. *Comput Stat Data Anal.* 2012;56(6):1869–1879.
- [17] Gavish M, Donoho DL. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Trans Inf Theory.* 2014;60(8):5040–5053.
- [18] Bai BY, Ng S. Determining the number of factors in approximate factor models. *Econometrica.* 2002;70(1):191–221 <http://www.jstor.org/stable/2692167>.
- [19] Passemier D, Li Z, Yao J. On estimation of the noise variance in high dimensional probabilistic principal component analysis. *J R Stat Soc.* 2017;79(1):51–67.
- [20] Tipping ME, Bishop CM. Probabilistic principal component analysis. *J R Stat Soc.* 1999; 61(3):611–622.
- [21] Bishop CM. Bayesian PCA. *Adv Neural Inf Process Syst.* 1999;11:382–388.
- [22] Hoff PD. Model averaging and dimension selection for the singular value decomposition. *J Amer Statist Assoc.* 2007;102(478):674–685.
- [23] Minka TP. Automatic choice of dimensionality for PCA. In: Dietterich T, Becker S, Ghahramani Z, editors. *Advances in Neural Information Processing Systems 14: Proceedings of the 2001 Conference.* Cambridge (MA): MIT Press; 2001. p. 598–604.
- [24] Bishop CM. Variational principal components. 07-10 September 1999, Edinburgh, UK. Ninth International Conference on Artificial Neural Networks (ICANN 99); 1999.
- [25] Nakajima S, Tomioka R, Sugiyama M, et al. Perfect dimensionality recovery by variational Bayesian PCA. In: Pereira F, Burges CJ, Bottou L, et al., editors. *Advances in Neural Information Processing Systems 25.* Cambridge (MA): MIT Press; 2012. p. 971–979.
- [26] Bouveyron C, Latouche P, Mattei PA. Exact dimensionality selection for Bayesian PCA. arXiv preprint arXiv:170302834. 2017.
- [27] Hoyle DC. Automatic PCA dimension selection for high dimensional data and small sample sizes. *J Mach Learn Res.* 2008;9(Dec):2733–2759.
- [28] Sobczyk P, Bogdan M, Josse J. Bayesian dimensionality reduction with pca using penalized semi-integrated likelihood. *J Comput Graph Stat.* 2017;26(4):826–839.
- [29] Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc.* 1996;58(1):267–288 <http://www.jstor.org/stable/2346178>.
- [30] Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. *J Comput Graph Stat.* 2006;15(2):265–286.
- [31] Bien J, Tibshirani RJ. Sparse estimation of a covariance matrix. *Biometrika.* 2011;98(4):807–820.
- [32] Wang H, Li B, Leng C. Shrinkage tuning parameter selection with a diverging number of parameters. *J R Stat Soc.* 2009;71(3):671–683 <http://dx.doi.org/10.1111/j.1467-9868.2008.00693.x>.
- [33] Kass RE, Raftery AE. Bayes factors. *J Amer Statist Assoc.* 1995;90(430):773–795.
- [34] Berthet Q, Rigollet P. Optimal detection of sparse principal components in high dimension. *Ann Stat.* 2013;41(4):1780–1815.
- [35] Everson R, Roberts S. Inferring the eigenvalues of covariance matrices from limited, noisy data. *IEEE Trans Signal Process.* 2000;48(7):2083–2091.
- [36] Rajan J, Rayner P. Model order selection for the singular value decomposition and the discrete Karhunen–Loeve transform using a Bayesian approach. *IEEE Proc Vis Image Signal Process.* 1997;144(2):116–123.
- [37] Ilin A, Raiko T. Practical approaches to principal component analysis in the presence of missing values. *J Mach Learn Res.* 2010;11:1957–2000.

[38] Passemier D, Yao J. Estimation of the number of spikes, possibly equal, in the high-dimensional case. *J Multivariate Anal.* **2014**;127:173–183.

[39] Lukk M, Kapushesky M, Nikkilä J, et al. A global map of human gene expression. *Nat Biotechnol.* **2010**;28(4):322–324.

[40] Heimberg G, Bhatnagar R, El-Samad H, et al. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst.* **2016**;2(4):239–250.

[41] Lenz M, Müller FJ, Zenke M, et al. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. *Sci Rep.* **2016**;6(1):1–11.

[42] Ding J, Condon A, Shah SP. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat Commun.* **2018**;9(1):1–13.

[43] Ross DT, Scherf U, Eisen MB, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet.* **2000 Mar**;24(3):227–235.

[44] Liu H, D’Andrade P, Fulmer-Smentek S, et al. mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. *Mol Cancer Ther.* **2010 May**;9(5):1080–1091.

[45] Su AI, Cooke MP, Ching KA, et al. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci USA.* **2002**;99(7):4465–4470.

[46] de Torrenté L, Zimmerman S, Suzuki M, et al. The shape of gene expression distributions matter: how incorporating distribution shape improves the interpretation of cancer transcriptomic data. *BMC Bioinform.* **2020**;21(21):1–18.

[47] Marko NF, Weil RJ. Non-Gaussian distributions affect identification of expression patterns, functional annotation, and prospective classification in human cancer genomes. *PLoS ONE.* **2012**;7(10):e46935.

[48] Ward Jr JH. Hierarchical grouping to optimize an objective function. *J Amer Statist Assoc.* **1963**;58(301):236–244.

[49] Hubert L, Arabie P. Comparing partitions. *J Classif.* **1985**;2(1):193–218.

Appendices

Appendix 1. Penalized PPCA and Minka’s criterion using Laplace’s method

Denote $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$, where $\lambda_i = \eta_i^2 + \zeta^2$. Intriguingly, if we use the following priors as suggested in Minka [23]:

$$p(U, \Lambda, V, \zeta^2) = p(\zeta^2)p(U)p(R)\prod_{i=1}^k p(\lambda_i); \tag{A1}$$

$$p(\zeta^2) = \frac{1}{\Gamma(\alpha(n-k)/2)\zeta^2} \left(\frac{\beta(n-k)}{2\zeta^2}\right)^{\alpha(n-k)/2} \exp\left(-\frac{\beta(n-k)}{2\zeta^2}\right); \tag{A2}$$

$$p(U) = 2^{-k}\prod_{i=1}^k \Gamma((n-i+1)/2)\pi^{-(n-i+1)/2}; \tag{A3}$$

$$p(\lambda_i) = \frac{1}{\Gamma(\alpha/2)\lambda_i} \left(\frac{\beta}{2\lambda_i}\right)^{\alpha/2} \exp\left(-\frac{\beta}{2\lambda_i}\right); \tag{A4}$$

and maximize the posterior with respect to (λ_i, ζ^2) at the maximum likelihood of U , we have:

$$\hat{\zeta}^2(\alpha, \beta) = \frac{m}{m-1+\alpha} \hat{\zeta}_k^2 + \frac{\beta}{m(n-k)}; \tag{A5}$$

$$\hat{\lambda}_i(\alpha, \beta) = \frac{m}{m-1+\alpha} \hat{\lambda}_i + \frac{\beta}{m}. \tag{A6}$$

The final approximated Laplace evidence removed any terms that do not depend strongly on k and simplified assuming m is large and (α, β) are small.

The penalized MLE from our proposed penalty function (7a) then corresponds to the hyperparameter values of:

$$\alpha = 1 - \frac{k}{n - k} \delta, \tag{A7}$$

$$\beta = 0; \tag{A8}$$

while the second penalty function (7b) corresponds to:

$$\alpha = 0, \tag{A9}$$

$$\beta = \frac{k}{n - k} \delta, \tag{A10}$$

that coincides with a prior using Levy distribution.

Appendix 2. Estimation of effective rank via penalized PPCA

A.1 Lemmas

Lemma A.1: Consider a sample $X \in \mathbb{R}^{n \times m}$ with each column following a multivariate Gaussian distribution $\mathcal{N}(0, WW^T + \zeta^2 I)$. Suppose W has rank k and further, the sample covariance matrix of X^T is positive semi-definite. Then, the penalized maximum log-likelihood at each fixed $q \in \{1, \dots, n - 1\}$ is a smooth function of $\tilde{\delta}$ on the interval $(0, 1/q - 1/n)$ and is monotonically decreasing on

$$(0, (1/q - 1/n)[1 - \hat{\zeta}_q^2]), \tag{A11}$$

where $\hat{\zeta}_q^2 = (\sum_{i=q+1}^n \hat{\lambda}_i)/(n - q)$.

Since the difference of two smooth functions is still a smooth function, the monotonicity of $l_p(q; \tilde{\delta}) - l_p(q + 1; \tilde{\delta})$ and $l_p(q; \tilde{\delta}) - l_p(q - 1; \tilde{\delta})$ can be established with respect to $\tilde{\delta}$.

Lemma A.2: Consider $\tilde{\delta} \in G(q + 1)$, where

$$G(q + 1) = \left(0, \frac{1}{n} \frac{(n - q - 1)(\hat{\lambda}_{q+1} - \hat{\zeta}_{q+1}^2)}{(q + 1)\hat{\lambda}_{q+1} + (n - q - 1)\hat{\zeta}_{q+1}^2} \right). \tag{A12}$$

Then, for any fixed $q \in \{2, 3, \dots, n - 2\}$, $l_p(q; \tilde{\delta}) - l_p(q + 1; \tilde{\delta})$ is a monotonically increasing and concave function of $\tilde{\delta} \in G(q + 1)$ and $l_p(q; \tilde{\delta}) - l_p(q - 1; \tilde{\delta})$ is a monotonically decreasing and convex function of $\tilde{\delta} \in G(q + 1)$.

Since $l_p(q; \tilde{\delta}_o)$ is a discrete function of q , the maximum can be at either the boundary points or interior points. Considering exclusively the interior points, for some $q \in \{2, \dots, n - 2\}$ to be the maximizer of $l_p(q; \tilde{\delta}_o)$ given $\tilde{\delta}_o$, $l_p(q; \tilde{\delta}_o) - l_p(q - 1; \tilde{\delta}_o) > 0$ and $l_p(q; \tilde{\delta}_o) - l_p(q + 1; \tilde{\delta}_o) > 0$ constitute a necessary but not sufficient condition. With the additional condition that $l_p(q; \tilde{\delta}_o)$ monotonically increases $\forall q < k$ and monotonically decreases $\forall q > k$, the condition A13 becomes necessary and sufficient. The following Lemma proves the sufficiency of the condition that guarantees the true dimension k to be the maximizer for some $\tilde{\delta}_o \in \cup_q G(q + 1)$.

Lemma A.3: Assume the same notation from Lemma A.2. For $k \in \{2, \dots, n - 2\}$, there exists $\tilde{\delta}_o \in \cup_q G(q + 1)$ such that $k = \operatorname{argmax}_q l_p(q; \tilde{\delta}_o)$ if and only if

$$\begin{cases} l_p(q; \tilde{\delta}_o) - l_p(q - 1; \tilde{\delta}_o) > 0 \\ l_p(q; \tilde{\delta}_o) - l_p(q + 1; \tilde{\delta}_o) > 0. \end{cases} \tag{A13}$$

It is convenient to define the sets that satisfy A13 for $2 \leq q \leq n - 2$:

$$\Delta_q = (a_q, b_q) \subset \cup_q G(q + 1), \tag{A14}$$

where

$$a_q = \min \left\{ \tilde{\delta} \in \cup_q G(q + 1); l_p(q; \tilde{\delta}) - l_p(q + 1; \tilde{\delta}) > 0 \right\} \tag{A15}$$

and

$$b_q = \max \left\{ \tilde{\delta} \in \cup_q G(q + 1); l_p(q; \tilde{\delta}) - l_p(q - 1; \tilde{\delta}) > 0 \right\}. \tag{A16}$$

Remark A.1: Note that an interesting result from Lemma A.3 showing $\hat{\lambda}_{q+1} = 1$ to be a sufficient condition for $l_p(q; \tilde{\delta}) - l_p(q + 1; \tilde{\delta}) > 0$ on $\tilde{\delta} \in G(q + 1)$. This coincides with with Kaiser’s rule for selecting k as the number of PCs to retain. Notice that as $m \rightarrow \infty$,

$$\hat{\lambda}_{k+1} \rightarrow \lambda_{k+1} = \zeta^2,$$

while the observed $\hat{\zeta}_k^2 < 1$, then $\hat{\lambda}_k > \hat{\lambda}_{k+1} = 1$ provides strong evidence that the true $\eta_k^2 = \lambda_k - \zeta^2 > 0$.

Lemma A.4: Consider

$$\Delta_q = \left\{ \tilde{\delta} \in G(q + 1); \text{conditions (A13) are satisfied} \right\} = (a_q, b_q),$$

whenever a_q exists. Then Δ_q can be approximated by $(u_a(q), u_b(q)) \subset \Delta_q \subset G(q + 1)$, where $u_a(q)$ represents an upper bound for a_k , and $u_b(q)$ a lower bound for b_q , such that $b_q/a_q > \frac{u_b(q)}{u_a(q)}$.

Remark A.2: Clearly, $\frac{b_q}{a_q} > \frac{u_b(q)}{u_a(q)}$ holds. If the ratio $\frac{u_b(q)}{u_a(q)}$ converges as $m \rightarrow \infty$, the ratio asymptotically reflects the amount of evidence for each possible dimension q . Thus, if $\frac{u_b(q)}{u_a(q)}$ were the largest for $q = k$, then a majority-voting strategy for estimating k is viable.

Lemma A.5: Suppose k is the true rank of W , then as $m \rightarrow \infty$,

- $u_b(k)/u_a(k) \rightarrow \infty$ in probability
- $|u_b(q) - u_a(q)| \rightarrow 0$ in probability for $q > k$.

Remark A.3: In theory, $u_b(k)/u_a(k) \rightarrow \infty$ in probability and the approximated ratio will be the largest as compared to other choices. However, in finite samples, the ratio $u_b(q)/u_a(q)$ for $q > k$ could also be quite large due to the numerical inaccuracy of the last $n - q$ sample eigenvalues as they approach the population values. In practice, the penalty tuning parameter $\tilde{\delta}$ needs to be calibrated such that $u_a(q)$ is not too close to 0.

Remark A.4: The proof of Proposition 3.1 implies that given any $\tilde{\delta} = \tilde{\delta}_o$, a non-boundary maximizer of (9), k , can be identified using the following conditions whenever $1 < q < k_{\max}(\tilde{\delta}_o)$:

$$\begin{cases} l_p(q; \tilde{\delta}_o) - l_p(q + 1; \tilde{\delta}_o) > 0; \\ l_p(q; \tilde{\delta}_o) - l_p(q - 1; \tilde{\delta}_o) > 0, \end{cases} \tag{A17}$$

where $k_{\max}(\tilde{\delta}_o)$ is the maximum value for the search space that ensures $\tilde{\zeta}_q^2$ is well-defined given $\tilde{\delta} = \tilde{\delta}_o$. In other words, $l_p(q; \tilde{\delta}_o)$ first increases with $q \leq k$ and then decreases with $q \geq k$, thus ensuring k maximizes $l_p(q; \tilde{\delta}_o)$ over $q \in \{2, \dots, k_{\max}(\tilde{\delta}_o)\}$.

Remark A.5: It is clear that Δ_q is an open interval for each q as the penalized likelihood in (9) is a continuous function of $\tilde{\delta} \in (0, 1/(q + 1) - 1/n)$ for any fixed q (Lemma A.1). Following Lemmas A.2 and A.3, for $q \neq q'$, Δ_q and $\Delta_{q'}$ are strictly non-overlapping sets. Therefore, the realized

range for $\tilde{\delta}$ is the union of all sets $\cup_{q=2}^{n-2} \Delta_q \subset [0, 1 - 1/n)$. But because of the restriction embedded in (8) and (A13), we must have $\tilde{\delta} \in (0, 1/(q + 1) - 1/n)$ for each examined value of q . Consequently, the restriction imposes a relationship whereby q is non-increasing in a_q (or $u_a(q)$) and b_q (or $u_b(q)$). For example, when $q = n - 1$, it must be that $a_{n-1} = 0 < b_{n-1} < \frac{(n-1)^{-1} - n^{-1}}{1 - \hat{\xi}_{n-1}^2}$, while for $q = 1$, $b_2 < a_1 < b_1 < (1 - n^{-1})(1 - \hat{\xi}_1^2)$.

Remark A.6: For $q = 1$ or $q = n - 1$, Δ_q can be defined such that only one of (A13) is satisfied. It is clear that Δ_q is an open interval for each q as the penalized likelihood function in (9) is a continuous function of $\delta \in G(q)$ for any fixed q . However, in this case, as $a_{n-1} = 0$ and b_1 is unbounded, results from Lemmas A.4 and A.5 no longer apply. Instead, a practical solution is to construct suitable probabilistic models for $q = 0$ and $q = n$ such that the boundary points become interior points.

Remark A.7: Since a_q and b_q are not analytically available, whenever possible, I obtained conservative upper and lower bounds for Δ_q using $u_a(q)$ and $u_b(q)$ such that $(u_a(q), u_b(q)) \subset \Delta_q$ (Lemma A.4). The proof of Lemma A.5 also demonstrates that $u_b(k)/u_a(k) > 1$ so that $(u_a(k), u_b(k)) \neq \emptyset$. Essentially, the number of votes provides a form of evidence for division between the first q and last $n - q$ sample eigenvalues relative to the first $q - 1$ and last $n - q + 1$ or the first $q + 1$ and last $n - q - 1$.

Lemmas A.1, A.2, A.3, A.4, and A.5 together imply: (1) there exists $\tilde{\delta}_o \in \Delta_k$ such that (9) is maximized at k ; (2) $\Delta_k = (a_k, b_k)$ can be approximated by $(u_a(k), u_b(k)) \subset \Delta_k$, satisfying

$$\lim_{m \rightarrow \infty} \frac{u_b(k)}{u_a(k)} = \infty, \tag{A18}$$

$$\lim_{m \rightarrow \infty} |u_b(q) - u_a(q)| \rightarrow 0, \quad \text{for } q > k, \tag{A19}$$

$$\lim_{m \rightarrow \infty} \frac{u_b(q)}{u_a(q)} < \infty, \quad \text{for } q < k. \tag{A20}$$

A.2 Establishing the range of plausible tuning parameter values

We first consider the search space for k . The smallest and the largest non-trivial choice for k is 1 and $n - 1$, respectively. Clearly, the largest possibly value that k can take depends on the actual rank of the sample covariance. We define $n_{\max} = \min\{i : \hat{\lambda}_i < \kappa\} - 1$ where κ is a tolerance value that can be set arbitrarily low to prevent digits over-floating in standard software (e.g. $\kappa = 0.001$). This, in effect, removes numerical uncertainty in the inverse of sample eigenvalues.

The construction of search grid is characterized by its range and the distance between adjacent grid values. Results from Lemma A.5 suggest a possible construction, $\tilde{\delta}_1, \dots, \tilde{\delta}_T$, using a sequence of T equidistant points on log scale. To determine $\tilde{\delta}_1$ and $\tilde{\delta}_T$, we need to bound the minimum and the maximum of $\tilde{\delta}$ values such that 1 and n_{\max} are the maximizer of the penalized profile log-likelihood (9). Since the exact relationship between q and a_q, b_q is not analytically available, we rely on conservative bounds obtained via Taylor series approximations to specify $\tilde{\delta}_1$ and $\tilde{\delta}_T$.

The maximum value $\tilde{\delta}_T$ is defined as the average of the two largest penalties that $\tilde{\zeta}_1^2 > \tilde{\zeta}_2^2$ and $\tilde{\zeta}_2^2 > \tilde{\zeta}_3^2$ hold, as above these values, the model is overwhelmed by the penalty and will always choose $k = 1$. The minimum value $\tilde{\delta}_1$ is chosen to be the value given by $u_a(n_{\max})$.

In practice, the boundary points $q = 1$ and $q = n - 1$ might be relevant, and have the interpretations of $n - 1$ independent error or signal components, respectively. To curb the definition of a maximizer according to A13, we propose to construct artificial boundary points for which the penalized profile log-likelihood is defined for $q = n$ or $q = n_{\max}$ and $q = 0$.

Define $\hat{\lambda}_0 = \sum_{i=1}^n \hat{\lambda}_i$ and the error variance is then $\hat{\zeta}_0^2 = 1$. This model corresponds to $W = \mathbf{0}$ and thus $q = 0$. By design, $l_p(q = 0; \tilde{\delta}) = l_p(q = 0)$ as there is no dimension to penalize. On the

other hand, since $\hat{\zeta}_{n-1} = \hat{\lambda}_n$ and $l_p(q = n - 1) = l_p(q = n)$, the construction must impose a small probabilistic component to the $q = n$ model (e.g. PCA) by introducing $\hat{\zeta}_n^2 = \frac{1}{2}\hat{\lambda}_n = \frac{1}{2}\hat{\zeta}_{n-1}^2$. This model corresponds to a dimension that is between $q = n - 1$ and $q = n$ and forces $l_p(q = n; \tilde{\delta}) = l_p(q = n) < l_p(q = n - 1)$.

These artificially constructed boundary points makes it possible to select a maximum for the penalized maximum likelihood by choosing $q = 1$ or $q = n - 1$ such that $l_p(q; \delta) - l_p(q - 1; \delta) > 0$ and $l_p(q; \delta) - l_p(q + 1; \delta) > 0$.

Appendix 3. Alternative methods

Details of the alternative methods considered in the simulation studies are included here and organized in alphabetical order.

A.3 Akaike information criterion (AIC)

The number of free parameters in the model is $nq + 1 - q(q - 1)/2$ and the model with the smallest AIC is selected:

$$AIC(q) = -2l_p(q) + nq + 1 - q(q - 1),$$

where $l_p(q)$ is defined as in (6).

$$\hat{k} = \arg \min_q AIC(q).$$

A.4 Bayesian information criterion (BIC)

A simplification to the Laplace’s method assuming $m \rightarrow \infty$ [33]:

$$\log p(D|q) = -\frac{m}{2} \left(\sum_{i=1}^q \log \hat{\lambda}_i \right) - \frac{m(n - q)}{2} \log \left(\frac{\sum_{i=q+1}^n \hat{\lambda}_i}{n - q} \right) - \frac{nq - (q + 1)q/2 + q}{2} \log(m), \tag{A21}$$

where any terms that do not depend on m are dropped. It can be shown that this simplifies to the likelihood under a model subtracted by a multiple of the number of free parameters, which is the usual BIC criterion $BIC(q) = -2l_p(q) + \frac{nq+1-q(q-1)}{2} \log m$.

$$\hat{k} = \arg \min_q BIC(q).$$

A.5 Bai and Ng’s criteria (BN)

Bai and Ng [18] developed six different criteria via a selection of penalty functions involving both m and n to identify the number of factors, where the errors are allowed to be correlated. The inference was performed jointly on (k, ζ^2) .

The three criteria applicable to PPCA models are:

$$\hat{k} = \arg \min_q V(q, \hat{F}^q) + q\hat{\zeta}_{BN}^2 g_j(m, n), \tag{A22}$$

where $\hat{\zeta}_{BN}^2 = \frac{1}{nm} \sum_{j=1}^m \sum_{i=1}^n E(x_{ij})^2$, $V(q, \hat{F}^q) = \frac{1}{nm} \sum_{j=1}^m E(X_j^T X_j)$ and the three penalty functions:

$$g_1(m, n) = \frac{m + n}{nm} \log \left(\frac{nm}{n + m} \right), \tag{A23}$$

$$g_2(m, n) = \frac{m + n}{nm} \log \min(n, m), \tag{A24}$$

and

$$g_3(m, n) = \frac{\log \min(n, m)}{\min(n, m)}. \tag{A25}$$

Following the PPCA model, the criterion reduces to

$$\hat{k} = \arg \min_q \hat{\zeta}_q^2 + q \hat{\zeta}_{k_o}^2 g_j(m, n), \tag{A26}$$

where k_o is the maximum number of PCs searched. Alternatively, the estimators $\hat{\zeta}_q^2$ can be replaced by the bias corrected estimators introduced in [19]. Thus, giving a total of 6 criteria used for comparison. For k_o , I chose $\lfloor \frac{n}{2} \rfloor$ as it gave the best performance across scenarios.

A.6 Empirical elbow approaches (Elbow)

I have also included in the comparison a few empirical approaches designed to detect an ‘elbow’ or a point of inflection in the scree plot produced by the sample eigenvalues:

- (1) The difference between log cumulative mean of the sample eigenvalues and the mean of the cumulative log sample eigenvalues (*cumlog*), defined by

$$\hat{k}_{cumlog} = \arg \min_q \log \frac{\sum_{i=1}^q \hat{\lambda}_i}{q} - \frac{1}{q} \sum_{i=1}^q \log \hat{\lambda}_i,$$

- (2) the variance of sample eigenvalues (*VarD*), defined by

$$\hat{k}_{VarD} = \arg \min_q \frac{\sum_{i=1}^q \hat{\lambda}_i^2}{q} - \left(\frac{\sum_{i=1}^q \hat{\lambda}_i}{q} \right)^2, \tag{A27}$$

- (3) the adjacent sample eigenvalues (*adjD*), defined by

$$\hat{k}_{adjD} = \arg \min_q \frac{\hat{\lambda}_q}{\hat{\lambda}_{q+1}}, \tag{A28}$$

- (4) and a criterion based on the log of estimated error variance (*log-var*), defined by

$$\hat{k}_{log-var} = \arg \min_q (n - q) \log \hat{\zeta}_q^2. \tag{A29}$$

A.7 A general cross-validation criterion (GCV)

This criterion is similar to the general cross-validation in regression to approximate the leave-one-out cross-validation, which is based on the relationship between prediction error and residual sum of squared via a weight matrix resulted from a projecting matrix. This enables a smoothing approximation to cross-validation criterion results in a general cross-validation (GCV) criterion that is computationally advantageous:

$$\hat{k}_{GCV} = \arg \min_q \frac{m^2 n \sum_{i=q+1}^n \hat{\lambda}_i}{[(m - 1)n - mq - nq + q^2 + q]^2}. \tag{A30}$$

To produce optional results, data would be transposed if the number of observations were smaller than sample size.

A.8 An approximation to the posterior likelihood using Laplace’s method (Laplace)

Laplace approximation [23,27] assumes the dimension of the parameter space is constant. Thus, Z is integrated out [23] and the resulting posterior likelihood is approximated using Laplace’s method [33], which requires the arg max of the parameters and the Hessian matrix at these values.

The log of the evidence is:

$$\begin{aligned}
 \log p(D|q) &= \log p(U) - m/2 \left(\sum_{i=1}^q \log \hat{\lambda}_i \right) - m(n-q)/2 \log \left(\frac{\sum_{i=q+1}^n \hat{\lambda}_i}{n-q} \right) \\
 &\quad + \frac{2nq - q^2 + q}{4} \log(2\pi) - q/2 \log(m) \\
 &\quad - 1/2 \sum_{i=1}^q \sum_{j=i+1}^n \left[\log \left(\frac{(\hat{\lambda}_i - \hat{\lambda}_j)^2}{\hat{\lambda}_i \hat{\lambda}_j} \right) + \log(m) \right] \\
 &= -m/2 \left(\sum_{i=1}^q \log \hat{\lambda}_i \right) - \frac{m(n-q)}{2} \log \left(\frac{\sum_{i=q+1}^n \hat{\lambda}_i}{n-q} \right) \\
 &\quad - 1/2 \sum_{i=1}^q \sum_{j=i+1}^n \left[\log \frac{(\hat{\lambda}_i - \hat{\lambda}_j)^2}{\hat{\lambda}_i \hat{\lambda}_j} \right] \\
 &\quad + \frac{2nq - 3q - q^2}{4} \log(2) - \frac{q^2 - 2nq + 3q}{4} \log(m) + \sum_{i=1}^q \log \Gamma \left(\frac{n-i+1}{2} \right),
 \end{aligned} \tag{A31}$$

where

$$\log p(U) = -q \log(2) + \sum_{i=1}^q \log \Gamma \left(\frac{n-i+1}{2} \right) - \frac{2nq + q - q^2}{4} \log(\pi) \tag{A32}$$

A.9 A hypothesis testing criterion for the equality of the last $n-k$ eigenvalues (Lawley)

The null hypothesis is $H_0 : \lambda_j = \lambda_{j+1} = \dots = \lambda_n$ against the alternative hypothesis that at least one is not equal to the remaining eigenvalues. The test statistic is given by Lawley [10]:

$$\chi^2 = (n-j)c \log \left(\frac{\sum_{i=j}^n \hat{\lambda}_i}{n-j} \right) - \sum_{i=j}^n \log \hat{\lambda}_i \tag{A33}$$

where

$$c = (n-j) - \frac{2(n-j) + 1 + 2/(n-j)}{6} + \left(\frac{\sum_{i=j}^n \hat{\lambda}_i}{n-j} \right)^2 \sum_{i=1}^j \left(\sum_{i=1}^j \hat{\lambda}_i - \frac{\sum_{i=j}^n \hat{\lambda}_i}{n-j} \right)^{-2},$$

with $\frac{(n-j)(n-j+1)}{2} - 1$ degrees of freedom.

A.10 A Penalized semi-integrated likelihood (PESEL)

The criteria proposed here [28] are inspired by BIC, which assumes the number of free parameters is independent of the number of observations, and clearly this is not always satisfied. The rationale is to integrate out some parameters from (2), either elements in Z so the model does not depend on m (i.e. $m \rightarrow \infty$) or to integrate out W so the model selection does not depend on n (i.e. $n \rightarrow \infty$). A total of four criteria are given under different asymptotics with respect to n and m while considering the first k eigenvalues are equal (homogeneous) or different (heterogeneous).

- Fixed m with $n \rightarrow \infty$: PESEL _{m}

- PESEL_{n,heter} is equivalent to the BIC approximation in [23].

$$\begin{aligned} \text{PESEL}_{n,\text{heter}}(q) &= \frac{-mn}{2} \log(2\pi) - \frac{n}{2} \sum_{j=1}^q \log(\hat{\lambda}_j) - \frac{n(m-q)}{2} \log(\hat{\zeta}_q^2) \\ &\quad - \frac{mn}{2} - \log(n) \frac{mq - \frac{q(q+1)}{2} + q + \mathbf{m} + 1}{2} \end{aligned} \quad (\text{A34})$$

- PESEL_{n,homo} assumes all PCs have the same variance (i.e. there is no dominant direction)

$$\begin{aligned} \text{PESEL}_{n,\text{homo}}(q) &= \frac{-mn}{2} \log(2\pi) - \frac{n\mathbf{q}}{2} \log\left(\frac{\sum_{j=1}^q \hat{\lambda}_j}{q}\right) - \frac{n(m-q)}{2} \log(\hat{\zeta}_q^2) \\ &\quad - \frac{mn}{2} - \log(n) \frac{mq - \frac{q(q+1)}{2} + q + \mathbf{1} + 1}{2} \end{aligned} \quad (\text{A35})$$

- Fixed n with $m \rightarrow \infty$: PESEL_m

- PESEL_{m,heter}

$$\begin{aligned} \text{PESEL}_{m,\text{heter}}(q) &= \frac{-mn}{2} \log(2\pi) - \frac{m}{2} \sum_{j=1}^q \log(\hat{\lambda}_j) - \frac{m(n-q)}{2} \log(\hat{\zeta}^2) \\ &\quad - \frac{mn}{2} - \log(m) \frac{nq - \frac{q(q+1)}{2} + q + \mathbf{n} + 1}{2} \end{aligned} \quad (\text{A36})$$

- PESEL_{m,homo}

$$\begin{aligned} \text{PESEL}_{m,\text{homo}}(q) &= \frac{-mn}{2} \log(2\pi) - \frac{m\mathbf{q}}{2} \log\left(\frac{\sum_{j=1}^q \hat{\lambda}_j}{q}\right) - \frac{m(n-q)}{2} \log(\hat{\zeta}^2) \\ &\quad - \frac{mn}{2} - \log(m) \frac{nq - \frac{q(q+1)}{2} + n + \mathbf{1} + 1}{2} \end{aligned} \quad (\text{A37})$$

A.11 A bias-corrected criterion (Passemier)

With the main asymptotic assumptions as follows:

$$n \rightarrow \infty \quad (\text{A38})$$

$$m \rightarrow \infty \quad (\text{A39})$$

$$c_n = \frac{m}{n-1} \rightarrow c > 0, \quad (\text{A40})$$

[19] proposed a plug-in estimator for ζ^2 using a bias correction that depends on q :

$$\hat{\zeta}_*^2 = \hat{\zeta}^2 + \frac{b(\hat{\zeta}^2)}{n-q} \hat{\zeta}^2 \sqrt{2c_n}. \quad (\text{A41})$$

where $b(\zeta^2) = \sqrt{c/2}\{q + \zeta^2 \sum_{i=1}^k (1/\lambda_i)\}$.

Without the correction, the noise variance is expected to have a downward bias as n increases relative to m . A consistent estimator for the true number of PCs (k) under $m \gg n$ is given, where it is assumed that $k \ll n$. The proposed criterion to select k requires a tuning parameter to be chosen and the default value is $b = 0.05$ for each q :

$$\hat{k}_{\text{Passemier}} = \arg \min_q \hat{\zeta}_{q*}^2 + q \hat{\zeta}_{k_o}^2 \frac{(c_n + 2\sqrt{c_n})(1 + m/n^{1+b})}{n}, \quad (\text{A42})$$

where k_o is the maximum number of PCs searched. In preliminary simulation results, I observed that b needs to be bigger than the default 0.05 to obtain the correct estimate in some cases, especially for the more difficult cases with smaller SNR. Thus, besides the default value of 0.05, I also included the 95% and 5% quantile values of $\{\hat{\lambda}_i\}_{i=1,\dots,n}$, and the best results from these choices are reported.

A.12 A profile likelihood-based criterion (ProfileL)

[8] proposed a simple profile likelihood-based criterion to detect the ‘elbow’ by separating the first q and last $n-q$ sample eigenvalues under the following models:

$$\hat{\lambda}_j \sim \mathcal{N}(\mu_1, \gamma); \quad j \leq q \tag{A43}$$

and

$$\hat{\lambda}_j \sim \mathcal{N}(\mu_2, \gamma); \quad j > q. \tag{A44}$$

The profile likelihood evaluates the evidence for a change-point by maximizing:

$$pL(q) = \sum_{j=1}^q \log \mathcal{N}(\hat{\lambda}_j | \mu_1(k), \gamma(q)) + \sum_{j=k+1}^n \log \mathcal{N}(\hat{\lambda}_j | \mu_2(k), \gamma(q)), \tag{A45}$$

where $\mu_1(q)$ and $\mu_2(q)$ are estimated by the mean sample eigenvalues in each partition separated by q , while $\gamma(q)$ is given by a pooled estimate using all sample eigenvalues.