

# Globaltest confidence regions and their application to ridge regression

Ningning Xu<sup>1</sup>  | Aldo Solari<sup>2</sup>  | Jelle Goeman<sup>1</sup> 

<sup>1</sup> Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

<sup>2</sup> Department of Economics, Management and Statistics, University of Milano-Bicocca, Milano, Italy

## Correspondence

Ningning Xu, Department of Biomedical Data Sciences, Leiden University Medical Center, Postzone S5-P, Postbus 9600, 2300 RC, Leiden, The Netherlands.  
Email: [n.xu@lumc.nl](mailto:n.xu@lumc.nl)

## Funding information

Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: 639.072.412



This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

## Abstract

We construct confidence regions in high dimensions by inverting the globaltest statistics, and use them to choose the tuning parameter for penalized regression. The selected model corresponds to the point in the confidence region of the parameters that minimizes the penalty, making it the least complex model that still has acceptable fit according to the test that defines the confidence region. As the globaltest is particularly powerful in the presence of many weak predictors, it connects well to ridge regression, and we thus focus on ridge penalties in this paper. The confidence region method is quick to calculate, intuitive, and gives decent predictive potential. As a tuning parameter selection method it may even outperform classical methods such as cross-validation in terms of mean squared error of prediction, especially when the signal is weak. We illustrate the method for linear models in simulation study and for Cox models in real gene expression data of breast cancer samples.

## KEYWORDS

confidence regions, high dimensional, tuning parameter selection

## 1 | INTRODUCTION

Confidence regions play a fundamental role in statistical inference. Points within a confidence region can be viewed as reasonable candidates for the true parameter. By distinguishing between acceptable and unacceptable values, confidence regions can be used to select the tuning parameter of penalized regression models.

The rationale for the confidence region approach to tuning parameter selection is as follows. If a model is not in the confidence region, it has significantly worse fit than the true model. It makes sense, therefore, to restrict attention to only models inside the confidence region. In the context of penalized methods, among all acceptable models we may prefer the model with the smallest penalty rather than the midpoint of the confidence region. This is the least complex, and therefore hopefully least overfitting model among all acceptable models.

There have been many confidence region approaches proposed for tuning parameter selection, for example, Obenchain (1977), McCabe (1978), and Oman (1981) proposed to use classical  $F$ -test to select the tuning parameter for ridge regression

-----  
This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

(Hoerl & Kennard, 1970). More recently, Gunes and Bondell (2012) applied the confidence region approach in variable selection with adaptive LASSO (Zou, 2006). A similar idea was proposed by Jiang et al. (2008) for model selection in linear mixed models. In one variant, their “fence” around all acceptable models is exactly the border of the likelihood ratio test confidence region. Within the fence they also select the model that minimizes the penalty. Note that the confidence regions used in these papers are all obtained by inverting the likelihood ratio test (i.e., the  $F$ -test for linear models), which is only applicable for low-dimensional data.

The purpose in this work is to extend the confidence region approach to high dimensions for a wide range of generalized linear models and Cox models, for which globaltest Goeman et al. (2004) can be used. We build confidence regions based on the globaltest. In the context of linear models, Goeman et al. (2006) showed that the globaltest is more powerful than the  $F$ -test when large variance principal components of the design matrix explain more of the variance of the outcome than the small variance ones. More importantly, the globaltest is powerful in high dimensions, especially when there are many predictors with weak effects.

In many biological data examples, it is common that good predictive ability can be obtained from the cumulative effect of many weak predictors even though they might be too weak to be identifiable individually. This is the scenario where both ridge regression and globaltest work well. Therefore, we concentrate on the combination of globaltest and ridge regression in this paper, that is, using the confidence region of globaltest to choose the tuning parameter for ridge regression.

The confidence region approach is more attractive than other criteria, such as, classically, cross-validation (CV) (Breiman & Spector, 1992) and information criteria, for several reasons. First, the confidence region approach can be viewed as a testimation procedure (Rahman & Gokhale, 1996), for which the resulting “testimator” is corresponding to the least overfitting estimator that is tested significant by globaltest at a prespecified significance level  $\alpha$ . Ridge regression selects either the full model or the null model. When the null model is false, the probability of choosing the full model converges to 1 for a fixed alternative because the global test is consistent (Goeman et al., 2006). When the null model is true, the confidence region method can guarantee that the probability of selecting the null model is asymptotically  $1 - \alpha$ . This can be an important property because it may prevent false predictive claims from entering the literature. Second, the significance level  $\alpha$  can take the role of the classical tuning parameter  $\lambda$ , of which the scale is arbitrary, making it difficult to interpret. The  $\alpha$  is well calibrated and, due to its direct interpretation as an error rate, may be chosen a priori at a reasonable level of acceptable type I error control. By linking tuning parameter selection to inferential theory in this way, tuning parameter selection becomes less of an algorithmic black box.

The classical choice of  $\alpha = 5\%$  is sensible if stringent error rate control is crucial, but this will lead to conservative model fits. In prediction modeling, many methods used in practice have type I error rate of around 50% (Gunes & Bondell, 2012). In contexts where weak type I error rate control is more desirable, the confidence region approach at level of 50% can therefore be expected to produce results very close to those of classical methods such as CV but is much faster than CV. Type I error at level 50% has also been recommended by Aitkin (1974) to prevent the conservativeness of a simultaneous variable selection procedure.

The structure of the paper is as follows. We will describe the confidence region of globaltest and revisit ridge regression, and then present our method in a general way in Section 2. The properties of the globaltest and ridge estimator for linear models will be discussed in more detail in Section 3, in which we compare the globaltest confidence region with the  $F$ -test confidence region. A numerical study compares our proposed method with other methods in Section 4, where we also perform a real data analysis based on three high-dimensional breast cancer data sets.

## 2 | THE GLOBALTEST CONFIDENCE REGION APPROACH

### 2.1 | The globaltest

Suppose we have data with  $n$  observations and  $p$  predictors.  $\mathbf{X}$  is an  $n \times p$  design matrix whose columns correspond to  $p$  predictors. A regression model relates the response to the predictors through the linear predictors  $\mathbf{x}_i^T \boldsymbol{\beta}$ , where  $\mathbf{x}_i^T = (x_{i1}, \dots, x_{ip})$  is the  $i$ th row of  $\mathbf{X}$  and  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  are the unknown model coefficients.

We assume a generalized linear model. Let  $\mathbf{y} = (y_1, \dots, y_n)^T$  be the vector of responses, where  $y_i$  follows a distribution in the exponential family. The model assumes the mean of the response and the linear predictors are related by  $g(E(y_i)) = \mathbf{x}_i^T \boldsymbol{\beta}$ , where  $g$  is a monotone link function, for example, the identity function for the linear model or the logit function for the logistic model. Extensions to the Cox proportional hazard model are straightforward and we come to those in Section 5.

Suppose that we are interested in testing the following null hypothesis:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$$

against the alternative  $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ . Goeman et al. (2011) derived the following globaltest statistic:

$$\hat{S}_{\boldsymbol{\beta}_0} = \mathbf{s}^\top \mathbf{s} - \text{trace}(\mathcal{I}),$$

where  $\mathbf{s} = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$  is the score of  $\boldsymbol{\beta}$  at  $\boldsymbol{\beta}_0$ ,  $\ell(\boldsymbol{\beta})$  is the log-likelihood of the model, and  $\mathcal{I} = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$  is the observed information matrix. Because  $\text{trace}(\mathcal{I})$  does not depend on the response (Goeman et al., 2011),  $\hat{S}_{\boldsymbol{\beta}_0}$  is equivalent to

$$S_{\boldsymbol{\beta}_0} = \mathbf{s}^\top \mathbf{s}.$$

Then by inverting the statistic  $S_{\boldsymbol{\beta}_0}$  we get the  $1 - \alpha$  confidence region of globaltest:

$$C_\alpha^{\text{gt}} = \{\boldsymbol{\beta}_0 \in \mathbb{R}^p : S_{\boldsymbol{\beta}_0} \leq c_\alpha\}. \quad (1)$$

Here  $c_\alpha$  is the  $1 - \alpha$  quantile of the null distribution of  $S_{\boldsymbol{\beta}_0}$ . Goeman et al. (2011) derived the exact null distribution of  $S_{\boldsymbol{\beta}_0}$  for linear models and asymptotic null distribution for other generalized linear models using the algorithms developed by Imhof (1961) and Robbins and Pitman (1949). The implementation of the globaltest can be referenced to the R package `globaltest` (Goeman et al., 2010).

## 2.2 | Ridge regression

Ridge regression, first proposed by Hoerl and Kennard (1970), is a useful technique for analyzing data that suffer from multicollinearity. In common with other shrinkage methods such as LASSO (Tibshirani, 1996) and the elastic net (Zou & Hastie, 2005), ridge regression aims at maximizing the likelihood function by adding a penalty to the model coefficients.

A general form for penalized regression is given by the following optimization problem:

$$\hat{\boldsymbol{\beta}}_\lambda = \arg \max_{\boldsymbol{\beta} \in \mathbb{R}^p} \{\ell(\boldsymbol{\beta}) - p_\lambda(\boldsymbol{\beta})\},$$

where  $p_\lambda(\boldsymbol{\beta}) = \lambda p(\boldsymbol{\beta})$  is the penalty term. For ridge regression,  $p(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$ , where  $\|\cdot\|_q$  is the  $L_q$  norm.

Therefore, ridge regression puts an additional penalty term on the parameters instead of just maximizing the log-likelihood function, where the penalty term is the tuning parameter  $\lambda$  times the square of the  $L_2$  norm of the coefficients vector  $\boldsymbol{\beta}$ . In the extreme cases, when  $\lambda = 0$ , the ridge estimator is simply the maximum likelihood estimation (MLE), while when  $\lambda$  approaches infinity, all the coefficients tend to zero. Consequently, the performance of ridge regression largely depends on the tuning parameter  $\lambda$  that balances the trade-off between bias and variance.

Interest in the applications of ridge regression has increased as high-dimensional data are increasingly common. Bøvelstad et al. (2007) compared several dimension reduction or parameter shrinkage methods for high-dimensional data and concluded that ridge regression has the overall best prediction performance. Based on the ridge estimation, Bühlmann et al. (2013) proposed a method for constructing  $p$ -values for general hypotheses in the high-dimensional linear model. An automatic method was derived by Cule and De Iorio (2013) to choose the ridge parameter for high-dimensional data. Van Wieringen and Peeters (2016) investigated the properties of the ridge estimation of the precision matrix for high-dimensional data. Recently, ridge regression was applied to VAR(1) models by Miok et al. (2017). Ridge regression has a good reputation in prediction for high-dimensional data.

## 2.3 | Choice of the tuning parameter

The main idea of the confidence region approach is to choose the  $L_2$ -sparsest solution  $\hat{\boldsymbol{\beta}}_\lambda$  contained in the confidence region for  $\boldsymbol{\beta}$ . The solution is the first time that the path of ridge estimator starting from  $\lambda = \infty$  to  $\lambda = 0$  reaches the

boundary of the confidence region. When the ridge path is completely included in the confidence region, the null model is chosen. Gunes and Bondell (2012) used the likelihood ratio test for low-dimensional data. However, the high dimensionality renders this test inapplicable. We replace it with the globaltest in this paper.

As a consequence, the tuning parameter selected by the globaltest confidence region at level  $1 - \alpha$  is

$$\lambda^{gt}(\alpha) = \sup\{\lambda \in [0, \infty) : \hat{\beta}_\lambda \in C_\alpha^{gt}\}. \tag{2}$$

Given a specific value of  $\alpha$ , the solution for  $\lambda$  in (2) is fully determined by  $\alpha$ , suggesting to use the penalized estimate  $\hat{\beta}_{\lambda^{gt}(\alpha)}$ . Similarly, for a given tuning parameter  $\lambda$ , it can be checked whether  $\hat{\beta}_\lambda$  lies in the  $1 - \alpha$  confidence region  $C_\alpha^{gt}$ , or which is the smallest level  $\alpha$  such that  $\hat{\beta}_\lambda \in C_\alpha^{gt}$ , that is,

$$\alpha^{gt}(\lambda) = \inf\{\alpha \in [0, 1] : \hat{\beta}_\lambda \in C_\alpha^{gt}\}.$$

A mapping, therefore, can be built between the tuning parameter  $\lambda$  and the confidence level parameter  $\alpha$ . Under the assumption that a smaller significance level  $\alpha$  corresponds to a larger confidence region, we have that  $\lambda(\alpha)$  is a nonincreasing function on  $\alpha$ , or equivalently,  $\alpha(\lambda)$  is a nonincreasing function on  $\lambda$ . Many of the commonly used tests satisfy this assumption, such as the likelihood ratio test, Wald test and globaltest: for  $\alpha_1 \leq \alpha_2$ ,  $c_{\alpha_1} \geq c_{\alpha_2}$  holds so that  $C_{\alpha_1} \supseteq C_{\alpha_2}$ , thereby  $\lambda(\alpha_1) \geq \lambda(\alpha_2)$ .

### 3 | LINEAR MODELS

For the specific case of linear models, we show more detail about the properties of the globaltest and the ridge estimator, and then compare the confidence regions of the globaltest and  $F$ -test.

#### 3.1 | Detectable regions for the globaltest

Consider a linear model

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n),$$

where  $\mathbf{I}_n$  is the  $n \times n$  identity matrix. Then the globaltest statistic for the linear model becomes

$$S_{\beta_0} = \frac{\|\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)\|_2^2}{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0\|_2^2}. \tag{3}$$

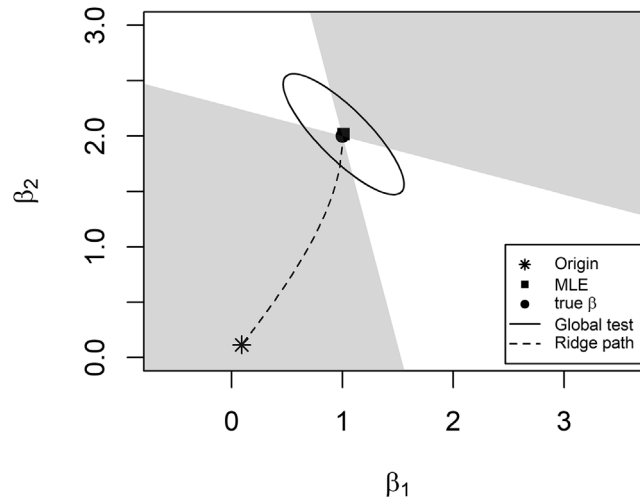
Goeman et al. (2006) proved that the globaltest is the locally most powerful test on average in a neighborhood of the null hypothesis. However, especially when  $p \gg n$ , there are points  $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$  for which the globaltest has negligible power. More specifically, let  $p > n$  and  $\mathbf{X}^\top\mathbf{X} = \sum_{i=1}^n \gamma_i \mathbf{V}_i$ , where  $\gamma_1 \geq \dots \geq \gamma_n \geq 0$  are the nonzero eigenvalues of  $\mathbf{X}^\top\mathbf{X}$  and  $\mathbf{V}_i$  is the  $p \times p$  projection matrix that projects onto the eigenvector of  $\mathbf{X}^\top\mathbf{X}$  corresponding to the eigenvalue  $\gamma_i$ . As detailed in Goeman et al. (2006), the globaltest is less powerful for the points whose expected test statistic under alternative hypothesis is smaller than that under the null hypothesis. The difference of the expectations under alternative and null hypotheses is approximately proportional to the covariance of  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_n)^\top$  and  $\mathbf{r}^2 = (r_1^2, \dots, r_n^2)^\top$ , where

$$r_i^2 = \frac{\gamma_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbf{V}_i(\boldsymbol{\beta} - \boldsymbol{\beta}_0)}{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbf{X}^\top\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + n\sigma^2}, \tag{4}$$

and  $r^2 = \sum_{i=1}^n r_i^2$  is the fraction of variance of  $\mathbf{y}$  explained by the alternative hypothesis. Thus, the detectable region for the globaltest is defined as

$$\mathcal{D} = \{\boldsymbol{\beta}_0 \in \mathbb{R}^p : \text{cov}(\boldsymbol{\gamma}, \mathbf{r}^2) > 0\}.$$

The globaltest has good power for testing the points inside  $\mathcal{D}$ , as opposed to the points outside.



**FIGURE 1** Detectable region (gray shaded area) of the globaltest  
 Note: \* denotes the origin of ridge path, and ■ is the end point, MLE. • represents the true coefficients. The ridge path is represented by the dashed line, and the solid line indicates the boundary of the globaltest confidence region

The ridge estimator for the linear model has the following form:

$$\hat{\beta}_\lambda = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{y}.$$

Based on the singular value decomposition of  $\mathbf{X} = \mathbf{U}\mathbf{\Gamma}^{1/2}\mathbf{V}^T$ ,  $\hat{\beta}_\lambda$  can be written as

$$\hat{\beta}_\lambda = \sum_{i=1}^n \frac{\gamma_i^{1/2}}{\gamma_i + \lambda} \mathbf{v}_i \mathbf{u}_i^T \mathbf{y}. \tag{5}$$

Here  $\mathbf{v}_i$  and  $\mathbf{u}_i$  are the  $i$ th columns of  $\mathbf{V}$  and  $\mathbf{U}$ , where  $\mathbf{v}_i$  is called the  $i$ th principal component direction of  $\mathbf{X}$ .

It can be seen from (5) that  $\frac{\gamma_i^{1/2}}{\gamma_i + \lambda}$  is an increasing function of  $\gamma_i$  when  $\lambda > \gamma_i$  and is a decreasing function of  $\gamma_i$  when  $\lambda < \gamma_i$ . In other words, the ridge estimator  $\hat{\beta}_\lambda$  will be more correlated with the large variance principal components of  $\mathbf{X}$  than with those with small variance when  $\lambda > \gamma_i$ ; it will be more correlated with the small variance principal components than with those with the large variance when  $\lambda < \gamma_{\min(n,p)}$ . Hence, the ridge path starting from  $\lambda = \infty$  to  $\lambda = 0$  would first move along the direction of strong principal components, and then change into the direction of small principal components until reaching the MLE.

Figure 1 illustrates the detectable region of the globaltest and the direction of ridge path for the Gaussian linear model with  $n = 50$  and  $p = 2$ . The true coefficients are  $\beta = (1, 2)^T$ , and the correlation between these two predictors is  $\rho = 0.6$ . It can be seen that the ridge path approaches to the MLE first in the direction of the strong principal component of the design matrix, which is the direction of minor axis of the ellipse, and then turns into the direction of the weak principal component, which is the direction of major axis.

### 3.2 | Comparisons with the Scheffé confidence region

The  $F$ -test statistic for testing  $H_0 : \beta = \beta_0$  against  $H_1 : \beta \neq \beta_0$  is given by

$$T_{\beta_0} = \frac{\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta_0\|_2^2/p}{\hat{\sigma}^2},$$

which follows an  $F$  distribution with degrees of freedom  $p$  and  $n - p$  under  $H_0$ , where  $\hat{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2/(n - p)$  and  $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ . The Scheffé confidence region obtained by inverting the  $F$ -test statistic is a hyperellipsoid centered at

the MLE:

$$C_{\alpha}^{\text{ft}} = \{\boldsymbol{\beta}_0 \in \mathbb{R}^p : (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})^{\top} \mathbf{X}^{\top} \mathbf{X} (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) \leq p \hat{\sigma}^2 f_{p, n-p}^{\alpha}\},$$

where  $f_{p, n-p}^{\alpha}$  is the  $1 - \alpha$  quantile of the  $F$  distribution with  $p$  and  $n - p$  degrees of freedom.

It is interesting to note that the confidence region of the globaltest is not always ellipsoid. Theoretically, for a given confidence level  $1 - \alpha$ , the border of the globaltest confidence region for linear models based on (1) and (3) is

$$\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)^{\top} \mathbf{X}\mathbf{X}^{\top} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)}{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)^{\top} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)} = c_{\alpha},$$

which is equivalent to

$$\boldsymbol{\beta}_0^{\top} (\mathbf{X}^{\top} \mathbf{X}\mathbf{X}^{\top} \mathbf{X} - c_{\alpha} \mathbf{X}^{\top} \mathbf{X}) \boldsymbol{\beta}_0 - 2\mathbf{y}^{\top} (\mathbf{X}\mathbf{X}^{\top} \mathbf{X} - c_{\alpha} \mathbf{X}) \boldsymbol{\beta}_0 + \mathbf{y}^{\top} (\mathbf{X}\mathbf{X}^{\top} - c_{\alpha} \mathbf{I}_n) \mathbf{y} = 0. \quad (6)$$

For  $p = 2$ , Equation (6) is exactly the general form of a conic section. The type of the conic section can be determined by the sign of  $\delta = \det(\mathbf{X}^{\top} \mathbf{X}\mathbf{X}^{\top} \mathbf{X} - c_{\alpha} \mathbf{X}^{\top} \mathbf{X})$  (Desgraupes, 2013). Then one has the following classifications based on  $\delta$ :

- if  $\delta > 0$ , (6) is an ellipse;
- if  $\delta = 0$ , (6) is a pair of parallel lines;
- if  $\delta < 0$ , (6) is a hyperbola.

Figure 2 shows the comparisons of the confidence regions of the globaltest and  $F$ -test for simulated data with  $n \in \{5, 50\}$  samples and two predictors, for which the correlation is  $\rho \in \{0, 0.9\}$ . It is shown that decreasing the sample size or increasing the correlation makes the confidence region of the globaltest become narrower than the Scheffé confidence region along the direction of the strong principal component. This results in smaller  $\lambda$  chosen by the confidence region of the globaltest than the  $F$ -test provided that the ridge path comes from the detectable region of globaltest. Note that it can happen that the whole ridge path is completely included in the confidence region, as the example of  $n = 5$  in Figure 2 demonstrates. In that case  $\lambda = \infty$  is chosen.

## 4 | SIMULATIONS

We conduct two simulations to illustrate the points raised in previous sections. One is to show the comparisons between the globaltest-based method and the  $F$ -based method. The other is to describe the predictive ability of the proposed method as compared with other methods, both for low- and high-dimensional data. The design matrix  $\mathbf{X}$  in the simulated data is based on a real gene expression data set from the breast cancer study published by Van't Veer et al. (2002) and Van De Vijver et al. (2002), including 14,318 gene features for 337 breast cancer patients (after removing the missing values). We take a low-dimensional setting with the first  $n = 300$  patients and the first  $p = 50$  gene features considered for the first simulation so that  $F$ -test can also be applied. A high-dimensional setting is added in the second simulation with the first  $n = 300$  patients and the first  $p = 1000$  gene features. We use the linear model  $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$  to generate the output variable. All simulation results shown below are based on 1000 replications.

### 4.1 | Comparison with the $F$ -based method

We use the same setup of the true coefficients as in Goeman et al. (2006) so that we can gain insights into the properties of the globaltest. In terms of singular value decomposition, we have  $\mathbf{X} = \mathbf{U}\boldsymbol{\Gamma}^{1/2}\mathbf{V}^{\top}$  with  $\mathbf{U}$  be an  $n \times \min(n, p)$  (semi)orthogonal matrix,  $\mathbf{V}$  a  $p \times \min(n, p)$  (semi)orthogonal matrix and  $\boldsymbol{\Gamma}$  is a  $\min(n, p) \times \min(n, p)$  diagonal matrix. It is shown in Goeman et al. (2006) that globaltest is powerful especially when the large principal components explain more of the variance of the response than the small ones. We therefore define the true model coefficients  $\boldsymbol{\beta}$  in a way that it can

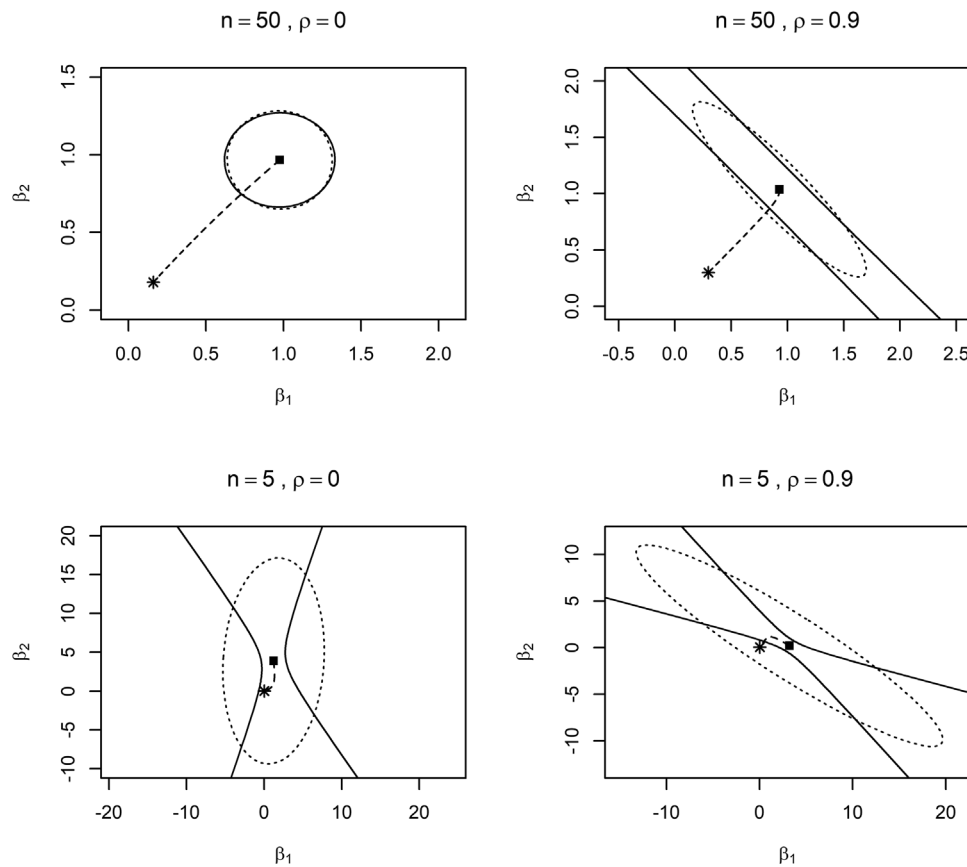


FIGURE 2 Confidence regions of the globaltest and  $F$ -test for  $n = 50, 5, p = 2$  with correlations  $\rho = 0, 0.9$

Note: The dashed line denotes the ridge path with \* is the origin and ■ is MLE. The solid line denotes the globaltest confidence region and the dotted line denotes the Scheffé confidence region

vary the amount of variance explained by the principal components by varying  $s$ :

$$\beta = \mathbf{V}\gamma^{s/2}, \tag{7}$$

where  $\gamma$  are the nonzero diagonals of  $\Gamma$ .

When  $s > 0$ , the large variance principal components will have large coefficients and also large  $r^2$  for fixed  $\sigma^2$ , in terms of (7) and (4). Positive correlations between  $\gamma$  and  $\mathbf{r}^2$  are thus obtained, leading to good power of the globaltest. When  $-1 < s < 0$ , the large variance principal components will have smaller coefficients but larger  $r^2$  than those small variance principal components. Although when  $s < -1$ ,  $\mathbf{y}$  is totally determined by the small variance principal components. Thus, when  $s$  becomes negative, globaltest tends to lose power due to the negative correlation between  $\gamma$  and  $\mathbf{r}^2$ .

Given the value of  $r^2$  and the true coefficients  $\beta$ , we can calculate  $\sigma^2$  based on Equation (4). We then use linear regression to generate the response  $\mathbf{y}$ . The larger  $r^2$ , the more variance of  $\mathbf{y}$  explained by the true model. The larger  $s$ , the more powerful the globaltest. Goeman et al. (2006) argued that  $s > 0$  is fortunately common in the real data, for which the globaltest has good power. For negative  $s$ , globaltest has negligible power even for large  $r^2$ .

Table 1 shows the proportion of times that the tuning parameter selected by the globaltest-based method is smaller than that by the  $F$ -based method. It is shown that the globaltest-based method would choose smaller, that is, less conservative, tuning parameters in comparison to the  $F$ -based method for large values of  $s$ , because the power of the globaltest in this case is better than that of  $F$ -test. For negative values of  $s$ , which cause negative correlations between  $\gamma$  and  $\mathbf{r}^2$ , the  $F$ -based method outperforms the globaltest-based method. For example, the proportion is 0.946 for  $s = 1.5$  and  $r^2 = 0.15$ , whereas for  $s = -1.5$  with the same  $r^2$ , the proportion is 0. This is consistent with the properties of the globaltest discussed in Goeman et al. (2006).

**TABLE 1** Proportion of times that  $\lambda^{\text{gt}} < \lambda^{\text{ft}}$ , where  $\lambda^{\text{gt}}$  and  $\lambda^{\text{ft}}$  denote the tuning parameters selected by the globaltest-based method and  $F$ -based method, respectively, with confidence level 95%

$s$	$r^2$			
	0.02	0.05	0.1	0.15
1.5	0.733	0.907	0.935	0.946
1	0.715	0.902	0.926	0.936
0.5	0.694	0.869	0.888	0.881
0	0.629	0.739	0.692	0.594
-0.5	0.486	0.407	0.237	0.109
-1	0.373	0.131	0.042	0.008
-1.5	0.291	0.056	0.005	0.000

**TABLE 2** Summary of penalties of information criteria used in the simulations

Information criterion	Penalty
AIC	$2 * mc$
AICc	$\{2 + \frac{2(mc+1)}{n-mc-1}\} \times mc$
BIC	$\log(n) \times mc$
mBIC	$\{\log(n) + 2 \log(\frac{p}{4} - 1)\} \times mc$
mBIC2	$\{\log(n) + 2 \log(\frac{p}{4})\} \times mc - 2 \log(mc!)$
GIC	$\{\log(\log(n)) \log(p)\} \times mc$
RIC	$2 \log(p) \times mc$

## 4.2 | Predictive ability

To investigate the predictive ability of the confidence region method, we calculate the mean squared error (MSE) of the predictions in terms of

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i^T \boldsymbol{\beta} - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\lambda)^2,$$

where  $\hat{\boldsymbol{\beta}}_\lambda$  is the ridge estimate. For the low-dimensional setting, we compare the globaltest-based method with CV (5-CV, LOOCV, and generalized CV (GCV), Golub et al., 1979), information criteria (AIC, Akaike, 1973; and its variant AICc, Cavanaugh et al., 1997; BIC, Schwarz et al., 1978; and its modified versions mBIC, mBIC2, Žak-Szatkowska and Bogdan, 2011; GIC, Fan and Tang, 2013; and RIC, Foster and George, 1994) and the  $F$ -based method.

The information criteria measure the balance between model fit and model complexity by minimizing the following expression (van Wieringen, 2020):

$$-2 \times \ell(\lambda) + \text{penalty on model complexity},$$

where  $\ell(\lambda)$  is the penalized log-likelihood and  $mc = \sum_{i=1}^{\min(n,p)} \frac{y_i}{y_i + \lambda}$  denotes the model complexity for ridge regression. We use the R package `penalized` (Goeman, 2012) to calculate  $\ell(\lambda)$ . Penalties used in all of the information criteria mentioned above are summarized in Table 2. The CV results are also calculated from R package `penalized`. For the high-dimensional setting, we exclude AIC and  $F$ -based method, as they totally break down. The results for both low and high dimensions are summarized in Figures 3 and 4, respectively.

FT50 and FT95 denote the  $F$ -based method with confidence levels 50% and 95%, respectively. Similarly, GT50 and GT95 are the globaltest-based method with confidence levels 50% and 95%, respectively. The reason why an alternative significance level 50% is considered is that the traditional methods like CV usually have a type I error rate that is close to 50%.



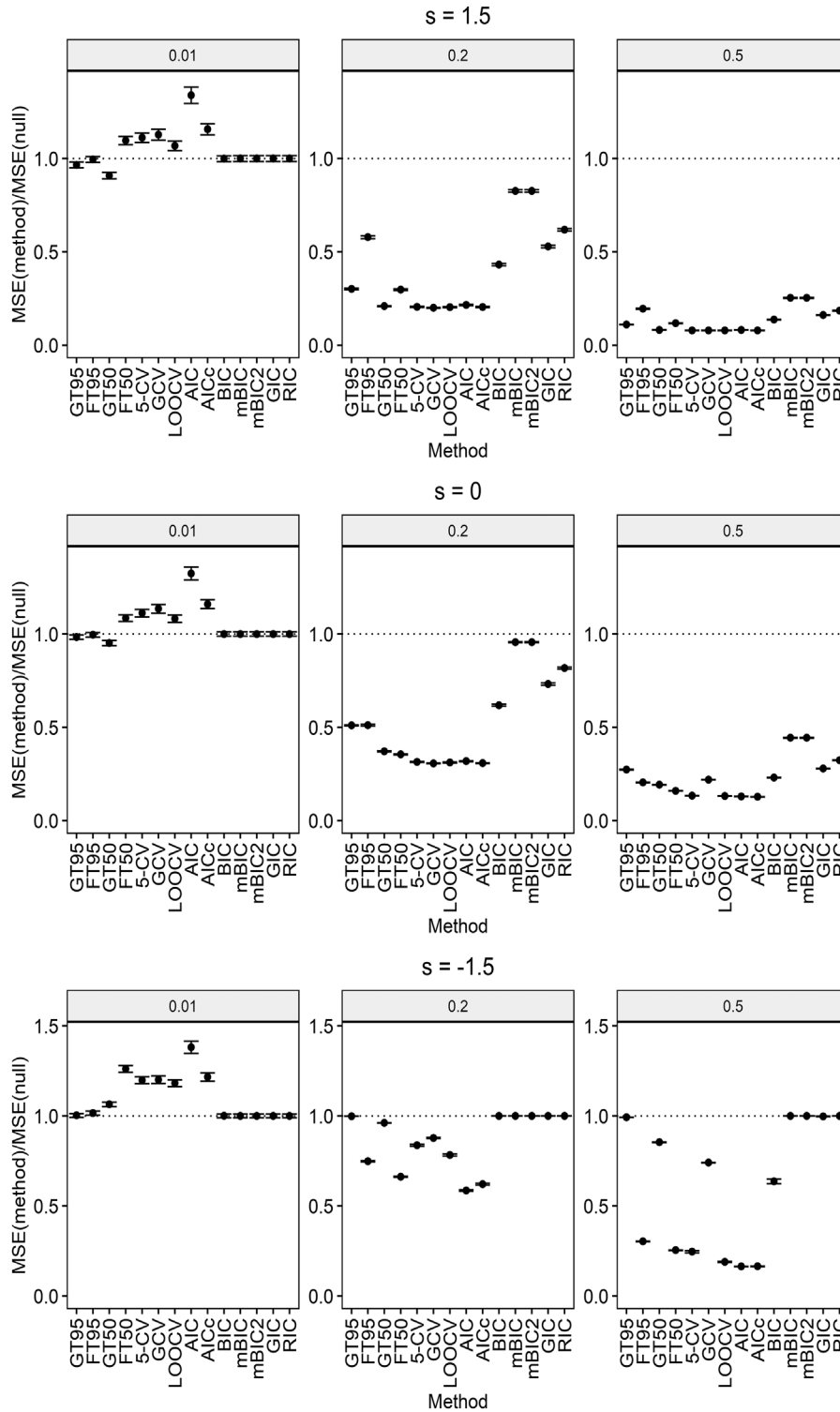
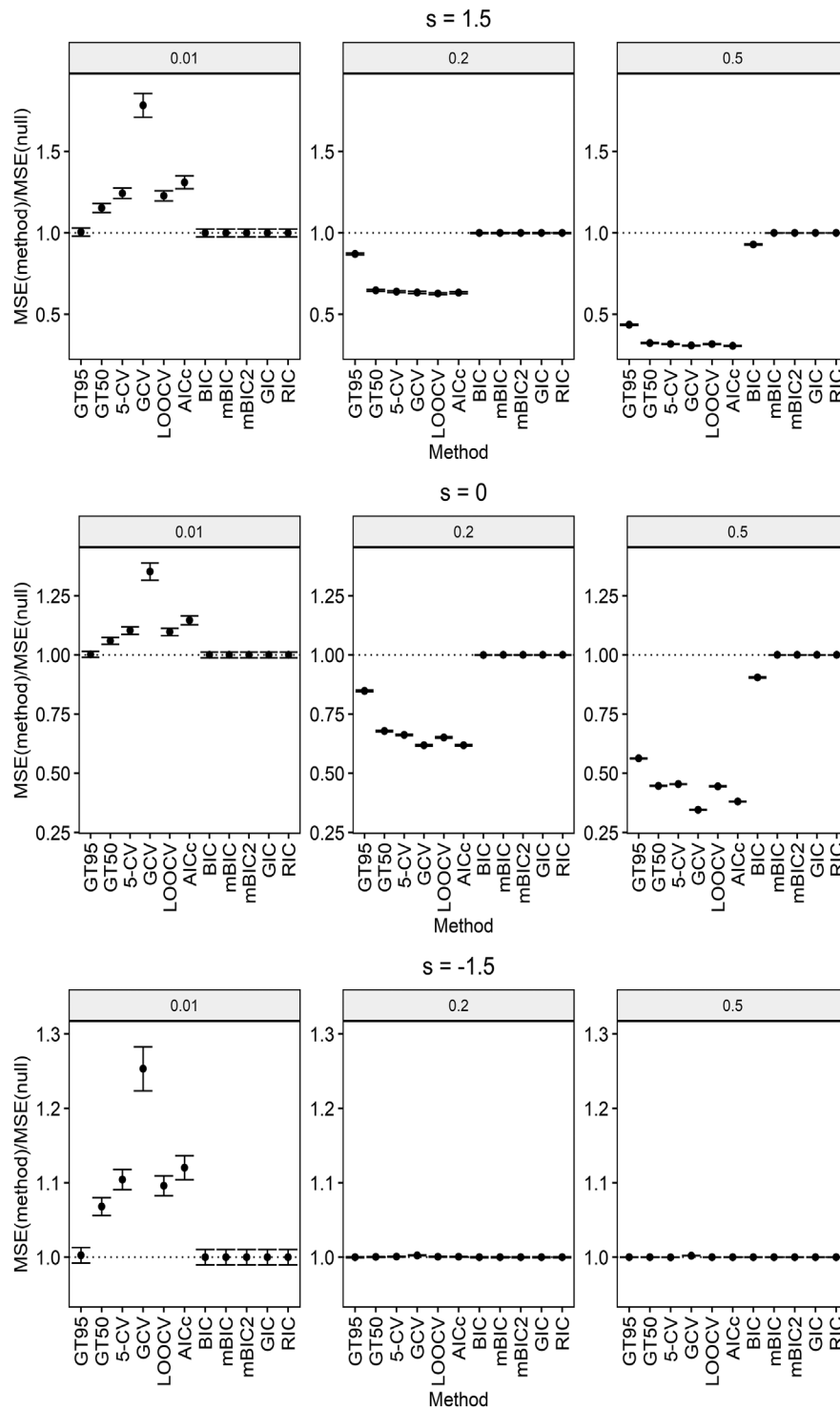


FIGURE 3 MSE relative to MSE of the null model:  $MSE(\text{method})/MSE(\text{null}) \pm$  standard errors in low dimensions for  $r^2 = 0.01, 0.2, 0$ . Note: The dotted line corresponds to MSE of the null model

Therefore, if strong type I error rate control is not desired, under similar type I error rate control, the 50% confidence region method becomes comparable to the traditional CV methods.

It can be seen from Figure 3 that the comparisons between the globaltest-based method and the  $F$ -based method is consistent with the result in Table 1. Moreover, it is shown in Figures 3 and 4 that, for the case with positive  $s$  where globaltest has good power, GT50 and the CV methods have similar performance in terms of MSE, making GT50 a good



**FIGURE 4** MSE relative to MSE of the null model:  $\text{MSE}(\text{method})/\text{MSE}(\text{null}) \pm$  standard errors in high dimensions for  $r^2 = 0.01, 0.2, 0.5$   
*Note:* The dotted line corresponds to MSE of the null model

alternative to CV but without high computational burden. When  $s$  becomes negative, the globaltest-based method falls behind the other method. However, negative  $s$  occurs only seldom in real-world data sets (Goeman et al., 2006).

For large  $r^2$  where the variance of the response is largely explained by predictors, AIC in low dimensions and AICc in both low and high dimensions achieve quite good prediction abilities in terms of MSE. This might be due to the small penalties of AIC and AICc compared to the large penalties of BIC, mBIC, mBIC2, GIC, and RIC, which result in underfitting models with an MSE very close to that of the null model, particularly in high dimensions.

TABLE 3 Probability of choosing  $\lambda = \infty$  when the null model is true

Methods	$p = 50$	$p = 1000$
AIC	0.729	0.000
AICc	0.748	0.792
BIC	0.998	1.000
mBIC	1.000	1.000
mBIC2	1.000	1.000
GIC	1.000	1.000
RIC	1.000	1.000
GCV	0.598	0.644
5-CV	0.579	0.618
LOOCV	0.583	0.650
GT95	0.953	0.973
GT50	0.508	0.607
FT95	0.947	–
FT50	0.507	–

Although for extremely small  $r^2$  where effects of the predictors on the response are extremely weak, globaltest has good power for testing groups of weak effects so that GT50 and GT95 have decent performance on prediction in both low and high dimensions, regardless of the sign of  $s$ . Models tuned by BIC, mBIC, mBIC2, GIC, and RIC also predict well in this case because of their large penalties on model complexity, which are mainly dominated by  $p$ , especially in high dimensions.

Additionally, we investigate the probability that the null model is chosen when it is true (see Table 3). The confidence region approach can guarantee that the null model is chosen with probability at least  $1 - \alpha$ , which is around 95% and 50%, respectively, for GT95 and GT50. We note that the probabilities computed by BIC and its variants and GIC and RIC are large because they adopt a large penalty on model complexity, causing an increasing risk of underfitting models, thereby a high probability that null model is chosen. Although for AICc, it is AIC with an additional penalty on model complexity that is depending on both the sample size and the model complexity itself and can avoid overfitting of AIC to some extent, which is probably the reason that the probability is 0.748 and 0.792 in low and high dimensions in our case.

## 5 | REAL DATA EXAMPLES

In the simulation study, we showed the application of confidence region method in linear models. In the real data analysis, we apply the method to Cox models. We consider three high-dimensional gene expression data sets on breast cancer study: MAINZ with 200 samples and 22,283 gene features (Schmidt et al., 2008); TRANSBIG with 198 samples and 22,283 features (Desmedt et al., 2007); UNT with 137 samples and 44,928 features (Sotiriou et al., 2006). We fit the data by the Cox proportional hazard model with a survival response, which is given by a vector of survival times  $\mathbf{t} = (t_1, \dots, t_n)^\top$  and a vector of status indicators  $\mathbf{d} = (d_1, \dots, d_n)^\top$ , where  $d_i = 1$  indicates that  $t_i$  is an observed survival time and  $d_i = 0$  indicates that the survival time is right-censored at  $t_i$ . Let  $h_i(t)$  denote the hazard function at time  $t$  for the  $i$ th subject. The Cox proportional hazards model assumes  $\log(h_i(t)/h_0(t)) = \mathbf{x}_i^\top \boldsymbol{\beta}$ , where  $h_0(t)$  is an unspecified underlying hazard. The globaltest confidence region for the Cox model can be obtained by inverting the Cox model version of the globaltest (Goeman et al., 2005).

Cross-validated partial likelihood (cvpl) can be used as a measure of the predictive ability of Cox models (Verweij & Van Houwelingen, 1993). We compare globaltest-based method with 5-fold CV by calculating cvpl, based on 5-fold CV and 10-fold CV, respectively. Note that the fold assigning used for estimating the tuning parameter is different from that is used for calculating cvpl. The results are listed in Table 4. The higher the cvpl, the better the predictive performance of the method. It can be seen from Table 4 that there is no large difference of cvpl between CV and confidence region method. CV outperforms the globaltest-based method in most cases, the globaltest-based method is, however, much easier to compute than CV without much loss of predictive accuracy.

The Brier score is another way to measure the predictive accuracy for survival analysis, which measures the mean squared difference between the predicted survival probability and the actual one (Van Houwelingen & Putter, 2011). It is

TABLE 4 Cross-validated partial likelihood by 5-fold CV and 10-fold CV

Data sets	Method	5-fold	10-fold
MAINZ	CV	−254.90	−257.16
	GT50	−256.71	−259.23
	GT95	−260.11	−262.76
TRANSBIG	CV	−353.71	−358.28
	GT50	−355.90	−359.31
	GT95	−356.66	−359.81
UNT	CV	−149.34	−151.75
	GT50	−149.99	−151.73
	GT95	−150.42	−151.90

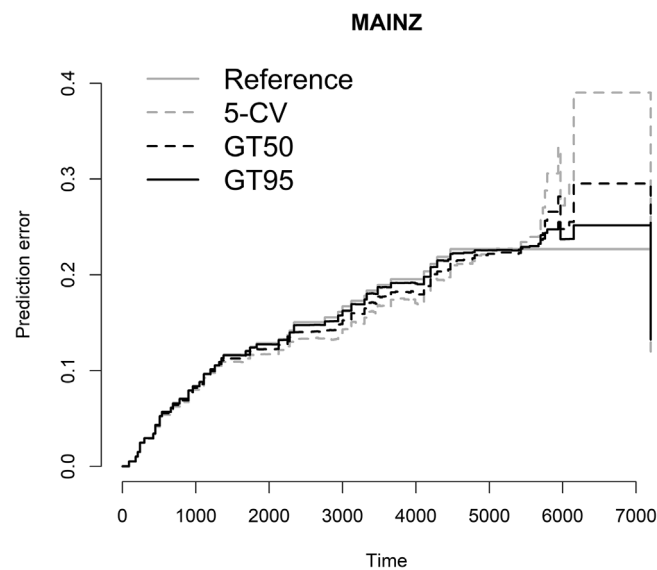


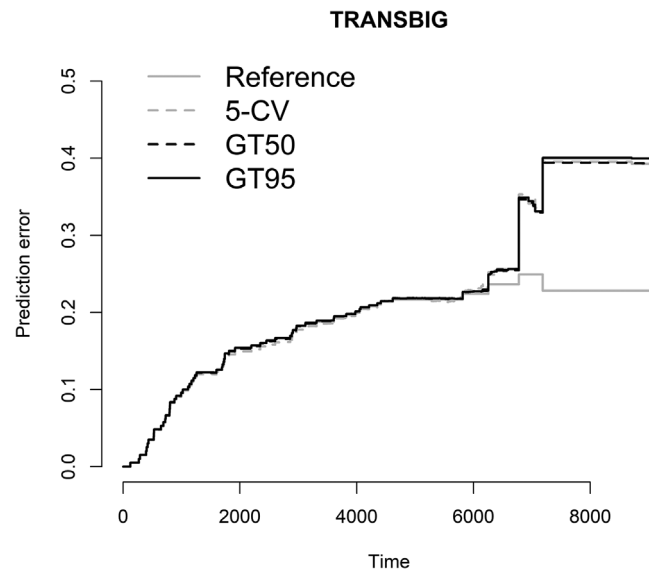
FIGURE 5 Brier score for MAINZ by the Kaplan–Meier procedure (solid gray line), 5-CV (dashed gray line), GT50 (dashed black line) and GT95 (solid black line)

an overall performance measurement that can be decomposed into two important characteristics of a prediction model, discrimination and calibration (Steyerberg et al., 2010). Figures 5–7 show the Brier score over time for the models obtained by 5-CV, GT50, and GT95 in data sets MAINZ, TRANSBIG, and UNT, respectively. The marginal Kaplan–Meier prediction model is presented as a reference to other models. The lower the Brier score, the better the prediction. The results in the figures confirm the conclusion obtained in terms of cvpl that both methods have similar prediction errors, especially at earlier time points. Some differences can be seen at later time points, where the globaltest-based method predicts the survival probability better than CV, especially for the MAINZ data.

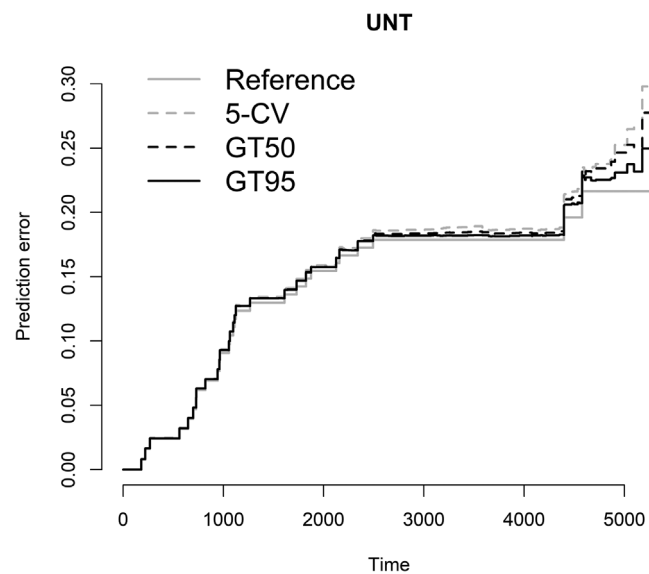
## 6 | DISCUSSION

In this work, we constructed the globaltest confidence region, which is powerful to test against high-dimensional alternatives especially when there are many weak effects, a setting also favorable for ridge regression. We thus proposed to use the globaltest confidence region to choose the tuning parameter of ridge regression, thereby extending the confidence region approach for tuning parameter selection in low to high dimensions by replacing the  $F$ -test with the globaltest. We argued that the globaltest has better power than the  $F$ -test when strong principal components of the design matrix explain more variance of the outcome than the weak ones, which is common in real-world applications.

The tuning parameter selected by the globaltest confidence region is the parameter corresponding to the first time that the ridge path reaches the boundary of the confidence region at a specified level  $\alpha$ , or is the infinity when the whole path



**FIGURE 6** Brier score for TRANSBIG



**FIGURE 7** Brier score for UNT

included in the region. It can be seen as the least complex model among all acceptable models. Tuning via confidence regions is computationally less demanding than CV. Compared with information criteria, the confidence region method has less dependence on the penalties on the model complexity. And, as a testimation procedure, it further guarantees that the null model is selected with a prespecified probability in the case that it is the true model. This can be linked to the weak family-wise error rate control from the perspective of multiple testing.

An important asset of the globaltest-based method is that it is known when this method is expected to perform well, that is, when the strong principal components dominate signals or when there are many weak signals. We focused on ridge regression because it is similar in spirit to the globaltest, but in principle our approach may be used for other penalized method for model selection as well. With regard to multiple testing corrections applied to model selection, such as family-wise error rate and false discovery rate, see [Žak-Szatkowska and Bogdan \(2011\)](#) for more detail.


## CONFLICT OF INTEREST

The authors have declared that there is no conflict of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in bioconductor with doi: <https://doi.org/10.18129/B9.BIOC.BREASTCANCERMAINZ>, <https://doi.org/10.18129/B9.BIOC.BREASTCANCERTRANSBIG>, <https://doi.org/10.18129/B9.BIOC.BREASTCANCERUNT>.

## OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

## ORCID

Ningning Xu  <https://orcid.org/0000-0002-8385-0670>

Aldo Solari  <https://orcid.org/0000-0003-1243-0385>

Jelle Goeman  <https://orcid.org/0000-0003-4283-0259>

## REFERENCES

- Aitkin, M. A. (1974). Simultaneous inference and the choice of variable subsets in multiple regression. *Technometrics*, 16(2), 221–227.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrow & F. Czaki (Eds.), *Proceedings of the 2nd international symposium on information*. Akademiai Kiado, Budapest.
- Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, Ø., Frigessi, A., & Lingjærde, O. C. (2007). Predicting survival from microarray data—a comparative study. *Bioinformatics*, 23(16), 2080–2087.
- Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression. The x-random case. *International Statistical Review*, 3, 291–319.
- Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4), 1212–1242.
- Cavanaugh, J. E. (1997). Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters*, 33(2), 201–208.
- Cule, E., & De Iorio, M. (2013). Ridge regression in prediction problems: Automatic choice of the ridge parameter. *Genetic Epidemiology*, 37(7), 704–714.
- Desgraupes, B. (2013). *conics: Plot conics*. R package version. <http://www.r-project.org>
- Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M. S., Bergh, J., Lidereau, R., Ellis, P., Harris, A. L., Klijn, J. G., Foekens, J. A., Cardoso, F., Piccart, M. J., Buyse, M., ... TRANSBIG Consortium, (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clinical Cancer Research*, 13(11), 3207–3214.
- Fan, Y., & Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3), 531–552.
- Foster, D. P., & George, E. I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics*, 22, 1947–1975.
- Goeman, J., Oosting, J., Finos, L., & Solari, A. (2010). The global test and the globaltest r package. *Bioconductor*. <http://www.bioconductor.org/packages/release/bioc/vignettes/globaltest/inst/doc/GlobalTest.pdf>
- Goeman, J. J. (2012). *Penalized R package*. R package version 09-41.
- Goeman, J. J., Oosting, J., Cleton-Jansen, A.-M., Anninga, J. K., & Van Houwelingen, H. C. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 21(9), 1950–1957.
- Goeman, J. J., Van De Geer, S. A., De Kort, F., & Van Houwelingen, H. C. (2004). A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics*, 20(1), 93–99.
- Goeman, J. J., Van De Geer, S. A., & Van Houwelingen, H. C. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3), 477–493.
- Goeman, J. J., Van Houwelingen, H. C., & Finos, L. (2011). Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika*, 98, 381–390.
- Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2), 215–223.
- Gunes, F., & Bondell, H. D. (2012). A confidence region approach to tuning for variable selection. *Journal of Computational and Graphical Statistics*, 21(2), 295–314.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.
- Imhof, J.-P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3/4), 419–426.
- Jiang, J., Rao, J. S., Gu, Z., & Nguyen, T., et al. (2008). Fence methods for mixed model selection. *Annals of Statistics*, 36(4), 1669–1692.
- McCabe, G. P. (1978). Evaluation of regression coefficient estimates using  $\alpha$ -acceptability. *Technometrics*, 20(2), 131–139.

- Miok, V., Wilting, S. M., & Wieringen, W. N. (2017). Ridge estimation of the var (1) model and its time series chain graph from multivariate time-course omics data. *Biometrical Journal*, 59(1), 172–191.
- Obenchain, R. (1977). Classical f-tests and confidence regions for ridge regression. *Technometrics*, 19(4), 429–439.
- Oman, S. D. (1981). A confidence bound approach to choosing the biasing parameter in ridge regression. *Journal of the American Statistical Association*, 76(374), 452–461.
- Rahman, M., & Gokhale, D. (1996). Testimation in regression parameter estimation. *Biometrical Journal*, 38(7), 809–817.
- Robbins, H., & Pitman, E. (1949). Application of the method of mixtures to quadratic forms in normal variates. *Annals of Mathematical Statistics*, 20, 552–560.
- Schmidt, M., Böhm, D., Von Törne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H. A., Hengstler, J. G., Kölbl, H., & Gehrmann, M. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Research*, 68(13), 5405–5413.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.
- Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M. J., Bergh, J., Piccart, M., & Delorenzi, M. (2006). Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, 98(4), 262–272.
- Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology*, 21(1), 128.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.
- Van De Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., ... Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25), 1999–2009.
- Van Houwelingen, H., & Putter, H. (2011). *Dynamic prediction in clinical survival analysis*. Boca Raton, FL: CRC Press.
- van Wieringen, W. N. (2020). Lecture notes on ridge regression. Preprint arXiv:1509.09169.
- Van Wieringen, W. N., & Peeters, C. F. (2016). Ridge estimation of inverse covariance matrices from high-dimensional data. *Computational Statistics & Data Analysis*, 103, 284–303.
- Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H., van der Kooy, K., Marton, M., Witteveen, A. T., Schreiber, G., Kerkhoven, R., Roberts, C., Linsley, P., Bernards, R., & Friend, S. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871), 530.
- Verweij, P. J., & Van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in Medicine*, 12(24), 2305–2314.
- Żak-Szatkowska, M., & Bogdan, M. (2011). Modified versions of the bayesian information criterion for sparse generalized linear models. *Computational Statistics & Data Analysis*, 55(11), 2908–2924.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Xu, N., Solari, A., & Goeman, J. (2021). Globaltest confidence regions and their application to ridge regression. *Biometrical Journal*, 63, 1351–1365. <https://doi.org/10.1002/bimj.202000063>