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 unsupervised estimation of the effective dimension of a sample of dependent random vectors. The proposed method uses the principal components (PC) decomposition of sample covariance to establish a low-rank approximation that helps uncover the hidden structure. The number of PCs to be included in the decomposition is determined via a Probabilistic Principal Components Analysis (PPCA) embedded in a penalized profile likelihood criterion. The choice of 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 PPCA 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: Effective dimension; Gene expression; Penalization; Principal component analysis; Probabilistic principal component analysis; Profile likelihood; Unsupervised clustering.

1 Introduction

Suppose we collect a sample \( \mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^m \times n \) from a population with hidden cluster structure. Of interest is the estimation of the number of clusters \( (K) \) in the sample \( \mathbf{x} \). An alternative formulation is to consider a parsimonious representation of the data

\[
\mathbf{x}^T = \mathbf{W}\mathbf{L} + \mathbf{\epsilon},
\]

where \( \mathbf{W} \in \mathbb{R}^{n \times K} \) is the loading matrix that characterizes dependence between the rows of \( \mathbf{x}^T \) representing measurements for individuals in the sample, \( \mathbf{L} = [\mathbf{L}_1, \ldots, \mathbf{L}_m] \in \mathbb{R}^{K \times m} \) is a

*Correspondence should be addressed to craiu@utstat.toronto.edu or deng@utstat.toronto.edu
matrix whose columns are arbitrary latent vectors $L_j \in \mathbb{R}^K, j = 1, 2, \ldots, m$, and $\epsilon \in \mathbb{R}^{n \times m}$ is the 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 in $L$, and each row corresponds to one of the clusters. Hence, the problem we are interested in can be translated to finding the minimal $1 \leq K \leq n$ such that the matrix $W$ has rank $K$, and the noise, $x^T - WL$, 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 $K$ using the probabilistic principal component analysis (PPCA) framework.

PPCA, 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 PPCA formulation, where a penalty is introduce to reduces the number of PCs 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 a mixture formulation that identifies the estimated effective dimension as the number of mixture components.

In Section 2, we will briefly review the key results of PPCA, 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 an $n$-dimensional random vector $y \in \mathbb{R}^n$, with mean vector $\mu \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n \times n}$, where $n$ can be considered the number of subjects to be clustered. Without loss of generality, we assume that $E(y) = \mu = 0$ so that the true covariance is $\Sigma = E(yy^T)$. Denote the number of unknown clusters for the $n$ subjects by $K^* \in \{1, 2, 3, \ldots, n\}$. Our aim is to determine $K^*$ over a large number of features or observations. We propose to solve this problem by finding the minimal rank, $K^*$, of $W$, such that for each observed instance of $y$, the rows of $y - Wl$, are independent and identically distributed, where $l \in \mathbb{R}^{K^*}$ is a latent vector with distribution $l \sim N(0, I_{K^*})$. In other words, the matrix $W$ completely accounts for the dependence structure in the data and the true covariance has the form

$$\Sigma = E(yy^T) = WW^T + \sigma^2 I_n,$$

where $\sigma^2 I_n$ refers to the covariance of the residual error vector $y - Wl$.

Given a sample $x = (x_1, \ldots, x_n)$ with each $x_i \in \mathbb{R}^m$, we denote the transpose $x^T = (y_1, \ldots, y_m)$ in terms of the $m$ features $y_j \in \mathbb{R}^n, 1 \leq j \leq m$, collected on the $n$ subjects. Usually, we standardize $y$ so that the sample covariance matrix of $(y_1, \ldots, y_m)$ is

$$S = \frac{x^T x}{m - 1} = \frac{\sum_{j=1}^m y_j y_j^T}{m - 1},$$

and has trace $n$. 

3
2.2 A Brief Review of PPCA

The probabilistic formulation of PCA decomposes the data to a low-rank approximation and residual noise that follows a multivariate Gaussian distribution, $\epsilon_j \sim N(0, \sigma^2 I_n)$:

$$x^T = WL + \epsilon,$$  \hspace{1cm} (4)

where we denote $K^* = \text{rank}(W)$ and $L = [l_1, \ldots, l_j, \ldots, l_m] \in \mathbb{R}^{K^* \times m}$ with each latent vector $l_j \sim N(0, I_{K^*})$, for $j = 1, 2, \ldots, m$.

The log-likelihood function with respect to $W$ and $\sigma^2$ is derived from the sampling density of $(y_1, \ldots, y_m)$ as

$$l(W, \sigma^2; x^T) = -\frac{m}{2} \left( n \log(2\pi) + \log |WW^T + \sigma^2 I| + \text{tr}[WW^T + \sigma^2 I^{-1}S] \right).$$  \hspace{1cm} (5)

Since the rank $K$ is intrinsic to $W$, we denote the maximum likelihood estimator (MLE) derived from (5) by $\hat{W}(K)$, which depends on the unknown residual variance $\sigma^2$. So we first maximized the resulting profile log-likelihood at $\hat{W}(K)$, and obtained an MLE of the residual variable, denoted by $\hat{\sigma}^2(K)$:

$$\hat{\sigma}^2(K) = \frac{\sum_{i=K+1}^{n} \hat{\lambda}_i}{n - K},$$  \hspace{1cm} (6)

where $\hat{\lambda}_i$'s are the sample eigenvalues of $S$ and $\sum_{i=1}^{n} \hat{\lambda}_i = n$.

Now, MLE of $W$ with rank $K$ is given in terms of a singular value decomposition:

$$\hat{W}(K) = U(K)D(K)V^T(K),$$  \hspace{1cm} (7)

where $U(K)$ is an $n \times K$ matrix with columns corresponding to the first $K$ eigenvectors of $S$, $D(K)$ is a diagonal matrix with $K$ non-zero entries each given by

$$\hat{d}_i(K) = \sqrt{\hat{\lambda}_i - \hat{\sigma}^2 K},$$  \hspace{1cm} (8)

and $V(K) \in \mathbb{R}^{K \times K}$ is an arbitrary orthogonal matrix [11].

Clearly, both MLEs depend on $K$. Thus, if we consider the parameter of interest to be $K$, and nuisance parameters to be $W(K)$ and $\sigma^2(K)$, the profile likelihood $l_p$ can be expressed as:

$$l_p(K) = -\frac{m}{2} [n \log(2\pi) + \sum_{i=1}^{K} \log \hat{\lambda}_i + (n - K) \log \hat{\sigma}(K)^2 + n]$$  \hspace{1cm} (9)

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 collection of samples \( \mathbf{x}^T \in \mathbb{R}^{n \times m} \) that follow a multivariate Gaussian distribution \( N(0, \mathbf{W} \mathbf{W}^T + \sigma^2 \mathbf{I}) \) with positive semi-definite sample covariance matrix \( \mathbf{S} \). If we consider \( K = \text{rank}(\mathbf{W} \mathbf{W}^T) \) as the parameter of interest, then the profile log-likelihood \( l_p(K) \) is non-decreasing in \( K \).

**Proof.** See Web Supplementary Material.

**Remark 1.** As a result of Proposition 1, the profile log-likelihood is positive infinity when \( K = n \). In this case, the effective dimension is the same as the sample size and sample covariance reveals no clustering structure.

The same conclusion can be reached by observing the proportion of variance explained by the PPCA model with rank \( K \):

\[
\text{tr}[\mathbf{W}(K) \mathbf{W}(K)^T] = \sum_{j=1}^{K} d_j(K)^2 = n[1 - \sigma^2(K)].
\]

(10)

When \( \sigma^2(K) \) is equal to 0 (or \( K = n \)), the model corresponds to PCA with a full-rank loading matrix; and when \( \sigma^2(K) \) is equal to 1, the model corresponds to an extreme case of the probabilistic model where the observations consist entirely of isotropic errors with an identity covariance matrix.

**Remark 2.** One may be tempted to use a classical model selection criterion, e.g. BIC, to identify the effective dimension. However, the scale of the log-likelihood dominates the penalty when the number of features \( m \) is large. This has been noticed previously by [2] in the context of factor analysis where BIC tended to overestimate \( K \) when \( m \) is much larger than \( n \).

**Remark 3.** Assume that the true dimension of the PPCA model is \( K^* \). The population eigenvalues of the true covariance matrix \( \mathbf{W}(K^*) \mathbf{W}(K^*)^T + \sigma^2(K^*) \mathbf{I} \) have the form

\[
\lambda_i = \begin{cases} 
   d_i^2(K^*) + \sigma^2(K^*), & \text{if } i \leq K^*; \\
   \sigma^2(K^*), & \text{otherwise},
\end{cases}
\]

(11)

and \( \lambda_{K^*+1} \) has multiplicity \( n - K^* \). When \( m \) is large, \( \hat{\sigma}^2(K^*) = \frac{\sum_{i=K^*+1}^{n} \hat{\lambda}_i}{n-K^*} \) is approximately equal to the MLE of the \( K \)th population eigenvalue for any \( K > K^* \) and \( \hat{\sigma}^2(K) \overset{p}{\rightarrow} \sigma^2(K^*) \) for all \( K > K^* \), and thus \( \hat{\sigma}^2(K) - \hat{\sigma}^2(K+1) \overset{p}{\rightarrow} 0 \) [1]. However, for finite samples, \( \hat{\sigma}^2(K) \) is always different from \( \hat{\sigma}^2(K+1) \), and this results in \( l_p(K) < l_p(K+1) \).

### 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 regulated via a tuning parameter which controls the trade-off between goodness-of-fit and complexity [40, 43, 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 the model $K$:

\begin{align}
\text{pen}_1(W(K), \sigma^2(K)) &= \text{rank}[W(K)]\text{tr}[W(K)W(K)^T] = nK[1 - \sigma^2(K)]; \\
\text{pen}_2(W(K), \sigma^2(K)) &= -\text{rank}[W(K)]\log \sigma^2(K) = -K \log \sigma^2(K); \\
\text{pen}_3(W(K), \sigma^2(K)) &= \text{rank}[W(K)] \frac{1}{\sigma^2(K)} = \frac{K}{\sigma^2(K)}. \\
\end{align}

We ultimately choose to use (12b) as it leads to simpler analytical derivations and intuitive heuristics. The natural guiding principle is to favour a parsimonious cluster structure by penalizing small values of $\sigma^2(K)$ as well as large values of $K$.

### 2.4 Construction of Penalized PPCA

The penalized log-likelihood has the form:

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

where the tuning parameter $\delta > 0$ controls the amount of penalty $\text{pen}(W, \sigma^2)$. 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 $S$.

Here we provide results for the penalized MLEs by maximizing the penalized log-likelihood under the penalty function (12b) with respect to $W$:

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

Similarly to the un-penalized profile log-likelihood, the penalized MLEs, $\hat{W}$ and $\hat{\sigma}^2$, are functions of $K$. Since $W$ is not present in the penalty function, the relationship between the penalized MLE of $W$ and $\sigma^2$ is the same as that between the unpenalized MLEs, but with different singular values due to the presence of a non-zero $\delta$ value. Given that the rank of $W$ is $K$, the penalized MLE of $\sigma^2(K)$ depends on $\delta$ and the unpenalized MLE $\hat{\sigma}^2(K)$:

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

Notice that 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 $\delta$. Thus, the theoretical range of $\delta$ has an upper bound at $\frac{n}{K} - 1$ since each $\hat{\sigma}^2(K)$ 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 $\tilde{\delta} = \frac{\delta}{n} \in [0, \frac{1}{K} - \frac{1}{n})$.

The penalized profile log-likelihood can be similarly written as a function of $K$ given each $\tilde{\delta}$:

$$l_p(K; \tilde{\delta}) = -\frac{m}{2} \sum_{i=1}^{K} \log(\hat{\lambda}_i) + (n - K) \log \hat{\sigma}^2(K) + K + \frac{\sum_{i=K+1}^{n} \hat{\lambda}_i}{\hat{\sigma}^2(K)} - nK\tilde{\delta} \log \hat{\sigma}^2(K) + n \log(2\pi)$$

$$= -\frac{m}{2} \sum_{i=1}^{K} \log(\hat{\lambda}_i) + (n - K) \log \hat{\sigma}^2(K) + \frac{\hat{\sigma}^2(K)}{\hat{\sigma}^2(K)}(n - K) + K - \tilde{\delta}nK \log \hat{\sigma}^2(K) + n \log(2\pi)$$

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

(15)

The introduction of penalty changes the monotonic property of the profile log-likelihood making it possible to select the correct dimension $K^*$ for appropriate choices of $\tilde{\delta}$.

**Proposition 2.** Consider a sample $x \in \mathbb{R}^{m \times n}$ with each row following a multivariate Gaussian distribution $N(0, WW^T + \sigma^2I)$. Suppose the sample covariance matrix of $x^T$ is positive semi-definite and $K^*$ is the effective dimension of $W$. Then there exist $\tilde{\delta}_o \in (0, 1 - \frac{1}{n})$ such that $l_p(K; \tilde{\delta}_o)$ is maximized at $K^*$.

Proof. See Web Supplementary Material.

**Remark 4.** The main result established in the proof of Proposition 2 is that for some $\tilde{\delta} = \tilde{\delta}_o$, the penalized profile log-likelihood is maximized at $K_o$. When $K_o \notin \{1, n - 1\}$, we identify it using the following sufficient conditions:

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

(16)

**Remark 5.** For every $\delta \in (0, \frac{1}{K} - \frac{1}{n})$ and $K \in \{2, 3, \ldots, n-2\}$ we can consider the conditions

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

(17)

For any $2 \leq K \leq n - 2$, define

$$\Delta_K = \{ \tilde{\delta} \in (0, \frac{1}{K} - \frac{1}{n}) : \text{ conditions (17) are satisfied} \}.$$

It is clear that $\Delta_K$ is an open interval for each $K$, since the penalized likelihood function in (21) is a continuous function of $\delta \in (0, \frac{1}{K} - \frac{1}{n})$ for any fixed $K$. Therefore, we let $\Delta_K = (a_K, b_K) \subset (0, \frac{1}{K} - \frac{1}{n})$. 7
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 $u_b(K)$, respectively. The proof of Proposition 2 also demonstrates that $\frac{u_b(K^*)}{u_a(K^*)} > 1$ so that $\emptyset \neq (u_a(K^*), u_b(K^*)) \subset \Delta_{K^*}$.

Remark 7. By definition, for $K \neq K'$, $\Delta_K$ and $\Delta_{K'}$ are non-overlapping sets. Therefore, the theoretical range for $\tilde{\delta}$ is the union of all sets $\bigcup_{K=1}^n \Delta_K = [0, 1 - \frac{1}{n})$. But because of the restriction embedded in (14) and (16), we must have $\tilde{\delta} < (\frac{1}{K} - \frac{1}{n})(1 - \tilde{\sigma}^2(K))$ for each examined value of $K$. Consequently, the restriction imposes a relationship whereby $K$ is non-increasing in $u_a(K)$ (or $a_K(K)$) and $u_b(K)$ (or $b_K(K)$). For example, when $K = n - 1$, we must have $a_{n-1} = 0 < b_{n-1} < (\frac{1}{n-1} - \frac{1}{n})(1 - \tilde{\sigma}^2(n - 1))$, while for $K = 1$, we must have $b_2 < a_1 < b_1 < (1 - \frac{1}{n})(1 - \tilde{\sigma}^2(1))$.

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. BIC, 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 like the BIC have produced 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 an estimate of $K$ can result from multiple $\tilde{\delta}$ values, we ultimately select the estimate that has been obtained most often as the $\delta$ values are varied. We provide analytical and simulation support for this “voting” procedure in the next subsection.

2.5 A Voting Strategy to Select $K^*$

The voting strategy stems from the proof (Web Supplementary Material) of Proposition 2 where we have shown: 1) there exists $\tilde{\delta} \in \Delta_{K^*}$ such that (21) 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 suggest a majority-voting strategy for estimating $K^*$. We establish a range of possible values for $K$, say $1 \leq K \leq K_{\text{max}}$, and consider a grid of tuning parameter values $\tilde{\delta}_1, \ldots, \tilde{\delta}_T$ where $T$ is a user-specified integer. Each $\tilde{\delta}_j$ will
result in a choice of $K$ as the maximizer of $l_p(K; \hat{\delta}_j)$. Then, the number of times that each possible $K = 1, 2, 3, \ldots, K_{\text{max}}$ maximizes the penalized profile log-likelihood is counted and the one with the highest vote count is selected and denoted by $\hat{K}$.

Below 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 $K = n$ it follows from (14) that $\Delta_n = \{0\}$ is the only possible choice. We define $K_{\text{max}} = \left(\frac{\delta}{1 - \hat{\sigma}^2(K_{\text{max}}')} + \frac{1}{n}\right)^{-1}$ and $K_{\text{max}'} = \left(\hat{\delta} + \frac{1}{n}\right)^{-1} > K_{\text{max}}$, since if $\hat{\sigma}^2(K)$ were extremely small, the numerical approximation of $u_a(K)$ and $u_b(K)$ could be ill-conditioned. The grid of $\delta$-values is constructed using a sequence of $T$ equidistant points on log scale from $\log u_a(K_{\text{max}})$ 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 empirically the validity of the procedure, we graphically illustrate in Figure 1 the behaviour of $l_p(K, \hat{\delta}) - l_p(K - 1, \hat{\delta})$ and $l_p(K, \hat{\delta}) - l_p(K + 1, \hat{\delta})$ as a function of log $\hat{\delta}$. For simplicity, we set $K^* = 10$, while fixing the number of observations at $m = 5, 000$ and 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 $\hat{\sigma}^2(K^*) = 0.3$.

Since $\hat{K}$ maximizes the penalized profile log-likelihood for a particular $\hat{\delta}$ value only when i) $l_p(K, \hat{\delta}) - l_p(\hat{K} + 1, \hat{\delta}) > 0$, and ii) $l_p(K, \hat{\delta}) - l_p(\hat{K} - 1, \hat{\delta}) > 0$ simultaneously, we can visualize this on a panel of plots where the differences in i) and ii) are plotted as functions of log $\hat{\delta}$ for a few possible $\hat{K}$’s including $K^*$. For $K^* = 10$, the penalized profile log-likelihood for $K = 9, 10$, and 11 were plotted in Figure 1. With little or no penalty, i.e., when log $\hat{\delta} < 10^{-5}$, $l_p(K, \hat{\delta}) - l_p(K + 1, \hat{\delta})$ can be shown to be negative and $l_p(K, \hat{\delta}) - l_p(K - 1, \hat{\delta})$ positive for all $K$. As log $\hat{\delta}$ 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. When both differences are positive, the numerically approximated $(\hat{a}_K, \hat{b}_K)$ for $K = 9, 10$, and 11 are formed and indicated by the vertical lines. It is easy to confirm visually that the approximations of the sets $\Delta_K$ are non-overlapping and $\hat{K}$ is indeed monotonically decreasing in $\hat{\delta}$. Notice that the log difference in $\hat{\delta}$ 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 $(\hat{a}_{K^*}, \hat{b}_{K^*})$.

Consequently, with the grid search for $\hat{\delta}$ constructed as proposed, the procedure would estimate the effective dimension by a majority vote, which is expected given the relationship between the log distance $\log(\hat{b}_K) - \log(\hat{a}_K)$ and the number of votes for that particular $\hat{K}$. 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 $\hat{\sigma}_1^2(K) = 0.7$. 

9
Figure 1: Difference in penalized profile log-likelihood as a function of log ($\tilde{\delta}$) with true $K^* = 10$. The difference in penalized log-likelihood for each choice of $K$ around $K^*$ for the two scenarios ($d^2(K^*) = 0.5$ and $d^2(K^*) = 1$) is illustrated in the panels. The solid and dashed curves represent the differences $l_p(k, \tilde{\delta}) - l_p(k-1, \tilde{\delta})$ and $l_p(k, \tilde{\delta}) - l_p(k+1, \tilde{\delta})$, respectively. The vertical lines capture the range of log ($\tilde{\delta}$) that both curves are positive, indicating the range for which $k = K^* - 1$, $k = K^*$, or $k = K^* + 1$ maximizes the penalized profile log-likelihood.
The scatterplots of the number of votes and $\log (\hat{b}_K) - \log (\hat{a}_K)$ shown in Figure 2 confirm that the effective dimension $K^*$ 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.

### 3 Simulation Studies

The purpose of our simulation studies is two-fold: first, to evaluate the performance of the proposed method as compared to alternative approaches; second, to explore limits of the proposed methods on simulated data that pose various challenges, including violations to normality and independence.

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 $\sigma^2(K^*) \in \{0.3, 0.7\}$ and $d_{K^*}^2 = \{0.5, 1\}$. 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 parameters $K^*$ and number of subjects to cluster ($n$) are often directly associated with the difficulty of the clustering problem. Thus, we kept $n = 100$ fixed while varying $K^*$. Lastly, the number of features or observations varies from $m = 50, 5,000$ to $10,000$.

#### 3.1 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 $\hat{K}_{\text{Cumlog}} = \arg \min_k \log \left( \frac{\sum_{i=1}^k \hat{\lambda}_i}{k} - \frac{1}{k} \sum_{i=1}^k \log \hat{\lambda}_i \right)$ and variance of sample eigenvalues (VarD), defined by $\hat{K}_{\text{VarD}} = \arg \max_k \frac{\sum_{i=1}^k \hat{\lambda}_i^2}{k} - \left( \frac{\sum_{i=1}^k \hat{\lambda}_i}{k} \right)^2$.

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.
Figure 2: Scatterplots of number of votes and $\log(\hat{b}_K) - \log(\hat{a}_K)$ with true $K^* = 10$. With the grid of $\delta$ values constructed as proposed, we approximated $a_K$ and $b_K$ for each potential choice of $K$. Each point represents the log difference in $\hat{a}_K$ and $\hat{b}_K$ plotted against the number of votes for that particular $K$. 
The performance was evaluated by the proportion of correct estimates over 100 replicates and the median of the estimates. Only methods with proportions of correct response of at least 50% in one of the scenarios were shown (Figure 3). When \( m = 5,000 \) or 10,000, even though PPPCA 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, PPPCA was more likely to underestimate slightly than overestimate across different scenarios. Most methods showed better performance under a larger number of observations \( m = 10,000 \) with the exception of BIC. In particular, Minka BIC outperformed BIC in all scenarios, but they both tended to overestimate and chose the maximum possible \( K \) when \( \sigma^2(K^*) = 0.3 \). Notably, among the “elbow” based approaches, Cumlog, was also adequate in the scenarios of a large \( d^2(K^*) \), which was expected considering that these methods were designed to capture clear trends in the size of eigenvalues. In addition, PESEL 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. However, we observed a large amount of variation in PESEL estimates across all scenarios. Lastly, GCV exhibited an overall strong performance except when \( \sigma^2(K^*) = 0.7 \) and \( d^2(K^*) = 0.5 \), in which it consistently underestimated \( K^* \) (Figure 4). When \( m = 50 \), all methods, including PPPCA, performed poorly (Web Supplementary C). In terms of computational time, PPPCA scales with \( n \) and thus was constant across the different numbers of observations \( Table \[ \text{II} \] \). The computational time for PESEL, GCV and PASSEMIER increased more quickly with \( m \) than PPPCA, BIC, Minka BIC or Cumlog, implying that these methods could become computationally expensive for large \( m \), which is not unusual in genome-wide data applications that motivated this study.

<table>
<thead>
<tr>
<th>( m )</th>
<th>PPPCA</th>
<th>BIC</th>
<th>Minka BIC</th>
<th>Cumlog</th>
<th>PESEL</th>
<th>PESEL1</th>
<th>PESEL2</th>
<th>PESEL3</th>
<th>GCV</th>
<th>PASSEMIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>106.254</td>
<td>0.107</td>
<td>0.416</td>
<td>0.011</td>
<td>0.459</td>
<td>0.225</td>
<td>0.752</td>
<td>0.773</td>
<td>1.793</td>
<td>0.487</td>
</tr>
<tr>
<td></td>
<td>(1.151)</td>
<td>(0.024)</td>
<td>(0.003)</td>
<td>(0.002)</td>
<td>(0.005)</td>
<td>(0.003)</td>
<td>(0.004)</td>
<td>(0.005)</td>
<td>(0.006)</td>
<td>(0.004)</td>
</tr>
<tr>
<td>5,000</td>
<td>128.192</td>
<td>0.124</td>
<td>0.418</td>
<td>0.012</td>
<td>30.661</td>
<td>N/A</td>
<td>3.162</td>
<td>3.292</td>
<td>149.434</td>
<td>19.735</td>
</tr>
<tr>
<td></td>
<td>(5.963)</td>
<td>(0.003)</td>
<td>(0.003)</td>
<td>(0.002)</td>
<td>(0.113)</td>
<td>N/A</td>
<td>(4.315)</td>
<td>(0.007)</td>
<td>(3.311)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>10,000</td>
<td>146.900</td>
<td>0.128</td>
<td>0.437</td>
<td>0.010</td>
<td>70.664</td>
<td>N/A</td>
<td>6.554</td>
<td>6.646</td>
<td>333.591</td>
<td>39.082</td>
</tr>
<tr>
<td></td>
<td>(4.845)</td>
<td>(0.006)</td>
<td>(0.003)</td>
<td>(1.551)</td>
<td>N/A</td>
<td>(0.003)</td>
<td>(0.137)</td>
<td>(0.118)</td>
<td>(9.298)</td>
<td>(0.616)</td>
</tr>
</tbody>
</table>

Table 1: Averaged computational time (in seconds) for each method on varying numbers of observations while fixing \( n = 100 \). Run time was recorded for analyses performed on a high-speed compute server with processor Intel(R) Xeon(R) CPU E7-8891 3.20GHz. The standard errors are shown in brackets.
Figure 3: Proportion of correctly estimated $K$ for $m = 10,000$. The shaded bars indicate the proportion of correctly estimated $K$ for each true $K^* = 5, 10$ or $20$ over 100 replicates against each method.
Figure 4: Boxplots of estimated \( K \) for \( m = 10,000 \). The box indicate the range of estimated \( K \) for each true \( K^* = 5, 10 \) or 20 over 100 replicates against each method.

### 3.2 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 \( d^2(K^*) = 0.5 \), \( \sigma^2(K^*) = 0.7 \), \( n = 100 \) and \( K^* = 10 \). We examined the estimated \( K \) as a function of the number of observations (\( m \)). The relationship between \( m \) and the estimated \( K^* \) over 10 replicates is shown in Figure 5. As \( m \) was increased, PPPCA, GCV and Cumlog estimates approached the correct value with PPPCA exhibiting faster convergence than Cumlog. In addition, our theoretical observation agrees with the simulation results that for a moderately sized \( m(>n) \), over-estimation of \( K^* \) was unlikely as shown by the solid line representing the median of PPPCA estimates over 10 replicates. Generally, GCV has more reliable performance and less variable estimates, but tended to underestimate \( K \), while
PPPCA estimates were slightly more variable at smaller values of $m$. The best “elbow” method Cumlog overestimated $K$ when $m$ was small, but approached the correct value at large values of $m$, which is expected since the average of last $n - K^*$ sample eigenvalues converges to $\sigma^2(K^*)$ when $m$ increases. In contrast, although PESEL has a similar profile to PPPCA, its estimates were visibly more variable at both smaller and larger values of $m$. The results also showed that BIC and Minka BIC had reasonable performance for moderate sizes of $m$, but, as the number of features increases, the estimates became unreliable for $m > 25,000$ with Minka BIC only slightly better than BIC. 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 $K = n - 1$.

![Figure 5: Estimated K as a function of the number of observations (m).](image)

3.3 Robustness

For many applications, the noise in the data might not be independent nor normally distributed. We first investigated cases where the observed error was drawn from a multivari-
ate normal distribution with correlation structure as typically observed in expressions data whereby the strength of correlation between pairs of genes depending on whether they are involved in similar sets of pathways. Since the features are order-invariant, we rearrange genes to have a block structure covariance driven by an autocorrelation process.

Let \( \nu \) denote the number of degrees of freedom for the \( t \)-distribution and \( \rho \) denote the autocorrelation coefficient. The error \( \epsilon_j \), for \( (j = 1, 2, \ldots, m) \), was simulated from

\[
\epsilon_j = \epsilon_{j-1} + \rho + r,
\]

where \( r \sim N(0, \sigma^2(K^*)I_n) \) or \( r \sim t(\nu, \sigma^2(K^*)I_n) \) was sampled independently for each \( j = 1, 2, \ldots, m \) from either a normal distribution or a Student’s \( t \)-distribution, respectively. In addition, we assumed the first \( K^* \) squared singular values decay at an exponential rate and were determined by varying \( d^2(K^*) = \{0.5, 1\} \), \( \nu \in \{3, 10\} \), \( \rho \in \{0, 0.1, 0.4, 0.7\} \) and \( \sigma^2(K^*) = \{0.3, 0.7\} \).

Intuitively, the \( t \)-distribution will induce fat tails in the distribution of the sample eigenvalues and concentrate most of the variance in the first a few eigenvalues. As the correlation coefficient between errors increased, the proportion of correctly estimated dimension over 10 replicates decreased for all methods (Figure 5), notably when \( \rho \) was greater than 0.4. For \( \rho = 0.1 \), the results were similar to those under the scenario without any autocorrelation. For the more difficult scenarios of \( \rho = 0.4 \) and \( \rho = 0.7 \), PPPCA and GCV had the best performance both in terms of the proportion of correctly identified \( K \) and the mean value of the estimated \( K \) across 10 replicates (additional results are included in Web Supplementary Material).
Figure 6: Proportion of correctly estimated $K$ for data with autocorrelated noise. The shaded bars indicate the proportion of correctly estimated $K$ for each true $K^* = 5, 10$ or $20$ over $10$ replicates against each method.

When the noise followed a multivariate $t$-distribution with degrees of freedom $\nu = 3$, all methods failed completely except PPPCA, PESEL, and GCV, with GCV being the most robust and PPPCA capturing only a small percentage of replicates (results are reported in Web Supplementary Material). When $\nu = 10$, PPPCA improved for large $d^2(K^*)$ and $K^* = 10$, but did less well under smaller $d^2(K^*)$ and $K^* = 5$ (Figure 7).
Figure 7: Proportion of correctly estimated $K$ for data with fat-tailed noise. The proportion of correct selection for each true $K^* = 5$ and 10 were shown in shaded bars against each method.

Both autocorrelation and fatter tails of the eigenvalues distribution present serious challenges for all approaches considered, with GCV and PESEL being the most robust. Naturally, the overestimation of $K$ resulted in an underestimation of the residual variance $\sigma^2(K^*)$, which can affect all methods under comparison. It is unsurprising to see that most methods failed completely when $\rho > 0.7$ and $\nu = 3$, with the exception of PPPCA, PESEL, and GCV in some of the scenarios considered (Figure 8). However, reasonable performances from PPPCA and GCV could be expected when $\nu > 10$ and $\rho < 0.4$ (Figure 8). Notably, even when PPPCA estimates were incorrect, they did not grossly overestimate the true $K^*$ as can be seen from the averaged estimates over 10 replicates (Web Supplementary Material C).
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 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-

Figure 8: Proportion of correctly estimated $K$ for data with autocorrelated and fat-tailed noise. The proportion of correct selection for each true $K^* = 5$ and 10 were shown in shaded bars against each method.
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 availability and quality control

The three gene expression datasets can be obtained publicly via the Gene Expression Omnibus (GEO) from NCBI or ArrayExpress from EBI (http://ebi.ac.uk). Pre-processed datasets were used in the present analyses. We have transformed measurements at each probe to a standard normal distribution and filtered based on a minimum probe variance of 0.3. Here each probe measures the levels of expression of a specific DNA sequence, and can be regarded as a feature or observation.

4.3 Data Analyses

Consider \( m \) expression measurements from \( n \) individuals \( \mathbf{x} = [x_1, \ldots, x_n] \in \mathbb{R}^{m \times n} \), where \( m \) is typically in the tens of thousands and the number of individuals \( n \) in the hundreds. The estimate \( \tilde{K} \) was then used in clustering algorithms. Alternative methods tested include BIC, 5-fold cross-validation (CV), Minka’s BIC version using Laplace approximation, PESEL, PASSEMIER, GCV, and Cumlog. For the 5-fold CV, computational time constraints limit the range to \( K \leq 30 \). Since PESEL has three variations derived under different asymptotics, we report the one derived under infinite \( m \) and fixed \( n \). The summary of results are shown in Table 2.

Example 1. Expression data from breast cancer subtypes: The expression measurements were collected on tissues from 153 patients with primary invasive breast cancer [13]. The four cancer subtypes defined clinically were (number in parenthesis indicate the number of patients with that type of cancer) Triple-negative (TN; 55), human epidermal growth factor receptor 2 (HER2; 39), Luminal A (29) and Luminal B (30) tumour patients, as well as two types of controls in the form of 11 normal tissue samples and 14 cell lines. A total of 29,874 genes probes were included as features, after applying the filtering criterion \( m = 23,897 \) gene probes remained. Our penalized approach estimated \( \tilde{K} = 8 \), with no other competing method giving any sensible estimates (Table 3).

It is expected that some of the subtypes of breast cancer would be clustered together for any choice of the clustering algorithm. Also, the normal tissue and the cell line derived...
Breast NCI60 Ovarian

Datasets

<table>
<thead>
<tr>
<th></th>
<th>Breast</th>
<th>NCI60</th>
<th>Ovarian</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>23,897</td>
<td>40,911</td>
<td>43,694</td>
</tr>
<tr>
<td>n</td>
<td>178</td>
<td>132</td>
<td>285</td>
</tr>
<tr>
<td>expert K</td>
<td>6</td>
<td>9</td>
<td>6</td>
</tr>
</tbody>
</table>

Results (K)

<table>
<thead>
<tr>
<th></th>
<th>PPPCA</th>
<th>CV5</th>
<th>BIC</th>
<th>Minka BIC</th>
<th>PESEL</th>
<th>Passemir</th>
<th>GCV</th>
<th>Cumlog</th>
<th>VarD</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>8</td>
<td>30</td>
<td>154</td>
<td>154</td>
<td>153</td>
<td>102</td>
<td>72</td>
<td>41</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2: A summary of gene expression datasets and estimated K by each method. The expert choice was determined by the relevant literature, usually estimated using a subset of the gene probes; the maximum K searched for CV5 was set to 30; the K reported for PESEL was based on the asymptotic assumption of large m.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Molecular Subtypes</th>
<th>Normal Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TN</td>
<td>HER2</td>
</tr>
<tr>
<td>1</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>34</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3: K-means Clustering using $\hat{K} = 8$. Clustering results using PPPCA estimated $\hat{K} = 8$ in 178 breast cancer patients over expression measured on 29,809 gene probes.
samples are distinctive enough that they would always be assigned their own clusters. A simple mixture clustering algorithm using EM reveals that two of the identified clusters belong to TN, one belong to HER2, one for Luminal A and one for Luminal B (Table 3). 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, HER2, 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 have no overlap. In other words, clustering these subtypes with gene 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 [34], and has been recently profiled using array technology for $n = 132$ individuals on $m = 41,000$ gene probes [26]. Our penalized approach estimated $\tilde{K} = 16$ for the $m = 40,911$ gene probes that remained after cleaning. No other competing methods gave reasonable estimates with the exception of an “elbow” based method - VarD, which suggested $K = 8$, which was likely an underestimate. Though the PPPCA estimate is larger than the number of organism cell lines, but is within reasonable range as most of the cancers have known subtypes. Cluster assignments using a Gaussian mixture model showed a dominant cluster that contains all cancer types (cluster size > 50). Out of the 16 clusters identified, three clusters of size 4, 2, and 2 uniquely correspond to Colon cancer, two clusters of size 4 and 4 uniquely correspond to Leukemia, three clusters of size 10, 4, and 4 uniquely correspond to Melanoma (Table 4).

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 [43]. The study discovered six subtypes defined by molecular pattern in $m = 8,732$ 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 [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 $m = 43,694$ gene probes, our PPPCA estimated $\tilde{K} = 9$ and a competing method VarD suggested $\hat{K} = 11$, while all other methods did not provide a reasonable estimate.

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
<table>
<thead>
<tr>
<th></th>
<th>Breast</th>
<th>CNS</th>
<th>Colon</th>
<th>Leukemia</th>
<th>Melanoma</th>
<th>NSCL</th>
<th>Ovarian</th>
<th>Prostate</th>
<th>Renal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>12</td>
<td>10</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 4: **K-means Clustering using $\tilde{K} = 16$**. Clustering results using PPPCA estimated $\tilde{K} = 16$ in 132 cancer patients over expression measured on 29,879 gene probes. CNS: Central nervous system; NSCL: Non-Small Cell Lung.
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 $\tilde{K}$ can be estimated based on important features with non-zero loadings.

**Acknowledgements**

The authors thank Andrey Feuerverger for helpful suggestions, Lei Sun and Stanislav Volgushev for a critical reading of the original version of the paper, and an anonymous referee for very useful comments that improved the presentation of the paper. Wei Q. Deng is supported by a NSERC doctoral graduate scholarship.

**Appendix**

5.1 Notations

- $n$: number of subjects or items to be clustered.
- $m$: number of observations or features.
- $K$: number of effective dimensions, $K \in \{1, 2, \ldots, n-1\}$.
- $p(k) = nk + 1 - \frac{k(k-1)}{2}$: number of free parameters associated with the model specified by $k$ principal components (PCs), where $k \in \{1, 2, \ldots, n-1\}$.
- $\sigma^2(k)$: the error variance associated with discarding the $n-k$ last PCs.
- $l_p(k)$: the profile log-likelihood conditional on $K = k$.
- $\delta$: the tuning parameter associated with the penalty in the profile log-likelihood with a theoretical range $[0, n-1)$.
- $\tilde{\delta} = \frac{\delta}{n}$: the scaled tuning parameter associated with the penalty in the profile log-likelihood with a theoretical range $[0, 1 - \frac{1}{n})$.
- $l_p(k; \tilde{\delta})$: the penalized profile log-likelihood conditional on $K = k$ with a scaled tuning parameter value $\tilde{\delta}$.
- $a_k = \min\{\tilde{\delta} : l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta}) \geq 0\}$, where $k \in \{1, 2, \ldots, n-1\}$.
- $b_k = \max\{\tilde{\delta} : l_p(k; \tilde{\delta}) - l_p(k-1; \tilde{\delta}) \geq 0\}$, where $k \in \{1, 2, \ldots, n-1\}$.
- $\Delta_k = (a_k, b_k)$: the range of scaled tuning parameter values that $k$ maximized the penalized profile log-likelihood.
- $\Delta_k \cap \Delta_k' = \emptyset$ for $k \neq k'$, i.e. the sets of scaled tuning parameters that maximized each $k$ are disjoint.
\[ \Delta = \bigcup_{k=1}^{n-1} \Delta_k \]

- \( u_b(k) \): the lower bound of \( b_k \) from solving the inequality, \( l_p(k; \tilde{\delta}) - l_p(k - 1; \tilde{\delta}) \geq 0 \), approximated by Taylor series.

- \( u_a(k) \): the upper bound of \( a_k \) from solving the inequality, \( l_p(k; \tilde{\delta}) - l_p(k + 1; \tilde{\delta}) \geq 0 \), approximated by Taylor series.

### 5.2 An Algorithm to Estimate the Effective Dimension

We give a summary of the steps to estimate the effective dimension given a dataset \( x_{n \times m} \):

**Step 1:** \( x_{n \times m} \) is standardized across all features for each observation \( i \) so that the sample covariance matrix \( s_n \) has diagonal variance 1.

**Step 2:** Compute the sample eigenvalues \( \hat{\lambda}_1, \ldots, \hat{\lambda}_n \) and eigenvectors \( u \) from the sample covariance matrix \( s_n \).

**Step 3:** For each possible choice of \( k = 1, \ldots, n - 1 \), compute the PPCA MLE of \( \sigma^2 \) according to the closed form representation \( \hat{\sigma}^2(k) = \frac{\sum_{i=k+1}^{n} \hat{\lambda}_i}{n-k} \).

**Step 4:** For each possible choice of \( k = 1, \ldots, n - 1 \), determine the upper and lower bound for the scaled tuning parameter as described, \((u_a(k_{\text{max}}), u_b(1)) \subset (0, (1 - \frac{1}{n})(1 - \hat{\sigma}^2(1)))\), where \( k_{\text{max}} \) is determined according to Lemma 3.

**Step 5:** For the interval of possible scaled tuning parameter values defined by \((u_a(k_{\text{max}}), u_b(1))\), a user-defined number of votes \( T \) is used to construct a geometric sequence of \( \tilde{\delta} \) values that form the search grid.

**Step 6:** For each \( \tilde{\delta} \) in the sequence, find \( k \in \{1, 2, 3, \ldots, n - 1\} \) that maximizes the corresponding penalized profile log-likelihood. The effective dimension is estimated by the \( k \) that has the highest vote counts over all values in the search grid, and denoted by \( \tilde{K} \).

Conditional on the estimated effective dimension \( \tilde{K} \), an EM algorithm has been proposed by [41] to estimate the PPCA MLE \( \hat{W}_{n \times \tilde{K}} \), which can help identify cluster memberships using the desired clustering algorithm, such as K-means or mixture models.

#### 5.2.1 Information Criteria

Let \( p = nK + 1 - \frac{K(K-1)}{2} \) denote the number of free parameters in the probabilistic model, we can select the rank by minimizing the following commonly used model selection criteria:

1. \( \text{AIC}(K) = -2l_p(K) + 2p \)
2. \( \text{BIC}(K) = -2l_p(K) + \log (m)p \).

### 5.3 Hypothesis Testings on the Equality of the Last \( n-K \) Eigenvalues

The null hypothesis of interest is \( H_0 : \lambda_j = \lambda_{j+1} = \ldots, = \lambda_n \) against the alternative hypothesis that at least one is not equal to the remaining eigenvalues. The test statistics
\[ \chi^2 = c[(n - j) \log \left( \frac{\sum_{i=j}^{n} \hat{\lambda}_i}{n - j} \right) - \sum_{i=j}^{n} \log \hat{\lambda}_i], \quad (19) \]

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

### 5.4 Various Measures for Detection of Elbows in Sample Eigenvalues

The sample eigenvalues from the covariance or the similarity matrix are denoted by \( \lambda_1, \lambda_2, \ldots, \lambda_n \). The “elbow” can be estimated by either maximizing or minimizing the following measures with estimates denoted by \( \hat{K} \):

- **AdjD (ratio of adjacent eigenvalues):** \( \hat{K} = \arg \min_j \frac{\lambda_j}{\lambda_j + 1} \).
- **CumD (ratio of cumulative mean of eigenvalues):** \( \hat{K} = \arg \max_j \frac{\sum_{i=1}^{j} \hat{\lambda}_i}{\sum_{i=1}^{n} \hat{\lambda}_i} \).
- **VarD (variance of cumulative eigenvalues):** \( \hat{K} = \arg \max_j \frac{\sum_{i=1}^{j} \hat{\lambda}_i^2}{\sum_{i=1}^{j} \hat{\lambda}_i} - (\frac{\sum_{i=1}^{j} \hat{\lambda}_i}{\sum_{i=1}^{n} \hat{\lambda}_i})^2 \).
- **Cumlog (log cumulative mean):** \( \hat{K} = \arg \min_j \log \frac{\sum_{i=1}^{j} \hat{\lambda}_i}{\sum_{i=1}^{n} \hat{\lambda}_i} - \sum_{i=1}^{j} \log \hat{\lambda}_i \).
- **Sigma2 (log of the PPCA MLE \( \hat{\sigma}_K^2 \)):** \( \hat{K} = \arg \max_j (n-j) \log \frac{\sum_{i=1}^{n} \hat{\lambda}_i}{n-j} \).

### 5.5 Proof of Proposition 1

**Proof.** Since the input data can be assumed to have been standardized, the sample eigenvalues sum to \( n \), i.e. \( \sum_{i=1}^{n} \hat{\lambda}_i = n \).

First, we observe the following relationships between the sample eigenvalues and the maximum likelihood estimate of \( \sigma^2 \) given \( K \in \{k, k+1, \ldots, n\} \) under the PPCA model [42]:

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

Notice that the sample eigenvalues are decreasing \( \hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \hat{\lambda}_k > \cdots > \hat{\lambda}_n > 0 \), we must have \( \hat{\sigma}^2(1) > \hat{\sigma}^2(2) > \cdots > \hat{\sigma}^2(k) > \cdots, \hat{\sigma}^2(n-1) \).
Put the two pieces together, we have
\[ \hat{\lambda}_k = (n-k+1)\hat{\sigma}^2(k-1) - (n-k)\hat{\sigma}^2(k) > \hat{\sigma}^2(k-1). \]

Then, we start with \( K = n - 1 \) since when \( K = n \), the model is of full rank and the profile log-likelihood based on \( \mathbf{W}_n \) results in \( \hat{\sigma}^2(n) = 0 \), and thus is ill-defined.

The profile log-likelihood at \( K = n - 1 \) can be simplified to:
\[ l_p(n - 1) = -\frac{m}{2} \sum_{i=1}^{n} \log \lambda_i. \]

We show the difference between \( l(n - 1) \) and \( l(n - 2) \) is non-negative:
\[
l_p(n - 1) - l_p(n - 2) = -\frac{m}{2} \left[ \log \hat{\lambda}_{n-1} + \log(\hat{\sigma}^2(n-1)) - 2\log(\hat{\sigma}^2(n-2)) \right]
\]
\[
= -\frac{m}{2} \left[ \log(2 - \frac{\hat{\sigma}^2(n-1)}{\hat{\sigma}^2(n-2)}) + \log(\frac{\hat{\sigma}^2(n-1)}{\hat{\sigma}^2(n-2)}) \right]
\]
\[
= -\frac{m}{2} \left[ \log(1 + 1 - \frac{\hat{\sigma}^2(n-1)}{\hat{\sigma}^2(n-2)}) + \log(1 + \frac{\hat{\sigma}^2(n-1)}{\hat{\sigma}^2(n-2)} - 1) \right]
\]
\[
> 0,
\]
by making use of the logarithmic inequality: \( \frac{x}{1+x} < \log(1 + x) < x \) for \( x > -1 \), with
\[
-\log(1 + 1 - \frac{\hat{\sigma}^2(n-1)}{\hat{\sigma}^2(n-2)}) > -1 + \frac{\hat{\sigma}^2(n-1)}{\hat{\sigma}^2(n-2)},
\]
and
\[
-\log(1 + \frac{\hat{\sigma}^2(n-1)}{\hat{\sigma}^2(n-2)} - 1) = \log(1 + \frac{\hat{\sigma}^2(n-2)}{\hat{\sigma}^2(n-1)} - 1) > 1 - \frac{\hat{\sigma}^2(n-1)}{\hat{\sigma}^2(n-2)}
\]

For an arbitrary choice of \( K = k \), where \( k \in \{1, 2, \ldots, n - 1\} \), we can similarly obtain the lower bound using the logarithmic inequalities:
\[
l_p(k) - l_p(k-1) = -\frac{m}{2} \left( \log \hat{\lambda}_k + (n-k) \log(\hat{\sigma}^2(k)) - (n-k+1) \log(\hat{\sigma}^2(k-1)) \right)
\]
\[
> -\frac{m}{2} \left( (n-k)(1 - \frac{\hat{\sigma}^2_k}{\hat{\sigma}^2_{k-1}}) + (n-k)(\frac{\hat{\sigma}^2_k}{\hat{\sigma}^2_{k-1}} - 1) \right)
\]
\[
= 0
\]

On the other hand, the difference between \( l_p(k) - l_p(k+1) \) is less or equal to zero
because of symmetry:

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

\[ < \frac{m}{2} \left( (n-k-1)(\frac{\hat{\sigma}^2(k+1)}{\hat{\sigma}^2(k)} - 1) + (n-k-1)(1 - \frac{\hat{\sigma}^2(k+1)}{\hat{\sigma}^2(k)}) \right) \]

\[ = 0 \]

Together, these results show that for any \( k = \{1, 2, \ldots, n-2\} \), \( l_p(k) - l_p(k+1) < 0 \); and for any \( k = \{2, 3, \ldots, n-1\} \), \( l_p(k) - l_p(k-1) > 0 \).

Since \( K = k \) is arbitrary, \( l_p(1) < l_p(2) < \ldots l_p(k) < l_p(k+1) < l_p(n-1) \), therefore the profile log-likelihood is a monotonically non-decreasing function of \( K \), with the maximum achieved at \( K = n-1 \).
6 Proof of Proposition 2

The true effective dimension can take on any values in \( \{1, 2, \ldots, n\} \), and we first discuss the case when \( K^* \) falls on the boundary of this search space, i.e. \( K^* = 1 \) and \( K^* = n - 1 \). Notice that we excluded the case when \( K = n \) for the same reason as discussion in proof of Proposition 1. Then, we discuss the second case of \( K^* \) falling in \( \{2, 3, \ldots, n - 2\} \).

First, the scaled tuning parameter can take any value on \([0, \frac{n - 1}{n})\) in theory. However, for each possible choice of \( K \), we would like to restrict it to a subset \((0, (\frac{1}{K} - \frac{1}{n})(1 - \tilde{\sigma}^2(K)))\) such that the penalized residual variance estimate \( \tilde{\sigma}^2(K) < 1 \). For ease of discussion, define \( g(K) = (\frac{1}{K} - \frac{1}{n})(1 - \tilde{\sigma}^2(K)) \). This condition turns out to be sufficient for the penalized profile log-likelihood \( l_p(k; \tilde{\delta}) \) to be a smooth and monotonically decreasing function of \( \tilde{\delta} \in (0, g(k)) \) at any \( K = k \) (Lemma 1).

Second, noticing the relationship between \( \tilde{\delta} \) and \( K \) and its implication on the penalized log-likelihood, we first focus on the discussion of \( K = \{2, 3, \ldots, n - 2\} \) and then move on to the boundary points of \( K = 1 \) and \( K = n - 1 \).

Restricting ourselves to \( \tilde{\delta} \in (0, g(K)) \cup (0, g(K - 1)) \cup (0, g(K + 1)) \), we can show that \( l_p(K; \tilde{\delta}) - l_p(K + 1; \tilde{\delta}) \) is a smooth and concave function of \( \tilde{\delta} \) and is increasing in the small interval near 0 (Lemma 2). Due to the symmetry in constructing \( l_p(K; \tilde{\delta}) - l_p(K - 1; \tilde{\delta}) \) and \( l_p(K; \tilde{\delta}) - l_p(K + 1; \tilde{\delta}) \), it is clear that \( l_p(K; \tilde{\delta}) - l_p(K - 1; \tilde{\delta}) \) is a smooth and convex function of \( \tilde{\delta} \) and is decreasing in the small interval near 0.

Since when \( \tilde{\delta} = 0 \), we have

\[
l_p(K; \tilde{\delta} = 0) - l_p(K + 1; \tilde{\delta} = 0) < 0,
\]

and

\[
l_p(K; \tilde{\delta} = 0) - l_p(K - 1; \tilde{\delta} = 0) > 0,
\]

combined with the property of convex and concave functions, for any \( K \), we can define

\[
\begin{align*}
\{ a_K & = \text{min}\{\tilde{\delta} : l(K; \tilde{\delta}) - l(K + 1; \tilde{\delta}) > 0\}; \\
b_K & = \text{max}\{\tilde{\delta} : l(K; \tilde{\delta}) - l(K - 1; \tilde{\delta}) > 0\}. \\
\end{align*}
\]

and \( \Delta_K = (a_K, b_K) \), whenever \( a_K < b_K \), otherwise \( \Delta_K = \emptyset \). Following this definition, for any \( \tilde{\delta}_o \in \Delta_K \), we must have \( l_p(K; \tilde{\delta}) - l_p(K + 1; \tilde{\delta}) > 0 \) and \( l_p(K; \tilde{\delta}) - l_p(K - 1; \tilde{\delta}) > 0 \).

When \( K = n - 1 \), the penalized profile log-likelihood \( l_p(K; \tilde{\delta} = 0) \) is clearly maximized at \( K = n - 1 \) and \( \tilde{\delta} = 0 \) following from Proposition 1. And we can take \( a_{n-1} = 0 \). For \( \tilde{\delta} \in (0, g(n-1)[\tilde{\lambda}_n + (n-1)\tilde{\lambda}_{n-1}]) \), we can show \( l_p(n-1; \tilde{\delta}) - l_p(n-2; \tilde{\delta}) \) to be a smooth and monotonically decreasing function of \( \tilde{\delta} \), which implies that \( l_p(n-1; \tilde{\delta}) - l_p(n-2; \tilde{\delta}_o) > 0 \) for some \( \tilde{\delta}_o \).

When \( K = 1 \), for \( \tilde{\delta} \in (0, (1 - \frac{1}{n})(1 - \tilde{\sigma}^2(1))) \), we can show \( l_p(1; \tilde{\delta}) - l_p(2; \tilde{\delta}) \) to be a smooth and monotonically increasing function of \( \tilde{\delta} \) as long as \( \tilde{\sigma}_1^2 > \tilde{\sigma}_2^2 \), which implies that \( l_p(1; \tilde{\delta}) - l_p(2; \tilde{\delta}_o) > 0 \) for some \( \tilde{\delta}_o \in (0, \frac{n\tilde{\lambda}_2}{(\tilde{\lambda}_2 - \tilde{\sigma}_2^2)(n-2)} - 1) \).
Third, for some \( \tilde{\delta}_o \), we show the sufficient conditions to claim \( l_p(K; \tilde{\delta}_o) \) is maximized at \( K' \). Again, we first focus on the discussion of \( K = \{2, 3, \ldots, n - 2\} \) and then move on to the boundary points of \( K = 1 \) and \( K = n - 1 \).

The condition for \( K = \{2, 3, \ldots, n - 2\} \) is summarized by two inequalities: for some \( \tilde{\delta}_o \), if \( l_p(K'; \tilde{\delta}_o) - l_p(K' + 1; \tilde{\delta}_o) \) and \( l_p(K'; \tilde{\delta}_o) - l_p(K' - 1; \tilde{\delta}_o) \), then \( K' \) maximizes \( l_p(K; \tilde{\delta}_o) \) (Lemma 3). In other words, the two inequalities are sufficient condition to ensure \( K_{kk} \) is not clear. \( K < K \) is unknown for \( K \) (Lemma 5), which implies that for data with a large number of features or observations, \( K \) always select showing a subset of \( \Delta K \). Again, we first focus on the discussion of the upper bound for \( \approx K \delta \) a the upper bound for \( \approx K \delta \) a

To further show that the search grid we proposed using a geometric sequence would always \( \max \) \( l_p(K; \tilde{\delta}) \) for some \( \tilde{\delta}_o \in \Delta n_1 \) whenever \( \Delta n_1 \neq \emptyset \), we use the telescoping sum to show,

\[
l_p(n - 1; \tilde{\delta}) - l_p(kk; \tilde{\delta}) = l_p(n - 1; \tilde{\delta}) - l_p(n - 2; \tilde{\delta}) + \cdots + l_p(kk + 1; \tilde{\delta}) - l_p(kk; \tilde{\delta}) > 0
\]

any \( kk < n - 2 \) (Lemma 3). Similarly, to show that \( K = 1 \) maximizes \( l_p(K; \tilde{\delta}) \) for some \( \tilde{\delta}_o \in \Delta 1 \), we use the telescoping sum to show

\[
l_p(1; \tilde{\delta}) - l_p(kk; \tilde{\delta}) = l_p(1; \tilde{\delta}) - l_p(2; \tilde{\delta}) + \cdots + l_p(kk - 1; \tilde{\delta}) - l_p(kk; \tilde{\delta}) > 0
\]

any \( kk > 1 \) (Lemma 3).

Finally, we just need to establish \( \Delta K \) when \( K = K^* \) and show that \( \Delta K^* \) is not empty by showing a subset of \( \Delta K^* \) is not empty (Lemma 4, 5) to complete the proof of Proposition 2.

To further show that the search grid we proposed using a geometric sequence would always select \( K^* \) with large enough \( \delta \), instead of \( K \), we use the telescoping sum to show

\[
l_p(K; \tilde{\delta}) - l_p(kk; \tilde{\delta}) = l_p(K; \tilde{\delta}) - l_p(2; \tilde{\delta}) + \cdots + l_p(kk - 1; \tilde{\delta}) - l_p(kk; \tilde{\delta}) > 0
\]

any \( kk > 1 \) (Lemma 3).

Since closed-form solutions to \( a_K \) and \( b_K \) are not analytically available for any \( K \), we approximated them by first order Taylor expansion at \( \tilde{\delta} = 0 \) (Lemma 4). Let \( u_a(K) \) denote the upper bound for \( a_K \), and \( u_b(K) \) the lower bound for \( b_K \). Then, we have \( \frac{b_K}{a_K} > \frac{u_b(K)}{u_a(K)} \).

We can instead show that \( \lim \max \arg \frac{u_b(K)}{u_a(K)} = \frac{u_b(K^*)}{u_a(K^*)} \) 1 (Lemma 5).

We also showed that for \( K > K^* \), \( |u_b(K) - u_a(K)| \) converges to 0 when \( m \to \infty \) (Lemma 5), which implies that for data with a large number of features or observations, over-estimation of \( K^* \) is unlikely. However, since the exact relationship of \( \frac{u_b(K)}{u_a(K)} \) at large \( m \) is unknown for \( K < K^* \), the impact of this on under-estimation of the effective dimension is not clear.
6.1 Proof of Lemma 1

**Lemma 1.** Consider a collection of samples \( \mathbf{x} \in \mathbb{R}^{m \times n} \) with each row following a multivariate Gaussian distribution \( N(0, \mathbf{W} \mathbf{W}^T + \sigma^2 \mathbf{I}) \). Suppose the sample covariance matrix of \( \mathbf{x}^T \) is positive semi-definite and \( K^* \) is the true number of PCs or rank of \( \mathbf{W} \). Then, the penalized profile log-likelihood at each possible choice of \( K \in \{1, 2, \ldots, n-1\} \) is a smooth function of \( \tilde{\delta} \) on the interval \((0, \frac{1}{K} - \frac{1}{n})\) and is monotonically decreasing on \((0, (\frac{1}{K} - \frac{1}{n})(1 - \tilde{\sigma}^2(K)))\).

**Proof.** To show that the penalized profile log-likelihood is a smooth function of the scaled tuning parameter \( \tilde{\delta} \) for each \( K \), we need to show that it is differentiable with respect to \( \tilde{\delta} \) and all derivatives of \( l_p(K, \tilde{\delta}) \) exist.

The penalized profile log-likelihood is
\[
l_p(K; \tilde{\delta}) = l_p(K) - \frac{m}{2} \{ (n-K) \log(1 - \frac{K}{n}) \\
- nK\tilde{\delta} \left[ 1 + \log(\tilde{\sigma}^2(K)) + \log(1 - \frac{K}{n}) \right] \\
+ \delta n K \log(1 - \frac{K}{n} - K\tilde{\delta}) - (n-K) \log(1 - \frac{K}{n} - K\tilde{\delta}) \} \quad (21)
\]
and the first order derivative with respect to \( \tilde{\delta} \) is:
\[
\frac{\partial l_p}{\partial \tilde{\delta}} = -\frac{mnK}{2} \log \left[ \frac{1 - \frac{K}{n} - K\tilde{\delta}}{(1 - \frac{K}{n})\tilde{\sigma}^2(K)} \right]
\]
the second order derivative with respect to \( \tilde{\delta} \) is:
\[
\frac{\partial^2 l_p}{\partial \tilde{\delta}^2} = \frac{mnK}{2} \frac{K}{1 - \frac{K}{n} - K\tilde{\delta}}
\]
The \( t \)th order derivative is then
\[
\frac{\partial^t l_p}{\partial \tilde{\delta}^t} = \frac{mnK}{2} (-1)^t \frac{(t-2)!K^{t-1}}{(1 - \frac{K}{n} - K\tilde{\delta})^{t-1}}. \quad (22)
\]
Since (22) is a rational function of \( \tilde{\delta} \) on \((0, \frac{1}{K} - \frac{1}{n})\), all derivatives of \( l_p(K, \tilde{\delta}) \) exist, and thus \( l_p(K, \tilde{\delta}) \) is a smooth function of \( \tilde{\delta} \).

In addition, we can show that \( l_p(K; \tilde{\delta}) \) for any \( K \in \{1, 2, \ldots, n-1\} \) is a monotonically decreasing function of \( \tilde{\delta} \) on \((0, (\frac{1}{K} - \frac{1}{n})(1 - \tilde{\sigma}^2(K)))\). The first order derivative is negative when:
\[
\log(1 - \frac{K}{n} - K\tilde{\delta}) - \log(1 - \frac{K}{n}) - \log \tilde{\sigma}^2(K) > 0 \\
\tilde{\delta} < (\frac{1}{K} - \frac{1}{n})(1 - \tilde{\sigma}^2(K)) \quad (23)
\]
and equals to zero when \( \tilde{\delta} = \left( \frac{1}{K} - \frac{1}{n} \right) \left( 1 - \tilde{\sigma}^2(K) \right) \). Thus, \( l_p(K; \tilde{\delta}) \) is a monotonically decreasing function of \( \tilde{\delta} \) for any \( K \in \{1, 2, \ldots, n - 1\} \) on \( (0, \left( \frac{1}{K} - \frac{1}{n} \right) (1 - \tilde{\sigma}^2(K))) \), this range also ensures that \( \tilde{\sigma}^2(K) < 1 \).

For ease of discussion, define \( g(K) = \left( \frac{1}{K} - \frac{1}{n} \right) (1 - \tilde{\sigma}^2(K)) \). And since the first part of \( g(K) \) decreases with increasing \( K \) and the second part increases with increasing \( K \), we do not assume any monotonic trend in \( g(K) \).

The penalization is based on the principal that a larger value of \( \tilde{\delta} \) is associated with a smaller \( l_p(K; \tilde{\delta}) \) for each \( K \), and instead of the theoretical range \( \cup_{i=1}^{n-1} (0, \frac{1}{K} - \frac{1}{n}) = (0, 1 - \frac{1}{n}) \), a practical range for \( \tilde{\delta} \) is taken to be \( \cup_{K=1}^{n-1} (0, g(K)) = (0, \max_K g(K)) \).

\[ \]
6.1.1 Proof of Lemma 2

Lemma 2. For $K \in \{2, 3, \ldots, n-2\}$, $l(K; \tilde{\delta}) - l(K+1; \tilde{\delta})$ and $l(K; \tilde{\delta}) - l(K-1; \tilde{\delta})$ are both smooth functions of $\tilde{\delta} \in (0, g(K-1)) \cup (0, g(K)) \cup (0, g(K+1))$, and are respectively concave and convex.

Proof. Lemma 1 suggests that $l_p(k; \tilde{\delta})$ is smooth and monotonically decreasing on $\tilde{\delta} \in (0, g(k))$. Since the difference of the two penalized profile log-likelihood at a fixed $K = k$ is a simple function, $l_p(k; \tilde{\delta}) - l_p(k-1; \tilde{\delta})$ is smooth on $(0, g(k)) \cup (0, g(k-1))$. And similarly, $l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta})$ is smooth on $(0, g(k)) \cup (0, g(k+1))$.

Next, we want to know the behaviour of $l_p(k; \tilde{\delta}) - l_p(k-1; \tilde{\delta})$ and $l_p(k; \tilde{\delta}) - l_p(k+1; \tilde{\delta})$ on $(0, g(k-1)) \cup (0, g(k)) \cup (0, g(k+1))$.

Condition 2.1. $l_p(k; \tilde{\delta}) - l_p(k-1; \tilde{\delta})$ is convex on $\tilde{\delta} \in (0, (\frac{1}{k} - \frac{1}{n})(1 - \hat{\sigma}^2(k-1))) \subset (0, g(k)) \cup (0, g(k-1))$.

Since the difference is a smooth function, we can compute the first order derivative with respect to $\tilde{\delta}$:

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

\[
= \frac{mnk}{2} \log \hat{\sigma}^2(k) - \frac{mn(k-1)}{2} \log \hat{\sigma}^2(k-1)
\]

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

(24)

Since $\log \hat{\sigma}^2(k) - \log \hat{\sigma}^2(k-1)$ could be either greater than or less than 0 depending on whether $\frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k-1)}$ is greater or equal to 1, we can look at the behaviour of

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

and it’s clear from its first order derivative $\frac{n}{(n-k-k\tilde{\delta})^2}$ constant $> 0$ that it is increasing in $\tilde{\delta}$, we can then bound it by

\[
\frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k-1)} \leq \left[ \frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k-1)} = \left(1 - \frac{1}{n} \right)(1 - \hat{\sigma}^2(k-1)) < 1
\]

where $(\frac{1}{k} - \frac{1}{n})(1 - \hat{\sigma}^2(k-1)) < \frac{n}{(n-k-k\tilde{\delta})^2} \text{ constant} > 0$ that it is increasing in $\tilde{\delta}$, we can then bound it by

\[\frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k-1)} \leq \left[ \frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k-1)} = \left(1 - \frac{1}{n} \right)(1 - \hat{\sigma}^2(k-1)) \right.\]

where $(\frac{1}{k} - \frac{1}{n})(1 - \hat{\sigma}^2(k-1)) < \frac{n}{(n-k-k\tilde{\delta})^2} \text{ constant} > 0$ that it is increasing in $\tilde{\delta}$, we can then bound it by

\[\frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k-1)} \leq \left[ \frac{\hat{\sigma}^2(k)}{\hat{\sigma}^2(k-1)} = \left(1 - \frac{1}{n} \right)(1 - \hat{\sigma}^2(k-1)) \right.\]
Thus, we can conclude \( \frac{\partial l_p(k; \delta)}{\partial \delta} - l_p(k-1; \delta) < 0 \) on \( \delta \in (0, (\frac{1}{k} - \frac{1}{n})(1 - \bar{\sigma}^2(k - 1))) \subset (0, g(k)) \cup (0, g(k - 1)) \).

The second order derivative can be computed easily and shown to be positive,

\[
\frac{\partial^2 l_p(K; \delta)}{\partial \delta^2} - l_p(K-1; \delta) = \frac{mnK}{2} \frac{K}{1 - \frac{K}{n} - K\delta} - \frac{mn(K - 1)}{2} \frac{K - 1}{1 - \frac{K - 1}{n} - (K - 1)\delta} \\
= \frac{mnK^2}{2} \left[ \frac{1}{1 - \frac{K}{n} - K\delta} - \frac{(K - 1)^2}{1 - \frac{K - 1}{n} - (K - 1)\delta} \right] > 0, \tag{25}
\]

indicating the first order derivative is monotonically non-decreasing on \( \delta \in (0, (\frac{1}{k} - \frac{1}{n})(1 - \bar{\sigma}^2(k - 1))) \subset (0, g(k)) \cup (0, g(k - 1)) \) and that \( l_p(K; \delta) \) is convex.

**Condition 2.2.** Similarly, \( l_p(k; \delta) - l_p(k+1; \delta) \) on \( \delta \in (0, (\frac{1}{k+1} - \frac{1}{n})(1 - \bar{\sigma}^2(k))) \subset (0, g(k)) \cup (0, g(k + 1)) \).

The first order derivative with respect to \( \delta \):

\[
\frac{\partial l_p(k; \delta)}{\partial \delta} - l_p(k + 1; \delta) = -\frac{mn(k + 1)}{2} \left\{ -\log \frac{1 - \frac{k}{n} - k\delta}{1 - \frac{k+1}{n} - (k + 1)\delta} - \log \bar{\sigma}^2(k + 1) \right\} \\
+ \frac{mn}{2} \log \left[ \frac{1 - \frac{k}{n} - k\delta}{(1 - \frac{k}{n})\bar{\sigma}^2(k)} \right] \\
= -\frac{mn}{2} \left( \log \frac{\bar{\sigma}^2(k + 1)}{\bar{\sigma}^2(k)} \right) + \frac{mn}{2} \log \bar{\sigma}^2(k + 1). \tag{26}
\]

It can be shown similarly to the first case due to symmetry by taking \( k' = k + 1 \). The interval here is constructed as \((\frac{1}{k+1} - \frac{1}{n})(1 - \bar{\sigma}^2(k)) < (\frac{1}{k+1} - \frac{1}{n})(1 - \bar{\sigma}^2(k + 1)) = g(k + 1)\) and \((\frac{1}{k+1} - \frac{1}{n})(1 - \bar{\sigma}^2(k)) < (\frac{1}{k} - \frac{1}{n})(1 - \bar{\sigma}^2(k)) = g(k)\).

The second order derivative can be computed easier and shown to be negative,

\[
\frac{\partial^2 l_p(k; \delta)}{\partial \delta^2} - l_p(k + 1; \delta) = \frac{mnK}{2} \frac{k}{1 - \frac{k}{n} - k\delta} - \frac{mn(k + 1)}{2} \frac{k + 1}{1 - \frac{k+1}{n} - (k + 1)\delta} \\
= \frac{mnK^2}{2} \left[ \frac{k}{1 - \frac{k}{n} - k\delta} - \frac{(k + 1)^2}{1 - \frac{k+1}{n} - (k + 1)\delta} \right] > 0, \tag{27}
\]

indicating the first order derivative is monotonically non-increasing on \((\frac{1}{k+1} - \frac{1}{n})(1 - \bar{\sigma}^2(k))\) and that \( l_p(k; \delta) - l_p(k + 1; \delta) \) is concave.

Finally, we conclude that for any fixed \( k \in \{2, 3, \ldots, n - 2\}, l_p(k; \delta) - l_p(k - 1; \delta) \) is monotonically decreasing and convex function on \((0, (\frac{1}{k} - \frac{1}{n})(1 - \bar{\sigma}^2(k - 1)))\), taking positive value when \( \delta = 0 \); \( l_p(k; \delta) - l_p(k + 1; \delta) \) is monotonically increasing and concave function on \((0, (\frac{1}{k+1} - \frac{1}{n})(1 - \bar{\sigma}^2(k)))\), taking negative value when \( \delta = 0 \).
Further, if we define $a_k = \min\{\tilde{\delta} : l(k; \tilde{\delta}) - l(k + 1; \tilde{\delta}) > 0\}$ and $b_k = \max\{\tilde{\delta} : l(k; \tilde{\delta}) - l(k - 1; \tilde{\delta}) > 0\}$, and if $a_k < a_k$, meaning both curves cross above the $y = 0$ line, then the interval $\Delta_k = (a_k, b_k)$ denotes the range of scaled tuning parameter values such that both $l_p(k; \tilde{\delta}) - l_p(k - 1; \tilde{\delta}) > 0$ and $l_p(k; \tilde{\delta}) - l_p(k + 1; \tilde{\delta}) > 0$ hold. However, we do not have sufficient condition to claim $a_k < b_k$ for all possible values of $k$. But for those that $a_k < b_k$ does hold, they can be shown to maximize $l_p(k; \tilde{\delta})$ as we show in the next lemma.
6.2 Proof of Lemma 3

Lemma 3. For every $k \in \{1, 2, \ldots, K_{\text{max}}\} = \mathcal{H}(\hat{\delta})$, where $K_{\text{max}} = \left\lfloor \frac{1}{1 - 2^2(\frac{1}{\delta} + \frac{1}{n})} \right\rfloor$ and $K_{\text{max}}' = \frac{1}{\delta + \frac{1}{n}} > K_{\text{max}}$, there exists $\hat{\delta}_o \in (0, \max K g(K))$ such that $k = \arg\max_{K \in \mathcal{H}(\hat{\delta}_o)} l_p(K; \hat{\delta}_o)$.

And if $K \in \{2, 3, \ldots, K_{\text{max}} - 1\}$, $k = \arg\max_{K \in \mathcal{H}(\hat{\delta}_o)} l_p(K; \hat{\delta}_o)$ if and only if

$$
\begin{cases}
  l_p(K; \hat{\delta}_o) - l_p(K - 1; \hat{\delta}_o) > 0 \\
  l_p(K; \hat{\delta}_o) - l_p(K + 1; \hat{\delta}_o) > 0,
\end{cases}
$$

(28)

for some $\hat{\delta} \in (0, \max K g(K))$.

For $K = 1$, we have $k = \arg\max_{K} l_p(K; \hat{\delta}_o)$, if and only if $l_p(K; \hat{\delta}_o) - l_p(K + 1; \hat{\delta}_o) > 0$; and for $K = K_{\text{max}} - 1$, $k = \arg\max_{K} l_p(K; \hat{\delta}_o)$ if and only if $l_p(K; \hat{\delta}_o) - l_p(K - 1; \hat{\delta}_o) > 0$.

Proof. Given any $\hat{\delta}$ in an interval where $l_p(K; \hat{\delta})$ is smooth, if $l_p(K; \hat{\delta})$ takes its maximum at $K'$ and $K'$ is not $n - 1$, then clearly $l_p(K'; \hat{\delta}) > l_p(K' - 1; \hat{\delta})$ and $l_p(K'; \hat{\delta}) - l_p(K' + 1; \hat{\delta})$. If $K' = 1$, then $l_p(K'; \hat{\delta}) - l_p(K' + 1; \hat{\delta})$ and if $K' = n - 1$ then $l_p(K'; \hat{\delta}) > l_p(K' + 1; \hat{\delta})$.

To show the reverse direction, if (28) holds, we must show that $l_p(K'; \hat{\delta}) > l_p(k; \hat{\delta})$, where $k < K'$ and $l_p(K'; \hat{\delta}) > l_p(kk; \hat{\delta})$, where $kk > K'$, where $kk \leq K_{\text{max}}$.

From Lemma 2, we know that $a_K = \min\{\hat{\delta} : l(K; \hat{\delta}) - l(K + 1; \hat{\delta}) \geq 0\}$ and $b_K = \max\{\hat{\delta} : l(K; \hat{\delta}) - l(K - 1; \hat{\delta}) \geq 0\}$. Thus, the interval $\Delta_{K'} = (a_{K'}, b_{K'})$ is not empty, then it gives the range of scaled tuning parameter values such that both $l_p(K'; \hat{\delta}) - l_p(K' - 1; \hat{\delta}) > 0$ and $l_p(K'; \hat{\delta}) - l_p(K' + 1; \hat{\delta}) > 0$ hold. We simply need to show that $l_p(K'; \hat{\delta}_o)$ is a maximum when $\hat{\delta}_o \in \Delta_{K'} = (a_{K'}, b_{K'})$ and $a_{K'} < b_{K'}$.

Condition 3.1. $l_p(K'; \hat{\delta}) - l_p(k; \hat{\delta}) > 0$ on $\hat{\delta}_o \in (a_{K'}, b_{K'})$:

First of all, we realize to show $l_p(K'; \hat{\delta}) > l_p(k; \hat{\delta}) > 0$ is equivalent to showing the telescoping sum greater than zero:

$$
l_p(K'; \hat{\delta}) - l_p(k; \hat{\delta}) = |l_p(K'; \hat{\delta}) - l_p(K' - 1; \hat{\delta})| + |l_p(K' - 1; \hat{\delta}) - l_p(K' - 2; \hat{\delta})| + \cdots + |l_p(k + 1; \hat{\delta}) - l_p(k; \hat{\delta})| > 0
$$

where we must have $l_p(K'; \hat{\delta}) - l_p(K' - 1; \hat{\delta}) > 0$ on $\Delta_{K'} = (a_{K'}, b_{K'})$. Now we just need to show each telescoping term is greater than zero for the overall sum to be positive.

Since we had shown that $l_p(K'; \hat{\delta}) - l_p(K' - 1; \hat{\delta})$ is a convex function, then $l_p(K; \hat{\delta}) - l_p(K - 1; \hat{\delta}) > 0 \forall \hat{\delta} \in (0, b_K)$ and $l_p(K; \hat{\delta}) - l_p(K - 1; \hat{\delta}) = 0$ when $\hat{\delta} = b_K$. Similarly, we have shown in Lemma 2 that for any $K \in \{2, \ldots, n - 2\}$, $l_p(K; \hat{\delta}) - l_p(K - 1; \hat{\delta})$ is monotonically decreasing and convex on $(0, (\frac{1}{K} - \frac{1}{n})[1 - \hat{\delta}^2(K - 1)])$.

We just need to show that the difference of the first order derivatives implies that neither $l_p(K' - 1; \hat{\delta}) - l_p(K' - 2; \hat{\delta})$ nor other telescoping term decreases as fast as $\frac{\partial l_p(K'; \hat{\delta}) - l_p(K' - 1; \hat{\delta})}{\partial \hat{\delta}}$, then $l_p(K' - 1; \hat{\delta}) - l_p(K' - 2; \hat{\delta}) > 0$ on $(0, b_{K'})$ where $l_p(K'; \hat{\delta}) - l_p(K' - 1; \hat{\delta}) > 0$. 

38
For $\hat{\delta}_o$ in $(a_{K'}, b_{K'})$, we must have $\hat{\sigma}^2(K') < \hat{\sigma}^2(K' - 1)$ and $\hat{\sigma}^2(k) < 1$ for all $k < K'$ as shown in Lemma 3. We also make use of the relationship:

\[
\begin{align*}
\frac{1 - \frac{K'}{n}}{1 - \frac{K'}{n} - K'\hat{\delta}_o} \cdot \frac{1 - \frac{K' - 2}{n}}{1 - \frac{K' - 2}{n} - (K' - 2)\hat{\delta}_o} > \left(\frac{1 - \frac{K' - 1}{n}}{1 - \frac{K' - 1}{n} - (K' - 1)\hat{\delta}_o}\right)^2 \\
1 - \left(\frac{1}{n + \hat{\delta}_o}\right)^2 > 1 - \left(\frac{1}{1 - \frac{K' - 1}{n} - (K' - 1)\hat{\delta}_o}\right)^2 \\
\left(\frac{1}{1 - \frac{K' - 1}{n} - (K' - 1)\hat{\delta}_o}\right)^2 > \left(\frac{1}{1 - \frac{K'}{n}}\right)^2,
\end{align*}
\]

and

\[
\begin{align*}
\frac{\frac{1}{n} + \hat{\delta}_o}{1 - \frac{K' - 1}{n} - (K' - 1)\hat{\delta}_o} = \frac{\frac{1}{n} - \frac{1}{n} + \frac{\hat{\delta}_o}{n}}{(1 - \frac{K' - 1}{n} - (K' - 1)\hat{\delta}_o)(1 - \frac{K' - 1}{n})},
\end{align*}
\]

which implies that $1 > \frac{\hat{\sigma}^2(K')}{\sigma^2(K' - 1)} > \frac{\hat{\sigma}^2(K')}{\sigma^2(K' - 1)}$ and $\frac{\hat{\sigma}^2(K')}{\sigma^2(K' - 1)} > \frac{\hat{\sigma}^2(K' - 1)}{\sigma^2(K' - 2)} > \frac{\hat{\sigma}^2(K' - 1)}{\sigma^2(K' - 2)}$.

The maximum of $\frac{\hat{\sigma}^2(K' - 1)}{\hat{\sigma}^2(K' - 2)}$ when $\hat{\delta}_o \in (a_{K'}, b_{K'})$ can be shown to be less than its value at $\tilde{\delta} = \frac{1}{n} - \frac{1}{n}:

\[
\begin{align*}
\frac{\hat{\sigma}^2(K' - 1)}{\hat{\sigma}^2(K' - 2)} &= \frac{\hat{\sigma}^2(K' - 1)}{\hat{\sigma}^2(K' - 2)} \cdot \frac{1 - \frac{K' - 1}{n}}{1 - \frac{K' - 1}{n} - (K' - 1)\hat{\delta}_o} \cdot \frac{1 - \frac{K' - 2}{n} - (K' - 2)\hat{\delta}_o}{1 - \frac{K' - 2}{n}} \\
&\leq \frac{\hat{\sigma}^2(K' - 1)}{\hat{\sigma}^2(K' - 2)} \cdot \frac{1 - \frac{K'}{n}}{1 - \frac{K'}{n} - (K')\hat{\delta}_o} \cdot \frac{1 - \frac{K' - 1}{n} - (K' - 1)\hat{\delta}_o}{1 - \frac{K' - 1}{n}} \\
&= \frac{\hat{\sigma}^2(K' - 1)}{\hat{\sigma}^2(K' - 2)} \cdot \left[1 + \frac{\delta_o - \frac{1}{n} - \frac{\hat{\delta}_o}{n}}{(1 - \frac{K' - 1}{n} - (K' - 1)\hat{\delta}_o)(1 - \frac{K' - 1}{n})}\right] \\
&\leq \frac{\hat{\sigma}^2(K' - 1)}{\hat{\sigma}^2(K' - 2)} \cdot \left[n - K' - 1\right] \\
&< 1.
\end{align*}
\]

Thus, the first order derivative must be negative on $(a_{K'}, b_{K'})$ where $b_{K'} < \frac{1}{n} - \frac{1}{n}$:

\[
\frac{\partial l_p(K' - 1; \hat{\delta}_o) - l_p(K' - 2; \hat{\delta}_o)}{\partial \hat{\delta}_o} = \frac{mn(K' - 2)}{2} \log \frac{\hat{\sigma}^2(K' - 1)}{\hat{\sigma}^2(K' - 2)} + \frac{mn}{2} \log \hat{\sigma}^2(K' - 1) < 0.
\]
where we must have Condition 3.2.

\[
\frac{\partial^2 l_p(K - 1; \tilde{\delta}_o) - l_p(K - 2; \tilde{\delta}_o)}{\partial \tilde{\delta}_o^2} = \frac{mn}{2} \left[ \frac{(K - 1)^2}{1 - \frac{K - 1}{n} - (K - 1)\tilde{\delta}_o} - \frac{K - 2}{1 - \frac{(K - 2)^2}{n} - (K - 2)\tilde{\delta}_o} \right] > 0,
\]

indicating the first order derivative is monotonically non-decreasing and that \( l_p(K - 1; \tilde{\delta}_o) - l_p(K - 2; \tilde{\delta}_o) \) or other telescoping term is convex.

Finally, to show directly the rate of decrease is slower for the remaining terms so that the sum remains positive, it suffice to show:

\[
\frac{\partial l_p(K - 1; \tilde{\delta}_o) - l_p(K - 2; \tilde{\delta}_o)}{\partial \tilde{\delta}_o} - \frac{\partial l_p(K; \tilde{\delta}_o) - l_p(K - 1; \tilde{\delta}_o)}{\partial \tilde{\delta}_o} = -\frac{\partial l_p(K; \tilde{\delta}_o) + l_p(K - 2; \tilde{\delta}_o)}{\partial \tilde{\delta}_o} = -\left[ \frac{mnK}{2} \log \tilde{\delta}^2(K) + \frac{mnK}{2} \log \tilde{\delta}^2(K - 2) \right] > 0,
\]

That is, the difference of the first order derivatives implies that neither \( l_p(K' - 1; \tilde{\delta}_o) - l_p(K' - 2; \tilde{\delta}_o) \) nor other telescoping term decreases as fast as \( l_p(K'; \tilde{\delta}_o) - l_p(K' - 1; \tilde{\delta}_o) \).

Summing up all the positive differences \( l_p(K' - 1; \tilde{\delta}_o) - l_p(K' - 2; \tilde{\delta}_o) > 0 \) on \((0, b'_K)\), \( l_p(K' - 1; \tilde{\delta}_o) - l_p(K' - 2; \tilde{\delta}_o) > 0 \) and etc., we have \( l_p(K'; \tilde{\delta}_o) - l_p(k; \tilde{\delta}_o) > 0 \) on \( \tilde{\delta}_o \in (0, b'_K) \subset (a'_K, b'_K) \).

**Condition 3.2.** \( l_p(K; \tilde{\delta}_o) - l_d(kk; \tilde{\delta}_o) > 0 \) on \( \tilde{\delta}_o \in (a'_K, b'_K) \):

We can mirror the proof for the other case. First of all, \( l_p(K'; \tilde{\delta}_o) > l_p(kk; \tilde{\delta}_o) > 0 \) is equivalent to showing the telescoping sum greater than zero:

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

where we must have \( l_p(K'; \tilde{\delta}_o) - l_p(K' + 1; \tilde{\delta}_o) > 0 \) on \( \Delta_K = (a'_K, b'_K) \). Now we just need to show each telescoping term is greater than zero for the overall sum to be positive.

Since we had shown that \( l_p(K'; \tilde{\delta}_o) - l_p(K' + 1; \tilde{\delta}_o) \) is a concave function on \((a'_K, b'_K)\), then \( l_p(K'; \tilde{\delta}_o) - l_p(K' + 1; \tilde{\delta}_o) > 0 \) for all \( \tilde{\delta}_o \in (a'_K, b'_K) \).
The remaining of this condition can be produced similarly to the previous one, we only show the last step where
\[
\frac{\partial l_p(K' + 1; \tilde{\delta}) - l_p(K' + 2; \tilde{\delta})}{\partial \tilde{\delta}} - \frac{\partial l_p(K'; \tilde{\delta}) - l_p(K' + 1; \tilde{\delta})}{\partial \tilde{\delta}} = \frac{\partial}{\partial \tilde{\delta}}(l_p(K' + 1; \tilde{\delta}) - l_p(K' + 2; \tilde{\delta})) = -\left[\frac{mnK}{2} \log \tilde{\sigma}^2(K') + \frac{mnK}{2} \log \tilde{\sigma}^2(K' + 2)\right] > 0,
\] (32)

That is, the difference of the first order derivatives implies that neither \(l_p(K' + 1; \tilde{\delta}) - l_p(K' + 2; \tilde{\delta})\) nor other telescoping term increases as fast as \(l_p(K'; \tilde{\delta}) - l_p(K' + 1; \tilde{\delta})\), then \(l_p(K' + 1; \tilde{\delta}) - l_p(K' + 2; \tilde{\delta}) > 0\) on \((a_K', b_K')\) where \(l_p(K; \tilde{\delta}) - l_p(K' + 1; \tilde{\delta}) > 0\).

When \(K' = n - 1\), the penalized profile log-likelihood \(l_p(K; \tilde{\delta} = 0)\) is clearly maximized at \(K = n - 1\) and \(\tilde{\delta} = 0\) following from Proposition 1. And we can take \(a_{n-1} = 0\). For \(\tilde{\delta} \in (0, g(n-1)(\tilde{\lambda}_n + (n-1)\tilde{\lambda}_{n-1}))\), we can show \(l_p(n - 1; \tilde{\delta}) - l_p(n - 2; \tilde{\delta})\) to be a smooth and monotonically decreasing function of \(\tilde{\delta}\), which implies that \(l_p(n - 1; \tilde{\delta}) - l_p(n - 2; \tilde{\delta}_o) > 0\) for some \(\tilde{\delta}_o \in (0, b_{n-1})\).

When \(K' = 1\), for \(\tilde{\delta} \in (0, (1 - \frac{1}{2})(1 - \tilde{\sigma}^2(1)))\), we can show \(l_p(1; \tilde{\delta}) - l_p(2; \tilde{\delta})\) to be a smooth and monotonically increasing function of \(\tilde{\delta}\) as long as \(\tilde{\sigma}_1^2 > \tilde{\sigma}_2^2\), which implies that \(l_p(1; \tilde{\delta}) - l_p(2; \tilde{\delta}_o) > 0\) for some \(\tilde{\delta}_o \in (\frac{n\tilde{\lambda} \tilde{\lambda}_2}{(\tilde{\lambda}_2 - \tilde{\sigma}_2^2)(n-2)} - 1, \max_K g(K))\).

Both can then be expressed using the telescoping sum similar to the other possible \(K'\) values. Finally, these two cases conclude the proof that if \((28)\) holds, we have that \(l_p(K'; \tilde{\delta}) > l_p(k; \tilde{\delta})\), where \(k < K'\) and \(l_p(K'; \tilde{\delta}) - l_p(kk; \tilde{\delta})\), where \(kk > K'\), where \(kk \leq K_{\text{max}}\) or equivalently for any \(\tilde{\delta} \in (a_K, b_K)\).
6.3 Proof of Lemma 4

Lemma 4. Suppose for any \( \tilde{\delta} \in (0, \max_K g(K)) \), the penalized profile log-likelihood is maximized at some \( K' \in \{1, 2, \ldots, K_{\max}\} \), where \( K_{\max} = \left\lfloor \frac{1}{\frac{1}{2} \frac{\delta}{\max_{P(K)}} + \frac{1}{n}} \right\rfloor \) and \( K_{\max}' = \frac{1}{\delta + \frac{1}{n}} > K_{\max} \). Then, the interval \( \Delta_K' = (a'_K, b'_K) \) denotes the range of scaled tuning parameter values such that both \( l_p(K'; \tilde{\delta}) - l_p(K' - 1; \tilde{\delta}) > 0 \) and \( l_p(K'; \tilde{\delta}) - l_p(K' + 1; \tilde{\delta}) > 0 \) hold and can be approximated by \( (u_a(K'), u_b(K')) \subset \Delta_K' = (a'_K, b'_K) \), where \( u_a(K') \) denote the upper bound for \( a'_K \), and \( u_b(K') \) the lower bound for \( b'_K \), such that \( \frac{b'_K}{a'_K} > \frac{u_b(K')}{u_a(K')} \).

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. \( K' \)

Condition 4.1. \( l_p(K'; \tilde{\delta}) - l_p(K' - 1; \tilde{\delta}) > 0 \):

Expanding the difference in penalized profile log-likelihoods:

\[
l_p(K'; \tilde{\delta}) - l_p(K' - 1; \tilde{\delta}) = l_p(K') - l_p(K') - 1 - \frac{m}{2} \left\{ (n - K') \log(1 - \frac{K'}{n}) - (n - K' + 1) \log(1 - \frac{K' - 1}{n}) \right. \\
- nK' \tilde{\delta} [1 + \log(\tilde{\sigma}^2(K')) + \log(1 - \frac{K'}{n})] \\
+ \tilde{\delta} nK' \log(1 - \frac{K'}{n} - K' \tilde{\delta}) - (n - K') \log(1 - \frac{K'}{n} - K' \tilde{\delta}) \\
+ n(K' - 1) \tilde{\delta} [1 + \log(\tilde{\sigma}^2(K' - 1)) + \log(1 - \frac{K' - 1}{n})] \\
- \tilde{\delta} n(K' - 1) \log(1 - \frac{K' - 1}{n} - (K' - 1) \tilde{\delta}) \\
+ (n - K' + 1) \log(1 - \frac{K' - 1}{n} - (K' - 1) \tilde{\delta}) \\
- \frac{m}{2} \left\{ \log \frac{\lambda_{K'}}{\tilde{\sigma}^2(K' - 1)} + (n - K') \log \frac{\tilde{\sigma}^2(K')}{\tilde{\sigma}^2(K' - 1)} \\
+ (n - K') \log(1 - \frac{K'}{n}) - (n - K' + 1) \log(1 - \frac{K' - 1}{n}) \\
- n \tilde{\delta} [K' \log \frac{\tilde{\sigma}^2(K')}{\tilde{\sigma}^2(K' - 1)} + K' \log (1 - \frac{K'}{n} + \frac{1}{1 - \frac{K' - 1}{n}}) \\
\log(\tilde{\sigma}^2(K' - 1)) + \log (1 - \frac{K' - 1}{n}) + 1] \\
+ (n - K' + 1) \log(1 - \frac{K' - 1}{n} - (K' - 1) \tilde{\delta}) - (n - K') \log(1 - \frac{K'}{n} - K' \tilde{\delta}) \\
+ nK' \tilde{\delta} \log(1 - \frac{K'}{n} - K' \tilde{\delta}) - n(K' - 1) \tilde{\delta} \log(1 - \frac{K' - 1}{n} - (K' - 1) \tilde{\delta}) \right\} \\
\text{(33)}
\]
For a given choice of $K'$, $l_p(K') - l_p(K' - 1)$ is a fixed positive value, and we have shown the difference $l_p(K'; \tilde{\delta}) - l_p(K' - 1; \tilde{\delta})$ is decreasing in $\tilde{\delta}$ for $\tilde{\delta} < b_K'$ (Lemma 2). Thus, for any $K'$, we need to search for the largest value of $\tilde{\delta}$ such that the condition holds, or equivalently, the smallest value such that the negation $l_p(K'; \tilde{\delta}) - l_p(K' - 1; \tilde{\delta}) \leq 0$ holds. Notice that for each $K'$, the theoretical range of $\tilde{\delta}$ is restricted to $(0, \frac{1}{K'} - \frac{1}{n})$, while the practical range depends on the estimated error variance: $(0, \frac{1}{K'} - \frac{1}{n})(1 - \hat{\sigma}^2(K'))$.

We approximate the following for any $K'$ by Taylor expansion at $\tilde{\delta} = 0$:

\[
\log \left(1 - \frac{K'}{n} - K'\tilde{\delta} \right) = \log \left(1 - \frac{K'}{n} \right) - \frac{(K')\tilde{\delta}}{1 - \frac{K'}{n}} + O(\tilde{\delta}^2) \tag{34a}
\]

\[
\log \left(1 - \frac{K'-1}{n} - (K' - 1)\tilde{\delta} \right) = \log \left(1 - \frac{K'-1}{n} \right) - \frac{(K' - 1)\tilde{\delta}}{1 - \frac{K'-1}{n}} + O(\tilde{\delta}^2) \tag{34b}
\]

and each converges if $\tilde{\delta} < \frac{1}{K'} - \frac{1}{n}$ and $\tilde{\delta} < \frac{1}{K'-1} - \frac{1}{n}$, respectively.

Merge all the terms and define $\zeta(\tilde{\delta})$, where

\[
l_p(K'; \tilde{\delta}) - l_p(K' - 1; \tilde{\delta}) = -\frac{m}{2} \left( \zeta(\tilde{\delta}) + O(\tilde{\delta}^2) \right)
\]

\[
= -\frac{m}{2} \left\{ \log \frac{\lambda_{K'}}{\hat{\sigma}^2(K' - 1)} + (n - K') \log \frac{\hat{\sigma}^2(K')}{\hat{\sigma}^2(K' - 1)} \right. \\
- n\tilde{\delta} \left[ K' \log \frac{\hat{\sigma}^2(K')}{\hat{\sigma}^2(K' - 1)} + \log \hat{\sigma}^2(K' - 1) \right] \\
+ \hat{\sigma}^2 \left[ \frac{n(K' - 1)^2}{1 - \frac{K'-1}{n}} - \frac{nK'^2}{1 - \frac{K'}{n}} \right] + O(\tilde{\delta}^3) \right\} \tag{35}
\]

So now we need to solve the inequality and find the smallest $\tilde{\delta}$ such that $\zeta(\tilde{\delta}) \geq 0$. 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 $\zeta(\tilde{\delta}) = a\tilde{\delta}^2 + b\tilde{\delta} + c$, where

\[
a = \frac{n(K' - 1)^2}{1 - \frac{K'-1}{n}} - \frac{nK'^2}{1 - \frac{K'}{n}} = \frac{n - K' - (2n - K')K'}{(n - K' + 1)(n - K')} < 0 \tag{36}
\]

\[
b = -n[(K' \log \frac{\hat{\sigma}^2(K')}{\hat{\sigma}^2(K' - 1)} + \log \hat{\sigma}^2(K' - 1))] > 0 \tag{37}
\]

\[
c = \log \frac{\lambda_{K'}}{\hat{\sigma}^2(K' - 1)} + (n - K') \log \frac{\hat{\sigma}^2(K')}{\hat{\sigma}^2(K' - 1)} < 0 \tag{38}
\]

Thus, if the discriminant $b^2 - 4ac > 0$ and we have $a < 0$, then there are two positive roots, $r_1(K')$ and $r_2(K')$, where

\[
\begin{cases}
\zeta(\tilde{\delta}) < 0 \text{ for } \tilde{\delta} \in (0, r_1(K')) \cup (r_2(K'), \frac{1}{K'} - \frac{1}{n}), \\
\zeta(\tilde{\delta}) > 0 \text{ for } \tilde{\delta} \in (r_1(K'), r_2(K')), \\
\end{cases} \tag{40}
\]
whereas if the discriminant $b^2 - 4ac < 0$, then $\zeta(\tilde{\delta}) < 0$ for any $\tilde{\delta} \in (0, \frac{1}{K'\hat{\rho}} - \frac{1}{n})$. In addition, the maximum of $\zeta(\tilde{\delta})$ is $\zeta(-\frac{b}{2a})$, where $r_1(K') < -\frac{b}{2a}$ and $r_2(K') > -\frac{b}{2a}$.

**Condition 4.2.** $l_p(K', \tilde{\delta}) - l_p(K' + 1; \tilde{\delta}) \geq 0$

Similarly, we can make the derivation for the other direction:

$$l_p(K', \tilde{\delta}) - l_p(K' + 1; \tilde{\delta}) = -\frac{m}{2} \left\{ -\log \frac{\lambda_{K' + 1}}{\sigma^2(K')} + (n - K' - 1) \log \frac{\hat{\sigma}^2(K')}{\sigma^2(K' + 1)} 
+ (n - K') \log(1 - \frac{K'}{n}) - (n - K' - 1) \log(1 - \frac{K' + 1}{n}) 
- n\tilde{\delta}[K' \log \frac{\hat{\sigma}^2(K')}{\sigma^2(K' + 1)} + K' \log \frac{1 - \frac{K'}{n}}{1 - \frac{K' + 1}{n}} 
- \log(\hat{\sigma}^2(K' + 1)) - \log(1 - \frac{K' + 1}{n}) - 1] 
+ (n - K' - 1) \log(1 - \frac{K' + 1}{n}) - (K' + 1) \tilde{\delta) - (n - K') \log(1 - \frac{K'}{n}) - K' \tilde{\delta) 
+ nK' \tilde{\delta} \log(1 - \frac{K'}{n}) - K' \tilde{\delta) - n(K' + 1) \tilde{\delta} \log(1 - \frac{K' + 1}{n}) - (K' + 1) \tilde{\delta)} \right\}$$

(41)

For a given choice of $K'$, $l_p(K') - l_p(K' + 1)$ is a fixed negative value, and we have shown the difference $l_p(K'; \tilde{\delta}) - l_p(K' + 1; \tilde{\delta})$ is increasing in $\tilde{\delta}$ for $\tilde{\delta} \in (a_K', b_K')$ (Lemma 2). Thus, for any $K'$, we need to search for the smallest value of $\tilde{\delta}$ such that the condition holds, or equivalently, the largest value such that the negation $l_p(K'; \tilde{\delta}) - l_p(K' + 1; \tilde{\delta}) \leq 0$ holds. Notice that for each $K'$, the theoretical range of $\tilde{\delta}$ is restricted to $(0, \frac{1}{K'\hat{\rho} + 1} - \frac{1}{n})$, while the practical range depends on the estimated error variance: $(0, (\frac{1}{K'\hat{\rho}} - \frac{1}{n})(1 - \hat{\sigma}^2(K' + 1)))$.

We approximate the following for any $K'$ by Taylor expansion at $\tilde{\delta} = 0$:

$$\log(1 - \frac{K'}{n}) - K' \tilde{\delta) = \log(1 - \frac{K'}{n}) - (\frac{K'}{1 - \frac{K'}{n}}) + O(\tilde{\delta}^2)$$

(42a)

$$\log(1 - \frac{K' + 1}{n}) - (K' + 1) \tilde{\delta) = \log(1 - \frac{K' + 1}{n}) - (\frac{K' + 1}{1 - \frac{K' + 1}{n}}) + O(\tilde{\delta}^2)$$

(42b)

and each converges if $\tilde{\delta} < \frac{1}{K'\hat{\rho} - \frac{1}{n}}$ and $\tilde{\delta} < \frac{1}{K'\hat{\rho} + 1} - \frac{1}{n}$, respectively.
Again, define \( \zeta'(\delta) \) such that

\[
l_p(K'; \delta) - l_p(K' + 1; \delta) = -\frac{m}{2} \left( \zeta'(\delta) + O(\delta^2) \right)
\]

\[
= -\frac{m}{2} \left\{ -\log \frac{\lambda_{K' + 1}}{\hat{\sigma}^2(K')} + (n - K' - 1) \log \frac{\tilde{\sigma}^2(K')}{\hat{\sigma}^2(K' + 1)} \right. \\
- n\delta[K' \log \frac{\hat{\sigma}^2(K')}{\hat{\sigma}^2(K' + 1)} - \log \hat{\sigma}^2(K' + 1)] \\
\left. + \tilde{\delta}^2 \left[ \frac{n(K' + 1)^2}{1 - \frac{K' + 1}{n}} - \frac{nK'^2}{1 - \frac{K'}{n}} \right] + O(\delta^3) \right\}
\]

(43)

Again, we can rewrite \( \zeta'(\delta) = a'\delta^2 + b'\delta + c' \), where

\[
a' = \frac{n(K' + 1)^2}{1 - \frac{K' + 1}{n}} - \frac{nK'^2}{1 - \frac{K'}{n}} = \frac{n - K' + (2n - K')K'}{(n - K' - 1)(n - K')} > 0
\]

(44)

\[
b' = n[(K' + 1) \log \frac{\hat{\sigma}^2(K' + 1)}{\hat{\sigma}^2(K')} + \log \hat{\sigma}^2(K')] < 0
\]

(45)

\[
c' = -\log \frac{\lambda_{K' + 1}}{\hat{\sigma}^2(K')} + (n - K' - 1) \log \frac{\tilde{\sigma}^2(K')}{\hat{\sigma}^2(K' + 1)} > 0
\]

(46)

Now solve the inequality and find the smallest \( \tilde{\delta} \) such that \( \zeta'(\tilde{\delta}) \leq 0 \). Clearly, when \( \tilde{\delta} = 0 \), \( \zeta'(\tilde{\delta}) > 0 \). Thus, if the discriminant \( b'^2 - 4a'c' > 0 \) and we have \( a' > 0 \), then there are two positive roots, \( r_3(K') \) and \( r_4(K') \), where

\[
\begin{aligned}
\zeta'(\tilde{\delta}) &< 0 \text{ for } \delta \in (r_3(K'), r_4(K')) \\
\zeta'(\tilde{\delta}) &> 0 \text{ for } \delta \in (0, r_3(K')) \cup (r_4(K'), \frac{1}{K' + 1} - \frac{1}{n})
\end{aligned}
\]

(47)

Note that if the discriminant \( b'^2 - 4a'c' < 0 \), then \( \zeta'(\tilde{\delta}) > 0 \) for any \( \delta \in (0, \frac{1}{K' + 1} - \frac{1}{n}) \).

Again, since \( \tilde{\delta} \) is bounded by \((1 - \frac{1}{n})(1 - \hat{\sigma}^2(1))\), the root near 0 can be approximated by the linear solution \( \xi \). So finally we have

\[
u_a(K') = -\log \frac{\lambda_{K'} + (n - K') \log \tilde{\sigma}^2(K')}{(K' - 1) \log \hat{\sigma}^2(K' - 1) - K' \log \hat{\sigma}^2(K')}
\]

and

\[
u_b(K') = -\log \frac{\lambda_{K' + 1} + (n - K' - 1) \log \tilde{\sigma}^2(K')}{(K' + 1) \log \hat{\sigma}^2(K' + 1) - K' \log \hat{\sigma}^2(K')}.
\]

Also notice that \( u_b(K') \) equals to \( u_a(K' - 1) \) due to the symmetry in \( K', K' - 1 \) and \( K', K' + 1 \), indicating that these approximated intervals are also non-overlapping.
6.4 Proof of Lemma 5

Lemma 5. Suppose the interval \( \Delta_K = (a_K, b_K) \) denotes the range of scaled tuning parameter values such that both \( l_p(K; \hat{\delta}) - l_p(K - 1; \hat{\delta}) > 0 \) and \( l_p(K; \hat{\delta}) - l_p(K + 1; \hat{\delta}) > 0 \) hold and can be approximated by \( (u_a(K), u_b(K)) \subset \Delta_K = (a_K, b_K) \), where \( u_a(K) \) denote the upper bound for \( a_K \), and \( u_b(K) \) the lower bound for \( b_K \), such that \( \frac{b_K}{a_K} > \frac{u_b(K)}{u_a(K)} \). As \( m \to \infty \), we have \( \frac{u_b(K^*)}{u_a(K^*)} \to \infty \), where \( K^* \) is the true rank; and \( |u_b(K) - u_a(K)| \to 0 \) for \( K > K^* \).

Proof. First of all, the numerators of \( u_b(K) \) and \( u_a(K) \) are smooth functions of \( \tilde{\sigma}^2(K) \) on \((0, 1)\), respectively, as both first derivatives with respect to \( \tilde{\sigma}^2(K) \) exist. In addition, it can be shown that \( c \) is a monotonically increasing function of \( \tilde{\sigma}^2(K) \), while \( c' \) is a monotonically decreasing function due to the symmetry between \( K, K - 1 \) and \( K + 1, K \).

\[
\frac{\partial c}{\partial \tilde{\sigma}^2(K)} = \frac{n - K}{\tilde{\sigma}^2(K)}\frac{[\lambda_K - \tilde{\sigma}^2(K)]}{[\tilde{\sigma}^2(K)]^2(n - K + \frac{\lambda_K}{\tilde{\sigma}^2})}
\]

\[
\frac{\partial c'}{\partial \tilde{\sigma}^2(K)} = \frac{n - K}{\tilde{\sigma}^2(K)}\frac{[\lambda_{K+1} - \tilde{\sigma}^2(K)]}{[\tilde{\sigma}^2(K)]^2(n - K + \frac{\lambda_{K+1}}{\tilde{\sigma}^2})}
\]

Secondly, the denominators of \( u_b(K) \) and \( u_a(K) \) are also smooth functions of \( \tilde{\sigma}^2(K) \) on \((0, 1)\), respectively, as both first derivatives with respect to \( \tilde{\sigma}^2(K) \) exist.

\[
\frac{\partial b}{\partial \tilde{\sigma}^2(K)} = -\frac{K\lambda_K + \tilde{\sigma}^2(K)(n - K)}{\tilde{\sigma}^2(K)[(n - K)\tilde{\sigma}^2(K) - \lambda_K]}
\]

\[
\frac{\partial b'}{\partial \tilde{\sigma}^2(K)} = -\frac{K\lambda_{K+1} + \tilde{\sigma}^2(K)(n - K)}{\tilde{\sigma}^2(K)[(n - K)\tilde{\sigma}^2(K) - \lambda_{K+1}]}
\]

Notice the following relationships for the MLEs of PPCA models:

\[
\tilde{\sigma}^2(K + 1) = \sum_{j=K+2}^{n} \frac{\lambda_j}{n - K - 1} = \frac{(n - K)\tilde{\sigma}^2(K) - \lambda_{K+1}}{n - K - 1} = \tilde{\sigma}^2(K) - \frac{\lambda_{K+1} - \tilde{\sigma}^2(K)}{n - K - 1} \quad (48)
\]

\[
\tilde{\sigma}^2(K - 1) = \sum_{j=K}^{n} \frac{\lambda_j}{n - K + 1} = \frac{(n - K)\tilde{\sigma}^2(K) + \lambda_K}{n - K + 1} = \tilde{\sigma}^2(K) + \frac{\lambda_K - \tilde{\sigma}^2(K)}{n - K + 1} \quad (49)
\]

The size of the interval depends on the relative ratio of each \( \lambda_K \) and \( \lambda_{K+1} \) relative to \( \tilde{\sigma}^2(K) \).

\[
\lambda_K - \tilde{\sigma}^2(K) = \tilde{d}_K > 0 \quad (50)
\]

\[
\lambda_{K+1} - \tilde{\sigma}^2(K) = (n - K)\tilde{\sigma}^2(K) - (n - K - 1)\tilde{\sigma}^2(K + 1) - \tilde{\sigma}^2(K)
= (n - K - 1)(\tilde{\sigma}^2(K) - \tilde{\sigma}^2(K + 1))
= \tilde{d}_{K+1} \quad (51)
\]
Since the true value of the \( n - K \) last eigenvalues are equal to \( \sigma^2(K) \), Theorem 8.3.2 (Mardia, 1979) implies that for normal data, the MLE of \( \hat{\sigma}^2(K) = \frac{\sum_{i=K+1}^{n} \lambda_i}{n-K} \) converges to \( \sigma^2(K) \). Similarly, MLE of \( \sigma^2(K+1) = \frac{\sum_{i=K+2}^{n} \lambda_i}{n-K-1} \) is \( \hat{\sigma}^2(K+1) \) 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 \( |\frac{\hat{\sigma}^2(K)}{\sigma^2(K+1)} - 1| \) converges to zero with probability 1. The difference \( \lambda_K - \hat{\sigma}^2(K) \) is always positive and converges to the true difference separating the signal and noise \( d^2_K \), while \( \lambda_{K+1} - \hat{\sigma}^2(K) \) diminishes to zero with probability 1 as \( m \) increases. In this case,

\[
\begin{align*}
c'(K) &= -\log \frac{\lambda_{K+1}}{\hat{\sigma}^2(K)} + (n - K - 1) \log \frac{\hat{\sigma}^2(K)}{\hat{\sigma}^2(K+1)} \\
&= (n - K - 1) \log \frac{\hat{\sigma}^2(K)}{\hat{\sigma}^2(K+1)} - \log \frac{(n - K)\hat{\sigma}^2(K) - (n - K - 1)\hat{\sigma}^2(K+1)}{\hat{\sigma}^2(K)} \\
&\to 0,
\end{align*}
\]

as \( m \to \infty \), implying that \( u_a(K) \to 0 \) if \( K = K^* \). On the other hand, \( u_b(K) \) is fixed positive number, 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 \( K > K^* \).
6.5 Behaviour of the penalized profile log-likelihood in simulated data

One dataset with $n = 100$ and $m = 10,000$ was simulated assuming the true effective
dimension to be $K^* = 10$, the residual variance $\sigma^2 = 0.5$, and the smallest singular value
at 0.5. The dataset was simulated under the assumption of normality and further stan-
dardized. The profile log-likelihood and profile log-likelihood were plotted as a function of
varying $K \in \{1, 2, \ldots, n - 1\}$ omitting any constants.

The top left figure 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 appears and approaches that
of the correct value at $K = 10$ marked by the red dashed line.

The bottom left figure 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 $(u_a(10), u_b(10))$ as
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 $(0, u_b(K))$. 
These two figures show the profile log-likelihood and the penalized profile log-likelihood as functions of the residual variance parameter $\sigma^2$. The profile log-likelihood is maximized at different $\sigma^2$ values as we fix at different $K$ values. In particular, as we increase the parameter $K$, the estimates moves closer to 0, and away from the true value at 0.5. On the other hand, if we fixed $K = 10$ at the correct value, the estimates from the penalized profile log-likelihood becomes smaller relative to the estimate when no penalty was enforced, and decreases as the penalty increases.
Figure 10: Profile log-likelihood as a function of $\sigma^2$ with and without introducing penalty.
Figure 11: **Proportion of correctly estimated \( K \) for \( m = 50 \).** The shaded bars indicate the proportion of correctly estimated \( K \) for each true \( K^* = 5, 10 \) or 20 over 100 replicates against each method.
Figure 12: Boxplots of estimated $K$ for $m = 50$. The box indicate the range of estimated $K$ for each true $K^* = 5, 10$ or $20$ over 100 replicates against each method.
Figure 13: Proportion of correctly estimated $K$ for $m = 5,000$. The shaded bars indicate the proportion of correctly estimated $K$ for each true $K^* = 5, 10$ or $20$ over 100 replicates against each method.
Figure 14: Boxplots of estimated $K$ for $m = 5000$. The box indicate the range of estimated $K$ for each true $K^* = 5, 10$ or 20 over 100 replicates against each method.
Figure 15: Proportion of correctly estimated $K$ for large $m$. The proportion of correct selection for each true $K^* = 5$ and 10 were shown in shaded bars against each method.
Figure 16: Proportion of correctly estimated $K$ for large $m$. The proportion of correct selection for each true $K^* = 5$ and 10 were shown in shaded bars against each method.
Figure 17: Proportion of correctly estimated $K$ for large $m$. The proportion of correct selection for each true $K^* = 5$ and 10 were shown in shaded bars against each method.
Figure 18: Proportion of correctly estimated $K$ for large $m$. The proportion of correct selection for each true $K^* = 5$ and 10 were shown in shaded bars against each method.
Figure 19: Proportion of correctly estimated $K$ for large $m$. The proportion of correct selection for each true $K^*=5$ and 10 were shown in shaded bars against each method.
Figure 20: **Proportion of correctly estimated $K$ for large $m$.** The proportion of correct selection for each true $K^* = 5$ and 10 were shown in shaded bars against each method.
Figure 21: Proportion of correctly estimated $K$ for data with autocorrelated and fat-tailed noise. The proportion of correct selection for each true $K^* = 5$ and 10 were shown in shaded bars against each method.
Figure 22: Proportion of correctly estimated K for large m. The proportion of correct selection for each true $K^* = 5$ and 10 were shown in shaded bars against each method.
<table>
<thead>
<tr>
<th>$K$</th>
<th>$\sigma^2(K)$</th>
<th>$d^2(K)$</th>
<th>$m$</th>
<th>PPPCA</th>
<th>BIC</th>
<th>Minka BIC</th>
<th>Cumlog</th>
<th>PESEL</th>
<th>GCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.30</td>
<td>0.50</td>
<td>50</td>
<td>2.71</td>
<td>49.00</td>
<td>49.00</td>
<td>18.95</td>
<td>3.18</td>
<td>10.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.76)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(1.51)</td>
</tr>
<tr>
<td>10</td>
<td>0.30</td>
<td>0.50</td>
<td>50</td>
<td>6.33</td>
<td>49.00</td>
<td>49.00</td>
<td>18.26</td>
<td>5.91</td>
<td>10.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.03)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(1.13)</td>
</tr>
<tr>
<td>20</td>
<td>0.30</td>
<td>0.50</td>
<td>50</td>
<td>10.92</td>
<td>49.00</td>
<td>49.00</td>
<td>21.12</td>
<td>6.97</td>
<td>14.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(2.79)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(1.01)</td>
</tr>
<tr>
<td>5</td>
<td>0.70</td>
<td>0.50</td>
<td>50</td>
<td>2.06</td>
<td>49.00</td>
<td>49.00</td>
<td>29.70</td>
<td>1.81</td>
<td>3.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.72)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.89)</td>
</tr>
<tr>
<td>10</td>
<td>0.70</td>
<td>0.50</td>
<td>50</td>
<td>6.81</td>
<td>49.00</td>
<td>49.00</td>
<td>30.17</td>
<td>1.33</td>
<td>4.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(5.75)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.83)</td>
</tr>
<tr>
<td>20</td>
<td>0.70</td>
<td>0.50</td>
<td>50</td>
<td>16.28</td>
<td>49.00</td>
<td>49.00</td>
<td>31.53</td>
<td>1.00</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(6.09)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.66)</td>
</tr>
<tr>
<td>5</td>
<td>0.30</td>
<td>1.00</td>
<td>50</td>
<td>3.48</td>
<td>49.00</td>
<td>49.00</td>
<td>18.17</td>
<td>3.82</td>
<td>9.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.83)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(1.45)</td>
</tr>
<tr>
<td>10</td>
<td>0.30</td>
<td>1.00</td>
<td>50</td>
<td>7.68</td>
<td>49.00</td>
<td>49.00</td>
<td>17.94</td>
<td>6.99</td>
<td>11.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.94)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(1.19)</td>
</tr>
<tr>
<td>20</td>
<td>0.30</td>
<td>1.00</td>
<td>50</td>
<td>12.94</td>
<td>49.00</td>
<td>49.00</td>
<td>21.63</td>
<td>6.54</td>
<td>16.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(2.39)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.79)</td>
</tr>
<tr>
<td>5</td>
<td>0.70</td>
<td>1.00</td>
<td>50</td>
<td>2.85</td>
<td>49.00</td>
<td>49.00</td>
<td>29.70</td>
<td>1.92</td>
<td>3.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(2.36)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.75)</td>
</tr>
<tr>
<td>10</td>
<td>0.70</td>
<td>1.00</td>
<td>50</td>
<td>10.08</td>
<td>49.00</td>
<td>49.00</td>
<td>30.36</td>
<td>1.05</td>
<td>4.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(6.31)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.76)</td>
</tr>
<tr>
<td>20</td>
<td>0.70</td>
<td>1.00</td>
<td>50</td>
<td>16.99</td>
<td>49.00</td>
<td>49.00</td>
<td>31.69</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(5.49)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.73)</td>
</tr>
</tbody>
</table>

Table 5: Mean and standard error of estimated $K$ when error is sampled from a standard normal distribution and the number of features is small ($m = 50$).
<table>
<thead>
<tr>
<th>$K$</th>
<th>$\sigma^2(K)$</th>
<th>$d^2(K)$</th>
<th>$m$</th>
<th>PPPCA</th>
<th>BIC</th>
<th>Minka BIC</th>
<th>Cumlog</th>
<th>PESEL</th>
<th>GCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.30</td>
<td>0.50</td>
<td>5000</td>
<td>3.90</td>
<td>98.70</td>
<td>95.40</td>
<td>5.00</td>
<td>42.60</td>
<td>5.10</td>
</tr>
<tr>
<td></td>
<td>(0.24)</td>
<td>(1.31)</td>
<td></td>
<td>(5.33)</td>
<td>(0.00)</td>
<td>(47.09)</td>
<td>(0.39)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.30</td>
<td>0.50</td>
<td>5000</td>
<td>9.22</td>
<td>96.77</td>
<td>29.57</td>
<td>9.46</td>
<td>56.28</td>
<td>10.27</td>
</tr>
<tr>
<td></td>
<td>(1.36)</td>
<td>(4.59)</td>
<td></td>
<td>(12.54)</td>
<td>(0.50)</td>
<td>(44.69)</td>
<td>(0.49)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.30</td>
<td>0.50</td>
<td>5000</td>
<td>20.00</td>
<td>38.67</td>
<td>21.94</td>
<td>18.04</td>
<td>62.66</td>
<td>20.00</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(7.80)</td>
<td></td>
<td>(1.29)</td>
<td>(0.35)</td>
<td>(39.57)</td>
<td>(0.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.70</td>
<td>0.50</td>
<td>5000</td>
<td>4.03</td>
<td>5.92</td>
<td>5.00</td>
<td>8.94</td>
<td>45.41</td>
<td>4.00</td>
</tr>
<tr>
<td></td>
<td>(0.17)</td>
<td>(0.88)</td>
<td></td>
<td>(0.00)</td>
<td>(1.14)</td>
<td>(46.78)</td>
<td>(0.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.70</td>
<td>0.50</td>
<td>5000</td>
<td>9.91</td>
<td>10.02</td>
<td>9.99</td>
<td>10.02</td>
<td>51.81</td>
<td>8.15</td>
</tr>
<tr>
<td></td>
<td>(0.40)</td>
<td>(0.14)</td>
<td></td>
<td>(0.14)</td>
<td>(0.14)</td>
<td>(44.66)</td>
<td>(0.36)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.70</td>
<td>0.50</td>
<td>5000</td>
<td>20.00</td>
<td>20.00</td>
<td>20.00</td>
<td>20.00</td>
<td>54.75</td>
<td>14.69</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(39.42)</td>
<td>(0.56)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.30</td>
<td>1.00</td>
<td>5000</td>
<td>4.92</td>
<td>98.67</td>
<td>94.57</td>
<td>5.00</td>
<td>49.18</td>
<td>5.32</td>
</tr>
<tr>
<td></td>
<td>(0.27)</td>
<td>(0.90)</td>
<td></td>
<td>(11.41)</td>
<td>(0.00)</td>
<td>(47.15)</td>
<td>(0.62)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.30</td>
<td>1.00</td>
<td>5000</td>
<td>10.00</td>
<td>91.42</td>
<td>23.83</td>
<td>10.00</td>
<td>51.83</td>
<td>10.18</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(14.47)</td>
<td></td>
<td>(8.66)</td>
<td>(0.00)</td>
<td>(44.64)</td>
<td>(0.44)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.30</td>
<td>1.00</td>
<td>5000</td>
<td>20.00</td>
<td>31.13</td>
<td>21.25</td>
<td>20.00</td>
<td>63.45</td>
<td>20.01</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(3.63)</td>
<td></td>
<td>(1.08)</td>
<td>(0.00)</td>
<td>(39.50)</td>
<td>(0.10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.70</td>
<td>1.00</td>
<td>5000</td>
<td>4.50</td>
<td>5.51</td>
<td>5.00</td>
<td>7.90</td>
<td>47.30</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td>(0.50)</td>
<td>(0.75)</td>
<td></td>
<td>(0.00)</td>
<td>(1.11)</td>
<td>(47.00)</td>
<td>(0.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.70</td>
<td>1.00</td>
<td>5000</td>
<td>10.00</td>
<td>10.01</td>
<td>10.00</td>
<td>10.01</td>
<td>58.95</td>
<td>10.00</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.10)</td>
<td></td>
<td>(0.00)</td>
<td>(0.10)</td>
<td>(44.50)</td>
<td>(0.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.70</td>
<td>1.00</td>
<td>5000</td>
<td>20.00</td>
<td>20.00</td>
<td>20.00</td>
<td>20.00</td>
<td>56.34</td>
<td>20.00</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(39.57)</td>
<td>(0.00)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Mean and standard error of estimated $K$ when error is sampled from a standard normal distribution and the number of features is big ($m = 5,000$).
<table>
<thead>
<tr>
<th>$K$</th>
<th>$\sigma^2(K)$</th>
<th>$d^2(K)$</th>
<th>$m$</th>
<th>PPPCA</th>
<th>BIC</th>
<th>Minka BIC</th>
<th>Cumlog</th>
<th>PESEL</th>
<th>GCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.30</td>
<td>0.50</td>
<td>10000</td>
<td>3.95</td>
<td>98.90</td>
<td>98.64</td>
<td>5.00</td>
<td>50.12</td>
<td>5.01</td>
</tr>
<tr>
<td></td>
<td>(0.30)</td>
<td>(0.39)</td>
<td></td>
<td>(0.84)</td>
<td>(0.00)</td>
<td>(47.20)</td>
<td>(0.10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.30</td>
<td>0.50</td>
<td>10000</td>
<td>9.32</td>
<td>98.55</td>
<td>91.59</td>
<td>9.42</td>
<td>56.28</td>
<td>10.20</td>
</tr>
<tr>
<td></td>
<td>(1.30)</td>
<td>(1.29)</td>
<td></td>
<td>(15.58)</td>
<td>(0.50)</td>
<td>(44.69)</td>
<td>(0.49)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.30</td>
<td>0.50</td>
<td>10000</td>
<td>20.01</td>
<td>87.95</td>
<td>30.24</td>
<td>18.02</td>
<td>62.66</td>
<td>20.01</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(15.93)</td>
<td></td>
<td>(4.43)</td>
<td>(0.35)</td>
<td>(39.57)</td>
<td>(0.10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.70</td>
<td>0.50</td>
<td>10000</td>
<td>4.08</td>
<td>96.82</td>
<td>5.00</td>
<td>5.53</td>
<td>51.06</td>
<td>4.00</td>
</tr>
<tr>
<td></td>
<td>(0.27)</td>
<td>(9.06)</td>
<td></td>
<td>(0.00)</td>
<td>(0.59)</td>
<td>(47.23)</td>
<td>(0.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.70</td>
<td>0.50</td>
<td>10000</td>
<td>10.00</td>
<td>12.75</td>
<td>10.00</td>
<td>10.00</td>
<td>50.94</td>
<td>8.13</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(1.82)</td>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(44.58)</td>
<td>(0.34)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.70</td>
<td>0.50</td>
<td>10000</td>
<td>20.00</td>
<td>20.03</td>
<td>20.00</td>
<td>20.00</td>
<td>57.13</td>
<td>14.76</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.17)</td>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(39.63)</td>
<td>(0.55)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.30</td>
<td>1.00</td>
<td>10000</td>
<td>4.85</td>
<td>98.88</td>
<td>98.46</td>
<td>5.00</td>
<td>54.82</td>
<td>5.18</td>
</tr>
<tr>
<td></td>
<td>(0.36)</td>
<td>(0.33)</td>
<td></td>
<td>(1.57)</td>
<td>(0.00)</td>
<td>(47.15)</td>
<td>(0.41)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.30</td>
<td>1.00</td>
<td>10000</td>
<td>10.00</td>
<td>98.46</td>
<td>87.42</td>
<td>10.00</td>
<td>55.39</td>
<td>10.20</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(1.37)</td>
<td></td>
<td>(19.81)</td>
<td>(0.00)</td>
<td>(44.72)</td>
<td>(0.47)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.30</td>
<td>1.00</td>
<td>10000</td>
<td>20.00</td>
<td>53.69</td>
<td>25.50</td>
<td>20.00</td>
<td>56.34</td>
<td>20.01</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(16.22)</td>
<td></td>
<td>(2.75)</td>
<td>(0.00)</td>
<td>(39.57)</td>
<td>(0.10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.70</td>
<td>1.00</td>
<td>10000</td>
<td>4.46</td>
<td>78.78</td>
<td>5.00</td>
<td>5.13</td>
<td>56.70</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td>(0.50)</td>
<td>(33.14)</td>
<td></td>
<td>(0.00)</td>
<td>(0.34)</td>
<td>(47.00)</td>
<td>(0.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.70</td>
<td>1.00</td>
<td>10000</td>
<td>10.00</td>
<td>11.81</td>
<td>10.00</td>
<td>10.00</td>
<td>54.50</td>
<td>10.00</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(1.44)</td>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(44.72)</td>
<td>(0.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.70</td>
<td>1.00</td>
<td>10000</td>
<td>20.00</td>
<td>20.01</td>
<td>20.00</td>
<td>20.00</td>
<td>65.82</td>
<td>20.00</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.10)</td>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(39.19)</td>
<td>(0.00)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7: Mean and standard error of estimated $K$ when error is sampled from a standard normal distribution and the number of features is large ($m = 10,000$).
Table 8: Mean and standard error of estimated $K$ when error is sampled from a $t$-distribution with degrees of freedom 3 and 10.

<table>
<thead>
<tr>
<th>$\sigma^2(K)$</th>
<th>$d^2(K)$</th>
<th>$\nu$</th>
<th>$K$</th>
<th>PPPCA</th>
<th>BIC</th>
<th>Minka BIC</th>
<th>Cumlog</th>
<th>PESEL</th>
<th>GCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>3.00</td>
<td>5.00</td>
<td>3.60</td>
<td>(1.43)</td>
<td>(0.42)</td>
<td>(1.85)</td>
<td>(1.23)</td>
<td>(29.73)</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>3.00</td>
<td>10.00</td>
<td>6.90</td>
<td>(0.74)</td>
<td>(1.26)</td>
<td>(27.14)</td>
<td>(1.23)</td>
<td>(45.96)</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>3.00</td>
<td>5.00</td>
<td>6.00</td>
<td>(4.27)</td>
<td>(3.24)</td>
<td>(1.63)</td>
<td>(0.79)</td>
<td>(45.41)</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>3.00</td>
<td>10.00</td>
<td>12.80</td>
<td>(3.85)</td>
<td>(6.57)</td>
<td>(3.60)</td>
<td>(1.14)</td>
<td>(28.14)</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>3.00</td>
<td>5.00</td>
<td>4.50</td>
<td>(1.90)</td>
<td>(0.71)</td>
<td>(1.89)</td>
<td>(1.25)</td>
<td>(48.54)</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>3.00</td>
<td>10.00</td>
<td>7.80</td>
<td>(2.25)</td>
<td>(6.92)</td>
<td>(23.69)</td>
<td>(0.82)</td>
<td>(45.96)</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>3.00</td>
<td>5.00</td>
<td>8.80</td>
<td>(3.55)</td>
<td>(7.09)</td>
<td>(3.86)</td>
<td>(2.18)</td>
<td>(48.54)</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>3.00</td>
<td>10.00</td>
<td>15.10</td>
<td>(1.97)</td>
<td>(5.41)</td>
<td>(3.62)</td>
<td>(2.55)</td>
<td>(42.99)</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>10.00</td>
<td>5.00</td>
<td>4.00</td>
<td>(0.00)</td>
<td>(0.67)</td>
<td>(1.60)</td>
<td>(0.00)</td>
<td>(39.63)</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>10.00</td>
<td>10.00</td>
<td>8.40</td>
<td>(1.78)</td>
<td>(0.48)</td>
<td>(15.97)</td>
<td>(0.48)</td>
<td>(45.96)</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>10.00</td>
<td>5.00</td>
<td>4.00</td>
<td>(0.00)</td>
<td>(44.06)</td>
<td>(0.00)</td>
<td>(1.14)</td>
<td>(39.63)</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>10.00</td>
<td>10.00</td>
<td>9.80</td>
<td>(0.63)</td>
<td>(1.48)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(42.99)</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>10.00</td>
<td>5.00</td>
<td>4.90</td>
<td>(0.32)</td>
<td>(1.26)</td>
<td>(1.99)</td>
<td>(0.00)</td>
<td>(49.54)</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>(0.00)</td>
<td>(0.97)</td>
<td>(28.28)</td>
<td>(0.00)</td>
<td>(46.91)</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>10.00</td>
<td>5.00</td>
<td>4.30</td>
<td>(0.48)</td>
<td>(2.75)</td>
<td>(0.00)</td>
<td>(0.57)</td>
<td>(39.63)</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>(0.00)</td>
<td>(0.67)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(45.96)</td>
</tr>
<tr>
<td>(\sigma^2(K))</td>
<td>(d^2(K))</td>
<td>(\rho)</td>
<td>(\nu)</td>
<td>(K)</td>
<td>(\text{PPPCA})</td>
<td>(\text{BIC})</td>
<td>(\text{Minka BIC})</td>
<td>(\text{Cumlog})</td>
<td>(\text{PESEL})</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>0.10</td>
<td>3.00</td>
<td>5.00</td>
<td>3.60</td>
<td>98.80</td>
<td>98.90</td>
<td>12.40</td>
<td>52.94</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>0.10</td>
<td>3.00</td>
<td>10.00</td>
<td>7.20</td>
<td>97.80</td>
<td>46.20</td>
<td>13.30</td>
<td>45.60</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>0.10</td>
<td>3.00</td>
<td>5.00</td>
<td>6.20</td>
<td>45.30</td>
<td>29.20</td>
<td>24.80</td>
<td>70.80</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>0.10</td>
<td>3.00</td>
<td>10.00</td>
<td>18.90</td>
<td>44.50</td>
<td>23.00</td>
<td>26.90</td>
<td>54.50</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.10</td>
<td>3.00</td>
<td>5.00</td>
<td>3.20</td>
<td>98.50</td>
<td>98.50</td>
<td>13.00</td>
<td>42.60</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.10</td>
<td>3.00</td>
<td>10.00</td>
<td>7.80</td>
<td>96.00</td>
<td>42.30</td>
<td>14.40</td>
<td>54.50</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>0.10</td>
<td>3.00</td>
<td>5.00</td>
<td>9.40</td>
<td>44.50</td>
<td>20.60</td>
<td>25.10</td>
<td>42.60</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>0.10</td>
<td>3.00</td>
<td>10.00</td>
<td>(3.75)</td>
<td>(2.76)</td>
<td>(2.17)</td>
<td>(1.60)</td>
<td>(48.54)</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>0.40</td>
<td>3.00</td>
<td>5.00</td>
<td>2.50</td>
<td>98.80</td>
<td>97.90</td>
<td>14.20</td>
<td>70.80</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>0.40</td>
<td>3.00</td>
<td>10.00</td>
<td>6.60</td>
<td>97.60</td>
<td>46.40</td>
<td>14.00</td>
<td>63.40</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>0.40</td>
<td>3.00</td>
<td>5.00</td>
<td>6.30</td>
<td>49.20</td>
<td>27.10</td>
<td>25.90</td>
<td>89.50</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>0.40</td>
<td>3.00</td>
<td>10.00</td>
<td>13.00</td>
<td>56.00</td>
<td>38.20</td>
<td>26.10</td>
<td>72.10</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.40</td>
<td>3.00</td>
<td>5.00</td>
<td>3.80</td>
<td>98.40</td>
<td>97.10</td>
<td>14.10</td>
<td>23.80</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.40</td>
<td>3.00</td>
<td>10.00</td>
<td>(1.03)</td>
<td>(0.84)</td>
<td>(2.60)</td>
<td>(1.37)</td>
<td>(39.63)</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>0.40</td>
<td>3.00</td>
<td>5.00</td>
<td>9.40</td>
<td>47.70</td>
<td>19.80</td>
<td>26.90</td>
<td>23.80</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>0.40</td>
<td>3.00</td>
<td>10.00</td>
<td>(4.95)</td>
<td>(6.52)</td>
<td>(2.35)</td>
<td>(3.07)</td>
<td>(39.63)</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>0.70</td>
<td>3.00</td>
<td>5.00</td>
<td>3.90</td>
<td>98.90</td>
<td>98.90</td>
<td>14.10</td>
<td>70.40</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>0.70</td>
<td>3.00</td>
<td>10.00</td>
<td>7.70</td>
<td>93.60</td>
<td>39.10</td>
<td>20.40</td>
<td>45.60</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>0.70</td>
<td>3.00</td>
<td>5.00</td>
<td>9.80</td>
<td>58.50</td>
<td>31.10</td>
<td>29.80</td>
<td>23.00</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>0.70</td>
<td>3.00</td>
<td>10.00</td>
<td>(6.97)</td>
<td>(14.44)</td>
<td>(23.69)</td>
<td>(4.66)</td>
<td>(40.06)</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>0.70</td>
<td>3.00</td>
<td>10.00</td>
<td>16.20</td>
<td>53.90</td>
<td>24.10</td>
<td>32.10</td>
<td>62.40</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.70</td>
<td>3.00</td>
<td>5.00</td>
<td>4.10</td>
<td>98.40</td>
<td>66.30</td>
<td>18.20</td>
<td>42.60</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.70</td>
<td>3.00</td>
<td>10.00</td>
<td>8.60</td>
<td>82.00</td>
<td>37.50</td>
<td>19.70</td>
<td>36.70</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>0.70</td>
<td>3.00</td>
<td>5.00</td>
<td>13.20</td>
<td>61.60</td>
<td>24.80</td>
<td>30.60</td>
<td>61.40</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>0.70</td>
<td>3.00</td>
<td>10.00</td>
<td>(7.97)</td>
<td>(13.98)</td>
<td>(4.05)</td>
<td>(3.34)</td>
<td>(48.54)</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>0.70</td>
<td>3.00</td>
<td>10.00</td>
<td>(7.73)</td>
<td>(10.67)</td>
<td>(7.66)</td>
<td>(4.16)</td>
<td>(46.91)</td>
</tr>
</tbody>
</table>

Table 9: Mean and standard error of estimated \(K\) when error is sampled from a \(t\)-distribution with 3 degrees of freedom and autocorrelation.
<table>
<thead>
<tr>
<th>$\sigma^2(K)$</th>
<th>$d^2(K)$</th>
<th>$\rho$</th>
<th>$\nu$</th>
<th>$K$</th>
<th>PPCA</th>
<th>BIC</th>
<th>Minka BIC</th>
<th>Cumlog</th>
<th>PESRE</th>
<th>GCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>0.10</td>
<td>10.00</td>
<td>5.00</td>
<td>4.00</td>
<td>99.0</td>
<td>98.70</td>
<td>5.00</td>
<td>42.60</td>
<td>5.00</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>0.10</td>
<td>10.00</td>
<td>10.00</td>
<td>9.50</td>
<td>98.60</td>
<td>87.40</td>
<td>9.30</td>
<td>45.60</td>
<td>10.00</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>0.10</td>
<td>10.00</td>
<td>5.00</td>
<td>4.00</td>
<td>81.20</td>
<td>5.00</td>
<td>8.30</td>
<td>70.80</td>
<td>4.00</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>0.10</td>
<td>10.00</td>
<td>10.00</td>
<td>9.80</td>
<td>12.00</td>
<td>10.00</td>
<td>10.20</td>
<td>54.50</td>
<td>8.00</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.10</td>
<td>10.00</td>
<td>5.00</td>
<td>4.00</td>
<td>98.90</td>
<td>98.70</td>
<td>5.00</td>
<td>42.60</td>
<td>5.00</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.10</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>98.10</td>
<td>65.80</td>
<td>10.00</td>
<td>54.50</td>
<td>10.00</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>0.10</td>
<td>10.00</td>
<td>5.00</td>
<td>4.00</td>
<td>13.50</td>
<td>5.00</td>
<td>7.10</td>
<td>52.00</td>
<td>5.00</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>0.10</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>63.40</td>
<td>10.00</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>0.40</td>
<td>10.00</td>
<td>5.00</td>
<td>4.00</td>
<td>99.00</td>
<td>98.70</td>
<td>5.00</td>
<td>42.60</td>
<td>5.00</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>0.40</td>
<td>10.00</td>
<td>10.00</td>
<td>7.00</td>
<td>98.70</td>
<td>92.80</td>
<td>9.40</td>
<td>63.40</td>
<td>10.00</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>0.40</td>
<td>10.00</td>
<td>5.00</td>
<td>4.00</td>
<td>42.40</td>
<td>5.00</td>
<td>12.00</td>
<td>61.40</td>
<td>4.00</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>0.40</td>
<td>10.00</td>
<td>10.00</td>
<td>9.40</td>
<td>11.40</td>
<td>10.00</td>
<td>10.80</td>
<td>54.50</td>
<td>7.30</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.40</td>
<td>10.00</td>
<td>5.00</td>
<td>4.00</td>
<td>98.80</td>
<td>98.70</td>
<td>5.00</td>
<td>70.80</td>
<td>5.00</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.40</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>96.90</td>
<td>56.20</td>
<td>10.00</td>
<td>72.30</td>
<td>10.00</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>0.40</td>
<td>10.00</td>
<td>5.00</td>
<td>4.00</td>
<td>19.80</td>
<td>5.00</td>
<td>10.80</td>
<td>61.40</td>
<td>4.90</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>0.40</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.90</td>
<td>54.50</td>
<td>9.70</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>0.70</td>
<td>10.00</td>
<td>5.00</td>
<td>4.00</td>
<td>98.30</td>
<td>92.10</td>
<td>7.40</td>
<td>48.70</td>
<td>4.00</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>0.70</td>
<td>10.00</td>
<td>10.00</td>
<td>6.20</td>
<td>98.30</td>
<td>47.70</td>
<td>9.70</td>
<td>36.70</td>
<td>9.20</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>0.70</td>
<td>10.00</td>
<td>5.00</td>
<td>3.30</td>
<td>28.00</td>
<td>4.20</td>
<td>24.90</td>
<td>51.50</td>
<td>3.00</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>0.70</td>
<td>10.00</td>
<td>10.00</td>
<td>8.40</td>
<td>21.80</td>
<td>8.80</td>
<td>5.70</td>
<td>81.00</td>
<td>5.70</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.70</td>
<td>10.00</td>
<td>5.00</td>
<td>4.00</td>
<td>98.50</td>
<td>98.20</td>
<td>6.10</td>
<td>42.60</td>
<td>5.00</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.70</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>97.50</td>
<td>28.70</td>
<td>10.00</td>
<td>72.30</td>
<td>10.00</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>0.70</td>
<td>10.00</td>
<td>5.00</td>
<td>4.00</td>
<td>23.00</td>
<td>5.00</td>
<td>23.00</td>
<td>42.60</td>
<td>4.00</td>
</tr>
<tr>
<td>0.70</td>
<td>1.00</td>
<td>0.70</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>48.54</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 10: Mean and standard error of estimated $K$ when error is sampled from a $t$-distribution with 10 degrees of freedom and autocorrelation.

References


