Electronic Journal of Statistics
Vol. 15 (2021) 2157–2178
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1837

Uncertainty quantification for principal component regression*

Suofei Wu[1] **Jan Hannig**[2] **Thomas C. M. Lee**[1]

[1]*Department of Statistics, University of California, Davis, One Shields Avenue, Davis, CA 95616, U.S.A.*
e-mail: swu@ucdavis.edu; tcmlee@ucdavis.edu

[2]*Department of Statistics & Operations Research, 318 Hanes Hall, University of North Carolina at Chapel Hill, NC 27599, U.S.A.*
e-mail: jan.hannig@unc.edu

Abstract: Principal component regression is an effective dimension reduction method for regression problems. To apply it in practice, one typically starts by selecting the number of principal components k, then estimates the corresponding regression parameters using say maximum likelihood, and finally obtains predictions with the fitted results. The success of this approach highly depends on the choice of k, and very often, due to the noisy nature of the data, it could be risky to just use one single value of k. Using the generalized fiducial inference framework, this paper develops a method for constructing a probability function on k, which provides an uncertainty measure on its value. In addition, this paper also constructs novel confidence intervals for the regression parameters and prediction intervals for future observations. The proposed methodology is backed up by theoretical results and is tested by simulation experiments and compared with other methods using real data. To the best of our knowledge, this is the first time that a full treatment for uncertainty quantification is formally considered for principal component regression.

Keywords and phrases: Confidence intervals, fiducial inference, high-dimensional data, model dimension selection.

Received September 2019.

1. Introduction

The high-dimensional regression problem has attracted enormous attention in recent years. A typical linear model can be expressed as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{u}$, where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is a vector of responses, $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^T$ is a design matrix of size $n \times p$. Also, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ is a vector of p unknown parameters, σ^2 is the unknown noise variance, and $\boldsymbol{u} = (u_1, ..., u_n)^T$ is a vector of n i.i.d. normal random variables with zero mean and variance one. It is assumed that $\boldsymbol{u}$ and $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are independent.

A high-dimensional model is the one with $p \gg n$. In many situations, the large number of predictors are often correlated. For such cases, principal com-

*This work was partially supported by the National Science Foundation under grants DMS-1512893, DMS-1512945, IIS-1633074, DMS-1811405, DMS-1811661, DMS-1916115. DMS-1916125 and CCF-1934568.

ponent regression, which uses principal components (PCs) instead of the original predictors, is a powerful method that reduces the dimension and orthogonalizes the regression problem (Jolliffe, 1982). The idea of applying principal component analysis in regression was first discussed by Hotelling (1957). Some recent engineering applications of PC regression can be found in Adusumilli *et al.* (2015); Huang and Yang (2012); Lipponen *et al.* (2010); Xu *et al.* (2013); Zhu *et al.* (2018). The additional structure, imposed by the assumption that the directions of biggest predictor variability are also important for explaining the dependent variable, allows for more specialized and efficient statistical methods than those that do not make this assumptions such as Lai *et al.* (2015).

Here is a brief description of PC regression. Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^T$ be the singular value decomposition of $\boldsymbol{X}$, where $\boldsymbol{L} = \text{diag}\{l_i\}$ are the non-negative singular values of $\boldsymbol{X}$. The columns of $\boldsymbol{U}_{n\times p}$ and $\boldsymbol{V}_{p\times p}$ are, respectively, orthonormal left and right singular vectors of $\boldsymbol{X}$. For any $k \in \{1, \ldots, p\}$, let $\boldsymbol{V}_k$ denote the $p \times k$ matrix having the first k columns of $\boldsymbol{V}$. Then $\boldsymbol{X}_k = \boldsymbol{X}\boldsymbol{V}_k$ is a $n \times k$ matrix having the first k principal components as its columns. In PC regression it is customary to assume that $\boldsymbol{Y}$ is generated by the first $k_0 < n$ principal components of $\boldsymbol{X}$ (k_0 unknown). That is, $\boldsymbol{Y}$ is generated from the model

$$\boldsymbol{Y} = \boldsymbol{X}_{k_0}\boldsymbol{\gamma}_{k_0} + \sigma\boldsymbol{u}, \tag{1}$$

where $\boldsymbol{\gamma}_{k_0} = (\gamma_1, ..., \gamma_{k_0})^T$ is a vector of k_0 unknown parameters and σ is the unknown standard deviation of the noise, and $\boldsymbol{u} = (u_1, ..., u_n)^T$ is a vector i.i.d. standard normal random variables. The unknown k_0 can be estimated by using a model selection criterion such as AIC or BIC; denote the resulting estimate as $\hat{k}_0$. Then one can estimate $\boldsymbol{\gamma}_{\hat{k}_0}$ for example by regressing $\boldsymbol{Y}$ on $\boldsymbol{X}_{\hat{k}_0}$. The final PC regression estimator of $\boldsymbol{\beta}$ based on the first $\hat{k}_0$ principal components is then given by $\widehat{\boldsymbol{\beta}}_{\hat{k}_0} = \boldsymbol{V}_{\hat{k}_0}\widehat{\boldsymbol{\gamma}}_{\hat{k}_0} \in \mathbb{R}^p$.

A top-down selection rule was used for deciding which PCs should be kept in the model (Xie and Kalivas, 1997). That is, one always selects the $\hat{k}_0$ PCs that correspond to the $\hat{k}_0$ largest singular values. Some alternative selection rules have also been proposed. For example, Sun (1995) suggested using correlation principal component regression in which the PCs are orderd by their correlations with the response, and Sutter *et al.* (1992) treated the PC selection as an optimization problem. However, Xie and Kalivas (1997) showed that the top-down approach described in the previous paragraph generates the most stable global model. In sequel, this paper will follow this top-down approach, although the methodology can be straightforwardly extended to other selection rules.

Although many researchers have worked on the problem of choosing the number of PCs (i.e., k_0, the model size) in PC regression, it seems that the issue of uncertainty quantification has received very little treatment. To fill this important gap, this paper develops a fully automatic method that quantifies the uncertainties in the estimates for the model parameters ($\boldsymbol{\beta}$ and σ^2), model size (k_0), and prediction error. The proposed method is based on *generalized fiducial inference* (Hannig *et al.*, 2016). To the best of our knowledge, this is the first time that a full treatment for uncertainty quantification is formally considered

in PC regression.

2. An introduction to generalized fiducial inference

Bayesian inference is an important methodology in statistics. It provides a "distribution estimate" for the unknown parameter, which has a wealthier information comparing to frequentist inference (Xie and Singh, 2013). However, the over-enthusiastic application of this methodology when prior information is unknown has also caused concerns (Efron, 2013). To avoid such potential issue, Fisher (1930) introduced fiducial inference. Instead of using uninformative prior as in Bayesian inference, Fisher considered a switching principle, which is similar to the idea of maximum likelihood, to assign a prior using the information from the observed data. However, for many years, Fisher's idea did not attract much attention from the majority of statisticians.

Recently, there has been a resurgence of interest in the variant of fiducial inference. The modern contributions include Dempster-Shafer theory (Dempster, 2008), its related work called inferential models (Martin and Liu, 2015), confidence distribution (Xie and Singh, 2013), and more generally fusion learning Cheng *et al.* (2014).

The particular variant of Fisher's fiducial idea that this paper considers is the so-called *generalized fiducial inference* (GFI). The success of GFI for conducting statistical inference has been demonstrated in many areas, including both traditional and modern problems; see Hannig *et al.* (2016) and reference therein.

Generalized fiducial inference begins with expressing the relationship between the data $\boldsymbol{y}$ and the parameter $\boldsymbol{\theta}$ as

$$\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{G}(\cdot, \cdot)$ denotes the so-called data generating algorithm, and $\boldsymbol{u}$ is the random component whose distribution is completely known; for example, an i.i.d. N(0,1) random vector. Similar to maximum likelihood estimation, in GFI the roles of $\boldsymbol{y}$ and $\boldsymbol{\theta}$ are switched: the random $\boldsymbol{y}$ is treated as deterministic in the likelihood function, while the deterministic $\boldsymbol{\theta}$ is treated as random. With this, we can define a set $\{\boldsymbol{\theta} : \boldsymbol{y} = \boldsymbol{G}(\boldsymbol{u}^*, \boldsymbol{\theta})\}$ as the inverse mapping of $\boldsymbol{G}$, where $\boldsymbol{u}^*$ is an independent copy of $\boldsymbol{u}$. Note that such an inverse does not always exist: there may be either no $\boldsymbol{\theta}$ or more than one $\boldsymbol{\theta}$ such that $\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{u}^*, \boldsymbol{\theta})$. For the first case, we remove the values of $\boldsymbol{u}$ for which there is no solution from the sample space and re-normalize the probability. For the second case, Hannig (2009) suggested randomly picking one element from $\{\boldsymbol{\theta} : \boldsymbol{y} = \boldsymbol{G}(\boldsymbol{u}^*, \boldsymbol{\theta})\}$.

A probability distribution on $\boldsymbol{\theta}$ can be defined through (2) in the following manner. Suppose for any observed $\boldsymbol{y}$ and all $\boldsymbol{u}$, there exists a unique $\boldsymbol{\theta}$ such that $\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})$. That is, the inverse

$$\boldsymbol{Q}_{\boldsymbol{y}}(\boldsymbol{u}) = \{\boldsymbol{\theta} : \boldsymbol{y} = \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})\} \tag{3}$$

always exists. Recall that the distribution of $\boldsymbol{u}$ is assumed to be completely known, one can always generate a random sample $\tilde{\boldsymbol{u}}_1$, $\tilde{\boldsymbol{u}}_2$, ..., and via (3), this sample can be transformed into a random sample of $\boldsymbol{\theta}$: $\tilde{\boldsymbol{\theta}}_1 = \boldsymbol{Q}_{\boldsymbol{y}}(\tilde{\boldsymbol{u}}_1), \tilde{\boldsymbol{\theta}}_2 = \boldsymbol{Q}_{\boldsymbol{y}}(\tilde{\boldsymbol{u}}_2), ...$ In below we shall call $\{\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \ldots\}$ a fiducial sample of $\boldsymbol{\theta}$. As with a posterior sample in the Bayesian context, we can use it to calculate the point estimates and also construct confidence intervals for $\boldsymbol{\theta}$. Meanwhile, we can also obtain the density $r(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$. Here $r(\boldsymbol{\theta})$ is called the generalized fiducial density for $\boldsymbol{\theta}$, and plays a similar role as the posterior density in the Bayesian context.

Observe that, for the PC regression problem the data generating algorithm is (1) and $\boldsymbol{\theta}$ contains three components: $\boldsymbol{\theta} = \{k, \sigma, \boldsymbol{\gamma}_k\}$, where k denotes the number of principal components, σ^2 is the noise variance, and $\boldsymbol{\beta}_k = V_k \boldsymbol{\gamma}_k$ is the vector of regression coefficients estimated using k principal components.

Next we will use the idea of GFI to obtain a generalized fiducial density for k and construct confidence intervals for $\boldsymbol{\beta}_k$, σ^2 and prediction for $\boldsymbol{y}$ (Wang *et al.*, 2012; Shen *et al.*, 2018).

3. GFI for principal component regression

While the above description for GFI seems conceptually simple and general, it may not be directly applicable in some situations. When the model dimension is known, Hannig (2013) derived a workable formula for $r(\boldsymbol{\theta})$ applicable in most situations where the data follows a continuous distribution. In what follows, we assume that for the observed $\boldsymbol{y}$ and all $\boldsymbol{\theta}$ (2) has a unique solution $\boldsymbol{u} = \boldsymbol{G}^{-1}(\boldsymbol{y}, \boldsymbol{\theta})$.

The next theorem derives fiducial density for a single model.

Theorem 3.1 (Theorem 1 of Hannig *et al.* (2016))**.** *Under some differentiability assumptions, the generalized fiducial distribution is absolutely continuous and has density*

$$r(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}, \boldsymbol{\theta}) J(\boldsymbol{y}, \boldsymbol{\theta})}{\int_\Theta f(\boldsymbol{y}, \boldsymbol{\theta}') J(\boldsymbol{y}, \boldsymbol{\theta}') d\boldsymbol{\theta}'}, \tag{4}$$

where $f(\boldsymbol{y}, \boldsymbol{\theta})$ is the likelihood and the function

$$J(\boldsymbol{y}, \boldsymbol{\theta}) = D\left\{\nabla_{\boldsymbol{\theta}} \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})|_{\boldsymbol{u}=\boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})}\right\} \text{ with } D(\boldsymbol{A}) = \{\det(\boldsymbol{A}^T \boldsymbol{A})\}^{\frac{1}{2}}. \tag{5}$$

However, for the current problem, dimension k_0 is unknown so it has to be included as a parameter in the data generating equation. Since k is a discrete parameter, Theorem 3.1 is not directly applicable. The following theorem gives fiducial probability in the context of model selection:

Theorem 3.2 (Theorem 4 of Hannig *et al.* (2016))**.** *Under identifiability and regularity assumptions (in particular the number of parameters $|M| \leq n$) the marginal generalized fiducial probability of model M is*

$$r(M) = \frac{q^{|M|} \int_{\boldsymbol{\Theta}_M} f_M(\boldsymbol{y}, \boldsymbol{\theta}_M) J_M(\boldsymbol{y}, \boldsymbol{\theta}_M)\, d\boldsymbol{\theta}_M}{\sum_{M' \in \mathcal{M}} q^{|M'|} \int_{\boldsymbol{\Theta}_{M'}} f_{M'}(\boldsymbol{y}, \boldsymbol{\theta}_{M'}) J_{M'}(\boldsymbol{y}, \boldsymbol{\theta}_{M'})\, d\boldsymbol{\theta}_{M'}}, \tag{6}$$

where $f_M(\boldsymbol{y}, \boldsymbol{\theta}_M)$ *is the likelihood under model* M *and* $J_M(\boldsymbol{y}, \boldsymbol{\theta}_M)$ *is the Jacobian* (5). *If* $|M| > n$ *then* $r(M) = 0$.

Notice that (6) bears some similarities to the Bayesian posterior probability of a model, with the integral $\int_{\boldsymbol{\Theta}_M} f_M(\boldsymbol{y}, \boldsymbol{\theta}_M) J_M(\boldsymbol{y}, \boldsymbol{\theta}_M)\, d\boldsymbol{\theta}_M$ playing the role of marginal probability of the data, and $q^{|M|}$playing a role of a prior model probability. However, (6) is not a result of a Bayes theorem. Rather, it is derived by inverting the data generating algorithm.

In our problem $\boldsymbol{\theta}_M = \{\sigma, \boldsymbol{\gamma}_k\}$ and $|M| = k + 1$, so in what follows we will simplify notation by replacing the subscript M with the subscript k. We will also follow the minimum description length principle (e.g., Barron *et al.*, 1998; Rissanen, 1996) and use $q = e^{-\frac{1}{2}\log n}$.

Let us first calculate $J_k(\boldsymbol{y}, \boldsymbol{\theta}_k)$ for a fixed k using the data generating algorithm (1) and formula (5). Denote the residual sum of squares as RSS_k when the corresponding $\boldsymbol{\gamma}_k$ is estimated by maximum likelihood using the first k principal components. Direct use of (5) and Cauchy-Binnet formula gives

$$J_k(\boldsymbol{y}, \sigma, \boldsymbol{\gamma}_k) = D\left(\boldsymbol{X}_k, \frac{\boldsymbol{y} - \boldsymbol{X}_k\boldsymbol{\gamma}_k}{\sigma}\right) = \sigma^{-1} |\det(\boldsymbol{X}_k^T \boldsymbol{X}_k)|^{\frac{1}{2}} \mathrm{RSS}_k^{\frac{1}{2}}.$$

Next we will calculate

$$\begin{aligned}
&\int f_k(\boldsymbol{y}, \sigma, \boldsymbol{\gamma}_k) J_k(\boldsymbol{y}, \sigma, \boldsymbol{\gamma}_k)\, d\sigma\, d\boldsymbol{\gamma}_k \\
&\qquad = \int \sigma^{-1} [\det(\boldsymbol{X}_k^T \boldsymbol{X}_k)]^{\frac{1}{2}} \mathrm{RSS}_k^{\frac{1}{2}} \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \\
&\qquad\quad \times \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}_k\boldsymbol{\gamma}_k)^T (\boldsymbol{y} - \boldsymbol{X}_k\boldsymbol{\gamma}_k)\right\} d\sigma\, d\gamma_k \\
&\qquad = \mathrm{RSS}_k^{-\frac{n-k-1}{2}} (\pi)^{-\frac{n-k}{2}} \Gamma\left(\frac{n-k}{2}\right).
\end{aligned}$$

Consequently, the marginal generalized fiducial probability (6) is proportional to

$$r(k) \propto n^{-\frac{k+1}{2}} \Gamma\left(\frac{n-k}{2}\right) (\pi \mathrm{RSS}_k)^{-\frac{n-k-1}{2}}. \tag{7}$$

3.1. Practical generation of fiducial sample

This subsection presents a practical procedure for generating a fiducial sample $\{\tilde{k}, \tilde{\sigma}^2, \tilde{\boldsymbol{\beta}}\}$. First, for any fixed k, it is straightforward to show that the generalized fiducial distribution of σ^2 conditional on k is

$$\sigma^2 \sim \mathrm{RSS}_k / \chi^2(n-k), \tag{8}$$

and the generalized fiducial distribution of $\boldsymbol{\gamma}_k$ conditional on (k, σ^2) is $\boldsymbol{\gamma}_k \sim N(\hat{\boldsymbol{\gamma}}_k^{\mathrm{ML}}, \sigma^2 (\boldsymbol{X}_k^T \boldsymbol{X}_k)^{-1})$, where $\hat{\boldsymbol{\gamma}}_k^{\mathrm{ML}}$ is the maximum likelihood estimate of $\boldsymbol{\gamma}_k$. Consequently, the generalized fiducial distribution of $\boldsymbol{\beta}_k = V_k \boldsymbol{\gamma}_k$ given (k, σ^2) is

$$\boldsymbol{\beta}_k \sim N(\boldsymbol{V}_k \hat{\boldsymbol{\gamma}}_k^{\mathrm{ML}}, \sigma^2 \boldsymbol{V}_k (\boldsymbol{X}_k^T \boldsymbol{X}_k)^{-1} \boldsymbol{V}_k^T). \tag{9}$$

Lastly, we will approximate the generalized fiducial density $r(k)$ for k in (7) as follows. For $k \in \{0, ..., n-1\}$, calculate

$$R(k) = \Gamma(\frac{n-k}{2})(\pi \mathrm{RSS}_k)^{-\frac{n-k-1}{2}} n^{-\frac{k+1}{2}},$$

and

$$r(k) = \frac{R(k)}{\sum_{k'=0}^{n-1} R(k')}. \tag{10}$$

With the above, we can generate a fiducial sample $\{\tilde{k}, \tilde{\sigma}^2, \tilde{\boldsymbol{\beta}}\}$ with the following steps:

1. Generate a $\tilde{k}$ from (10).
2. With $\tilde{k}$, generate a $\tilde{\sigma}^2$ from (8).
3. Obtain the maximum likelihood estimate $\hat{\boldsymbol{\gamma}}_k^{\mathrm{ML}} = (\boldsymbol{X}_k^T \boldsymbol{X}_k)^{-1} \boldsymbol{X}_k^T \boldsymbol{y}$ where $k = \tilde{k}$.
4. With $\tilde{k}$, $\tilde{\sigma}^2$ and $\hat{\boldsymbol{\gamma}}_k^{\mathrm{ML}}$, generate a $\tilde{\boldsymbol{\beta}}_k$ from (9).

Repeating the above steps one can obtain multiple copies of $\{\tilde{k}, \tilde{\sigma}^2, \tilde{\boldsymbol{\beta}}_k\}$. With these one can form point estimates and confidence intervals for the unknown parameters in a similar manner as with a Bayesian posterior sample. For example, The average of all $\tilde{\sigma}^2$ can be used as an estimate for σ^2, while the 2.5% smallest and 2.5% largest $\tilde{\sigma}^2$ values can be used as, respectively, the lower and upper limits for a 95% confidence interval for σ^2. Steps for obtaining estimates and confidence intervals for $\boldsymbol{\beta}$ and prediction intervals for $\boldsymbol{y}$ are similar.

4. Theoretical properties

We have the following theorem for which the proof is delayed to the appendix.

Theorem 4.1. *Let $\boldsymbol{H}_k$ be the projection matrix of $\boldsymbol{X}_k$; i.e., $\boldsymbol{H}_k = \boldsymbol{X}_k(\boldsymbol{X}_k^T \times \boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^T$. Let $\Delta_k = ||\boldsymbol{\mu} - \boldsymbol{H}_k\boldsymbol{\mu}||^2$, where $\boldsymbol{\mu} = E(\boldsymbol{y}) = \boldsymbol{X}_{k_0}\boldsymbol{\gamma}_{k_0}$. Assume*

$$\lim_{p\to\infty} \min\left\{\frac{\Delta_k}{k_0 \log(p)} : k < k_0\right\} = \infty. \tag{11}$$

As $n \to \infty$, $p \to \infty$, $\log p = o(n)$ and $k_0 = o(\log(n))$, for $K = o(n)$, we have

$$R(k_0)/\sum_{k=1}^{K} R(k) \xrightarrow{p} 1.$$

We remark that the Assumption (11) ensures that the true model is identifiable. We also remark that Theorem 4.1 implies that the confidence intervals constructed using the generalized fiducial density (7) will have correct asymptotic coverage, and the generalized fiducial distribution and the derived point estimators are consistent.

5. Simulation results

Simulation experiments are conducted to evaluate the practical performance of the proposed GFI method. We set $n = 100$ and test different combinations of p, k_0 and σ^2. For any fixed combination of (p, k_0, σ^2), we first generated $\boldsymbol{X}$ (of size $n \times p$) where its elements are i.i.d. $N(0,1)$. Let $\boldsymbol{X}_{k_0}$ be the $n \times k_0$ matrix having the first k_0 principal components of $\boldsymbol{X}$ as its columns. Then the noisy *training data* were generated by $\boldsymbol{y} = \boldsymbol{X}_{k_0}\gamma_{k_0} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a vector of n i.i.d. $N(0, \sigma^2)$ noise random variables. We then applied the proposed generalized fiducial procedure to obtain 1,000 fiducial samples $\{\tilde{k}, \tilde{\sigma}^2, \tilde{\boldsymbol{\beta}}\}$, from which the generalized fiducial confidence intervals for σ^2 and the regression coefficients $\boldsymbol{\beta}_k$ can be computed.

We use 2 values of $p = (200, 1000)$, 3 values of $k_0 = (2, 5, 10)$, 2 values of $\sigma = (0.5, 1)$ and 2 values of $\boldsymbol{\gamma}_{k_0} = (\{1, ..., 1\}, \{5, ..., 5\})$; thus a total of $2 \times 3 \times 2 \times 2 = 24$ experimental configurations are considered. For each experimental configuration, we simulated $1,000$ datasets of size $n = 100$, and for each data set generalized fiducial confidence intervals are obtained for σ^2 and β_1. For comparison, we also report results obtained from two methods, Oracle and BIC. For Oracle the true k_0 is used and the confidence intervals are calculated using classical linear model theory, while for BIC, k_0 is estimated using BIC and the confidence intervals are calculated using the same classical linear model theory. The results are summarized in the Tables 1, 2, 5 and 6. One can see that the performance of the proposed method is very close to Oracle, and is superior to BIC.

Additional *testing data* $(\boldsymbol{X}^*, \boldsymbol{y}^*)$ are generated to assess the qualities of the confidence intervals for $E(\boldsymbol{y}^*|\boldsymbol{X}^*)$ and the prediction intervals for $\boldsymbol{y}^*$ obtained by the proposed method. These testing data are generated as follows to ensure $\boldsymbol{X}$ and $\boldsymbol{X}^*$ to have the same PC structure. First we set $n = 100$ and generated a $n \times p$ matrix $\boldsymbol{X}_*$ with its elements as i.i.d. $N(0,1)$. Then we apply the singular value decomposition to both $\boldsymbol{X}$ and $\boldsymbol{X}_*$ to obtain $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^T$ and $\boldsymbol{X}_* = \boldsymbol{U}_*\boldsymbol{L}_*\boldsymbol{V}_*^T$, respectively. Lastly the design matrix is calculated as $\boldsymbol{X}^* = \boldsymbol{U}_*\boldsymbol{L}_*\boldsymbol{V}^T$, and the response vector as $\boldsymbol{y}^* = \boldsymbol{X}^*_{k_0}\boldsymbol{\gamma}_{k_0} + \boldsymbol{\epsilon}$, where similarly $\boldsymbol{X}^*_{k_0}$ contains the first k_0 PCs of $\boldsymbol{X}^*$. We use the fiducial samples $\{\tilde{k}, \tilde{\sigma}^2, \tilde{\boldsymbol{\beta}}\}$ obtained from the *training data* to construct confidence intervals for the first test data $E(y_1^*|\boldsymbol{X})$ and prediction intervals for y_1^*. The empirical coverage rates are reported in Tables 3, 4, 7 and 8. As before, the proposed method performs very well.

Lastly, Tables 9 and 10 summarize how well the proposed method selects the correct model. In particular we show the percentage of times the correct model selected, the coverage and average size of 95% high probability confidence interval for k_0. These intervals were obtained by sorting the candidate models according to their fiducial probability and taking the smallest number that adds up to at least 95% fiducial probability. The tables show that the correct model usually has the highest fiducial probability. Even if the true model does not have the highest fiducial probability, it is almost always in the 95% confidence interval. Moreover, the average size of the CIs for $n = 1000$ is very close to 1. This agrees with Theorem 4.1, that states that the fiducial distribution concentrates on the true model as n increases.

TABLE 1

Empirical coverage rates for the confidence intervals for σ^2 obtained by different methods when $\boldsymbol{\gamma}_{k_0} = \{1, ..., 1\}$ and $n = 100$. The numbers in parentheses are the averaged width of the intervals.

$\sigma = 0.5$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	87.9 (0.15)	92.7 (0.15)	98.6 (0.19)
	BIC	91.4 (51.82)	95.1 (51.82)	99.7 (1291.77)
	Oracle	88.1 (0.15)	92.8 (0.15)	98.5 (0.2)
p=200, $k_0 = 5$	proposed	91.2 (0.15)	95.5 (0.15)	99.4 (0.2)
	BIC	91.6 (51.44)	94.9 (51.44)	99.3 (1282.21)
	Oracle	90.8 (0.15)	95.5 (0.15)	99.5 (0.2)
p=200, $k_0 = 10$	proposed	90.3 (0.15)	95.8 (0.15)	98.5 (0.2)
	BIC	91.4 (69.22)	96.1 (69.22)	99.2 (1726.27)
	Oracle	90.6 (0.15)	95.7 (0.15)	98.9 (0.2)
p=1000, $k_0 = 2$	proposed	90.5 (0.14)	95.2 (0.14)	98.4 (0.19)
	BIC	92.2 (52.17)	95.9 (52.17)	99.1 (1301.91)
	Oracle	90.6 (0.15)	95.4 (0.15)	98.9 (0.2)
p=1000, $k_0 = 5$	proposed	89.5 (0.15)	94.7 (0.15)	98.8 (0.2)
	BIC	91.0 (59.33)	95.2 (59.33)	98.9 (1479.82)
	Oracle	89.4 (0.15)	94.3 (0.15)	99.0 (0.2)
p=1000, $k_0 = 10$	proposed	90.5 (0.15)	94.4 (0.15)	99.1 (0.2)
	BIC	92.6 (66.18)	96.4 (66.18)	99.6 (1650.3)
	Oracle	90.2 (0.15)	94.9 (0.15)	99.6 (0.21)
$\sigma = 1$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	89.6 (0.58)	95.1 (0.58)	98.8 (0.77)
	BIC	92.1 (198.86)	96.9 (198.86)	99.6 (4957.77)
	Oracle	90.0 (0.58)	95.6 (0.58)	98.8 (0.78)
p=200, $k_0 = 5$	proposed	90.9 (0.59)	95.8 (0.59)	99.0 (0.79)
	BIC	91.5 (234.3)	95.3 (234.3)	99.6 (5840.47)
	Oracle	90.9 (0.59)	95.3 (0.59)	99.3 (0.8)
p=200, $k_0 = 10$	proposed	90.7 (0.61)	94.7 (0.61)	99.0 (0.82)
	BIC	92.2 (285.27)	96.3 (285.27)	98.7 (7116.35)
	Oracle	90.2 (0.61)	95.4 (0.61)	98.7 (0.82)
p=1000, $k_0 = 2$	proposed	89.6 (0.57)	94.3 (0.57)	98.3 (0.76)
	BIC	90.1 (209.06)	95.1 (209.06)	99.0 (5219.15)
	Oracle	90.1 (0.57)	94.3 (0.57)	99.2 (0.77)
p=1000, $k_0 = 5$	proposed	90.0 (0.59)	95.1 (0.59)	99.2 (0.79)
	BIC	90.9 (206.92)	95.3 (206.92)	99.0 (5157.02)
	Oracle	89.9 (0.59)	95.7 (0.59)	99.1 (0.8)
p=1000, $k_0 = 10$	proposed	90.4 (0.61)	95.3 (0.61)	99.2 (0.81)
	BIC	91.6 (262.44)	96.2 (262.44)	99.2 (6544.73)
	Oracle	91.3 (0.61)	95.6 (0.61)	99.2 (0.82)

6. Real data example

Lan *et al.* (2006) conducted an experiment to examine the genetics of two inbred mouse populations (B6 and BTBR). Expression levels of 22575 genes of 31 female and 29 male mice were recorded. Some physiological phenotypes, includ-

TABLE 2
Similar to Table 1 but for β_1.

$\sigma = 0.5$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	92.1 (0.02)	96.0 (0.02)	99.2 (0.03)
	BIC	87.4 (2.84)	93.6 (2.84)	98.6 (14.03)
	Oracle	89.8 (0.02)	95.1 (0.02)	98.8 (0.02)
p=200, $k_0 = 5$	proposed	91.9 (0.03)	95.9 (0.03)	99.5 (0.04)
	BIC	85.2 (3.08)	92.0 (3.08)	98.4 (15.22)
	Oracle	90.7 (0.03)	95.1 (0.03)	99.3 (0.04)
p=200, $k_0 = 10$	proposed	90.3 (0.04)	94.9 (0.04)	99.0 (0.06)
	BIC	85.1 (3.41)	91.8 (3.41)	98.7 (16.8)
	Oracle	90.4 (0.04)	94.9 (0.04)	99.3 (0.05)
p=1000, $k_0 = 2$	proposed	91.3 (0.0)	95.8 (0.0)	99.1 (0.0)
	BIC	85.8 (0.05)	92.5 (0.05)	98.3 (0.25)
	Oracle	89.9 (0.0)	94.8 (0.0)	98.4 (0.0)
p=1000, $k_0 = 5$	proposed	90.4 (0.0)	95.4 (0.0)	98.5 (0.01)
	BIC	86.3 (0.06)	93.4 (0.06)	98.9 (0.27)
	Oracle	90.1 (0.0)	95.1 (0.0)	98.8 (0.01)
p=1000, $k_0 = 10$	proposed	91.7 (0.01)	96.2 (0.01)	98.8 (0.01)
	BIC	87.5 (0.06)	93.2 (0.06)	99.0 (0.29)
	Oracle	91.8 (0.01)	96.0 (0.01)	99.0 (0.01)
$\sigma = 1$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	93.1 (0.05)	96.3 (0.05)	98.9 (0.06)
	BIC	85.3 (6.63)	93.0 (6.63)	98.9 (32.85)
	Oracle	90.2 (0.03)	94.7 (0.03)	98.6 (0.04)
p=200, $k_0 = 5$	proposed	90.2 (0.06)	95.4 (0.06)	99.4 (0.09)
	BIC	85.3 (6.63)	92.7 (6.63)	98.6 (32.68)
	Oracle	88.7 (0.05)	95.2 (0.05)	99.2 (0.07)
p=200, $k_0 = 10$	proposed	91.7 (0.09)	96.2 (0.09)	98.8 (0.12)
	BIC	85.3 (6.96)	93.0 (6.96)	98.7 (34.36)
	Oracle	90.1 (0.08)	95.6 (0.08)	99.1 (0.11)
p=1000, $k_0 = 2$	proposed	93.2 (0.01)	97.3 (0.01)	99.5 (0.01)
	BIC	84.6 (0.1)	92.5 (0.1)	98.4 (0.5)
	Oracle	90.4 (0.01)	95.9 (0.01)	98.8 (0.01)
p=1000, $k_0 = 5$	proposed	92.3 (0.01)	96.5 (0.01)	99.0 (0.01)
	BIC	85.5 (0.1)	92.1 (0.1)	98.0 (0.49)
	Oracle	90.2 (0.01)	94.8 (0.01)	99.0 (0.01)
p=1000, $k_0 = 10$	proposed	89.9 (0.01)	95.1 (0.01)	98.6 (0.02)
	BIC	86.0 (0.12)	93.1 (0.12)	99.0 (0.56)
	Oracle	89.1 (0.01)	94.8 (0.01)	98.7 (0.02)

ing numbers of phosphoenopyruvate carboxykinase (PEPCK) and glycerol-3-phosphate acyltransferase (GPAT) were also measured by quantitative real-time polymerase chain reaction. Using the credible set approach, Bondell and Reich (2012) derived two methods to predict each of these two phenotypes based on the gene expression data. They also compared their results with those from the LASSO estimator Tibshirani (1996).

TABLE 3
Similar to Table 1 but for $E[Y_1^*|\boldsymbol{X}^*]$.

$\sigma = 0.5$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	91.4 (0.34)	95.4 (0.34)	99.4 (0.46)
	BIC	87.5 (4.03)	93.0 (4.03)	98.6 (19.46)
	Oracle	89.2 (0.3)	94.8 (0.3)	99.3 (0.39)
p=200, $k_0 = 5$	proposed	91.9 (0.48)	96.0 (0.48)	99.3 (0.63)
	BIC	86.9 (4.15)	94.8 (4.15)	98.4 (19.93)
	Oracle	90.4 (0.45)	95.6 (0.45)	99.1 (0.6)
p=200, $k_0 = 10$	proposed	90.8 (0.65)	95.8 (0.65)	99.3 (0.86)
	BIC	85.6 (4.9)	92.8 (4.9)	98.5 (23.55)
	Oracle	90.8 (0.64)	95.9 (0.64)	99.3 (0.85)
p=1000, $k_0 = 2$	proposed	89.7 (0.55)	94.7 (0.55)	98.8 (0.73)
	BIC	85.9 (4.18)	92.7 (4.18)	98.7 (20.08)
	Oracle	89.5 (0.53)	93.9 (0.53)	99.1 (0.7)
p=1000, $k_0 = 5$	proposed	90.1 (0.64)	94.6 (0.64)	98.7 (0.84)
	BIC	87.1 (4.53)	93.0 (4.53)	98.5 (21.74)
	Oracle	89.8 (0.62)	94.9 (0.62)	98.8 (0.83)
p=1000, $k_0 = 10$	proposed	91.6 (0.77)	95.7 (0.77)	99.5 (1.01)
	BIC	86.5 (4.88)	93.3 (4.88)	98.6 (23.37)
	Oracle	90.6 (0.76)	96.3 (0.76)	99.1 (1.0)
$\sigma = 1$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	93.0 (0.78)	96.2 (0.78)	99.1 (1.07)
	BIC	84.9 (7.99)	92.2 (7.99)	99.2 (38.67)
	Oracle	88.7 (0.61)	94.2 (0.61)	98.1 (0.81)
p=200, $k_0 = 5$	proposed	92.0 (1.03)	97.0 (1.03)	99.3 (1.38)
	BIC	86.4 (8.78)	92.9 (8.78)	98.4 (42.19)
	Oracle	91.9 (0.91)	96.4 (0.91)	99.2 (1.2)
p=200, $k_0 = 10$	proposed	92.0 (1.35)	95.7 (1.35)	98.5 (1.78)
	BIC	87.3 (9.94)	93.3 (9.94)	98.3 (47.95)
	Oracle	92.1 (1.26)	96.1 (1.26)	98.6 (1.67)
p=1000, $k_0 = 2$	proposed	91.5 (1.14)	95.6 (1.14)	98.9 (1.51)
	BIC	85.9 (8.39)	93.1 (8.39)	98.1 (40.55)
	Oracle	89.7 (1.04)	95.4 (1.04)	98.0 (1.38)
p=1000, $k_0 = 5$	proposed	90.1 (1.33)	95.1 (1.33)	99.1 (1.76)
	BIC	84.6 (8.4)	91.6 (8.4)	98.3 (40.0)
	Oracle	89.7 (1.24)	94.4 (1.24)	99.3 (1.65)
p=1000, $k_0 = 10$	proposed	89.0 (1.6)	94.3 (1.6)	98.7 (2.1)
	BIC	85.6 (9.5)	92.6 (9.5)	98.4 (45.35)
	Oracle	88.6 (1.53)	94.3 (1.53)	98.7 (2.03)

To reduce the number of candidate predictors, we first apply a screening procedure to remove insignificant predictors and kept the 1999 genes having largest marginal correlation with response. Then the proposed method along with the Joint Sets and Marginal Sets methods in Bondell and Reich (2012) and the LASSO estimator are applied to the data set with $p = 2000$ predictors (1999 genes along with gender) and $n = 60$ observations.

TABLE 4
Similar to Table 1 but prediction intervals for Y_1^*.

$\sigma = 0.5$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	89.8 (2.0)	95.1 (2.0)	99.4 (2.62)
	BIC	86.6 (6.19)	92.3 (6.19)	98.9 (28.23)
	Oracle	89.9 (2.01)	94.5 (2.01)	99.6 (2.66)
p=200, $k_0 = 5$	proposed	90.2 (2.03)	94.5 (2.03)	98.5 (2.66)
	BIC	86.7 (6.26)	94.0 (6.26)	98.3 (28.74)
	Oracle	89.8 (2.03)	95.2 (2.03)	98.7 (2.69)
p=200, $k_0 = 10$	proposed	92.3 (2.07)	96.2 (2.07)	98.9 (2.72)
	BIC	85.6 (7.18)	92.6 (7.18)	98.4 (33.7)
	Oracle	92.5 (2.08)	96.3 (2.08)	99.1 (2.76)
p=1000, $k_0 = 2$	proposed	90.3 (2.05)	95.8 (2.05)	99.3 (2.68)
	BIC	88.1 (6.3)	93.6 (6.3)	98.6 (28.98)
	Oracle	90.9 (2.06)	96.0 (2.06)	99.6 (2.72)
p=1000, $k_0 = 5$	proposed	89.5 (2.08)	94.2 (2.08)	98.6 (2.72)
	BIC	85.5 (6.73)	92.2 (6.73)	98.4 (31.22)
	Oracle	90.2 (2.08)	94.8 (2.08)	98.6 (2.76)
p=1000, $k_0 = 10$	proposed	90.2 (2.12)	95.6 (2.12)	99.0 (2.78)
	BIC	88.0 (7.16)	94.1 (7.16)	98.4 (33.45)
	Oracle	89.9 (2.13)	95.8 (2.13)	99.2 (2.82)
$\sigma = 1$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	89.4 (4.01)	95.0 (4.01)	98.8 (5.26)
	BIC	85.3 (12.19)	92.6 (12.19)	98.8 (55.95)
	Oracle	89.6 (4.03)	95.5 (4.03)	99.0 (5.33)
p=200, $k_0 = 5$	proposed	91.5 (4.07)	95.9 (4.07)	99.3 (5.36)
	BIC	86.5 (13.22)	93.4 (13.22)	98.9 (60.84)
	Oracle	91.3 (4.08)	96.0 (4.08)	99.4 (5.4)
p=200, $k_0 = 10$	proposed	89.7 (4.17)	94.9 (4.17)	98.6 (5.48)
	BIC	86.4 (14.63)	93.4 (14.63)	98.9 (68.71)
	Oracle	90.0 (4.18)	95.3 (4.18)	98.9 (5.54)
p=1000, $k_0 = 2$	proposed	88.7 (4.07)	94.5 (4.07)	99.1 (5.34)
	BIC	86.5 (12.62)	92.6 (12.62)	98.7 (58.44)
	Oracle	89.2 (4.08)	94.5 (4.08)	99.2 (5.4)
p=1000, $k_0 = 5$	proposed	91.3 (4.16)	95.7 (4.16)	98.6 (5.47)
	BIC	86.2 (12.56)	92.6 (12.56)	98.7 (57.57)
	Oracle	91.5 (4.17)	95.8 (4.17)	98.8 (5.52)
p=1000, $k_0 = 10$	proposed	90.1 (4.26)	94.5 (4.26)	98.4 (5.58)
	BIC	85.7 (13.93)	92.9 (13.93)	99.1 (64.92)
	Oracle	89.6 (4.26)	94.9 (4.26)	98.9 (5.65)

To compare the performance of these methods, we randomly split the sample into a training set with size 55 and a test set with size 5. The four methods are applied to predict the 5 observations in the test set. We repeat this process for 100 times to compare the mean squared prediction errors (MSPEs) and the model sizes of the final models obtained by the methods. The results are summarized in Table 11. Notice that there are two responses: PEPCK and

TABLE 5

Empirical coverage rates for the confidence intervals for σ^2 obtained by different methods when $\boldsymbol{\gamma}_{k_0} = \{5, ..., 5\}$ and $n = 100$. The numbers in parentheses are the averaged width of the intervals.

$\sigma = 0.5$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	90.5 (0.12)	94.9 (0.14)	99.0 (0.19)
	BIC	91.1 (2.45)	95.6 (4.90)	99.3 (23.75)
	Oracle	90.3 (0.12)	95 (0.14)	99.1 (0.19)
p=200, $k_0 = 5$	proposed	89.3 (0.12)	93.7 (0.14)	98.0 (0.19)
	BIC	90.6 (2.60)	95.6 (5.19)	98.6 (25.35)
	Oracle	88.7 (0.12)	94.4 (0.14)	98.3 (0.19)
p=200, $k_0 = 10$	proposed	89.8 (0.12)	94.5 (0.15)	98.8 (0.20)
	BIC	91.2 (2.80)	96.1 (5.65)	99.2 (28.07)
	Oracle	90.1 (0.12)	95.2 (0.15)	98.9 (0.20)
p=1000, $k_0 = 2$	proposed	89.3 (0.12)	94.6 (0.14)	99.0 (0.19)
	BIC	90.4 (2.38)	95.8 (4.78)	99.1 (23.28)
	Oracle	90.1 (0.12)	94.9 (0.14)	99.3 (0.19)
p=1000, $k_0 = 5$	proposed	88.1 (0.12)	94.2 (0.14)	98.3 (0.19)
	BIC	91.6 (2.52)	95.4 (5.10)	99.2 (25.02)
	Oracle	88.5 (0.12)	94.3 (0.14)	98.6 (0.19)
p=1000, $k_0 = 10$	proposed	89.3 (0.12)	94.6 (0.15)	99.1 (0.19)
	BIC	91.6 (2.98)	95.9 (5.98)	99.2 (29.97)
	Oracle	89.9 (0.12)	94.8 (0.15)	99.2 (0.19)
$\sigma = 1$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	89.3 (0.58)	94.7 (0.58)	98.9 (0.77)
	BIC	91.4 (207.02)	95.9 (207.02)	99.1 (5165.94)
	Oracle	89.7 (0.58)	94.6 (0.58)	99.1 (0.78)
p=200, $k_0 = 5$	proposed	90.7 (0.59)	95.2 (0.59)	98.4 (0.79)
	BIC	92.1 (232.48)	95.6 (232.48)	99.1 (5798.24)
	Oracle	90.8 (0.6)	95.4 (0.6)	98.7 (0.8)
p=200, $k_0 = 10$	proposed	89.9 (0.61)	94.8 (0.61)	99.1 (0.81)
	BIC	91.5 (278.65)	94.8 (278.65)	99.2 (6950.64)
	Oracle	88.7 (0.61)	94.7 (0.61)	99.2 (0.82)
p=1000, $k_0 = 2$	proposed	89.9 (0.58)	95.7 (0.58)	99.0 (0.77)
	BIC	91.7 (212.87)	96.0 (212.87)	99.1 (5311.31)
	Oracle	90.1 (0.58)	94.9 (0.58)	99.5 (0.78)
p=1000, $k_0 = 5$	proposed	90.1 (0.59)	95.0 (0.59)	98.7 (0.78)
	BIC	90.9 (250.73)	95.5 (250.73)	99.4 (6257.91)
	Oracle	89.4 (0.59)	94.8 (0.59)	98.7 (0.79)
p=1000, $k_0 = 10$	proposed	89.2 (0.61)	93.7 (0.61)	97.9 (0.81)
	BIC	93.8 (288.72)	97.4 (288.72)	99.7 (7199.96)
	Oracle	89.0 (0.61)	93.9 (0.61)	98.2 (0.82)

GPAT. The results show that the proposed method performs well. Although the MSPEs of the proposed method are slightly larger than the methods by Bondell and Reich (2012), the standard errors of the MSPEs and the sizes of the models selected by the proposed method are smaller. These suggest that the proposed method gives a more stable performance, and is less prone to

TABLE 6
Similar to Table 5 but for β_1.

$\sigma = 0.5$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	88.6 (0.0063)	93.4 (0.0075)	99.3 (0.0063)
	BIC	85.4 (0.14)	92.3 (0.28)	98.3 (1.35)
	Oracle	90.1 (0.17)	95.1 (0.20)	99.4 (0.27)
p=200, $k_0 = 5$	proposed	88.8 (0.011)	93.9 (0.013)	98.5 (0.011)
	BIC	84.6 (0.15)	91.6 (0.30)	97.9 (1.42)
	Oracle	89.7 (0.17)	94.5 (0.20)	98.9 (0.27)
p=200, $k_0 = 10$	proposed	88.9 (0.017)	95.0 (0.020)	98.4 (0.017)
	BIC	85.4 (0.16)	92.3 (0.32)	98.3 (1.53)
	Oracle	90.2 (0.18)	95.3 (0.21)	99.3 (0.28)
p=1000, $k_0 = 2$	proposed	89.5 (0.0016)	94.5 (0.0019)	98.7 (0.0016)
	BIC	85.0 (0.021)	93.7 (0.041)	98.2 (0.20)
	Oracle	89.4 (0.17)	94.7 (0.20)	99.0 (0.27)
p=1000, $k_0 = 5$	proposed	90.1 (0.0028)	94.8 (0.0033)	98.7 (0.0028)
	BIC	85.1 (0.023)	92.5 (0.044)	98.4 (0.21)
	Oracle	90.7 (0.17)	94.6 (0.20)	99.2 (0.27)
p=1000, $k_0 = 10$	proposed	90.7 (0.0041)	95.8 (0.0049)	98.7 (0.0041)
	BIC	85.7 (0.027)	93.9 (0.052)	98.8 (0.25)
	Oracle	90.3 (0.17)	95.6 (0.21)	99.0 (0.28)
$\sigma = 1$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	92.9 (0.04)	96.5 (0.04)	99.2 (0.06)
	BIC	86.4 (5.85)	92.4 (5.85)	98.6 (29.05)
	Oracle	91.5 (0.03)	95.0 (0.03)	99.0 (0.04)
p=200, $k_0 = 5$	proposed	91.8 (0.06)	95.8 (0.06)	99.4 (0.09)
	BIC	88.1 (6.73)	93.7 (6.73)	98.4 (33.25)
	Oracle	90.9 (0.05)	95.7 (0.05)	99.3 (0.07)
p=200, $k_0 = 10$	proposed	91.0 (0.09)	95.3 (0.09)	99.6 (0.12)
	BIC	87.1 (7.06)	92.7 (7.06)	98.0 (34.9)
	Oracle	90.6 (0.08)	94.9 (0.08)	98.9 (0.11)
p=1000, $k_0 = 2$	proposed	93.6 (0.01)	96.5 (0.01)	99.4 (0.01)
	BIC	86.5 (0.1)	93.4 (0.1)	98.5 (0.49)
	Oracle	91.8 (0.01)	94.9 (0.01)	99.3 (0.01)
p=1000, $k_0 = 5$	proposed	90.6 (0.01)	95.2 (0.01)	99.3 (0.01)
	BIC	84.8 (0.11)	92.2 (0.11)	98.7 (0.55)
	Oracle	88.9 (0.01)	94.5 (0.01)	98.7 (0.01)
p=1000, $k_0 = 10$	proposed	90.6 (0.01)	94.7 (0.01)	98.3 (0.02)
	BIC	86.1 (0.13)	93.7 (0.13)	99.4 (0.61)
	Oracle	89.3 (0.01)	94.7 (0.01)	98.2 (0.02)

overfitting.

Next, the following experiment was carried out to evaluate the empirical coverage rates of the proposed method on this data set. For both responses (GPAT and PEPCK), we left out the first observation of the data set and used the remaining 59 observations to construct a 95% and a 99% prediction interval for this first observation. We repeated this leave-one-out process for the remaining

TABLE 7
Similar to Table 5 but for $E[Y_1^*|\boldsymbol{X}_1^*]$.

$\sigma = 0.5$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	93.1 (0.35)	96.7 (0.35)	99.1 (0.47)
	BIC	85.2 (4.05)	92.2 (4.05)	98.1 (19.59)
	Oracle	90.8 (0.31)	95.9 (0.31)	99.3 (0.41)
p=200, $k_0 = 5$	proposed	90.9 (0.49)	96.5 (0.49)	99.3 (0.64)
	BIC	86.5 (4.23)	93.9 (4.23)	98.9 (20.43)
	Oracle	90.3 (0.46)	95.4 (0.46)	99.4 (0.61)
p=200, $k_0 = 10$	proposed	91.1 (0.66)	94.9 (0.66)	99.2 (0.86)
	BIC	86.9 (5.14)	92.3 (5.14)	98.6 (24.81)
	Oracle	91.3 (0.64)	95.0 (0.64)	99.1 (0.85)
p=1000, $k_0 = 2$	proposed	91.6 (0.53)	95.3 (0.53)	99.2 (0.7)
	BIC	87.1 (4.15)	92.9 (4.15)	98.7 (19.87)
	Oracle	90.7 (0.51)	95.4 (0.51)	99.1 (0.68)
p=1000, $k_0 = 5$	proposed	90.5 (0.64)	94.9 (0.64)	98.8 (0.85)
	BIC	85.7 (4.41)	93.7 (4.41)	98.7 (21.18)
	Oracle	89.7 (0.63)	94.5 (0.63)	99.1 (0.83)
p=1000, $k_0 = 10$	proposed	89.5 (0.78)	95.5 (0.78)	98.6 (1.02)
	BIC	85.8 (4.7)	93.0 (4.7)	98.4 (22.37)
	Oracle	90.0 (0.76)	95.2 (0.76)	99.0 (1.01)
$\sigma = 1$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	92.3 (0.77)	96.0 (0.77)	99.3 (1.06)
	BIC	84.6 (8.12)	91.8 (8.12)	98.6 (39.52)
	Oracle	90.0 (0.59)	95.9 (0.59)	99.3 (0.79)
p=200, $k_0 = 5$	proposed	89.5 (1.05)	94.5 (1.05)	98.7 (1.4)
	BIC	85.4 (8.8)	92.2 (8.8)	98.3 (42.45)
	Oracle	88.9 (0.93)	94.2 (0.93)	98.9 (1.23)
p=200, $k_0 = 10$	proposed	92.7 (1.36)	96.4 (1.36)	99.3 (1.79)
	BIC	87.1 (9.87)	93.0 (9.87)	98.6 (47.51)
	Oracle	91.9 (1.28)	96.1 (1.28)	99.2 (1.69)
p=1000, $k_0 = 2$	proposed	92.3 (1.14)	96.8 (1.14)	99.7 (1.52)
	BIC	87.5 (8.35)	92.9 (8.35)	98.7 (40.09)
	Oracle	90.6 (1.03)	96.3 (1.03)	99.6 (1.37)
p=1000, $k_0 = 5$	proposed	89.7 (1.31)	95.2 (1.31)	98.8 (1.74)
	BIC	86.2 (9.33)	92.5 (9.33)	98.9 (44.94)
	Oracle	89.4 (1.23)	95.0 (1.23)	99.0 (1.63)
p=1000, $k_0 = 10$	proposed	90.6 (1.59)	94.7 (1.59)	99.1 (2.09)
	BIC	87.7 (10.28)	93.9 (10.28)	99.0 (49.31)
	Oracle	90.9 (1.52)	94.3 (1.52)	99.1 (2.01)

59 observations. The resulting prediction intervals are summarized in Figure 1. For both PEPCK and GPAT, the coverage rates of the 99% prediction intervals are both 95%, while the coverage rates of the 95% prediction intervals are both 90%. Since the fiducial intervals are well calibrated when the PC regression assumption is appropriate, these under-coverage rates suggest that the assumption might not be entirely suitable for these data sets.

TABLE 8
Similar to Table 5 but prediction intervals for Y_1^*.

$\sigma = 0.5$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	89.6 (1.65)	94.0 (2.0)	98.8 (1.65)
	BIC	84.9 (3.25)	92.1 (6.12)	98.4 (28.09)
	Oracle	90.3 (1.68)	94.5 (2.0061)	99.3 (2.66)
p=200, $k_0 = 5$	proposed	88.8 (1.67)	94.4 (1.99)	99.2 (1.67)
	BIC	85.1 (3.40)	91.6 (6.44)	98.1 (29.90)
	Oracle	89.8 (1.71)	95.6 (2.04)	99.4 (2.71)
p=200, $k_0 = 10$	proposed	88.7 (1.68)	94.8 (2.0077)	98.7 (1.68)
	BIC	86.3 (3.63)	92.9 (6.89)	98.4 (32.32)
	Oracle	90.9 (1.75)	95.0 (2.09)	99.2 (2.77)
p=1000, $k_0 = 2$	proposed	88.5 (1.65)	95.2 (1.98)	98.6 (1.65)
	BIC	78.8 (2.52)	88.5 (4.67)	97.3 (20.89)
	Oracle	89.9 (1.68)	95.3 (2.01)	98.6 (2.66)
p=1000, $k_0 = 5$	proposed	89.6 (1.66)	94.1 (1.98)	98.4 (1.66)
	BIC	82.0 (2.65)	91.4 (4.90)	98.1 (21.95)
	Oracle	89.4 (1.70)	94.5 (2.04)	98.3 (2.70)
p=1000, $k_0 = 10$	proposed	87.7 (1.67)	94.0 (2.00)	98.2 (1.67)
	BIC	81.5 (2.97)	89.0 (5.61)	98.0 (26.16)
	Oracle	89.9 (1.75)	95.4 (2.09)	99.0 (2.76)
$\sigma = 1$	method	90% CI	95% CI	99% CI
p=200, $k_0 = 2$	proposed	89.2 (4.01)	93.5 (4.01)	98.3 (5.26)
	BIC	84.7 (12.44)	92.1 (12.44)	98.4 (57.27)
	Oracle	89.9 (4.02)	94.2 (4.02)	98.6 (5.32)
p=200, $k_0 = 5$	proposed	88.7 (4.09)	94.7 (4.09)	98.7 (5.36)
	BIC	85.0 (13.22)	92.3 (13.22)	98.9 (61.16)
	Oracle	88.9 (4.09)	94.8 (4.09)	99.0 (5.41)
p=200, $k_0 = 10$	proposed	90.8 (4.16)	95.0 (4.16)	99.2 (5.46)
	BIC	86.7 (14.54)	92.8 (14.54)	98.4 (68.09)
	Oracle	90.9 (4.16)	95.5 (4.16)	99.5 (5.52)
p=1000, $k_0 = 2$	proposed	90.0 (4.12)	95.5 (4.12)	99.1 (5.39)
	BIC	85.8 (12.65)	93.3 (12.65)	98.8 (57.91)
	Oracle	90.4 (4.12)	95.6 (4.12)	99.2 (5.45)
p=1000, $k_0 = 5$	proposed	89.5 (4.14)	94.3 (4.14)	98.9 (5.43)
	BIC	86.2 (13.85)	93.0 (13.85)	98.1 (64.53)
	Oracle	89.8 (4.14)	94.9 (4.14)	99.0 (5.49)
p=1000, $k_0 = 10$	proposed	89.8 (4.25)	95.0 (4.25)	99.4 (5.58)
	BIC	88.4 (15.01)	95.4 (15.01)	99.1 (70.51)
	Oracle	90.3 (4.26)	95.9 (4.26)	99.6 (5.65)

7. Conclusion

This paper developed a new approach for variable selection and uncertainty quantification for sparse high-dimensional principal component regression based on *generalized fiducial inference*. The consistency properties of the proposed method was established and was verified by simulation experiments. The pro-

TABLE 9. *Percentages of times k_0 has the highest fiducial probability/empirical coverages of 95% confidence intervals for k_0 when $\boldsymbol{\gamma}_{k_0} = \{1, ..., 1\}$. Numbers in parentheses are the average numbers of the k's in the confidence intervals.*

	$n = 100$, $p = 200$			$n = 100$, $p = 1000$		
σ	$k_0 = 2$	$k_0 = 5$	$k_0 = 10$	$k_0 = 2$	$k_0 = 5$	$k_0 = 10$
0.1	99.1/100.0 (1.3)	99.1/100.0 (1.3)	99.5/100.0 (1.3)	99.3/100.0 (1.3)	99.7/100.0 (1.3)	98.9/100.0 (1.3)
0.5	94.4/99.9 (2.4)	93.5/99.9 (2.4)	96.1/99.9 (2.4)	94.4/99.8 (2.5)	94.8/99.7 (2.4)	96.1/99.8 (2.4)
1	85.6/99.4 (4.1)	85.7/99.6 (4.1)	86.1/99.3 (4.1)	87.1/99.7 (4.1)	88.0/99.2 (4.0)	85.3/99.3 (4.1)
	$n = 1000$, $p = 200$			$n = 1000$, $p = 1000$		
σ	$k_0 = 2$	$k_0 = 5$	$k_0 = 10$	$k_0 = 2$	$k_0 = 5$	$k_0 = 10$
0.1	99.9/100.0 (1.1)	99.8/99.9 (1.1)	99.7/100.0 (1.1)	99.7/100.0 (1.1)	100.0/100.0 (1.1)	99.9/100.0 (1.1)
0.5	99.3/100.0 (1.6)	99.0/100.0 (1.7)	98.8/100.0 (1.6)	98.5/100.0 (1.6)	98.7/100.0 (1.6)	98.7/100.0 (1.6)
1	97.5/99.9 (2.2)	97.0/99.8 (2.1)	97.2/99.8 (2.1)	97.7/100.0 (2.1)	97.5/100.0 (2.1)	97.4/100.0 (2.1)

TABLE 10. *Similar to Table 9 but for $\boldsymbol{\gamma}_{k_0} = \{5, ..., 5\}$.*

	$n = 100$, $p = 200$			$n = 100$, $p = 1000$		
σ	$k_0 = 2$	$k_0 = 5$	$k_0 = 10$	$k_0 = 2$	$k_0 = 5$	$k_0 = 10$
0.1	99.0/100.0 (1.3)	99.5/100.0 (1.3)	99.1/100.0 (1.3)	99.3/100.0 (1.3)	99.4/100.0 (1.3)	99.5/100.0 (1.3)
0.5	94.0/99.8 (2.4)	95.2/100.0 (2.4)	94.6/99.7 (2.4)	95.0/100.0 (2.4)	95.8/100.0 (2.4)	95.8/99.8 (2.4)
1	84.9/99.7 (4.1)	83.7/99.1 (4.0)	84.3/99.6 (4.1)	84.6/99.6 (4.0)	84.6/99.2 (4.1)	84.8/99.2 (4.1)
	$n = 1000$, $p = 200$			$n = 1000$, $p = 1000$		
σ	$k_0 = 2$	$k_0 = 5$	$k_0 = 10$	$k_0 = 2$	$k_0 = 5$	$k_0 = 10$
0.1	99.8/100.0 (1.1)	99.9/100.0 (1.1)	99.8/99.9 (1.1)	99.9/100.0 (1.1)	100.0/100.0 (1.1)	99.7/100.0 (1.1)
0.5	98.0/100.0 (1.6)	99.3/100.0 (1.6)	99.2/100.0 (1.6)	98.3/100.0 (1.6)	98.2/100.0 (1.6)	98.9/100.0 (1.6)
1	97.3/100.0 (2.1)	96.8/100.0 (2.1)	97.2/100.0 (2.1)	97.7/99.9 (2.1)	97.5/100.0 (2.1)	97.1/99.8 (2.1)

TABLE 11
MSPEs and averaged model sizes obtained from various methods based on 100 *random splits of the real data set. Numbers in parentheses are standard errors.*

	Response PEPCK		Response GPAT	
Method	MSPE	Model Size	MSPE	Model Size
Joint Sets	2.03 (0.14)	9.60 (0.46)	3.83 (0.34)	4.20 (0.43)
Marginal Sets	1.84 (0.14)	23.30 (0.67)	5.33 (0.41)	21.80 (0.72)
LASSO	3.03 (0.19)	7.70 (0.96)	5.03 (0.42)	3.30 (0.79)
proposed	2.30 (0.11)	5.50 (0.13)	4.87 (0.42)	3.26 (0.04)

posed method was also compared with other existing methods using a mice phenotype data. It was shown that the proposed method was competitive for a smaller MSPE, model size and standard error.

In the above the following restriction is imposed: if one wants to include the k-th principal component in the final model, the first $k-1$ principal components should also be included. To relax this restriction, one needs to derive a new generalized fiducial density, adopt a new penalty as now the number of possible models is increased to $\binom{n}{k}$, and develop a new algorithm for generating fiducial samples from the enlarged model space. We plan to explore this possibility in the future.

Acknowledgments

The authors are most grateful to the reviewers for their most constructive and helpful comments, which led to a much improved version of the paper.

References

Adusumilli, S., Bhatt, D., Wang, H., Devabhaktuni, V. and Bhattacharya, P. (2015) A novel hybrid approach utilizing principal component regression and random forest regression to bridge the period of GPS outages. *Neurocomputing*, **166**, 185–192.

Barron, A. R., Rissanen, J. and Yu, B. (1998) The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, **44**, 2743–2760. MR1658898

Bondell, H. D. and Reich, B. J. (2012) Consistent high-dimensional bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, **107**, 1610–1624. MR3036420

Cheng, J. Q., Liu, R. Y. and Xie, M.-g. (2014) Fusion learning. *Wiley StatsRef: Statistics Reference Online*, 1–8.

Dempster, A. P. (2008) The Dempster–Shafer calculus for statisticians. *International Journal of approximate reasoning*, **48**, 365–377. MR2419025

Efron, B. (2013) Bayes' theorem in the 21st century. *Science*, **340**, 1177–1178. MR3087705

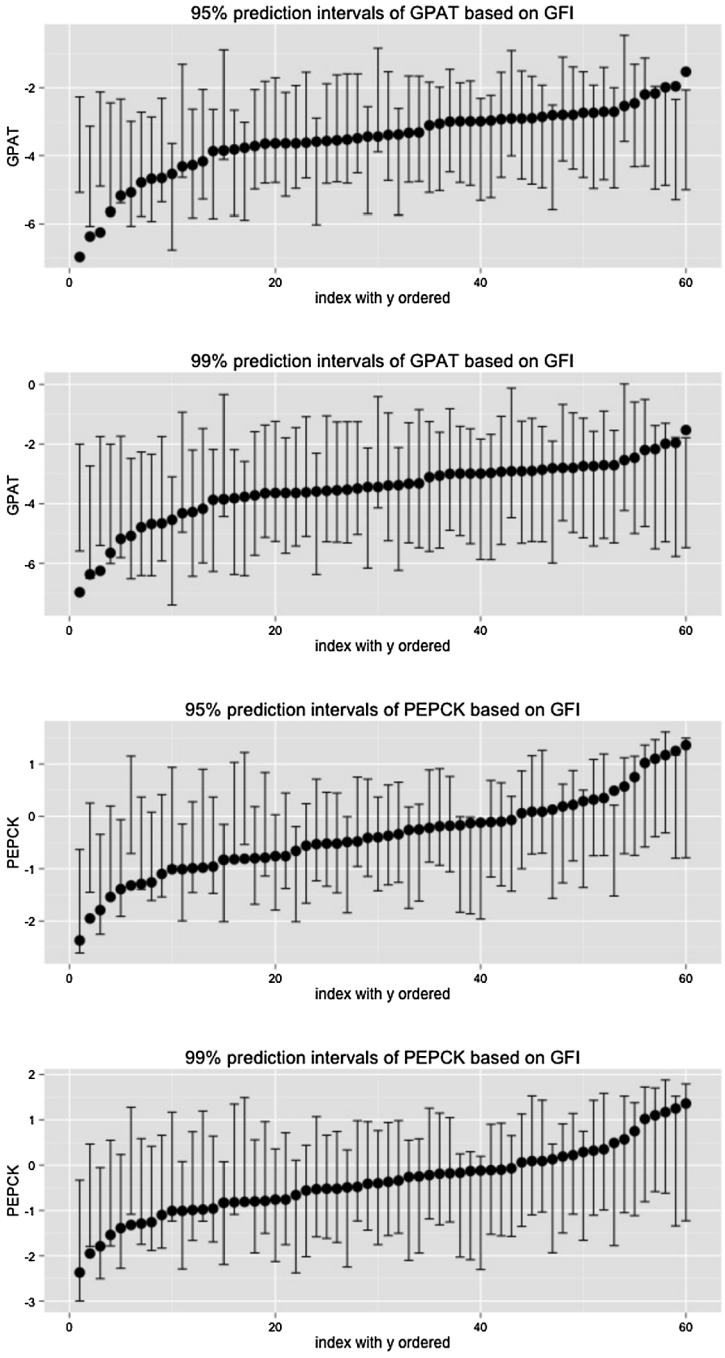

FIG 1. *In each panel the dots represent the responses y_i's, while the vertical error bars are the corresponding prediction intervals. For clarity the y_i's are sorted in ascending order.*

Fisher, R. A. (1930) Inverse probability. In *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 26, 528–535. Cambridge Univ Press.

Hannig, J. (2009) On generalized fiducial inference. *Statistica Sinica*, 491–544. MR2514173

Hannig, J. (2013) Generalized fiducial inference via discretization. *Statistica Sinica*, **23**, 489–514. MR3086644

Hannig, J., Iyer, H. K., Lai, R. C. S. and Lee, T. C. M. (2016) Generalized Fiducial Inference: A Review and New Results. *Journal of the American Statistical Association*, **111**, 1346–1361. MR3561954

Hotelling, H. (1957) The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Mathematical and Statistical Psychology*, **10**, 69–79.

Huang, S.-M. and Yang, J.-F. (2012) Improved principal component regression for face recognition under illumination variations. *IEEE Signal Processing Letters*, **19**, 179–182.

Jolliffe, I. T. (1982) A note on the use of principal components in regression. *Applied Statistics*, 300–303. MR0841268

Lai, R. C., Hannig, J. and Lee, T. C. M. (2015) Generalized fiducial inference for ultrahigh-dimensional regression. *Journal of the American Statistical Association*, **110**, 760–772. MR3367262

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T.-K., Flowers, M. T., Schueler, K. L., Manly, K. F. *et al.* (2006) Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genet*, **2**, e6.

Lipponen, J. A., Tarvainen, M. P., Laitinen, T., Lyyra-Laitinen, T. and Karjalainen, P. A. (2010) A principal component regression approach for estimation of ventricular repolarization characteristics. *IEEE Transactions on Biomedical Engineering*, **57**, 1062–1069.

Martin, R. and Liu, C. (2015) *Inferential Models: Reasoning with uncertainty*, vol. 145. CRC Press. MR3618727

Rissanen, J. (1996) Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, **42**, 40–47. MR1375327

Shen, J., Liu, R. Y. and Xie, M.-g. (2018) Prediction with confidence—a general framework for predictive inference. *Journal of Statistical Planning and Inference*, **195**, 126–140. MR3760843

Sun, J. (1995) A correlation principal component regression analysis of NIR data. *Journal of Chemometrics*, **9**, 21–29.

Sutter, J. M., Kalivas, J. H. and Lang, P. M. (1992) Which principal components to utilize for principal component regression. *Journal of chemometrics*, **6**, 217–225.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288. MR1379242

Wang, J. C.-M., Hannig, J. and Iyer, H. K. a. (2012) Fiducial Prediction Intervals. *Journal of Statistical Planning and Inference*, **142**, 1980–1990. MR2903406

Xie, M.-g. and Singh, K. (2013) Confidence distribution, the frequentist distri-

bution estimator of a parameter: A review. *International Statistical Review*, **81**, 3–39. MR3047496

Xie, Y.-L. and Kalivas, J. H. (1997) Evaluation of principal component selection methods to form a global prediction model by principal component regression. *Analytica chimica acta*, **348**, 19–27.

Xu, J., Myodo, E. and Sakazawa, S. (2013) Motion synthesis for affective agents using piecewise principal component regression. In *IEEE International Conference on Multimedia and Expo (ICME)*, 1–7.

Zhu, Y., Zhu, C. and Li, X. (2018) Improved principal component analysis and linear regression classification for face recognition. *Signal Processing*, **145**, 175–182.

Appendix A: Proof of Theorem 4.1

Proof. First we prove that

$$\frac{R(k)}{R(k_0)} \xrightarrow{P} 0, \quad \forall k \neq k_0.$$

Without loss of generality, we assume that $\sigma^2 = 1$. Rewrite $\frac{R(k)}{R(k_0)} = \exp\{-T_1 - T_2\}$, where

$$T_1 = \frac{n-k-1}{2} \log\left(\frac{\mathrm{RSS}_k}{\mathrm{RSS}_{k_0}}\right)$$

and

$$T_2 = \frac{k-k_0}{2}(\log n - \log(\pi \mathrm{RSS}_{k_0})) + \log\left(\frac{\Gamma(\frac{n-k_0}{2})}{\Gamma(\frac{n-k}{2})}\right).$$

Case 1: $k < k_0$
Now calculate

$$\begin{aligned}\mathrm{RSS}_k - \mathrm{RSS}_{k_0} &= \Delta_k + 2\boldsymbol{\mu}^T(\boldsymbol{I} - \boldsymbol{H}_k)\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T\boldsymbol{H}_k\boldsymbol{\varepsilon} \\ &\quad - (\Delta_{k_0} + 2\boldsymbol{\mu}^T(\boldsymbol{I} - \boldsymbol{H}_{k_0})\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T\boldsymbol{H}_{k_0}\boldsymbol{\varepsilon}) \\ &= \Delta_k + 2\boldsymbol{\mu}^T(\boldsymbol{I} - \boldsymbol{H}_k)\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T(\boldsymbol{H}_{k_0} - \boldsymbol{H}_k)\boldsymbol{\varepsilon}. \qquad (12)\end{aligned}$$

For the last term in (12), first notice that $\boldsymbol{\varepsilon}^T(\boldsymbol{H}_{k_0} - \boldsymbol{H}_k)\boldsymbol{\varepsilon} = -\chi^2_{k_0-k}$. Then let $c_k = k \log\log p$ and calculate

$$\begin{aligned}\sum_{k=1}^{+\infty} P(\chi^2_k > c_k) &\leq \sum_{k=1}^{+\infty} (\frac{c_k}{k} e^{1-\frac{c_k}{k}})^{k/2} \quad \text{by Chernoff bound since} \quad c_k > 1 \\ &= \sum_{k=1}^{+\infty} (\frac{e \log\log p}{\log p})^{k/2} \longrightarrow 0 \qquad \text{as} \quad p \longrightarrow \infty.\end{aligned}$$

Therefore,

$$P(\boldsymbol{\varepsilon}^T(\boldsymbol{H}_{k_0} - \boldsymbol{H}_k)\boldsymbol{\varepsilon} < -c_{k_0-k}) \longrightarrow 0 \quad \text{as} \quad p \longrightarrow \infty, \quad \forall 0 \leq k < k_0.$$

For the second term in (12), denote $Z_k = \boldsymbol{\mu}^T(\boldsymbol{I}-\boldsymbol{H}_k)\boldsymbol{\varepsilon}/\sqrt{\Delta_k}$, we have $\mu^T(\boldsymbol{I}-\boldsymbol{H}_k)\boldsymbol{\varepsilon} = \sqrt{\Delta_k}Z_k$ and $Z_k \sim N(0,1)$ as $var(Z_k) = \frac{\boldsymbol{\mu}^T(\boldsymbol{I}-\boldsymbol{H}_k)(\boldsymbol{I}-\boldsymbol{H}_k)^T\boldsymbol{\mu}}{\Delta_k} = 1$. Furthermore,

$$\begin{aligned} P(\max_{k\in(0,+\infty)}|Z_k/\sqrt{c_k}|>1) &\le \sum_{k=1}^{+\infty}P(Z_k^2>c_k) = \sum_{k=1}^{+\infty}P(\chi_1^2>c_k) \le \sum_{k=1}^{+\infty}P(\chi_2^2>c_k) \\ &= \sum_{k=1}^{+\infty}\exp(-\frac{c_k}{2}) = \frac{\frac{1}{\sqrt{\log p}}}{1-\frac{1}{\sqrt{\log p}}} \longrightarrow 0 \text{ as } p\longrightarrow\infty. \end{aligned}$$

Therefore, $P(|\boldsymbol{\mu}^T(\boldsymbol{I}-\boldsymbol{H}_k)\boldsymbol{\varepsilon}| > \sqrt{\Delta_k c_k}) \longrightarrow 0$ as $p\longrightarrow\infty$. Thus, we have

$$P(|\mathrm{RSS}_k - \mathrm{RSS}_{k_0}| < 0.5\Delta_k) \longrightarrow 0 \quad \text{as} \quad p\longrightarrow\infty.$$

In addition, $P(\chi^2_{n-K} < \frac{n}{4}) \le P(\chi^2_{n-K} < \frac{n-K}{2}) \le (\frac{\sqrt{e}}{2})^{\frac{n-K}{2}} \longrightarrow 0$ as $n\longrightarrow\infty$, which means $P(\min_{0<k\le K}\chi^2_{n-k} < \frac{n}{4}) \longrightarrow 0$ as $n\longrightarrow\infty$. Thus,

$$\begin{aligned} T_1 &= \frac{n-k-1}{2}\log(\frac{\mathrm{RSS}_k}{\mathrm{RSS}_{k_0}}) = -\frac{n-k-1}{2}\log(1+\frac{\mathrm{RSS}_{k_0}-\mathrm{RSS}_k}{\mathrm{RSS}_k}) \\ &\ge \frac{n-k-1}{2}\frac{\mathrm{RSS}_k-\mathrm{RSS}_{k_0}}{\mathrm{RSS}_k} = \Omega_p(\Delta_k) \end{aligned}$$

and

$$T_2 = \frac{k_0-k}{2}\log(\pi\mathrm{RSS}_{k_0}) + \log\{\Gamma(\frac{n-k_0}{2})/\Gamma(\frac{n-k}{2})\} + \frac{k-k_0}{2}\log n.$$

For the first term of T_2 notice that $\log(\mathrm{RSS}_{k_0}) = \Omega_p(\log(n-k_0))$, while for the second term, $\log(\Gamma(\frac{n-k_0}{2})/\Gamma(\frac{n-k}{2})) = \Omega_p(\frac{k-k_0}{2}\log n)$, for n large enough.

Case 2: $k > k_0$

In this case, $\mathrm{RSS}_k - \mathrm{RSS}_{k_0} = -\chi^2_{k-k_0}$. By Chernoff bound,

$$\begin{aligned} P(\max_{k_0<k<K}\chi^2_{k-k_0} > c_{k-k_0}) &\le \sum_{k=k_0+1}^{K}P(\chi^2_{k-k_0}>c_{k-k_0}) \\ &\le \sum_{k=k_0+1}^{K}(\frac{c_{k-k_0}}{k-k_0}e^{1-\frac{c_{k-k_0}}{k-k_0}})^{(k-k_0)/2} \\ &= \sum_{k=k_0+1}^{K}(\frac{e\log\log p}{\log p})^{(k-k_0)/2} \longrightarrow 0 \text{ as } p\longrightarrow\infty. \end{aligned}$$

Furthermore,

$$P(\chi^2_{n-K} < \frac{n}{4}) \le P(\chi^2_{n-K} < \frac{n-K}{2}) \le (\frac{\sqrt{e}}{2})^{\frac{n-K}{2}} \longrightarrow 0 \quad \text{as} \quad n\longrightarrow\infty,$$

which means

$$P(\max_{k_0<k<K}\chi^2_{n-k} < \frac{n}{4}) \longrightarrow 0 \quad \text{as} \quad n\longrightarrow\infty.$$

Thus,

$$T_1 = \frac{n-k-1}{2}\log\left(\frac{\mathrm{RSS}_k}{\mathrm{RSS}_{k_0}}\right) = -\frac{n-k-1}{2}\log\left(1+\frac{\chi^2_{k-k_0}}{\chi^2_{n-k}}\right)$$
$$\geq -\frac{n-k-1}{2}\frac{\chi^2_{k-k_0}}{\chi^2_{n-k}} = \Omega_p(-c_{k-k_0}).$$

The discussion of T_2 is similar to Case 1; i.e., $T_2 = \Theta(\frac{k-k_0}{2}\log n)$.

Thus

$$\sum_{k\neq k_0} R(k)/R(k_0) = \sum_{k=1}^{k_0-1} R(k)/R(k_0) + \sum_{k=k_0+1}^{K} R(k)/R(k_0)$$
$$\leq \sum_{k=1}^{k_0-1} e^{-\frac{1}{8}n} + \sum_{k=k_0+1}^{K} e^{-\frac{k-k_0}{4}\log n}$$
$$\leq k_0 e^{-\frac{1}{4}n} + \sum_{k=1}^{K} n^{-\frac{k}{4}} \longrightarrow 0.$$

Equivalently,

$$r(k_0)/\sum_{k=1}^{K} r(k) \xrightarrow{P} 1. \qquad \square$$