Data Distillery: Effective Dimension Estimation via Penalized Probabilistic PCA

Wei Q. Deng and Radu V. Craiu

Department of Statistical Sciences, University of Toronto, Canada.

Abstract

The paper tackles the estimation of the effective dimension of a sample of dependent random vectors. The proposed method uses the principal component decomposition of sample covariance to establish a low-rank approximation that helps uncover the hidden structure. The number of principal components to be included in the decomposition is determined via a probabilistic principal components analysis embedded in a penalized profile likelihood criterion. The choice of the penalty parameter is guided by a data-driven procedure that is justified via analytical derivations and extensive finite sample simulations. Application of the proposed penalized approach is illustrated with three gene expression datasets in which the number of cancer subtypes is estimated from all expression measurements. The analyses point towards hidden structures in the data, e.g. additional subgroups, that could be of scientific interest.

Keywords: Cancer molecular subtype; Clustering; Effective dimension; Gene expression; Penalization; Principal component analysis; Probabilistic principal component analysis; Profile likelihood.

1 Introduction

Suppose we collect a sample $X = [X_1, \ldots, X_n] \in \mathbb{R}^{m \times n}$ from a population with hidden cluster structure. Of interest is the estimation of the number of clusters ($k$) in the sample $X$. An alternative formulation is to consider a parsimonious representation of the data

$$X^T = WZ + \epsilon,$$

where $W \in \mathbb{R}^{n \times k}$ is the loading matrix that characterizes dependence between the rows of $X^T$, which represent measurements for each sample unit, $Z = [Z_1, \ldots, Z_m]$ is a matrix...
whose columns are arbitrary latent vectors \(Z_j \in \mathbb{R}^k, j = 1, \ldots, m\), and \(\epsilon \in \mathbb{R}^{n \times m}\) represents noise in the data. When the dependence structure in the sample corresponds to \(k\) clusters, the above representation suggests that \(X_i\) is a linear combination of the \(k\) rows of \(Z\), and each row corresponds to one of the clusters. Hence, the problem we are interested in can be translated to finding the minimal \(1 \leq k < n\) such that the matrix \(W\) has rank \(k\), and the noise, \(X^T - WZ\), has independent and identically distributed rows. Henceforth, we refer to \(k\) as the effective dimension of the data because, intuitively, a clustering structure in the sample reduces the data dimension from \(n\) to \(k\). We will approach the estimation of \(k\) within the probabilistic principal component analysis framework.

Probabilistic principal component analysis, introduced in the seminal paper of Tipping and Bishop \[41\], casts the estimation of PCs as a likelihood optimization problem. Although it is possible to use a Bayesian approach to estimate the number of PCs \((6)\), the full Bayesian estimation using Markov Chain Monte Carlo (MCMC) can be computationally prohibitive for large datasets \([17]\) and approximations are needed, such as variational inference \([7, 29]\). Alternatively, Minka \[28\] implemented Laplace’s method to approximate the posterior likelihood and showed it to be often superior to cross-validation and variational inference with the benefit of fast computation. For high-dimensional data with a small number of observations, \([18]\) 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, \([37]\) approximated the Bayesian models for both \(x\) and \(x^T\), and thus proposed a unifying criterion that works well under divergence of either the number of features, \(m\), or sample size, \(n\).

In the context of isotropic factor analysis (FA), Bai and Ng \[2\] proposed to estimate the number of factors by finding some threshold to separate large and small eigenvalues of the data covariance matrix, but the approach depends on correct estimation of error variance. Using a different strategy, Passemier et al. \[31\] tackled the estimation of the noise variance, which led to a bias-corrected criterion for estimating \(k\). Alternatively, Gavish and Donoho \[12\] proposed to remove the underlying noise in the sample eigenvalues 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 dimension of the low-rank approximation.

The literature devoted to determining the number of principal components (PC) to retain is too extensive to be comprehensively described here, so we refer to two classic texts \([21, 20]\) on principal component analysis (PCA), an excellent review by Jackson \[19\], as well as the state-of-the-art probabilistic methods discussed in Sobczyk et al. \[37\].

For small datasets, cross validation is also frequently used \([27]\) with a general cross-validation (GCV) criterion proposed by \([22]\) that also works well with large datasets. Another class of methods rely on asymptotic tests such as the likelihood ratio test (LRT) for equality of eigenvalues \([3, 11, 23, 24, 35]\) and differ according to asymptotic conditions on data dimensions. Instead of an asymptotic test, Choi et al. \([8]\) 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.

The approach we propose here spearheads the use of penalized likelihood methods within the probabilistic formulation, where a penalty is introduce to reduce the number of principal components identified via profile likelihood maximization. The aim is to automatically select the effective dimension without manual tuning nor feature selection. Subsequent estimation of cluster structure can be achieved using either k-means or a mixture formulation that identifies the estimated effective dimension as the number of clusters or mixture components.

In Section 2, we will briefly review the key results of probabilistic principal component analysis, motivate the use of a penalty function, and propose a data-driven strategy to choose $k$ automatically. In Section 3, the method’s performance in finite samples is empirically explored via simulations performed under a number of scenarios, and is compared with alternative methods. In Section 4, we illustrate the use of our approach to estimate the number of cancer subtypes in three gene expression datasets. Finally, Section 5 discusses the advantages and limitations of our approach for estimating the number of clusters in gene expression and other potential applications.

2 Penalized PPCA

2.1 Notation

Consider a sample $X = (X_1, \ldots, X_n)$ with each $X_i \in \mathbb{R}^m$ representing the measurements for the $i$th sample unit. Denote the effective dimension for the $n$ sample units by $k^* \in \{1, \ldots, n\}$. Our aim is to determine $k^*$ using a large number of features, $m$.

Define the transpose $X^T = Y = (Y_1, \ldots, Y_m)$ where each $Y_j$ is the random vector of the $j$th feature measured for the sample. We assume that $\{Y_j : 1 \leq j \leq m\}$ are independent and identically distributed random vectors from a population with mean $\mu \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n \times n}$. Without loss of generality, we assume that $E(Y_j) = \mu = 0$.

We propose to solve this problem by finding a lower dimensional representation for $Y$ of the form

$$Y = WZ + \epsilon$$

where $W \in \mathbb{R}^{n \times k^*}$ is the loading matrix of rank $k^* < n$, $Z \in \mathbb{R}^{k^* \times m}$ with each column independently following the $N(0, I_{k^*})$ and $\epsilon \in \mathbb{R}^{n \times m}$ is the noise matrix with independent and identically distributed columns sampled from $N(0, \sigma^2 I_n)$. The representation in (1) induces the following partition of the covariance matrix

$$\Sigma = WW^T + \sigma^2 I_n.$$

In other words, the matrix $W$ completely accounts for the dependence structure between units in the sample and $k^*$ is the minimum rank of $W$ in the range $\{1, \ldots, n\}$ such that the representation (1) holds.
2.2 A Brief Review of PPCA

The probabilistic formulation of principal component analysis decomposes the data \( y = (y_1, \ldots, y_m) \) to a low-rank approximation and residual noise that follows a multivariate Gaussian distribution, \( \varepsilon_j \sim N(0, \sigma^2 I_n), j = 1, \ldots, m \). The log-likelihood function with respect to \( W \) and \( \sigma^2 \) is

\[
l(W, \sigma^2; y) = -\frac{m}{2}[n \log(2\pi) + \log|WW^T + \sigma^2 I| + \text{tr}\{(WW^T + \sigma^2 I)^{-1} S\}],
\]

where \( S \) is the sample covariance matrix of \( y \). The maximum likelihood estimators can be derived via an eigendecomposition of \( S \) and depend on the number of retained principal components (\( k \)):

\[
\hat{W}(k) = U(k)\hat{D}(k)V^T(k), \quad \hat{\sigma}^2(k) = \frac{\sum_{i=k+1}^n \hat{\lambda}_i}{n-k},
\]

where \( \hat{\lambda}_i \)'s are the sample eigenvalues of \( S \) and \( \sum_{i=1}^n \hat{\lambda}_i = n \). \( U(k) \) is an \( n \times k \) matrix with columns corresponding to the first \( k \) eigenvectors of \( S \), \( \hat{D}(k) \) is a diagonal matrix with \( k \) non-zero entries each given by \( \hat{d}_i(k) = \sqrt{\hat{\lambda}_i - \hat{\sigma}^2_i} \), and \( V(k) \in \mathbb{R}^{k \times k} \) is an arbitrary orthogonal matrix \[11\]. Since we consider the parameter of interest to be \( k \), and nuisance parameters to be \( W \) and \( \sigma^2 \), the profile likelihood \( l_p \) can be expressed as:

\[
l_p(k) = -\frac{m}{2}\{n \log(2\pi) + \sum_{i=1}^k \log \hat{\lambda}_i + (n - k) \log \hat{\sigma}(k)^2 + n\}
\]

The following result shows that the profile log-likelihood is monotonically non-decreasing in \( k \), suggesting that it can not be used as a criterion to select \( k \) in finite samples.

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

**Remark 1.** For \( k = n \) the effective dimension is equal to the sample size and \( \hat{\psi} \) is ill-defined due to the degeneracy of the noise distribution which becomes identically zero. In this case, \( W \) is of rank \( n \) and does not reveal any clustering structure.

**Remark 2.** One may be tempted to use a classical model selection criterion, e.g. Bayesian information criterion, to identify the effective dimension. However, the scale of the log-likelihood dominates the penalty when the number of features \( m \) is large. A similar phenomenon has been reported in factor analysis \[2\].

**Remark 3.** Assume that the true dimension of the probabilistic principal component model is \( k^* \). The population eigenvalues of the true covariance matrix \( W(k^*)W(k^*)^T + \sigma^2(k^*)I \) have the form

\[
\lambda_i = \begin{cases} 
d_i^2(k^*) + \sigma^2(k^*), & \text{if } i \leq k^*; \\ 
\sigma^2(k^*), & \text{otherwise},
\end{cases}
\]
and \( \lambda_{k^*+1} \) has multiplicity \( n - k^* \). When \( m \) is large, the maximum likelihood estimator \( \hat{\sigma}^2(k^*) \) is approximately equal to that of the \( k \)th population eigenvalue for any \( k > k^* \) and \( \hat{\sigma}^2(k) \overset{p}{\to} \sigma^2(k^*) \), and thus \( \hat{\sigma}^2(k) - \hat{\sigma}^2(k+1) \overset{p}{\to} 0 \) for all \( k > k^* \). However, for finite samples, \( \hat{\sigma}^2(k) \) is always different from \( \hat{\sigma}^2(k+1) \) which implies \( l_p(k) < l_p(k+1) \) for all \( 1 \leq k < n - 1 \).

### 2.3 Motivation for the Penalty

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 [40, 45, 5]. In the problem considered here, the model complexity is directly related to the effective dimension selected, while the fit corresponds to the amount of variance explained, i.e. \( 1 - \sigma^2(k) \).

We have investigated the use of the following penalty functions as they all account for both the amount of variance explained and the complexity of model with effective dimension \( k \):

1. \[
pen_1(W(k), \sigma^2(k)) = \text{rank}[W(k)] \text{tr}[W(k)W(k)^T] = nK[1 - \sigma^2(k)];
\]
2. \[
pen_2(W(k), \sigma^2(k)) = \text{rank}[W(k)] \log \sigma^2(k) = K \log \sigma^2(k);
\]
3. \[
pen_3(W(k), \sigma^2(k)) = \text{rank}[W(k)] \frac{1}{\sigma^2(k)} = \frac{k}{\sigma^2(k)}.
\]

We ultimately choose to use (4b) as it leads to simpler analytical derivations and intuitive heuristics. The natural guiding principle is to favor a parsimonious cluster structure by penalizing small values of \( \sigma^2(k) \) as well as large values of \( k \).

### 2.4 Construction of penalized probabilistic principal component analysis

The penalized log-likelihood has the form:

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

where the tuning parameter \( \delta > 0 \) controls the amount of penalty due to \( \text{pen}(W, \sigma^2) \). Notice that the number of features \( m \) is a scaling factor and does not affect the maximization other than through the computation of the sample covariance matrix \( S \).

Similarly to (2), the penalized maximum likelihood estimators, \( \hat{W} \) and \( \hat{\sigma}^2 \), are functions of \( k \). Since \( W \) is not present in the penalty function, the relationship between the penalized maximum likelihood estimator of \( W \) and \( \sigma^2 \) remains the same, but the resulting singular values are different due to the presence of a non-zero \( \delta \)-value. Given that the rank of \( W \) is \( k \), the penalized profile likelihood estimator can be expressed in terms of \( \delta \) and \( \hat{\sigma}^2(k) \):

\[
\hat{\sigma}^2(k) = \frac{\sum_{i=k+1}^n \hat{\lambda}_i}{n - k - \delta k} = \hat{\sigma}^2(k) \frac{n - k}{n - k - \delta k}.
\]

(5)
For a fixed $k$, $\hat{\sigma}^2(k)$ is unbounded as $n-k-\delta k$ can be very close to 0 or even negative for large $\delta$-values. Thus, the theoretical range of $\delta$ has an upper bound at $n/k - 1$ such that $\hat{\sigma}^2(k)$ is positive. This implies that the choice of $k$ poses a restriction of the range of $\delta$, and vice versa. Henceforth, we will work with the scaled tuning parameter $\tilde{\delta} = \delta/n \in [0, 1/k - 1/n)$.

Similarly, the penalized profile log-likelihood can be written as a function of $k$ given each $\tilde{\delta}$:

$$l_p(k; \tilde{\delta}) = l_p(k) - \frac{m}{2} \left[ n \left( 1 - \frac{k}{n} - \tilde{\delta} k \right) \log \frac{n-k}{n-k-nk\tilde{\delta}} - \tilde{\delta} nk \{ \log \hat{\sigma}^2(k) + 1 \} \right]$$

(6)

The introduction of penalty changes the monotonicity property of (5), and thus making it possible to select the correct dimension $k^*$ for appropriate choices of $\tilde{\delta}$-value.

Proposition 2. Consider a sample $Y \in \mathbb{R}^{n \times m}$ with each column following a multivariate Gaussian distribution $N(0, WW^T + \sigma^2 I)$. If the sample covariance matrix of $Y$ is positive semi-definite and $k^*$ is the rank of $W$, then there exist $\tilde{\delta}_0 \in (0, 1 - 1/n)$ such that $l_p(k; \tilde{\delta}_0)$ is maximized at $k^*$.

Proof. Please see the Appendix.

Remark 4. The proof of Proposition 2 implies that for $\tilde{\delta} = \tilde{\delta}_0$, the maximizer of (5), $k_0$, can be identified using the following sufficient conditions whenever $2 \leq k \leq n - 2$:

$$\begin{align*}
l_p(k; \tilde{\delta}_o) - l_p(k + 1; \tilde{\delta}_o) > 0; \\
l_p(k; \tilde{\delta}_o) - l_p(k - 1; \tilde{\delta}_o) > 0,
\end{align*}$$

(7)

Remark 5. Define $G(k) = (0, \{1/k - 1/n\}\{1 - \hat{\sigma}^2(k)\})$. For every $\tilde{\delta} \in G(k)$, we can consider the conditions in (7). For any $2 \leq k \leq n - 2$, define

$$\Delta_k = \{ \tilde{\delta} \in G(k); \text{conditions (7) are satisfied} \}.$$ 

It is clear that $\Delta_k$ is an open interval for each $k$ as the penalized likelihood function in (5) is a continuous function of $\delta \in G(k)$ for any fixed $k$. Following Lemma 3, we have $\Delta_k = (a_k, b_k) \subseteq G(k)$.

Remark 6. Since $a_k$ and $b_k$ are not analytically available, we obtained conservative upper and lower bounds for $\Delta_k$ using $u_a(k)$ and $u_b(k)$ such that $(u_a(k), u_b(k)) \subseteq \Delta_k$. The proof of Proposition 2 also demonstrates that $u_b(k^*)/u_a(k^*) > 1$ so that $(u_a(k^*), u_b(k^*)) \neq \emptyset$.

Remark 7. By definition, for $k \neq k'$, $\Delta_k$ and $\Delta_{k'}$ are non-overlapping sets. Therefore, the realized range for $\tilde{\delta}$ is the union of all sets $\bigcup_{i=1}^{n} \Delta_i \subseteq [0, 1 - 1/n)$. But because of the restriction embedded in (5) and (7), we must have $\tilde{\delta} \in G(k)$ for each examined value of $k$. Consequently, the restriction imposes a relationship whereby $k$ is non-increasing in $a_k$ (or $u_a(k)$) and $b_k$ (or $u_b(k)$). For example, when $k = n - 1$, we must have $a_{n-1} = 0 < b_{n-1} < \{1/(n-1) - 1/n\}\{1 - \hat{\sigma}^2(n-1)\}$, while for $k = 1$, we must have $b_2 < a_1 < b_1 < \{1 - 1/n\}\{1 - \hat{\sigma}^2(1)\}$. 


The penalized approach requires proper calibration of \( \tilde{\delta} \) so that \( k^* \) is the maximizer of \( l_p(k; \tilde{\delta}) \). The selection of appropriate tuning parameter values in other well-known problems, such as the selection of shrinkage tuning parameter in lasso [40, 44], uses either a model selection criterion, e.g. Bayesian information criterion, or cross-validation. However, the use of a cross-validation approach is based on optimizing a certain objective function which can be analytically expressed, a task that is difficult when of interest is determining the effective dimension. Our attempts at using an off-the-shelf information criterion have produced only modest results for finite samples.

So far, the best performance is obtained through a “voting” strategy in which each value of \( \tilde{\delta} \) will lead to a vote for a particular estimate of \( k \). Since the same estimated \( k \) can result from multiple \( \tilde{\delta} \)-values, we ultimately select the estimate that has been obtained most often. We provide analytical and simulation support for this “voting” procedure in the next subsection.

2.5 A voting strategy to estimate \( k^* \)

The voting strategy stems from the proof of Proposition 2, where we have shown: 1) there exists \( \tilde{\delta} \in \Delta_{k^*} \) such that (11) 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

\[
\lim_{m \to \infty} \frac{u_b(k^*)}{u_a(k^*)} = \infty, \\
\lim_{m \to \infty} |u_b(k) - u_a(k)| \to 0, \text{ for } k > k^*, \\
\lim_{m \to \infty} \frac{u_b(k^*)}{u_a(k^*)} < \infty, \text{ for } k < k^*.
\]

Therefore, the ratio of the boundaries for non-overlapping sets \((u_a(k), u_b(k))\) is asymptotically the largest for \( k = k^* \) and this suggests a majority-voting strategy for estimating \( k^* \). We establish a grid of tuning parameter values \( \tilde{\delta}_1, \ldots, \tilde{\delta}_T \) on \( \cup_{i=1}^{n-1} \Delta_i \) where \( T \) is a user-specified integer, \( \tilde{\delta}_1 = u_a(n-1) \) and \( \tilde{\delta}_T = u_b(1) \). Each \( \tilde{\delta}_i \) will result in (11) having support on a range of possible values for \( k \), say \( 1 \leq k \leq k_{\text{max}} \), and some choice of \( k \) will be the maximizer of \( l_p(k; \tilde{\delta}_i) \). Then, the number of times that each possible \( k \) maximizes the penalized profile log-likelihood is counted and the one with the highest vote count is selected and denoted by \( \tilde{k} \).

Here we detail the construction of the search grid characterized by its range and the distance between adjacent grid values. Since the exact relationship between \( k \) and \( a_k, b_k \) is not analytically available, we rely on conservative bounds obtained via Taylor series approximations. The largest non-trivial choice for \( k \) is \( n-1 \), as when \( k = n \) the log-likelihood is not defined. The grid of \( \tilde{\delta} \)-values is constructed using a sequence of \( T \) equidistant points on log scale from \( \log u_a(n-1) \) to \( \log u_b(1) \). The value of \( T \) needs to be large enough to identify a mode and in our simulations we used \( T = 50 \times n \).
To demonstrate the validity of this voting procedure empirically, we illustrate in Figure 1 the behavior of $l_p(k, \tilde{\delta}) - l_p(k - 1, \tilde{\delta})$ and $l_p(k, \tilde{\delta}) - l_p(k + 1, \tilde{\delta})$ as a function of $\log \delta$. For simplicity, we set $k^* = 10$, while fixing the number of subjects to cluster at $n = 100$. The covariance structure was generated according to a singular value decomposition using simulated orthogonal matrices and specified squared singular values equal to $(20, 15, 9, 7, 6, 5, 4, 2.5, 1, 0.5)$ or $(15, 12, 11, 8, 7, 6, 5, 3, 2, 1)$, both corresponding to $\sigma^2(k^*) = 0.3$.

Since $k = \tilde{k}$ maximizes the penalized profile log-likelihood for a particular $\tilde{\delta}$ value only when i) $l_p(k, \tilde{\delta}) - l_p(k - 1, \tilde{\delta}) > 0$, and ii) $l_p(k, \tilde{\delta}) - l_p(k + 1, \tilde{\delta}) > 0$ simultaneously, we can visualize this on a panel of plots where the differences in i) and ii) are plotted as functions of $\log \delta$ for a few possible $k$’s including $k^*$. With little or no penalty, i.e., when $\log(\delta) < 10^{-5}$, $l_p(k, \tilde{\delta}) - l_p(k + 1, \tilde{\delta})$ can be shown to be negative and $l_p(k, \tilde{\delta}) - l_p(k - 1, \tilde{\delta})$ positive for all $k$. As $\log(\delta)$ increases, the difference in penalized profile log-likelihood between $k$ and $k + 1$ increases and becomes positive while the difference between $k$ and $k - 1$ eventually drops below zero. It is easy to confirm visually that the approximations of the sets $\Delta_k$ are non-overlapping and $k$ is indeed monotonically decreasing in $\delta$. Notice that the log difference in $\delta$ is the largest for $k = k^* = 10$ as compared to $k = 9$ or $k = 11$, suggesting that if we took a grid-set of values that are equally spaced on $\log$ scale, the majority would fall in $(\tilde{a}_{k^*}, \tilde{b}_{k^*})$.

Consequently, with the search grid constructed as proposed, the procedure would estimate the effective dimension by a majority vote, which is expected given the relationship between the log distance $\log (\tilde{b}_k) - \log (\tilde{a}_k)$ and the number of votes for that particular $k$. We simulated two additional scenarios characterized by the singular values $(6, 5, 4, 3, 3, 2.5, 2, 2, 1.5, 1)$ and $(6, 5, 4, 3.5, 3.5, 3.25, 2, 1.5, 0.75, 0.5)$ such that $\sigma^2(k^*) = 0.7$. The scatterplots of the number of votes and $\log (\tilde{b}_k) - \log (\tilde{a}_k)$ shown in Figure 2 confirm the effective dimension $k^* = 10$ can be recovered by a majority vote.

The readers interested in implementing this procedure on their data can use the R package available at [https://github.com/WeiAkaneDeng/SPAC2](https://github.com/WeiAkaneDeng/SPAC2)

### 3 Simulation Studies

#### 3.1 Comparison with Alternative Methods

To mimic realistic scenarios, we assumed the first $k^*$ squared singular values decay at an exponential rate with their values determined by varying the two parameters $\sigma^2(k^*) \in \{0.2, 0.5, 0.9\}$ and $d^2_k = \{0.2, 0.5\}$. The combinations yield different degrees of difficulty for recovering the effective dimension. The residual noise was assumed to have a multivariate Gaussian distribution with mean vector zero and $\Sigma = \sigma^2(k^*) I_n$. The number of subjects to cluster ($n$) is often directly associated with the difficulty of recovering the effective dimension. Thus, we examined $n = 100$, but varying $k^* \in \{5, 10, 20\}$. Lastly, the number of features or observations was set to $m = 10,000$. 
Figure 1: The differences $l_p(k, \tilde{\delta}) - l_p(k - 1, \tilde{\delta})$ (solid) and $l_p(k, \tilde{\delta}) - l_p(k + 1, \tilde{\delta})$ (dashes) for each choice of $k$ around $k^* = 10$ is illustrated under two scenarios ($d^2(k^*) = 0.5$ and $d^2(k^*) = 1$). When both differences are positive, the numerically approximated $(\hat{a}_k, \hat{b}_k)$ for $k = 9, 10, \text{and } 11$ are formed and indicated by the vertical lines, which capture the range of $\log(\tilde{\delta})$ for which $k = k^* - 1$, $k = k^*$, or $k = k^* + 1$ maximizes the penalized profile log-likelihood.
Figure 2: Scatterplots of number of votes and $\log(\tilde{b}_k) - \log(\tilde{a}_k)$ with true $k^* = 10$. With the grid of $\tilde{\delta}$ values constructed as proposed, we approximated $a_k$ and $b_k$ for each potential choice of $k$. Each point represents the log difference in $\tilde{a}_k$ and $\tilde{b}_k$ plotted against the number of votes for that particular $k$. 
3.2 Comparison with Alternative Methods

We compared the performance of our approach with alternative methods, for which details are provided in the Appendix. We included BIC, the Bayesian model selection approach proposed in [28] for PPCA, denoted Minka BIC, the best performer from a class of Bayesian criteria using PEnalized Semi-integrated Likelihood (PESEL; [37]), and the bias-corrected criterion for estimating $k$ (PASSEMIER; [31]). We have also included in the comparison a couple of empirical approaches designed to detect an “elbow” in the distribution of the sample eigenvalues: the difference between log cumulative mean of the sample eigenvalues and the mean of the cumulative log sample eigenvalues (Cumlog), defined by $\hat{K}_{\text{Cumlog}} = \arg\min_k \log \sum_{i=1}^k \hat{\lambda}_i - \frac{1}{k} \sum_{i=1}^k \log \hat{\lambda}_i$ and variance of sample eigenvalues (VarD), defined by $\hat{K}_{\text{VarD}} = \arg\max_k \sum_{i=1}^k \hat{\lambda}_i^2 - \left( \frac{\sum_{i=1}^k \hat{\lambda}_i}{k} \right)^2$.

Some of the methods we do not consider in our comparison are Bishop’s ARD [6] and related methods followed it [10, 33] as they have been shown to be outperformed by methods using the Laplace approximation [28]. We have also not included Bayesian methods that rely on MCMC sampling [17] as they become computationally prohibitive when $m$ is large, e.g., $m > 1000$. For the same reason, we excluded cross-validation, but included the general cross-validation (GCV) criterion of [22] that has better scalability properties.

The performance was evaluated by the mean (Figure 3) and proportion (Figure 4) of correct estimates over 100 replicates. Even though our method was not the “best” in every scenario, the overall performance was strong, capturing the effective dimension for most scenarios when either $\sigma^2(k^*)$ or $d^2(k^*)$ was large. As expected, our penalized criterion was more likely to slightly underestimate than overestimate across different scenarios. Notably, all the likelihood based methods performed well when $\sigma^2(k^*)$ is large. In particular, BIC approximated by Laplace’s method outperformed BIC in all scenarios, while BIC tended to overestimate and chose the maximum possible $k$ for smaller $k^*$ when $\sigma^2(k^*) = 0.2$. In addition, the penalized semi-integrated likelihood criterion [37] maintained consistent performance in all scenarios and showed a clear preference to small $\sigma^2(k^*)$ and large $d^2(k^*)$, i.e. a scenario characterized by a high signal-to-noise ratio (Figure 3).

3.3 Scalability

One of the data attributes encountered in real world applications is a larger number of features, $m$, than number of subjects to cluster, $n$. To capture this characteristic, we fixed $d^2(k^*) = \{0.2, 0.5\}$, $\sigma^2(k^*) = \{0.5, 0.9\}$, $n = 100$ and $k^* = \{10, 20\}$. The relationship between $m$ and the estimated $k$ over 10 replicates is shown in Figure 5 and 6. As $m$ was increased, all methods except BIC approach the correct value. Notice that when $d^2(k^*)$ and $\sigma^2(k^*)$ are both small, our method tended to under-estimation of $k^*$ as shown by the solid line representing the median of our estimates over 10 replicates. In contrast, although the penalized semi-integrated likelihood criterion and Laplace approximated BIC have a
<table>
<thead>
<tr>
<th>Methods</th>
<th>$d_K^2 = 0.2$</th>
<th>$d_K^2 = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MinkaBIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PESEL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPCA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: Distribution of the estimated $k$ on data with sample size $n = 100$ over 100 replicates against each method. The true $k^* = \{5, 10\}$ values are indicated by horizontal lines.
Figure 4: The shaded bars represent the proportion of correctly estimated $k$ for each true $k^* = \{5, 10\}$ with $n = 100$ over 100 replicates against each method.
similar outlook as our penalized profile likelihood criterion, their estimates approach the correct value slower than our method in most of the scenarios considered. The results also showed that BIC had reasonable performance for moderate sizes of $m$, but, as the number of features increases, their estimates became unreliable for $m > 10,000$ whenever $\sigma^2(k^*)$ is small. This reaffirms our observation that for large values of $m$, the leading difference in BIC between choices of $k$, is of order $O(m)$ and dominates the difference between respective profile log-likelihoods. In other words, for large $m$, BIC behaves just like the profile log-likelihood and always prefers the full model with $k = n - 1$.

### 4 Application to sample-based Gene Expression Clustering

#### 4.1 Sample-based gene clustering

Gene expression is an intermediary measurement of how genetic information is translated to the observed phenotype through production of mRNA and protein. As a consequence of the
Figure 6: The estimated $k$'s over 10 replicates by each method were plotted against the number of observations ($m$) and are represented by different shapes as shown in the legend. The dotted line represents the true dimension $k^* = 10$, while the solid line represents the median of estimates using penalized profile log-likelihood criterion over 10 replicates.
natural fluctuation in biochemical reactions, the measured gene expression can be seen as a single instance of a complex stochastic process and is often quite noisy [32]. These unique characteristics of gene expression data make it possible to cluster either genes or samples, commonly referred to as gene-based clustering and sample-based clustering. For example, gene-based clustering aims to find genes that might share a common regulator [4] or co-express [39], or to find a group of genes that are enriched for specific functions in the same pathway [36]. On the other hand, sample-based clustering relies on the correlation over a large number of measured expression levels between any pairs of individuals. Correlations could possibly lead to similarities between individuals in terms of cell compositions and reveal subtypes of disease.

Here we demonstrate the utility of our procedure on three gene expression datasets and discuss its potential use as a basis for clustering algorithms. It has been shown that the PCs are the continuous solution to the assigned membership of K-means clustering [9], implying an utility that goes beyond noise reduction. For convenience, we applied both K-means and a mixture model to cluster using the estimated number of PCs.

### 4.2 Data Analyses

Consider $m$ expression measurements from $n$ individuals $X = [X_1, \ldots, X_n] \in \mathbb{R}^{m \times n}$, where $m$, the number of features, is typically in the tens of thousands and $n$, the number of individuals, in the hundreds. A summary of results on three gene expression data applications is shown in Table 1.

**Example 1.** Expression data from breast cancer subtypes: The expression mea-
Table 2: **Confusion matrix between known breast cancer subtypes and k-means clustering results with $\tilde{k} = 9$.** Clustering results of 178 breast cancer patients over expression measured on 29,809 gene probes using the penalized profile log-likelihood criterion estimate of $\tilde{k} = 9$.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Triple negative</th>
<th>HER2</th>
<th>Luminal A</th>
<th>Luminal B</th>
<th>Normal tissue</th>
<th>Normal cell line</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>20</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>14</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>28</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
</tr>
</tbody>
</table>

measurements were collected on tissues from 153 patients with primary invasive breast cancer. The four cancer subtypes defined clinically are (number in parenthesis indicates patients with that type of breast cancer) Triple-negative (TN; 55), human epidermal growth factor receptor 2 (HER2; 39), Luminal A (29) and Luminal B tumour (30), and the two types of controls are in the form of 11 normal tissue samples and 14 normal cell lines. A total of 29,874 genes probes were included as features. Our penalized approach estimated $\tilde{k} = 9$, with no other competing method giving any sensible estimates (Table 1).

It is expected that some of the subtypes would be clustered together for any choice of the clustering algorithm as a result of the original molecular classification for these subtypes of breast cancer. However, the normal tissue and the cell line derived samples are distinctive enough that they would always be assigned their own clusters. K-means clustering algorithm reveals that two of the identified clusters belong to TN, two belong to HER2, two for Luminal A and one for Luminal B (Table 2). Since triple-negative tumours are known to be a more heterogeneous group that contains mostly Basal type and also encompass all other subtypes, it is unsurprising to observe these results. Meanwhile, Luminal A and Luminal B have more homogenous expression signature. In particular, TN and HER2 are hormone-receptor negative, while Luminal A and Luminal B are hormone-receptor positive, and these two classes should have very little overlap. In other words, clustering cancer subtypes using expression from all genome-wide gene probes did not blur the subgroup structure, but rather introduced clarity to specific cases that could be further characterized.

**Example 2. NCI60 cancer cell line data:** This dataset originally contained gene
expression of nine types of cancer cell line \cite{34}, and has been recently profiled using array technology for $n = 132$ individuals on $m = 41,000$ gene probes \cite{26}. Our penalized approach estimated $\hat{k} = 16$. Though our estimate is larger than the number of cancer cell lines, it is within a reasonable range as most of the cancers have known subtypes.

**Example 3. Human Ovarian Gene Expression Data**: This experiment profiled expression levels of 285 ovarian cancer tumours at various stages with the objective to identify molecular subtypes \cite{43}. The study discovered six subtypes defined by molecular pattern in a subset of $m = 8,732$ probe sets. It is common in the literature to use only a subset of biologically relevant genes to cluster the subjects so that the clusters would be better separated and results of hierarchical clustering would be stable \cite{38}. It is not clear how many genes should be included since the relevant pathways are usually determined by experts and two separate sets of genes are usually used to ensure the robustness of the clusters. On $m = 54,621$ gene probes, our method estimated $\hat{k} = 21$, while none of the other methods provided a reasonable estimate (Table 1).

5 Discussion

In this article, we propose to estimate the number of principal components via a penalized profile log-likelihood derived from the probabilistic principal components analysis formulation. The main aim is to simplify the data structure and provide a lower dimensional representation. Specifically, the focus of the method proposed here is to recover the effective dimension and thus to inform the cluster structure in the observed noisy data.

Under simulated scenarios, there was no universally best method, but PPPCA had the best performance when averaging over scenarios. This is advantageous in applications when only one opportunity is given to estimate the number of principal components. The successful recovery of the latter facilitates subsequent use of clustering algorithms that require a value for the number of clustered.

The main difficulty of clustering is the unsupervised nature, and mapping of complex data structure to independent latent space is usually not directly interpretable as each dimension might not necessarily correspond to, for example, the type of cancer, but possibly a common basis of several cancer types. For example, several clinically distinct cancer types might share similar physiological mechanism that influences the expression profile of the same subset of gene. The advantage of our method is that all features could be incorporated without assuming a homogenous feature space, so gene expression, f-MRI and environmental data could all be combined to improve estimation of the number of clusters. The large dimension of the feature space is also 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. if the number of clusters is local to a subset of the features. In addition, this type of data almost always present certain violations to model assumptions, such as highly correlated features,
non-normality, or a combination of the two. But as we have shown with simulated data, PPPCA is reasonably robust to dependent observations and moderately fat-tailed data.

In addition, we would like to point out that our method might not be suited to any application that hinges on minimizing reconstruction error, such as digit recognition or image processing. In addition, since principal components are essentially a low-rank approximation of the data by linear projection into subspace, we are restricted to signals in the linear space. Nevertheless, this is a reasonable assumption for gene expression data. Due to the nature of gene expression levels, both row and column correlations are present. Our method does not explicitly account for correlation in the columns (genes), but as we demonstrated in simulation with autocorrelated features that PPPCA was reasonably robust as long as there was a good separation between signal and noise, as well as a large number of observations \( m \). This has been observed by [30] and others [14, 16, 15, 25] for clustering of high-dimensional data that correlation does not strongly impact the results as long as the cluster centres are sufficiently separated.

Finally, we should mention that sparse PCA [45], where certain loadings within the top PCs are shrunk to zero to produce sparse loadings, is different from our approach since in our case, only loadings for the last \( n-k \) PCs are shrunk to zero. The penalty function in PPPCA effectively shrinks the sample eigenvalues of the last \( n-k \) PCs relative to the amount of noise variance to select the first \( k \) PCs sufficient to account for the data structure, and not to select the features that are important for the first \( k \) PCs. However, it would be interesting to combine the two methods to induce sparsity from both sides such that \( \hat{k} \) can be estimated based on important features with non-zero loadings.

Acknowledgements

The authors thank Andrey Feuerverger for helpful suggestions, Dehan Kong, Lei Sun, Qiang Sun, Stanislav Volgushev, and Fang Yao, for a critical reading of the original version of the paper. Wei Q. Deng is supported by Alexander Graham Bell Canada Graduate Scholarships-Doctoral Program from the Natural Sciences and Engineering Research Council of Canada.

Appendix

Alternative methods

**Information Criteria:** Let \( p(k) = nk + 1 - k(k-1)/2 \) denote the number of free parameters in the probabilistic model, we can select the rank by minimizing the following commonly used model selection criteria: \( AIC(k) = -2l_p(k) + 2p(k) \) and \( BIC(k) = -2l_p(k) + \log (m)p(k) \).

**Hypothesis Testings on the Equality of the Last \( n-k \) Eigenvalues:** The null hypothesis of interest is \( H_0 : \lambda_j = \cdots = \lambda_n \) against the alternative hypothesis that at
least one is not equal to the remaining eigenvalues. The test statistics given by [23]:

\[
\chi^2 = c \left\{ (n - j) \log \left( \frac{\sum_{i=j}^{n} \hat{\lambda}_i}{n-j} \right) - \sum_{i=j}^{n} \log \hat{\lambda}_i \right\},
\]

where \( c = (n - j) - \frac{2(n-j)+1}{6} + \left( \frac{\sum_{i=j}^{n} \hat{\lambda}_i}{n-j} \right)^2 \sum_{i=1}^{j} (\hat{\lambda}_j - \frac{\sum_{i=j}^{n} \hat{\lambda}_i}{n-j})^2 \) with \( \frac{(n-j)(n-j+1)}{2} - 1 \) degrees of freedom.

**Data availability**

The three gene expression datasets can be obtained publicly via the Gene Expression Omnibus from National Center for Biotechnology Information or ArrayExpress from the European Bioinformatics Institute (http://www.ebi.ac.uk). Pre-processed datasets were used in the present analyses. We have transformed measurements at each probe to a standard normal distribution but no further cleaning nor filtering was performed. Here each probe measures the levels of expression of a specific DNA sequence, and can be regarded as a feature or observation.

**Proof of Proposition 1**

*Proof.* Following [41] we have:

\[
\hat{\lambda}_n = \hat{\sigma}^2(n-1) \\
\hat{\lambda}_{n-1} = 2\hat{\sigma}^2(n-2) - \hat{\lambda}_n = 2\hat{\sigma}^2(n-2) - \hat{\sigma}^2(n-1) \\
\vdots \\
\hat{\lambda}_k = (n-k+1)\hat{\sigma}^2(k-1) - (n-k)\hat{\sigma}^2(k)
\]

Notice that the sample eigenvalues are decreasing \( \hat{\lambda}_1 > \cdots > \hat{\lambda}_n > 0 \) thus implying \( \hat{\sigma}^2(1) > \cdots > \hat{\sigma}^2(n-1) \).

Therefore,

\[
\hat{\lambda}_k = (n-k+1)\hat{\sigma}^2(k-1) - (n-k)\hat{\sigma}^2(k) > \hat{\sigma}^2(k-1).
\]

For \( k \in \{2, \ldots, n-1\} \), based on the inequality \( \frac{x}{1+x} < \log(1+x) < x \) for any \( x > -1 \) we obtain

\[
l_p(k) - l_p(k-1) = -\frac{m}{2} \left[ \log \hat{\lambda}_k + (n-k) \log(\hat{\sigma}^2(k)) - (n-k+1) \log(\hat{\sigma}^2(k-1)) \right]
\]

\[
> -\frac{m}{2} \left\{ (n-k)(1 - \frac{\hat{\sigma}^2_k}{\hat{\sigma}^2_{k-1}}) + (n-k) \left( \frac{\hat{\sigma}^2_k}{\hat{\sigma}^2_{k-1}} - 1 \right) \right\} = 0
\]
Proof of Proposition 2

The proof is based on the following lemmas whose proofs are included in the web appendix of the paper.

**Lemma 1.** Consider a collection of samples $x \in \mathbb{R}^{m \times n}$ with each row following a multivariate Gaussian distribution $N(0, WW^T + \sigma^2 I)$. Suppose $k^*$ is the rank of $W$ and further, the sample covariance matrix of $x^T$ is positive semi-definite. Then, the penalized profile log-likelihood at each fixed $k \in \{1, \ldots, n-1\}$ is a smooth function of $\delta$ on the interval $(0, 1/k - 1/n)$ and is monotonically decreasing on $(0, (1/k - 1/n)\{1 - \tilde{\sigma}^2(k)\})$, where $\tilde{\sigma}^2(k) = (\sum_{i=k+1}^n \hat{\lambda}_i)/(n - k)$.

**Lemma 2.** Suppose we consider $\tilde{\delta} \in G(k) = (0, \{1/(k + 1) - 1/n\}\{1 - \tilde{\sigma}^2(k - 1)\})$, then for any fixed $k \in \{2, 3, \ldots, n - 2\}$, $l_p(k; \tilde{\delta}) - l_p(k + 1; \tilde{\delta})$ is a monotonically increasing and concave function of $\tilde{\delta}$ and $l_p(k; \tilde{\delta}) - l_p(k - 1; \tilde{\delta})$ is a monotonically decreasing and convex function of $\tilde{\delta}$.

**Lemma 3.** For every $k \in \{1, \ldots, n-1\}$, there exists $\tilde{\delta}_o \in G(k)$ such that $k = \text{argmax}_{k'} l_p(k'; \tilde{\delta}_o)$. And if $k \in \{2, 3, \ldots, n - 1\}$, $k = \text{argmax}_{k'} l_p(k'; \tilde{\delta}_o)$ if and only if

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

For $k = 1$, we have $k = \text{argmax}_{k'} l_p(k'; \tilde{\delta}_o)$, if and only if $l_p(k; \tilde{\delta}_o) - l_p(k + 1; \tilde{\delta}_o) > 0$; and for $k = n - 1$, $k = \text{argmax}_{k'} l_p(k'; \tilde{\delta}_o)$ if and only if $l_p(k; \tilde{\delta}_o) - l_p(k - 1; \tilde{\delta}_o) > 0$.

**Lemma 4.** Consider

$$
\Delta_k = \left\{ \tilde{\delta} \in G(k); \text{ conditions } \text{ are satisfied} \right\}.
$$

Then $\Delta_k$ can be approximated by $(u_\omega(k), u_b(k)) \subset \Delta_k \subset G(k)$, where $u_\omega(k)$ denote an upper bound for $a_k$, and $u_b(k)$ a lower bound for $b_k$, such that $b_k/a_k > u_b(k)/u_\omega(k)$.

**Lemma 5.** When $k = k^*$ is at the true rank, as $m \to \infty$, we have $u_b(k^*)/u_\omega(k^*) \to \infty$ and $|u_b(k) - u_\omega(k)| \to 0$ for $k > k^*$.

of Proposition 2. We prove the result for $\tilde{\delta}$ in the range

$$
G(k) = (0, \{1/(k + 1) - 1/n\}\{1 - \tilde{\sigma}^2(k)\}),
$$

where $G(k) \subset (0, 1 - 1/n)$.

**A:** Case $2 \leq k \leq n - 2$
Following Lemma 1 and Lemma 2 when \( \tilde{\delta} \in G(k) \), the difference \( l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta}) \) is a smooth and concave function of \( \tilde{\delta} \) and is increasing. Similarly, \( l_p(k; \tilde{\delta}) - l_p(k-1; \tilde{\delta}) \) is a smooth and convex function of \( \tilde{\delta} \) and is decreasing. This enables us to define an interval

\[
\Delta_k = \left\{ \tilde{\delta} \in G(k); \text{ conditions (9) are satisfied} \right\} = (a_k, b_k)
\]

for any \( k \in \{2, \ldots, n-2\} \). Lemma 3 implies that on this interval, there exists \( \tilde{\delta}_o \in G(k) \) such that \( k = \arg\max_{k'} l_p(k'; \tilde{\delta}_o) \), including \( k = k^* \).

Following from Lemma 3 and 4 since \( a_k^* < u(a_k^*) < u(b_k^*) \) or \( \Delta_k^* \neq \emptyset \), there must exist \( \tilde{\delta}_o \in \Delta_k^* \), such that \( l_p(k^*; \tilde{\delta}_o) - l_p(k^* + 1; \tilde{\delta}_o) > 0 \) and \( l_p(k^*; \tilde{\delta}_o) - l_p(k^* - 1; \tilde{\delta}_o) > 0 \). And finally, as \( m \to \infty \), we have \( \frac{u(a^*_k)}{u(a^*_k)} \to \infty \) and \( |u_b(k) - u_o(k)| \to 0 \) for \( k > k^* \). Thus, there must exist some

\[
\tilde{\delta}_o \in (u(a_k^*), u(b_k^*)) \subset (a_k^*, b_k^*) \subset G(k^*) \subset (0, 1 - 1/n),
\]

such that \( k^* = \arg\max_{k'} l_p(k'; \tilde{\delta}_o) \).

\[\text{B: Case } k = 1\]

Following Lemma 1 and Lemma 2 when \( \tilde{\delta} \in G(1) \), the difference \( l_p(1; \tilde{\delta}) - l_p(2; \tilde{\delta}) \) is a smooth and concave function of \( \tilde{\delta} \) and is increasing. This enables us to define an interval

\[
\Delta_k = \left\{ \tilde{\delta} \in G(k); \ l_p(1; \tilde{\delta}_o) - l_p(2; \tilde{\delta}_o) > 0 \right\} = (a_1, 1 - 1/n).
\]

Lemma 3 implies that on this interval, there exists \( \tilde{\delta}_o \in G(1) \) such that \( \arg\max_{k'} l_p(k'; \tilde{\delta}_o) = 1 \).

\[\text{C: Case } k = n - 1\]

Following Lemma 1 and Lemma 2 when \( \tilde{\delta} \in G(n - 1) \), the difference \( l_p(n - 1; \tilde{\delta}) - l_p(n - 2; \tilde{\delta}) \) is a smooth and convex function of \( \tilde{\delta} \) and is decreasing. This enables us to define an interval

\[
\Delta_{n-1} = \left\{ \tilde{\delta} \in G(n - 1); \ l_p(n - 1; \tilde{\delta}_o) - l_p(n - 2; \tilde{\delta}_o) > 0 \right\} = (0, b_{n-1}).
\]

Lemma 3 implies that on this interval, there exists \( \tilde{\delta}_o \in G(n-1) \) such that \( \arg\max_{k'} l_p(k'; \tilde{\delta}_o) = n - 1 \).

\[\square\]

**Proof of Lemmas**

**Proof of Lemma 1**

*Proof.* The penalized profile log-likelihood is a smooth function of the scaled tuning parameter \( \tilde{\delta} \) for each \( k \), if it is differentiable with respect to \( \tilde{\delta} \) and all derivatives of \( l_p(k, \tilde{\delta}) \) exist on \( \tilde{\delta} \in (0, 1/k - 1/n) \).
The penalized profile log-likelihood at \( k = k' \) is
\[
l_p(k'; \tilde{\delta}) = l_p(k') - \frac{m}{2} \left\{ (n - k') \log \left( 1 - \frac{k'}{n} \right) - nk' \tilde{\delta} \left[ 1 + \log \hat{\sigma}^2(k') + \log \left( 1 - \frac{k'}{n} \right) \right] + \tilde{\delta} nk' \log \left( 1 - \frac{k'}{n} - k\tilde{\delta} \right) - (n - k') \log \left( 1 - \frac{k'}{n} - k'\tilde{\delta} \right) \right\}
\] (11)
and the first order derivative with respect to \( \tilde{\delta} \) is given by:
\[
\frac{\partial l_p}{\partial \tilde{\delta}} = -\frac{mnk'}{2} \log \left\{ \frac{n - k' - nk'\tilde{\delta}}{(n - k')\hat{\sigma}^2(k')} \right\},
\] (12)
the second order derivative with respect to \( \tilde{\delta} \) is:
\[
\frac{\partial^2 l_p}{\partial \tilde{\delta}^2} = \frac{mnk'}{2} \frac{nk'}{n - k' - nk'\tilde{\delta}}.
\]
Consequently, the \( t \)th order derivative is then
\[
\frac{\partial^t l_p}{\partial \tilde{\delta}^t} = \frac{mnk'}{2} (-1)^t \frac{(t - 2)!t!}{(n - k' - nk'\tilde{\delta})^t-1}.
\] (13)
Since (13) is a rational function of \( \tilde{\delta} \) on \((0, 1/k' - 1/n)\), all derivatives of \( l_p(k', \tilde{\delta}) \) exist, and thus \( l_p(k, \tilde{\delta}) \) is a smooth function of \( \tilde{\delta} \) on the same range.

Observe the first order derivative (12) is negative for any \( k \in \{1, 2, \ldots, n - 1\} \) when:
\[
\log \left( 1 - \frac{k}{n} - k\tilde{\delta} \right) - \log \left( 1 - \frac{k}{n} \right) - \log \hat{\sigma}^2(k) > 0
\]
\[
\tilde{\delta} < \left( \frac{1}{k} - \frac{1}{n} \right) \{1 - \hat{\sigma}^2(k)\}
\] (14)
and equals to zero when \( \tilde{\delta} = (1/k - 1/n)\{1 - \hat{\sigma}^2(k)\} \). Thus, \( l_p(k; \tilde{\delta}) \) is a monotonically decreasing function of \( \tilde{\delta} \) for any \( k \in \{1, 2, \ldots, n - 1\} \) on \((0, (1/k - 1/n)\{1 - \hat{\sigma}^2(k)\})\), and notice that on this range we also have \( \hat{\sigma}^2(k) < 1 \).

**Proof of Lemma 2**

*Proof.* Lemma 1 proves that \( l_p(k; \tilde{\delta}) \) is smooth and monotonically decreasing on \( \tilde{\delta} \in (0, (1/k - 1/n)\{1 - \hat{\sigma}^2(k)\}) \) for any fixed \( k \in \{2, \ldots, n - 1\} \). From the definition, it is clear that \( l_p(k; \tilde{\delta}) \) and \( l_p(k + 1; \tilde{\delta}) \), as well as \( l_p(k - 1; \tilde{\delta}) \) are monotonically decreasing functions of \( \tilde{\delta} \in G(k) \) since \( G(k) \subset (0, (1/k - 1/n)\{1 - \hat{\sigma}^2(k)\}) \), \( G(k) \subset (0, \{1/(k - 1) - 1/n\}\{1 - \hat{\sigma}^2(k - 1)\}) \), and \( G(k) \subset (0, \{1/(k + 1) - 1/n\}\{1 - \hat{\sigma}^2(k + 1)\}) \).
Since the difference of the two penalized profile log-likelihood at a fixed \( k \) is a simple function, then \( l_p(k; \tilde{\delta}) - l_p(k - 1; \tilde{\delta}) \) is smooth on \( G(k) \). And similarly, \( l_p(k; \tilde{\delta}) - l_p(k + 1; \tilde{\delta}) \) is also smooth on \( G(k) \).

Next, we investigate the behaviours of these two functions separately:

\( A: \) Show that \( l_p(k; \tilde{\delta}) - l_p(k - 1; \tilde{\delta}) \) is convex on \( G(k) \)

The first order derivative with respect to \( \tilde{\delta} \), which is the difference of the first order derivative of \( l_p(k; \tilde{\delta}) \) and \( l_p(k - 1; \tilde{\delta}) \), is

\[
\frac{\partial l_p(k; \tilde{\delta}) - l_p(k - 1; \tilde{\delta})}{\partial \tilde{\delta}} = -\frac{mn}{2} \log \left\{ \frac{1 - \frac{k}{n} - k \tilde{\delta}}{(1 - \frac{k}{n})\tilde{\sigma}^2(k)} \right\} + \frac{mn(k - 1)}{2} \log \left\{ \frac{1 - \frac{k-1}{n} - (k - 1)\tilde{\delta}}{(1 - \frac{k-1}{n})\tilde{\sigma}^2(k - 1)} \right\}
\]

\[
= \frac{mn(k - 1)}{2} \left\{ \log \frac{\tilde{\sigma}^2(k)}{\tilde{\sigma}^2(k - 1)} \right\} + \frac{mn}{2} \log \tilde{\sigma}^2(k). \quad (15)
\]

Notice that \( \frac{\tilde{\sigma}^2(k)}{\tilde{\sigma}^2(k - 1)} \) can be greater or less than 1 depending on \( \tilde{\delta} \):

\[
\frac{\tilde{\sigma}^2(k)}{\tilde{\sigma}^2(k - 1)} = \frac{n - k + 1 - n(k - 1)\tilde{\delta}}{n - k - nk\tilde{\delta}} \frac{n - k}{n - k + 1 \tilde{\sigma}^2(k - 1)}
\]

However, its first order derivative

\[
\frac{n^2}{(n - k - nk\tilde{\delta})^2 (n - k + 1 \tilde{\sigma}^2(k - 1))} \frac{(n - k - (n - k - nk\tilde{\delta})) \tilde{\sigma}^2(k)}{(n - k - 1 \tilde{\sigma}^2(k - 1))} > 0,
\]

which indicates that it is increasing in \( \tilde{\delta} \). We can find its maximum at

\[
\left\{ \frac{\tilde{\sigma}^2(k)}{\tilde{\sigma}^2(k - 1)} \right\}_{\tilde{\delta} = (1/(k + 1) - 1/n)(1 - \tilde{\sigma}^2(k - 1))} = \frac{(n - k + 1)\tilde{\sigma}^2(k - 1) + 2 n \frac{1 - \tilde{\sigma}^2(k - 1)}{k + 1}}{(n - k)\tilde{\sigma}^2(k) + \frac{n \tilde{\sigma}^2(k - 1)}{k + 1}} \frac{(n - k)\tilde{\sigma}^2(k)}{(n - k + 1)\tilde{\sigma}^2(k - 1)} < 1
\]

Since \( G(k) \subset (0, (1/(k - 1) - 1/n)(1 - \tilde{\sigma}^2(k))) \), we can conclude \( \frac{\partial l_p(k; \tilde{\delta}) - l_p(k - 1; \tilde{\delta})}{\partial \tilde{\delta}} < 0 \) on \( \tilde{\delta} \in G(k) \). The second order derivative can be shown to be positive,

\[
\frac{\partial^2 l_p(k; \tilde{\delta}) - l_p(k - 1; \tilde{\delta})}{\partial \tilde{\delta}^2} = \frac{mnk}{2} \frac{k}{1 - \frac{k}{n} - k \tilde{\delta}} - \frac{mn(k - 1)}{2} \frac{k - 1}{1 - \frac{k-1}{n} - (k - 1)\tilde{\delta}}
\]

\[
= \frac{mn}{2} \left\{ \frac{nk^2}{n - k - nk\tilde{\delta}} - \frac{n(k - 1)^2}{n - (k - 1) - (k - 1)\tilde{\delta}} \right\} > 0, \quad (16)
\]

and we conclude that \( l_p(k; \tilde{\delta}) \) is a convex function of \( \tilde{\delta} \) for any fixed \( k \).
Further, since the first order derivative is negative on \( \tilde{\delta} \in G(k) \), combined with the following
\[
\left\{ \frac{l_p(k; \tilde{\delta}) - l_p(k-1; \tilde{\delta})}{\partial \tilde{\delta}} \right\}_{\tilde{\delta}=(\frac{1}{k+1} - \frac{1}{n}) \{1 - \sigma^2(k-1)} < 0
\]
and
\[
\left\{ \frac{l_p(k; \tilde{\delta}) - l_p(k-1; \tilde{\delta})}{\partial \tilde{\delta}} \right\}_{\tilde{\delta}=0} > 0,
\]
indicating a solution to \( l_p(k; \tilde{\delta}) - l_p(k-1; \tilde{\delta}) = 0 \) exists on \( G(k) \).

B: Show that \( l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta}) \) is concave on \( G(k) \) The first order derivative with respect to \( \tilde{\delta} \):
\[
\frac{\partial l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta})}{\partial \tilde{\delta}} = -\frac{mn(k+1)}{2} \left\{ -\log \frac{n - (k + 1)}{n - (k + 1) - n(k + 1)\tilde{\delta}} - \log \sigma^2(k + 1) \right\}
\]
\[
+ \frac{mnk}{2} \log \left\{ \frac{n - k - nk\tilde{\delta}}{(n - k)\sigma^2(k)} \right\}
\]
\[
= \frac{mnk}{2} \left\{ \log \frac{\sigma^2(k + 1)}{\sigma^2(k)} \right\} + \frac{mn}{2} \log \sigma^2(k + 1).
\]
(17)

It can be shown similarly to the first case by taking \( k' = k + 1 \) and taking the opposite sign, we can conclude \( \frac{\partial l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta})}{\partial \tilde{\delta}} > 0 \) on \( \tilde{\delta} \in G(k) \). The second order derivative can be shown to be negative,
\[
\frac{\partial^2 l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta})}{\partial \tilde{\delta}^2} = \frac{mnk}{2} \left\{ \frac{k}{1 - k - n\tilde{\delta}} - \frac{mn(k + 1)}{2} \frac{k + 1}{1 - \frac{k+1}{n} - (k + 1)\tilde{\delta}} \right\}
\]
\[
\quad \quad = \frac{mn}{2} \left\{ \frac{nk^2}{n - k - nk\tilde{\delta}} - \frac{n(k+1)^2}{n - k - 1 - n(k + 1)\tilde{\delta}} \right\} < 0,
\]
(18)
and that \( l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta}) \) is a concave function of \( \tilde{\delta} \) for any fixed \( k \).

Because the first order derivative is positive on \( \tilde{\delta} \in G(k) \), combined with the following
\[
\left\{ \frac{l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta})}{\partial \tilde{\delta}} \right\}_{\tilde{\delta}=(\frac{1}{k+1} - \frac{1}{n}) \{1 - \sigma^2(k-1)} > 0
\]
and
\[
\left\{ \frac{l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta})}{\partial \tilde{\delta}} \right\}_{\tilde{\delta}=0} < 0,
\]
indicating a solution to \( l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta}) = 0 \) exists on \( G(k) \).

Finally, we conclude that for any fixed \( k \in \{2, 3, \ldots, n-2\} \), \( l_p(k; \tilde{\delta}) - l_p(k-1; \tilde{\delta}) \) is a monotonically decreasing and convex function of \( \tilde{\delta} \in G(k) \), taking positive value when \( \tilde{\delta} = 0; l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta}) \) is a monotonically increasing and concave function of \( \tilde{\delta} \in G(k) \), taking negative value when \( \tilde{\delta} = 0 \).
Proof of Lemma 3

Proof. Following Lemma 2, \( l_p(k; \tilde{\delta}) > l_p(k-1; \tilde{\delta}) \) and \( l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta}) \) are smooth and monotone functions of \( \tilde{\delta} \in G(k) \). Define

\[
a_k = \min\{\tilde{\delta} \in G(k) : l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta}) \geq 0\}
\]

and

\[
b_k = \max\{\tilde{\delta} \in G(k) : l_p(k; \tilde{\delta}) - l_p(k-1; \tilde{\delta}) \geq 0\}.
\]

The existence of such values follow from the proof of Lemma 2. And according to this definition, \( a_{k-1} = b_k \), and combined with the fact that the first order derivative of \( l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta}) \) is smaller than that of \( -\{l_p(k; \tilde{\delta}) - l_p(k-1; \tilde{\delta})\} \), implying that \( a_k < a_{k-1} = b_k \). Thus, the interval \( \Delta_k = (a_k, b_k) \subset G(k) \) is not empty and gives the range of \( \tilde{\delta} \)-values in \( G(k) \) such that conditions (9) hold.

Consider \( k \in \{2, 3, \ldots, n-2\} \) and suppose for any \( \tilde{\delta}_o \in \Delta_k = (a_k, b_k) \subset G(k) \), \( l_p(k; \tilde{\delta}) \) takes its maximum at \( k = k' \). Then clearly \( l_p(k'; \tilde{\delta}) > l_p(k'-1; \tilde{\delta}) \) and \( l_p(k'; \tilde{\delta}) - l_p(k'+1; \tilde{\delta}) \).

On the other hand, if conditions (9) hold for some \( \tilde{\delta}_o \in \Delta_k = (a_k, b_k) \), we need to show \( l_p(k'; \tilde{\delta}) - l_p(k''; \tilde{\delta}) > 0 \) for any \( k'' < k' \) by showing the following sum is greater than zero:

\[
l_p(k'; \tilde{\delta}) - l_p(k''; \tilde{\delta}) = [l_p(k'; \tilde{\delta}) - l_p(k'-1; \tilde{\delta})] + [l_p(k'-1; \tilde{\delta}) - l_p(k'-2; \tilde{\delta})] + \cdots + [l_p(k''+1; \tilde{\delta}) - l_p(k''; \tilde{\delta})] > 0, \tag{19}
\]

Since \( l_p(k'; \tilde{\delta}) - l_p(k'-1; \tilde{\delta}) \) is positive on \( (a_k', b_k') \), we only need to show \( l_p(k'-1; \tilde{\delta}) - l_p(k'-2; \tilde{\delta}) \) and all other telescoping terms are positive on \( (a_k', b_k') \). This is straightforward to show according to the definitions of \( (a_k'-1, b_k'-1) \) and since \( a_{k'-1} = b_k' \). This implies that for each \( k'' < k' \), \( l_p(k''; \tilde{\delta}) - l_p(k''-1; \tilde{\delta}) \) is positive on \( (0, b_{k''}) \subset (0, b_k'). \)

Similarly, \( l_p(k'; \tilde{\delta}_o) - l_p(k''; \tilde{\delta}_o) > 0 \) is equivalent to the telescoping sum being greater than zero:

\[
l_p(k'; \tilde{\delta}_o) - l_p(k''; \tilde{\delta}_o) = [l_p(k'; \tilde{\delta}_o) - l_p(k'+1; \tilde{\delta}_o)] + [l_p(k'+1; \tilde{\delta}_o) - l_p(k'+2; \tilde{\delta}_o)] + \cdots + [l_p(k''-1; \tilde{\delta}_o) - l_p(k''; \tilde{\delta}_o)] > 0 \tag{20}
\]

where we must have \( l_p(k'; \tilde{\delta}_o) - l_p(k'+1; \tilde{\delta}_o) > 0 \) on \( \Delta_k = (a_k', b_k') \).

It is sufficient to show \( l_p(k'+1; \tilde{\delta}) - l_p(k'+2; \tilde{\delta}) \) and all other telescoping terms are positive on \( (a_k', b_k') \). According to the definitions of \( (a_{k'+1}, b_{k'+1}) \) and the result that \( a_{k'} = b_{k'+1} \). This implies that for each \( k'' > k' \), \( l_p(k''; \tilde{\delta}) - l_p(k''+1; \tilde{\delta}) \) is positive when \( \tilde{\delta} > a_{k''} \), and \( (a_{k''}, b_{k''}) \subset (a_k', b_k') \).

When \( k = n-1 \) or \( k = 1 \), they fall in the general case of \( k'' = n-1 \) or \( k'' = 1 \). \( \square \)
Proof of Lemma 4

Proof. Since \(a_k\) and \(b_k\) are not analytically tractable, we find these values by approximating the differences using Taylor series, and then find roots to the Taylor approximations.

A: Case \(l_p(k; \tilde{\delta}) - l_p(k - 1; \tilde{\delta}) > 0\)

Expanding the difference in penalized profile log-likelihoods at \(k\) in terms of \(\hat{\sigma}^2(k - 1)\), \(\hat{\sigma}^2(k)\), \(\hat{\lambda}_k\) and \(\lambda_k\): 

\[
l_p(k; \tilde{\delta}) - l_p(k - 1; \tilde{\delta}) = -\frac{m}{2} \left[ \log \frac{\hat{\lambda}_k}{\hat{\sigma}^2(k - 1)} + (n - k) \log \frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k - 1)} + (n - k) \log \left(1 - \frac{k}{n}\right) - (n - k + 1) \log \left(1 - \frac{k - 1}{n}\right) - n\delta \left\{ k \log \frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k - 1)} + k \left( \frac{n - k}{n - (k - 1)} \right) \right\} \log(\hat{\sigma}^2(k - 1)) + \log \left(1 - \frac{k - 1}{n}\right) + 1 \right\} + \frac{n k \delta}{n} \log \left(1 - \frac{k - 1}{n} - (k - 1)\tilde{\delta}\right) - n(k - 1)\delta \log \left(1 - \frac{k - 1}{n} - (k - 1)\tilde{\delta}\right) \right] \]

We approximate the following for any \(k\) by Taylor expansion at \(\tilde{\delta} = 0\):

\[
\log \left(1 - \frac{k}{n} - k\tilde{\delta}\right) = \log \left(1 - \frac{k}{n}\right) - \frac{n k \delta}{n - k} + O(\tilde{\delta}^2) \quad (21a)
\]

\[
\log \left(1 - \frac{k - 1}{n} - (k - 1)\tilde{\delta}\right) = \log \left(1 - \frac{k - 1}{n}\right) - \frac{n(k - 1)\delta}{n - k + 1} + O(\tilde{\delta}^2), \quad (21b)
\]

and they respectively converge if \(\tilde{\delta} < 1/k - 1/n\) and \(\tilde{\delta} < 1/(k - 1) - 1/n\), respectively. The difference is approximated by

\[
l_p(k; \tilde{\delta}) - l_p(k - 1; \tilde{\delta}) = -\frac{m}{2} \left\{ \zeta(\tilde{\delta}) + O(\tilde{\delta}^2) \right\} = -\frac{m}{2} \left[ \log \frac{\hat{\lambda}_k}{\hat{\sigma}^2(k - 1)} + (n - k) \log \frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k - 1)} + n\delta \left\{ k \log \frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k - 1)} + \log \hat{\sigma}^2(k - 1) \right\} + \frac{n^2(k - 1)^2}{n - k + 1} - \frac{n^2 k^2}{n - k} \right\} + O(\tilde{\delta}^3) \]

So now we need to solve the inequality and find the smallest \(\tilde{\delta}\) such that \(\zeta(\tilde{\delta}) \leq 0\). Clearly, \(\zeta(\tilde{\delta})\) is a quadratic function of \(\tilde{\delta}\) for fixed \(k\) and when \(\tilde{\delta} = 0\), \(\zeta(\tilde{\delta}) < 0\). In quadratic
equation representation, we can rewrite \( \zeta(\tilde{\delta}) = c_2 \tilde{\delta}^2 + c_1 \tilde{\delta} + c_0 \), where

\[
\begin{align*}
c_2 &= \frac{n^2(k-1)^2}{n-k+1} - \frac{n^2k^2}{n-k} = \frac{n(1 - 2k) - k(1-k)}{(n-k+1)(n-k)} < 0 \quad (23) \\
c_1 &= -n\left\{ k \log \frac{\tilde{\sigma}^2(k)}{\tilde{\sigma}^2(k-1)} + \log \tilde{\sigma}^2(k-1) \right\} > 0 \quad (24) \\
c_0 &= \log \frac{\hat{\lambda}_k}{\tilde{\sigma}^2(k-1)} + (n-k) \log \frac{\tilde{\sigma}^2(k)}{\tilde{\sigma}^2(k-1)} < 0 \quad (25)
\end{align*}
\]

Clearly, the discriminant \( c_1^2 - 4c_2c_0 > 0 \) and we have \( c_2 < 0 \), then there are two positive roots, \( r_1(k) \) and \( r_2(k) \), where

\[
\begin{align*}
\{ \zeta(\tilde{\delta}) < 0 & \text{ for } \tilde{\delta} \in (0, r_1(k)) \cup (r_2(k), 1/k - 1/n), \\
\zeta(\tilde{\delta}) > 0 & \text{ for } \tilde{\delta} \in (r_1(k), r_2(k)).
\}
\end{align*}
\]

\( B: \text{Case } l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta}) > 0 \)

Similarly, expanding the difference for \( k \) and \( k+1 \):

\[
\begin{align*}
l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta}) &= -\frac{m}{2} \left\{ \frac{\hat{\lambda}_k+1}{\tilde{\sigma}^2(k+1)} - \frac{\tilde{\sigma}^2(k)}{\tilde{\sigma}^2(k+1)} \right\} \\
&\quad + (n-k) \log \left( 1 - \frac{k}{n} \right) - (n-k-1) \log \left( 1 - \frac{k+1}{n} \right) \\
&\quad - n\tilde{\delta} \left\{ k \log \frac{\tilde{\sigma}^2(k)}{\tilde{\sigma}^2(k+1)} + k \log \frac{n-k}{n-k-1} \\
&\quad - \log \tilde{\sigma}^2(k+1) - \log \left( 1 - \frac{k+1}{n} \right) - 1 \right\} \\
&\quad + (n-k-1) \log \left( 1 - \frac{k+1}{n} - (k+1)\tilde{\delta} \right) - (n-k) \log \left( 1 - \frac{k}{n} \right) \\
&\quad + nk\tilde{\delta} \log \left( 1 - \frac{k}{n} - k\tilde{\delta} \right) - n(k+1)\tilde{\delta} \log \left\{ 1 - \frac{k}{n} - (k+1)\tilde{\delta} \right\}
\end{align*}
\]

We approximate the following for any \( k \) by Taylor expansion at \( \tilde{\delta} = 0 \):

\[
\log \left( 1 - \frac{k+1}{n} - (k+1)\tilde{\delta} \right) = \log \left( 1 - \frac{k'+1}{n} \right) - \frac{n(k+1)\tilde{\delta}}{n-k+1} + O(\tilde{\delta}^2), \quad (27a)
\]

and it converges if \( \tilde{\delta} < 1/(k+1) - 1/n \). Again, define \( \zeta'(\tilde{\delta}) \) and the approximated
difference is:
\[
\ell_p(k; \tilde{\delta}) - \ell_p(k + 1; \tilde{\delta}) = \frac{m}{2} \left\{ \zeta'(\delta) + O(\tilde{\delta}^3) \right\}
\]
\[
= -\frac{m}{2} \left[ -\log \frac{\hat{\lambda}_{k+1}}{\hat{\sigma}^2(k')} + (n - k' - 1) \log \frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k + 1)} \\
- n\tilde{\delta}\left\{ k \log \frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k + 1)} - \log \hat{\sigma}^2(k + 1) \right\} \\
+ \tilde{\delta}^2 \left\{ \frac{n^2(k + 1)^2}{n - k - 1} - \frac{n^2k^2}{n - k} \right\} + O(\tilde{\delta}^3) \right]
\] (28)

Again, we can rewrite \( \zeta'(\tilde{\delta}) = \epsilon_2 \tilde{\delta}^2 + \epsilon_1 \tilde{\delta} + \epsilon_0 \), where
\[
\epsilon_2 = \frac{n^2(k + 1)^2}{n - k - 1} - \frac{n^2k^2}{n - k} = \frac{n - k + (2n - k)k}{(n - k - 1)(n - k)} > 0
\]
\[
\epsilon_1 = n \left\{ (k + 1) \log \frac{\hat{\sigma}^2(k + 1)}{\hat{\sigma}^2(k)} + \log \hat{\sigma}^2(k) \right\} < 0
\]
\[
\epsilon_0 = -\log \frac{\hat{\lambda}_{k+1}}{\hat{\sigma}^2(k)} + (n - k - 1) \log \frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k + 1)} > 0
\]

Now solve the inequality and find the smallest \( \tilde{\delta} \) such that \( \zeta'(\tilde{\delta}) \leq 0 \). Clearly, when \( \tilde{\delta} = 0 \), \( \zeta'(\tilde{\delta}) > 0 \). Clearly, the discriminant \( \epsilon_1^2 - 4\epsilon_2\epsilon_0 > 0 \) and we have \( \epsilon_2 > 0 \), then there are two positive roots, \( r_3(k) \) and \( r_4(k) \), where
\[
\left\{ \begin{aligned}
\zeta'(\tilde{\delta}) &< 0 \text{ for } \delta \in (r_3(k), r_4(k)) \\
\zeta'(\tilde{\delta}) &> 0 \text{ for } \delta \in (0, r_3(k)) \cup (r_4(k), 1/(k + 1) - 1/n).
\end{aligned} \right.
\]

Since \( \tilde{\delta} \in (0, 1/(k + 1) - 1/n) \), the root near \( 0 \) can be approximated by the Vieta’s solution \( -\frac{\epsilon_0}{\epsilon_1} \) and \( -\frac{\epsilon_1'}{\epsilon_1} \). So finally we have
\[
\ell_u(k) = -\frac{\log \frac{\hat{\lambda}_k}{\hat{\sigma}^2(k - 1)} + (n - k) \log \frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k - 1)}}{n \left\{ k \log \frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k - 1)} - (k - 1) \log \hat{\sigma}^2(k - 1) \right\}}
\]
and
\[
\ell_d(k) = -\frac{\log \frac{\hat{\lambda}_{k+1}}{\hat{\sigma}^2(k)} + (n - k - 1) \log \frac{\hat{\sigma}^2(k + 1)}{\hat{\sigma}^2(k + 1)}}{n \left\{ (k + 1) \log \frac{\hat{\sigma}^2(k + 1)}{\hat{\sigma}^2(k + 1)} - k \log \hat{\sigma}^2(k) \right\}}.
\]

Also notice that \( \ell_u(k) \) is always greater than \( \ell_d(k - 1) \) due to the difference in the numerators while the denominators remains the same, and we conclude that these approximated intervals are also non-overlapping. \( \square \)
Proof of Lemma 5

Proof. Note that the numerators of \( u_b(k) \) and \( u_a(k) \) are smooth functions of \( \hat{\sigma}^2(k) \) on \((0, 1)\), respectively, as both first derivatives with respect to \( \hat{\sigma}^2(k) \) exist. In addition, it can be shown that the quadratic coefficient \( c_0 \) is a monotonically increasing function of \( \hat{\sigma}^2(k) \), while \( c_0' \) is a monotonically decreasing function.

The size of the interval depends on the relative ratio of each \( \hat{\lambda}_k \) and \( \hat{\lambda}_{k+1} \) relative to \( \hat{\sigma}^2(k) \).

\[
\lambda_k - \hat{\sigma}^2(k) = a_k^2 > 0
\]

\[
\lambda_{k+1} - \hat{\sigma}^2(k) = (n - k)\hat{\sigma}^2(k) - (n - k - 1)\hat{\sigma}^2(k + 1) - \hat{\sigma}^2(k)
= (n - k - 1)(\hat{\sigma}^2(k) - \hat{\sigma}^2(k + 1))
= a_{k+1}^2
\]

Since the true value of the \( n - k \) last eigenvalues are equal to \( \sigma^2(k) \), Theorem 8.3.2 [27] implies that for normal data, the maximum likelihood estimator of \( \hat{\sigma}^2(k) = (\sum_{i=k+1}^{n} \hat{\lambda}_i)/(n-k) \) converges to \( \sigma^2(k) \). Similarly, maximum likelihood estimator of \( \sigma^2(k + 1) \) is \( \hat{\sigma}^2(k + 1) = (\sum_{i=k+1}^{n} \hat{\lambda}_i)/(n-k-1) \) and also converges to \( \sigma^2(k) \), which implies that \(|\hat{\sigma}^2(k) - \hat{\sigma}^2(k + 1)|\) converges to zero with probability 1 or \(|\hat{\sigma}^2(k) - \hat{\sigma}^2(k + 1)|\) converges to zero with probability 1.

The difference \( \hat{\lambda}_k - \hat{\sigma}^2(k) \) is always positive and converges to the true difference separating the signal and noise \( d_k^2 \), while \( \hat{\lambda}_{k+1} - \hat{\sigma}^2(k) \) diminishes to zero with probability 1 as \( m \) increases. In this case, at \( k = k^* \), the denominator of \( u_a(k^*) \) is

\[
c_0 = - \log \frac{\hat{\lambda}_{k+1}}{\hat{\sigma}^2(k)} + (n - k - 1) \log \frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k + 1)}
= (n - k - 1) \log \frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k + 1)} - \log \frac{(n - k)\hat{\sigma}^2(k) - (n - k - 1)\hat{\sigma}^2(k + 1)}{\hat{\sigma}^2(k)}.
\]

and it approaches 0 if \( m \to \infty \), implying that \( u_a(k^*) \to 0 \). On the other hand, \( u_b(k^*) \) is a positive constant, thus \( \frac{u_b(k^*)}{u_a(k^*)} \to \infty \).

For \( k > k^* \), due to the symmetry in \( u_b(k) \) and \( u_a(k) \), and \( u_b(k + 1) = u_a(k) \to 0 \) as well as \( u_a(k + 1) \to 0 \) as \( m \to \infty \). Thus, we must have \(|u_b(k) - u_a(k)| \to 0 \) for all \( k > k^* \) as \( m \to \infty \). \( \square \)

An R package

An R package available at https://github.com/WeiAkaneDeng/SPAC2 implements our penalized profile log-likelihood criterion to estimate the number of effective dimensions of a data matrix. The package also provides various functions to simulate either the sample eigenvalues or sample data under specific covariance structures and possibly with violation to normality or independence assumption.
Additional results on the behaviour of penalized profile log-likelihood

One dataset with \( n = 100 \) and \( m = 10,000 \) was simulated assuming the true effective dimension to be \( k^* = 10 \). We similarly assumed the first \( k^* \) squared singular values decay at an exponential rate with their values determined by the residual variance \( \sigma^2 = 0.5 \) and the smallest squared singular value \( d^2(k^*) \) at 0.5. The profile log-likelihood and penalized profile log-likelihood are plotted as a function of varying \( k \in \{1, \ldots, n-1\} \) omitting any constants or scaling constants.

The top left plot of Figure 7 shows \( l_p(k, \tilde{\delta}) \) at various \( \tilde{\delta} \) values as a function of \( k \), and clearly as the penalty tuning parameter value increases, the maximum emerges and approaches that of the correct value of \( k^* = 10 \) as marked by the red dashed line.

The bottom left plot of Figure 7 is a depiction of the voting scheme that we implemented, for a grid of \( \tilde{\delta} \) values following a geometric sequence on the interval \((u_a(n-1), u_b(1))\), we estimated the \( k \) that maximized \( l_p(k, \tilde{\delta}_o) \) for each possible \( \tilde{\delta}_o \). The correct value is marked by the red dashed line at \( k^* = 10 \) and gives an empirical estimate of \((u_a(10), u_b(10))\) and its range marked by the orange dashed lines.

The top right figure shows \( l_p(k, \tilde{\delta}) \) at various \( k \) values as a function of \( \log(\tilde{\delta}) \) values. The orange dashed lines marks the empirical values of \((u_a(10), u_b(10))\) as estimated from the grid of \( \tilde{\delta} \) values. It is clear that on this interval \((u_a(10), u_b(10))\), \( l_p(k = 10, \tilde{\delta}_o) \) is the largest.

The bottom right figure illustrates the difference between two adjacent penalized log-likelihoods as a function of \( \tilde{\delta} \) for various values of \( k \). We see that the difference between smaller \( k \) values generally dominates those between larger \( k \) values, and strictly so if we restrict ourselves to the interval \((0, u_b(k))\).

References


Figure 7: Behaviour of penalized profile log-likelihood as a function of \( k \) and \( \log_{10}(\tilde{\delta}) \).


[43] Richard W Tothill, Anna V Tinker, Joshy George, Robert Brown, Stephen B Fox, Stephen Lade, Daryl S Johnson, Melanie K Trivett, Darish Etemadmoghadam, Bianca Locandro, Nadia Traficante, Sian Fereday, Jillian A Hung, Yoke-Eng Chiew,
