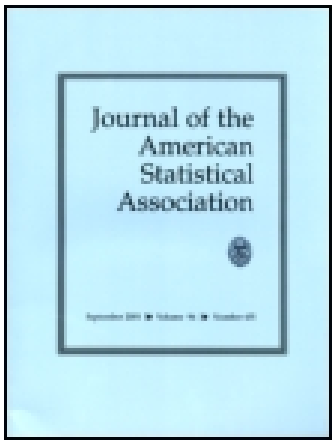

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Generalized Fiducial Inference for Ultrahigh-Dimensional Regression

Randy C. S. Lai, Jan Hannig & Thomas C. M. Lee

Accepted author version posted online: 12 Jun 2014.Published online: 06 Jul 2015.

To cite this article: Randy C. S. Lai, Jan Hannig & Thomas C. M. Lee (2015) Generalized Fiducial Inference for Ultrahigh-Dimensional Regression, Journal of the American Statistical Association, 110:510, 760-772, DOI: 10.1080/01621459.2014.931237

To link to this article: http://dx.doi.org/10.1080/01621459.2014.931237

Generalized Fiducial Inference for Ultrahigh-Dimensional Regression

Randy C. S. LAI, Jan HANNIG, and Thomas C. M. LEE

In recent years, the ultrahigh-dimensional linear regression problem has attracted enormous attention from the research community. Under the sparsity assumption, most of the published work is devoted to the selection and estimation of the predictor variables with nonzero coefficients. This article studies a different but fundamentally important aspect of this problem: uncertainty quantification for parameter estimates and model choices. To be more specific, this article proposes methods for deriving a probability density function on the set of all possible models, and also for constructing confidence intervals for the corresponding parameters. These proposed methods are developed using the generalized fiducial methodology, which is a variant of Fisher's controversial fiducial idea. Theoretical properties of the proposed methods are studied, and in particular it is shown that statistical inference based on the proposed methods will have correct asymptotic frequentist property. In terms of empirical performance, the proposed methods are tested by simulation experiments and an application to a real dataset. Finally, this work can also be seen as an interesting and successful application of Fisher's fiducial idea to an important and contemporary problem. To the best of the authors' knowledge, this is the first time that the fiducial idea is being applied to a so-called "large p small n" problem. A connection to objective Bayesian model selection is also discussed.

KEY WORDS: Confidence intervals; Large p small n; Minimum description length principle; Uncertainty quantification; Variability estimation.

1. INTRODUCTION

The ultrahigh-dimensional linear regression problem has attracted enormous attentions in recent years. A typical description of the problem begins with the usual linear model

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \text{or equivalently} \quad \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a vector of n responses, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is a design matrix of size $n \times p$ with iid variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of p parameters, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ is a vector of n iid random errors with zero mean and unknown variance σ^2. It is assumed that $\boldsymbol{\epsilon}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent, and that p is larger than n and grows at an exponential rate as n increases. It is this last assumption that makes the ultrahigh-dimensional regression problem different from the classical multiple regression problem, for which $p < n$.

When $p \gg n$, it is customary to assume that the number of nonzero coefficient predictors in the true model is small, that is, the true model is sparse. The problem is then to identify which β_j's are nonzero, and to estimate their values. To solve this variable selection problem, one common strategy is to first apply a so-called screening procedure to remove a large number of insignificant predictors, and then apply a penalized method such as the least absolute shrinkage and selection operator (LASSO) method of Tibshirani (1996) or the smoothly clipped absolute deviation (SCAD) method of Fan and Li (2001) to the surviving predictors to select the final set of variables. For screening procedures, one of the earliest is the sure independence screening (SIS) procedure of Fan and Lv (2008). Since then various screening procedures have been proposed: Wang (2009) developed a consistent screening procedure that combines forward regression and the extended Bayesian information criterion (BIC) of Chen and Chen (2008), Bühlmann, Kalisch, and Maathuis (2010) proposed a screening procedure that is based on conditional partial corrections, and Cho and Fryzlewicz (2011) constructed a screening procedure that uses information from both marginal correlation and tilted correlation. Also, other screening procedures are developed for more complicated settings, including generalized linear models and nonparametric additive modeling, for example, Meier, Van De Geer, and Bühlmann (2009), Ravikumar et al. (2009), Fan and Lv (2011), and Fan, Feng, and Song (2011). For an overview of variable selection for high-dimensional problems, see Fan and Lv (2010).

While much efforts have been spent on model selection and parameter estimation for the ultrahigh-dimensional regression problem, virtually no published work is devoted to quantify the uncertainty in the chosen models and their parameter estimates. A notable exception is the pioneering work of Fan, Guo, and Hao (2012), where a cross-validation-based method is proposed to estimate the error variance σ^2. Given such an estimate and a final model, confidence intervals for β_j's can be constructed using classical linear model theory. However, this approach does not account for the additional variability contributed by the need of selecting a final model.

The goal of this article is to investigate the use of Fisher's fiducial idea (Fisher 1930) in the ultrahigh-dimensional regression problem. In particular, a new procedure is developed for constructing confidence intervals for all the parameters

Randy C. S. Lai (E-mail: *rcslai@ucdavis.edu*) is Ph.D. student, and Thomas C. M. Lee (E-mail: *tcmlee@ucdavis.edu*) is Professor, Department of Statistics, University of California at Davis, Davis, CA 95616. Jan Hannig is Professor, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3260 (E-mail: *jan.hannig@unc.edu*).
The authors are most grateful to the constructive comments from the reviewers, which have led to a much improved version of the article. They also thank the associate editor and editor for their editorial efforts in handling this article. Finally, they also thank Professors Jianqing Fan and Ning Hao for sharing the housing price appreciation dataset. The work of Hannig was supported in part by the National Science Foundation under Grants 1007543 and 1016441. The work of Lee was supported in part by the National Science Foundation under Grants 1007520, 1209226 and 1209232.

Journal of the American Statistical Association
June 2015, Vol. 110, No. 510, Theory and Methods
DOI: 10.1080/01621459.2014.931237

(including σ) in the final selected model. This procedure automatically accounts for the variability introduced by model selection. To the best of our knowledge, this is the first time that Fisher's fiducial idea is being applied to the so-called "large p small n" problem.

Fisher (1930) introduced fiducial inference to define a statistically meaningful distribution on the parameter space in cases when one cannot use a Bayes' theorem due to the lack of prior information. While never formally defined, fiducial inference has a long and storied history. We refer an interested reader to Hannig (2009) and Salome (1998) where a wealth of references can be found.

Ideas related to fiducial inference have experienced an exciting resurgence in the last decade. Some of these modern ideas are Dempster–Shafer calculus and its generalizations (Dempster 2008; Martin, Zhang, and Liu 2010; Zhang and Liu 2011; Martin and Liu 2013), confidence distributions (Singh, Xie, and Strawderman 2005; Xie, Singh, and Strawderman 2011; Xie and Singh 2013), generalized inference (Weerahandi 1993, 1995), and reference priors in objective Bayesian inference (Berger, Bernardo, and Sun 2009). There has also been a wealth of successful applications of these methods to practical problems. For selected examples, see McNally, Iyer, and Mathew (2003), Wang and Iyer (2005), Lidong, Hannig, and Iyer (2008), Edlefsen, Liu, and Dempster (2009), Hannig and Lee (2009), and Cisewski and Hannig (2012).

The particular variant of Fisher's fiducial idea that this article considers is the so-called *generalized fiducial inference*. Some early ideas were developed by Hannig, Iyer, and Patterson (2006), and later Hannig (2009) used these ideas to formally define a *generalized fiducial distribution*. A brief description of generalized fiducial inference is given below. The word "fiducial" often brings up disagreements about foundations of statistics. To keep the article focused, we choose not to discuss philosophical issues in detail. We only remark that our outlook is purely frequentist; we view generalized fiducial inference as a tool for proposing statistical procedures that then are evaluated on their own merits. Due to the particular structure of our model, this article also brings a contribution to the objective Bayesian model selection discussion. We discuss this issue in detail at the end of Appendix A.

The rest of this article is organized as follows. Section 2 provides some background material on generalized fiducial inference, and applies the methodology to the ultrahigh-dimensional regression problem. The theoretical properties of the proposed solution are examined in Section 3, while its empirical properties are illustrated in Section 4. Finally, concluding remarks are offered in Section 5 and technical details are delayed to the Appendices.

2. METHODOLOGY

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

$$\boldsymbol{Y} = \boldsymbol{G}(\boldsymbol{U}, \boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{G}(\cdot, \cdot)$ is sometimes known as the *structural equation*, and $\boldsymbol{U}$ is the random component of the relation whose distribution is *completely known*, for example, a vector of iid U(0,1)'s. Recall that in the definition of the celebrated maximum likelihood estimator, Fisher "switched" the roles of $\boldsymbol{Y}$ and $\boldsymbol{\theta}$: the random $\boldsymbol{Y}$ is treated as fixed in the likelihood function, while the deterministic $\boldsymbol{\theta}$ is treated as variable. Through (1), generalized fiducial inference uses this "switching principle" to define a valid probability distribution on $\boldsymbol{\theta}$.

This switching principle proceeds as follows. For the moment suppose for any given realization $\boldsymbol{y}$ of $\boldsymbol{y}$, the inverse

$$\boldsymbol{\theta} = \tilde{\boldsymbol{G}}^{-1}(\boldsymbol{y}, \boldsymbol{u}) \tag{2}$$

always exists for any realization $\boldsymbol{u}$ of $\boldsymbol{U}$. Since the distribution of $\boldsymbol{U}$ is assumed known, one can always generate a random sample $\tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{u}}_2, \ldots$, and via (2) a random sample of $\boldsymbol{\theta}$ can be obtained by $\tilde{\boldsymbol{\theta}}_1 = \tilde{\boldsymbol{G}}^{-1}(\boldsymbol{y}, \tilde{\boldsymbol{u}}_1), \tilde{\boldsymbol{\theta}}_2 = \tilde{\boldsymbol{G}}^{-1}(\boldsymbol{y}, \tilde{\boldsymbol{u}}_2), \ldots$ This is called a fiducial sample of $\boldsymbol{\theta}$, which can be used to calculate estimates and construct confidence intervals for $\boldsymbol{\theta}$ in a similar fashion as with a bootstrap sample, the CD-random variable described in Xie and Singh (2013). This process is also similar to obtaining credible sets from a Bayesian posterior sample. Through the above switching and the inverse operations, one can see that a density function $r(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ is implicitly defined. We term $r(\boldsymbol{\theta})$ the *generalized fiducial density* for $\boldsymbol{\theta}$, and the corresponding distribution the *generalized fiducial distribution* for $\boldsymbol{\theta}$. An illustrative example of applying this idea to simple linear regression can be found in Hannig and Lee (2009), and a formal mathematical definition of generalized fiducial inference is described in detail in Hannig (2009). The latter work also provides strategies to ensure the existence of the inverse (2).

Observe that for the ultrahigh-dimensional regression problem that this article considers, $\boldsymbol{\theta}$ can be decomposed into three components: $\boldsymbol{\theta} = \{M, \sigma, \boldsymbol{\beta}_M\}$, where M denotes a candidate model and can be seen as a sequence of p binary variables indicating which predictors have nonzero coefficients, σ is the noise standard deviation, and $\boldsymbol{\beta}_M$ is the parameter values of the nonzero coefficients predictors. In the next section, we derive the generalized fiducial density $r(M)$ for M, and then we will demonstrate how to generate a fiducial sample $\{\tilde{M}, \tilde{\sigma}, \tilde{\boldsymbol{\beta}}\}$ using $r(M)$.

2.1 Generalized Fiducial Density for Ultrahigh-Dimensional Regression

While the above formal definition of generalized fiducial inference is conceptually simple and very general, it may not be easily applicable in some practical situations. When the model dimension is known, Hannig (2013) derived a workable formula for $r(\boldsymbol{\theta})$ for many practical situations. Assume that the parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^m$ is m-dimensional and that the inverse $\boldsymbol{G}^{-1}(\boldsymbol{y}, \boldsymbol{\theta}) = \boldsymbol{u}$ to (1) exists. This assumption is satisfied for many natural structural equations, provided that $\boldsymbol{y}$ and $\boldsymbol{u}$ have the same dimension and $\boldsymbol{G}$ is smooth. Note that this inverse is different from the inverse $\tilde{\boldsymbol{G}}^{-1}$ in (2). Then under some differentiability assumptions, Hannig (2013) showed that the generalized fiducial distribution is absolutely continuous with density

$$r(\boldsymbol{\theta}) = \frac{f(\boldsymbol{y}, \boldsymbol{\theta}) J(\boldsymbol{y}, \boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} f(\boldsymbol{y}, \boldsymbol{\theta}') J(\boldsymbol{y}, \boldsymbol{\theta}') \, d\boldsymbol{\theta}'}, \tag{3}$$

where

$$J(\boldsymbol{y},\boldsymbol{\theta}) = \sum_{\substack{\boldsymbol{i}=(i_1,\ldots,i_m)\\ 1\le i_1<\cdots<i_m\le n}} \left|\det\left[\left\{\frac{d}{d\boldsymbol{y}}\boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})\right\}^{-1} \times \frac{d}{d(\boldsymbol{\theta},\boldsymbol{y}_{\boldsymbol{i}}^{\complement})}\boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})\right]\right|. \quad (4)$$

In the above, $f(\boldsymbol{y},\boldsymbol{\theta})$ is the likelihood and the sum goes over all m-tuples of indices $\boldsymbol{i} = (1 \le i_1 < \cdots < i_m \le n) \subset \{1,\ldots,n\}$. Also, for each $\boldsymbol{i}$, we denoted the list of unused indices by $\boldsymbol{i}^{\complement} = \{1,\ldots,n\} \setminus \boldsymbol{i}$, the collection of variables indexed by $\boldsymbol{i}$ by $\boldsymbol{y}_{\boldsymbol{i}} = (y_{i_1},\ldots,y_{i_m})$, and its complement by $\boldsymbol{y}_{\boldsymbol{i}}^{\complement} = (y_i : i \in \boldsymbol{i}^{\complement})$. The formula $\frac{d}{d(\boldsymbol{\theta},\boldsymbol{y}_{\boldsymbol{i}}^{\complement})}\boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})$ stood for the Jacobian matrix computed with respect to all parameters $\boldsymbol{\theta}$ and the observations $\boldsymbol{y}_{\boldsymbol{i}^{\complement}}$. Similarly $\frac{d}{d\boldsymbol{y}}\boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})$ stood for the Jacobian matrix computed with respect to the observations $\boldsymbol{y}$.

Recall that the formula (3) was derived for situations where the model dimension is known, and hence it cannot be directly applied to the current problem. When model selection is required, Hannig and Lee (2009) proposed adding extra penalty structural equations to (3). This is similar to adding a penalty term to the likelihood function to account for model complexity. In particular, their derivation shows that the fiducial probability of each candidate model M is proportional to

$$r(M) \propto \int_{\Theta} f_M(\boldsymbol{y},\boldsymbol{\theta}) J_M(\boldsymbol{y},\boldsymbol{\theta})\, d\boldsymbol{\theta}\, e^{-q(M)}, \quad (5)$$

where $f_M(\boldsymbol{y},\boldsymbol{\theta})$ is the likelihood, $J_M(\boldsymbol{y},\boldsymbol{\theta})$ is the Jacobian (4), and $q(M)$ is the penalty associated with the model M. In the context of wavelet regression, they recommended using the minimum description length (MDL) principle (Rissanen 1989, 2007) to derive the penalty $q(M)$, which is shown to possess attractive theoretical and empirical properties.

Given the success of Hannig and Lee (2009), we also attempted to use the MDL principle to derive a penalty $q(M)$ for the current problem, which gives $q(M) = 0.5|M| \log n$ with $|M|$ being the number of nonzero parameters in M. However, this form of $q(M)$ fails here, as the classical MDL principle was not designed to handle ultrahigh-dimensional problems. In what follows, we denote the size of the full model by p and the number of nonzero coefficients of the true model by d. Typically, $p \gg n \gg d$. We will also sometimes use $m = |M|$ to denote the dimension of a candidate model.

In Appendix A, we rederive the penalized fiducial distribution in the present ultrahigh-dimensional setup. Denote the residual sum of squares of M as RSS_M, when the corresponding $\boldsymbol{\beta}$ is estimated with maximum likelihood. It is shown in Appendix A that the fiducial probability for model M of dimension $m = |M| < n$ is

$$r(M) \propto \Gamma\left(\frac{n-|M|}{2}\right)(\pi \mathrm{RSS}_M)^{-\frac{n-|M|-1}{2}} n^{-\frac{|M|+1}{2}} \binom{p}{|M|}^{-\gamma}. \quad (6)$$

In particular, the penalty used is

$$q(M) = \frac{|M|}{2}\log n + \log_{e^{1/\gamma}}\binom{p}{|M|}, \quad (7)$$

where γ is a tuning parameter. The most natural choice is $\gamma = 1$ but we allow other choices. In all our numerical work, we use $\gamma = 1$. We note that the second term of (7) is similar to the extended BIC (EBIC) penalty of Chen and Chen (2008). In sequel we write $r(M)$ as $r_\gamma(M)$.

2.2 Practical Generation of Fiducial Sample

In this section, we propose a practical procedure for generating a fiducial sample $\{\tilde{M}, \tilde{\sigma}, \tilde{\boldsymbol{\beta}}\}$ using (6). First note that even for a moderate p, the total number of models 2^p is huge and hence any method that is exhaustive in nature is computationally not feasible. Moreover, the current generalized fiducial paradigm can only assign meaningful probabilities to models of dimension smaller than sample size. Bigger models can usually fit the data perfectly, which means that the penalized fiducial distribution will differ only due to the penalty term.

The proposed procedure therefore begins with constructing a class of candidate models, denoted as $\mathcal{M}'$. This $\mathcal{M}'$ should satisfy the following two properties: the number of models in $\mathcal{M}'$ is small and it contains the true model and models that have nonnegligible values of $r_\gamma(M)$. To construct $\mathcal{M}'$, we first apply the SIS procedure of Fan and Lv (2008) to reduce the number of predictors from p to p', where p' is of order $O(n)$. To further reduce the number of possible models (which is $2^{p'}$), we apply LASSO to those p' predictors that survived SIS, and take all those models that lie on the LASSO solution path as $\mathcal{M}'$. Note that the LASSO solution path can be quickly obtained via the least angle regression method (Efron et al. 2004), and that constructing $\mathcal{M}'$ in this way will ensure the true model is captured in $\mathcal{M}'$ with high probability (Fan and Lv 2008).

Once $\mathcal{M}'$ is obtained, for each $M \in \mathcal{M}'$, calculate

$$R_\gamma(M) = \Gamma\left(\frac{n-|M|}{2}\right)(\pi \mathrm{RSS}_M)^{-\frac{n-|M|-1}{2}} n^{-\frac{|M|+1}{2}} \binom{p}{|M|}^{-\gamma},$$

and approximate the generalized fiducial probability (6) by

$$\hat{r}_\gamma(M) = R_\gamma(M) / \sum_{M' \in \mathcal{M}'} R_\gamma(M'), \quad \text{for } M \in \mathcal{M}'. \quad (8)$$

Next for σ and $\boldsymbol{\beta}_M$. For any given M, it is straightforward to show that the generalized fiducial distribution of σ conditional on M is

$$\mathrm{RSS}_M/\sigma^2 \sim \chi^2(n-|M|) \quad (9)$$

and that of $\boldsymbol{\beta}_M$ conditional on M and σ is

$$\boldsymbol{\beta}_M \sim N\left(\boldsymbol{\beta}_M^{\mathrm{ML}}, \sigma^2\left(\boldsymbol{X}_M^T \boldsymbol{X}_M\right)^{-1}\right), \quad (10)$$

where $\boldsymbol{\beta}_M^{\mathrm{ML}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}_M$ for model M, and $\boldsymbol{X}_M$ is the design matrix for model M.

Thus to generate $\{\tilde{M}, \tilde{\sigma}, \tilde{\boldsymbol{\beta}}\}$, we first draw a model $\tilde{M} \in \mathcal{M}'$ from (8), then $\tilde{\sigma}$ from (9) given $\tilde{M}$, and finally, $\tilde{\boldsymbol{\beta}}$ from (10) given $\{\tilde{M}, \tilde{\sigma}\}$.

2.3 Point Estimates and Confidence Intervals

Applying the above procedure repeatedly, one can obtain multiple copies of $\{\tilde{M}, \tilde{\sigma}, \tilde{\boldsymbol{\beta}}\}$ that form a fiducial sample for $\{M, \sigma, \boldsymbol{\beta}_M\}$. This fiducial sample can be used to form estimates and confidence intervals for σ in a similar manner as with a Bayesian posterior sample. For example, the average of all $\tilde{\sigma}$'s

can be used as an estimate of σ, while the 2.5% smallest and 2.5% largest $\tilde{\sigma}$ values can be used, respectively, as the lower and upper limits for a 95% confidence interval for σ.

Obtaining estimates and confidence intervals for $\boldsymbol{\beta}$ is, however, less straightforward. It is because for any β_j, it is possible that it is included in some but not all $\tilde{M}$'s. In other words, some of the generated fiducial values for β_j are zeros, some are not.

We use the following simple procedure to deal with this issue. For each β_j, we count the percentage of zero fiducial sample values. If it is more than 50%, we declare that the value of this particular β_j is zero. Otherwise, we treat β_j as a nonzero parameter, and use all the nonzero fiducial sample values to obtain estimates and confidence intervals for it, in the same way as for σ. Although a similar idea has been used by Barbieri and Berger (2004) to determine the significance of a parameter in the Bayesian context, we note that constructing confidence intervals for β_j's in this manner may not lead to precise results.

3. THEORETICAL PROPERTIES

This section investigates the theoretical properties of the above-proposed generalized fiducial-based method, under the situation that p is diverging and the size of true model is either fixed or diverging. For similar results in the classical situations where p is fixed, see Hannig (2009, 2013).

First, some notations. Let M be any model, M_0 be the true model, and $\boldsymbol{H}_M$ be the projection matrix of $\boldsymbol{X}_M$, that is, $\boldsymbol{H}_M = \boldsymbol{X}_M(\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^T$. Define

$$\Delta_M = ||\boldsymbol{\mu} - \boldsymbol{H}_M\boldsymbol{\mu}||^2,$$

where $\boldsymbol{\mu} = E(\boldsymbol{Y}) = \boldsymbol{X}_{M_0}\boldsymbol{\beta}_{M_0}$. Throughout this section, we assume the following identifiability condition holds:

$$\lim_{n\to\infty} \min\left\{\frac{\Delta_M}{|M_0|\log p} : M_0 \not\subset M, |M| \le k|M_0|\right\} = \infty \quad (11)$$

for some fixed $k > 1$. This condition ensures that the true model is identifiable and has been used, for example, by Luo and Chen (2013). Condition (11) is closely related to the sparse Riesz condition (Zhang and Huang 2008). The sparse Riesz condition states that the eigenvalues of $X_M^TX_M/n$ are bounded away from 0 and ∞ uniformly for all models, that is,

$$0 < c_1 < \lambda_{\min}\left(\frac{1}{n}X_M^TX_M\right) \le \lambda_{\max}\left(\frac{1}{n}X_M^TX_M\right) < c_2 < \infty, \text{ for all } M \in \mathcal{M}, \quad (12)$$

where c_1, c_2 are constants, $\lambda_{\min}$ and $\lambda_{\min}$ are the smallest and largest eigenvalues, respectively, and $\mathcal{M}$, the class of candidate models, is defined in next paragraph. It can be shown that, under the sparse Riesz condition (12) and the minimum signal condition

$$\sqrt{\frac{n}{|M_0|\log p}}\min\{|\beta_j|; j \in M_0\} \to \infty, \quad (13)$$

the identifiability condition (11) holds. However, the converse does not hold in general.

Let $\mathcal{M}$ be the collection of models such that $\mathcal{M} = \{M : |M| \le k|M_0|\}$ for some fixed k. The restriction $|M| \le k|M_0|$ is imposed because in practice we only consider models with size comparable with the true model.

If p is large, the size of $\mathcal{M}$ could still be too large in practice. In this situation, we could use a variable screening procedure to reduce the size. This variable screening procedure should result in a class of candidate models $\mathcal{M}'$, which satisfies

$$P(M_0 \in \mathcal{M}') \to 1 \quad \text{and} \quad \log(m_j') = o(j\log n), \quad (14)$$

where $\mathcal{M}_j'$ contains all models in $\mathcal{M}'$ that are of size j, and m_j' is the number of models in $\mathcal{M}_j'$. The first condition in (14) guarantees the model class contains the true model, at least asymptotically. The second condition in (14) ensures that the size of the model class is not too large. These two conditions are satisfied by the practical algorithm presented in Section 2.2.

In Appendix B, the following theorem is established.

Theorem 3.1. Under (11), as $n \to \infty$, $p \to \infty$, $|M_0|\log(p) = o(n)$, $\log(|M_0|)/\log(p) \to \delta$, and $\log(n)/\log(p) \to \eta$, then there exists $\gamma > \frac{1+\delta}{1-\delta} - \frac{3\eta}{2(1-\delta)}$ such that

$$\max_{M\neq M_0, M\in\mathcal{M}} r_\gamma(M)/r_\gamma(M_0) = \max_{M\neq M_0, M\in\mathcal{M}} R_\gamma(M)/R_\gamma(M_0) \xrightarrow{P} 0. \quad (15)$$

Furthermore, if (14) holds, with the same γ,

$$\hat{r}_\gamma(M) = R_\gamma(M_0)/\sum_{M\in\mathcal{M}'} R_\gamma(M) \xrightarrow{P} 1. \quad (16)$$

Equation (15) states that the true model has the highest generalized fiducial probability among all the models in $\mathcal{M}$. However, it does not imply Equation (16) in general because the class of candidate models can be very large. If we constrain the class of models being considered in such a way that (14) holds, then Equation (16) states that, with probability tending to 1, the true model will be selected. If $\gamma = 1$, as in our simulation study, Theorem 3.1 provides a relationship between the rates of growth for $d = |M_0|$, n, and p. From Theorem 3.1 and Bernstein–von Mises theorem for generalized fiducial distribution (Hannig 2009; Sonderegger and Hannig 2014), one can conclude the following important corollary.

Corollary 3.1. One sided and equal tailed confidence intervals constructed using the generalized fiducial density (6) will have correct asymptotic coverage. Consequently, the generalized fiducial distribution and derived point estimators are consistent.

Remark 3.1. Theorem 3.1 and its corollary show that statistical inference based on (6) will have correct asymptotic frequentist property under the assumptions of Theorem 3.1 and conditions (14). Particularly, it is mentioned that conditions (14) are satisfied by the practical algorithm presented in Section 2.2. The rest of this remark discusses these conditions in more details and comment on their impacts to the practical situations. To begin with, recall there are two steps to construct the class of candidate models. The first step is to apply the SIS procedure of Fan and Lv (2008). This SIS procedure only selects predictors with high marginal correlations with the response variable and reduces the number of predictors to the order $p' = O(n)$. We used $\mathcal{M}$ to denote the class of all possible models after applying SIS (and there are $2^{p'}$ models in $\mathcal{M}$). Under fairly weak assumptions, Fan and Lv (2008) (Theorem 1) showed the probability that the true

Table 1. Bias of the various estimates of σ^2

		$(n, p, d) =$ (200, 2000, 3)	$(n, p, d) =$ (300, 8000, 5)	$(n, p, d) =$ (500, 50,000, 8)
$b = 1/\sqrt{d}$	Proposed	−0.180 (0.323)	−0.166 (0.271)	0.230 (0.219)
$\rho = 0$	RCV	1.507 (0.488)	−16.749 (0.330)	−27.287 (0.221)
	Oracle	−0.018 (0.317)	−0.115 (0.263)	−0.031 (0.200)
$b = 2/\sqrt{d}$	Proposed	−0.511 (0.327)	−0.455 (0.259)	−0.089 (0.202)
$\rho = 0$	RCV	−0.297 (0.465)	−7.932 (0.353)	−13.909 (0.255)
	Oracle	−0.383 (0.321)	−0.474 (0.260)	−0.151 (0.200)
$b = 3/\sqrt{d}$	Proposed	−0.457 (0.332)	−0.112 (0.256)	0.103 (0.203)
$\rho = 0$	RCV	−0.495 (0.451)	−4.303 (0.362)	−7.245 (0.286)
	Oracle	−0.316 (0.328)	−0.283 (0.254)	−0.021 (0.201)
$b = 1/\sqrt{d}$	Proposed	0.352 (0.335)	0.271 (0.285)	1.046 (0.227)
$\rho = 0.5$	RCV	0.455 (0.467)	−10.333 (0.334)	−17.287 (0.247)
	Oracle	0.367 (0.329)	−0.548 (0.258)	−0.406 (0.205)
$b = 2/\sqrt{d}$	Proposed	−0.505 (0.328)	−0.092 (0.263)	−0.302 (0.199)
$\rho = 0.5$	RCV	−0.533 (0.442)	−3.046 (0.357)	−6.73 (0.257)
	Oracle	−0.103 (0.325)	−0.160 (0.261)	−0.483 (0.198)
$b = 3/\sqrt{d}$	Proposed	−1.585 (0.304)	0.135 (0.259)	−0.080 (0.198)
$\rho = 0.5$	RCV	−1.404 (0.430)	−2.275 (0.342)	−3.279 (0.274)
	Oracle	−1.251 (0.302)	−0.188 (0.258)	−0.355 (0.197)

NOTES: Numbers in parentheses are standard errors. All numbers are multiplied by 100.

model will be in $\mathcal{M}$ is

$$P(M_0 \in \mathcal{M}) = 1 - O[\exp\{-Cn^{1-2\kappa}\}/\log n], \qquad (17)$$

where $C > 0$ and κ is a constant depending on a minimum signal condition similar to condition (13).

The second step is to apply LASSO on the p' remained predictors on $\mathcal{M}$ and construct the model class $\mathcal{M}'$ from the models that lie on the LASSO solution path. From this, it is easily seen that the second condition of (14) is satisfied and it remains to verify the validity of the first condition. Theoretical properties of LASSO have been well studied in the literature (e.g., Meinshausen and Buhlmann 2006; Zhao and Yu 2006; Huang, Ma, and Zhang 2008; Zhang and Huang 2008). In Zhao and Yu (2006), the authors defined the notion of irrepresentable condition, an important requirement of the relevant and irrelevant predictors for the consistency of LASSO. Under this condition together with some other weak assumptions, the authors showed that LASSO is model selection consistent. For our purpose, we only need a weaker result of their Theorem 4, namely,

$$P(M_0 \in \mathcal{M}' | M_0 \in \mathcal{M}) \geq 1 - o\left(e^{-n^c}\right), \qquad (18)$$

where $c > 0$ is again a constant depending on a minimum signal condition similar to condition (13). And (18) is sufficient for our purpose because we only require the true model lies in the LASSO solution path but not necessarily to be selected by LASSO procedure.

Combining the results of (17) and (18), it is sufficient to conclude the validity of the first condition of (14). Moreover, the probability that the true model would be in the set of candidate models goes to 1 exponentially fast. It suggests that irrelevant models (i.e., models that do not contain the true model) would only affect the fiducial probability function in an exponentially small size. This observation helps to fill the gap between the asymptotic theory and the practical performance of our purposed methodology.

Remark 3.2. One of the accepted techniques to address the issue of inference after model selection is model averaging (e.g., Dong 2007). Our method could be viewed as a new proposal for frequentist model averaging.

As a first step, our method uses SIS-LASSO screening to select a set of candidate models. As shown above, this step will not have a big impact on the inference. Indeed Shen, Huang, and Ye (2004) stated:

> It is worth noticing that inference after model selection may be less problematic for certain model selection routines. For instance, LASSO (Tibshirani 1996) and SCAD (Fan and Li 2001; Fan and Peng 2004) only need to take into account modeling bias for an estimated tuning parameter, whose modeling variability is expected to be less than that of selecting variables via CV.

Since we do not select a single value of the tuning parameter, but instead are using model averaging over the models associated with a range of tuning parameters, we argue that the uncertainty due to model selection is considered. Our simulation study results support this assertion.

Remark 3.3. When comparing to Leeb and Pötscher (2008), our theoretical results are addressing a simpler problem of a pointwise asymptotic. Investigation of uniform asymptotics, like in Pötscher and Leeb (2009) who showed that the LASSO and SCAD estimates are uniformly consistent at a rate different from $n^{1/2}$, for the fiducial model averaged confidence sets is an interesting direction for future work.

Table 2. Empirical coverage rates for various confidence intervals for σ^2

			90%	95%	99%
	$b = 1/\sqrt{3}$	Proposed	0.895 (0.338)	0.949 (0.405)	0.985 (0.537)
	$\rho = 0$	Oracle	0.896 (0.336)	0.948 (0.402)	0.985 (0.534)
	$b = 2/\sqrt{3}$	Proposed	0.892 (0.337)	0.937 (0.404)	0.987 (0.535)
	$\rho = 0$	Oracle	0.892 (0.335)	0.941 (0.401)	0.988 (0.532)
	$b = 3/\sqrt{3}$	Proposed	0.884 (0.338)	0.941 (0.404)	0.986 (0.536)
$(n, p, d) = (200, 2000, 3)$	$\rho = 0$	Oracle	0.886 (0.335)	0.943 (0.401)	0.986 (0.533)
	$b = 1/\sqrt{3}$	Proposed	0.895 (0.344)	0.945 (0.412)	0.988 (0.547)
	$\rho = 0.5$	Oracle	0.896 (0.338)	0.946 (0.404)	0.988 (0.536)
	$b = 2/\sqrt{3}$	Proposed	0.889 (0.339)	0.939 (0.405)	0.991 (0.538)
	$\rho = 0.5$	Oracle	0.891 (0.336)	0.94 (0.402)	0.991 (0.534)
	$b = 3/\sqrt{3}$	Proposed	0.906 (0.335)	0.955 (0.401)	0.993 (0.532)
	$\rho = 0.5$	Oracle	0.908 (0.332)	0.957 (0.397)	0.992 (0.528)
	$b = 1/\sqrt{5}$	Proposed	0.891 (0.277)	0.948 (0.331)	0.985 (0.438)
	$\rho = 0$	Oracle	0.898 (0.273)	0.948 (0.326)	0.987 (0.432)
	$b = 2/\sqrt{5}$	Proposed	0.909 (0.275)	0.951 (0.328)	0.987 (0.434)
	$\rho = 0$	Oracle	0.904 (0.272)	0.95 (0.325)	0.985 (0.43)
	$b = 3/\sqrt{5}$	Proposed	0.913 (0.274)	0.953 (0.328)	0.993 (0.433)
$(n, p, d) = (300, 8000, 5)$	$\rho = 0$	Oracle	0.907 (0.273)	0.955 (0.326)	0.993 (0.431)
	$b = 1/\sqrt{5}$	Proposed	0.887 (0.286)	0.936 (0.342)	0.984 (0.453)
	$\rho = 0.5$	Oracle	0.898 (0.272)	0.948 (0.325)	0.992 (0.43)
	$b = 2/\sqrt{5}$	Proposed	0.894 (0.275)	0.947 (0.328)	0.99 (0.434)
	$\rho = 0.5$	Oracle	0.893 (0.273)	0.946 (0.326)	0.992 (0.432)
	$b = 3/\sqrt{5}$	Proposed	0.906 (0.274)	0.954 (0.328)	0.99 (0.433)
	$\rho = 0.5$	Oracle	0.906 (0.273)	0.952 (0.326)	0.99 (0.432)
	$b = 1/\sqrt{8}$	Proposed	0.88 (0.215)	0.939 (0.257)	0.989 (0.339)
	$\rho = 0$	Oracle	0.909 (0.211)	0.952 (0.252)	0.99 (0.332)
	$b = 2/\sqrt{8}$	Proposed	0.898 (0.212)	0.942 (0.253)	0.991 (0.333)
	$\rho = 0$	Oracle	0.899 (0.211)	0.942 (0.251)	0.991 (0.332)
	$b = 3/\sqrt{8}$	Proposed	0.901 (0.212)	0.952 (0.253)	0.991 (0.333)
	$\rho = 0$	Oracle	0.9 (0.211)	0.953 (0.252)	0.992 (0.332)
$(n, p, d) = (500, 50000, 8)$	$b = 1/\sqrt{8}$	Proposed	0.865 (0.224)	0.935 (0.267)	0.985 (0.352)
	$\rho = 0.5$	Oracle	0.9 (0.21)	0.94 (0.251)	0.99 (0.331)
	$b = 2/\sqrt{8}$	Proposed	0.895 (0.211)	0.95 (0.252)	0.993 (0.332)
	$\rho = 0.5$	Oracle	0.895 (0.21)	0.949 (0.251)	0.992 (0.331)
	$b = 3/\sqrt{8}$	Proposed	0.905 (0.211)	0.947 (0.251)	0.989 (0.331)
	$\rho = 0.5$	Oracle	0.903 (0.21)	0.945 (0.251)	0.99 (0.331)

NOTE: Numbers in parentheses are averaged widths of the confidence intervals.

4. FINITE SAMPLE PROPERTIES

4.1 Simulations

A simulation study was conducted to evaluate the practical performance of the proposed procedure. The following model from Fan, Guo, and Hao (2012) was used to generate the noisy data

$$Y = b(X_1 + \cdots + X_d) + \epsilon,$$

where ϵ is iid standard normal error, d is the number of predictors with nonzero coefficients, and the value b controls the signal-to-noise ratio. All the covariates are standard normal variables with correlation $\text{cor}(X_i, X_j) = \rho^{|i-j|}$. Three combinations of (n, p, d) were used: (200, 2000, 3), (300, 8000, 5), and (500, 50,000, 8). For each of these three combinations, three choices of b and two choices of ρ were used: $b = 1/\sqrt{d}, 2/\sqrt{d}$, and $3/\sqrt{d}$, and $\rho = 0$ and 0.5. Therefore, a total of $3 \times 3 \times 2 = 18$ experimental configurations were considered. The number of repetitions for each experimental configuration was 1000. For $\rho = 0$, the cases $b = 1/\sqrt{d}, 2/\sqrt{d}$, and $3/\sqrt{d}$ correspond to the cases when the signal-to-noise ratios are 1, 2, and 3, respectively.

For each generated dataset, we applied the proposed generalized fiducial procedure described in Section 2.2 to obtain a fiducial sample of size 10,000 for $\{M, \sigma, \boldsymbol{\beta}\}$, and from this we computed the generalized fiducial estimate for σ^2. We also obtained two other estimates for σ^2: the first one from the refitted cross-validation (RCV) method of Fan, Guo, and Hao (2012), while the second one is the classical maximum likelihood estimate for σ^2 obtained from the true model. Of course, the last estimate cannot be obtained in practice, but it is computed here for benchmark comparisons. In consequence, it is termed as the oracle estimate. Also, for RCV, the particular version we compared with is RCV-LASSO.

The bias of these three estimates for σ^2 is summarized in Table 1. From this table, one can see that the bias of the fiducial estimates is usually not much larger than the bias from the

Table 3. Empirical coverage rates for the confidence intervals for β_1

			90%	95%	99%
	$b = 1/\sqrt{3}$	Proposed	0.888 (0.236)	0.946 (0.283)	0.987 (0.377)
	$\rho = 0$	RCV	0.869 (0.250)	0.915 (0.298)	0.956 (0.392)
		Oracle	0.897 (0.235)	0.946 (0.279)	0.988 (0.367)
	$b = 2/\sqrt{3}$	Proposed	0.884 (0.235)	0.948 (0.282)	0.991 (0.376)
	$\rho = 0$	RCV	0.887 (0.238)	0.945 (0.284)	0.988 (0.373)
		Oracle	0.889 (0.234)	0.946 (0.279)	0.990 (0.367)
	$b = 3/\sqrt{3}$	Proposed	0.892 (0.236)	0.947 (0.282)	0.987 (0.376)
	$\rho = 0$	RCV	0.896 (0.238)	0.95 (0.284)	0.99 (0.373)
$(n, p, d) = (200, 2000, 3)$		Oracle	0.897 (0.234)	0.952 (0.279)	0.987 (0.367)
	$b = 1/\sqrt{3}$	Proposed	0.886 (0.282)	0.936 (0.338)	0.985 (0.454)
	$\rho = 0.5$	RCV	0.814 (0.289)	0.849 (0.345)	0.902 (0.453)
		Oracle	0.894 (0.271)	0.943 (0.323)	0.988 (0.424)
	$b = 2/\sqrt{3}$	Proposed	0.898 (0.271)	0.944 (0.325)	0.987 (0.433)
	$\rho = 0.5$	RCV	0.903 (0.274)	0.945 (0.326)	0.988 (0.429)
		Oracle	0.894 (0.270)	0.949 (0.322)	0.986 (0.423)
	$b = 3/\sqrt{3}$	Proposed	0.901 (0.269)	0.948 (0.322)	0.989 (0.429)
	$\rho = 0.5$	RCV	0.899 (0.271)	0.953 (0.323)	0.988 (0.424)
		Oracle	0.897 (0.269)	0.955 (0.321)	0.99 (0.422)
	$b = 1/\sqrt{5}$	Proposed	0.810 (0.191)	0.896 (0.229)	0.976 (0.303)
	$\rho = 0$	RCV	0.903 (0.204)	0.935 (0.243)	0.956 (0.320)
		Oracle	0.900 (0.192)	0.948 (0.229)	0.992 (0.301)
	$b = 2/\sqrt{5}$	Proposed	0.871 (0.189)	0.936 (0.226)	0.984 (0.300)
	$\rho = 0$	RCV	0.897 (0.201)	0.936 (0.239)	0.981 (0.315)
		Oracle	0.907 (0.191)	0.959 (0.228)	0.989 (0.300)
	$b = 3/\sqrt{5}$	Proposed	0.888 (0.19)	0.934 (0.227)	0.984 (0.301)
	$\rho = 0$	RCV	0.900 (0.197)	0.945 (0.235)	0.979 (0.309)
$(n, p, d) = (300, 8000, 5)$		Oracle	0.879 (0.192)	0.941 (0.228)	0.991 (0.300)
	$b = 1/\sqrt{5}$	Proposed	0.812 (0.269)	0.887 (0.322)	0.963 (0.427)
	$\rho = 0.5$	RCV	0.871 (0.236)	0.915 (0.281)	0.960 (0.369)
		Oracle	0.912 (0.221)	0.954 (0.264)	0.992 (0.346)
	$b = 2/\sqrt{5}$	Proposed	0.895 (0.250)	0.949 (0.299)	0.989 (0.396)
	$\rho = 0.5$	RCV	0.864 (0.224)	0.922 (0.266)	0.975 (0.350)
		Oracle	0.891 (0.222)	0.950 (0.264)	0.991 (0.347)
	$b = 3/\sqrt{5}$	Proposed	0.908 (0.250)	0.950 (0.299)	0.990 (0.397)
	$\rho = 0.5$	RCV	0.852 (0.220)	0.917 (0.262)	0.975 (0.344)
		Oracle	0.904 (0.222)	0.949 (0.264)	0.983 (0.347)
	$b = 1/\sqrt{8}$	Proposed	0.781 (0.148)	0.875 (0.177)	0.978 (0.233)
	$\rho = 0$	RCV	0.813 (0.151)	0.857 (0.180)	0.884 (0.237)
		Oracle	0.910 (0.149)	0.954 (0.177)	0.993 (0.233)
	$b = 2/\sqrt{8}$	Proposed	0.853 (0.147)	0.919 (0.176)	0.980 (0.232)
	$\rho = 0$	RCV	0.804 (0.156)	0.878 (0.186)	0.965 (0.244)
		Oracle	0.902 (0.148)	0.947 (0.177)	0.988 (0.232)
	$b = 3/\sqrt{8}$	Proposed	0.873 (0.147)	0.925 (0.176)	0.986 (0.232)
	$\rho = 0$	RCV	0.841 (0.155)	0.911 (0.184)	0.981 (0.242)
$(n, p, d) = (500, 50{,}000, 8)$		Oracle	0.897 (0.149)	0.944 (0.177)	0.988 (0.233)
	$b = 1/\sqrt{8}$	Proposed	0.820 (0.206)	0.885 (0.246)	0.950 (0.324)
	$\rho = 0.5$	RCV	0.895 (0.179)	0.935 (0.213)	0.965 (0.280)
		Oracle	0.925 (0.172)	0.965 (0.204)	0.995 (0.269)
	$b = 2/\sqrt{8}$	Proposed	0.897 (0.193)	0.949 (0.230)	0.988 (0.304)
	$\rho = 0.5$	RCV	0.861 (0.169)	0.922 (0.202)	0.976 (0.265)
		Oracle	0.893 (0.171)	0.944 (0.204)	0.989 (0.268)
	$b = 3/\sqrt{8}$	Proposed	0.888 (0.193)	0.945 (0.230)	0.989 (0.304)
	$\rho = 0.5$	RCV	0.840 (0.168)	0.909 (0.201)	0.968 (0.264)
		Oracle	0.899 (0.171)	0.942 (0.204)	0.987 (0.268)

NOTE: The numbers in the parentheses are the averaged widths of the corresponding confidence intervals.

Table 4. Empirical coverage rates for the confidence intervals for $E(Y_i|\boldsymbol{x}_i)$

			90%	95%	99%
	$b = 1/\sqrt{3}$	Proposed	0.899 (0.421)	0.948 (0.511)	0.988 (0.696)
	$\rho = 0$	RCV	0.966 (1.160)	0.981 (1.382)	0.993 (1.817)
		Oracle	0.896 (0.343)	0.947 (0.409)	0.989 (0.538)
	$b = 2/\sqrt{3}$	Proposed	0.903 (0.424)	0.953 (0.516)	0.990 (0.704)
	$\rho = 0$	RCV	0.857 (0.603)	0.910 (0.718)	0.966 (0.944)
		Oracle	0.888 (0.342)	0.944 (0.408)	0.988 (0.536)
	$b = 3/\sqrt{3}$	Proposed	0.911 (0.428)	0.956 (0.519)	0.991 (0.709)
	$\rho = 0$	RCV	0.931 (0.605)	0.965 (0.720)	0.992 (0.947)
$(n, p, d) = (200, 2000, 3)$		Oracle	0.897 (0.343)	0.947 (0.409)	0.987 (0.537)
	$b = 1/\sqrt{3}$	Proposed	0.903 (0.452)	0.948 (0.549)	0.987 (0.748)
	$\rho = 0.5$	RCV	0.925 (1.281)	0.943 (1.526)	0.964 (2.005)
		Oracle	0.892 (0.344)	0.944 (0.410)	0.987 (0.538)
	$b = 2/\sqrt{3}$	Proposed	0.910 (0.444)	0.955 (0.538)	0.990 (0.733)
	$\rho = 0.5$	RCV	0.855 (0.583)	0.907 (0.695)	0.963 (0.914)
		Oracle	0.896 (0.343)	0.948 (0.408)	0.988 (0.536)
	$b = 3/\sqrt{3}$	Proposed	0.913 (0.438)	0.959 (0.532)	0.993 (0.725)
	$\rho = 0.5$	RCV	0.925 (0.492)	0.961 (0.587)	0.993 (0.771)
		Oracle	0.899 (0.342)	0.947 (0.408)	0.989 (0.536)
	$b = 1/\sqrt{5}$	Proposed	0.888 (0.444)	0.938 (0.536)	0.981 (0.725)
	$\rho = 0$	RCV	0.951 (1.864)	0.973 (2.221)	0.99 (2.919)
		Oracle	0.898 (0.388)	0.950 (0.462)	0.990 (0.607)
	$b = 2/\sqrt{5}$	Proposed	0.909 (0.439)	0.956 (0.531)	0.992 (0.724)
	$\rho = 0$	RCV	0.949 (1.291)	0.977 (1.538)	0.995 (2.022)
		Oracle	0.900 (0.386)	0.949 (0.46)	0.990 (0.605)
	$b = 3/\sqrt{5}$	Proposed	0.909 (0.429)	0.957 (0.519)	0.992 (0.708)
	$\rho = 0$	RCV	0.942 (0.915)	0.973 (1.090)	0.995 (1.432)
$(n, p, d) = (300, 8000, 5)$		Oracle	0.897 (0.387)	0.948 (0.461)	0.990 (0.606)
	$b = 1/\sqrt{5}$	Proposed	0.871 (0.496)	0.925 (0.602)	0.975 (0.820)
	$\rho = 0.5$	RCV	0.953 (1.641)	0.978 (1.956)	0.996 (2.570)
		Oracle	0.898 (0.387)	0.947 (0.461)	0.988 (0.606)
	$b = 2/\sqrt{5}$	Proposed	0.914 (0.437)	0.962 (0.531)	0.994 (0.728)
	$\rho = 0.5$	RCV	0.947 (0.741)	0.977 (0.883)	0.996 (1.160)
		Oracle	0.901 (0.387)	0.954 (0.461)	0.991 (0.606)
	$b = 3/\sqrt{5}$	Proposed	0.914 (0.422)	0.960 (0.512)	0.993 (0.701)
	$\rho = 0.5$	RCV	0.914 (0.431)	0.958 (0.514)	0.992 (0.676)
		Oracle	0.900 (0.388)	0.951 (0.462)	0.991 (0.607)
	$b = 1/\sqrt{8}$	Proposed	0.841 (0.445)	0.896 (0.534)	0.951 (0.711)
	$\rho = 0$	RCV	0.934 (1.889)	0.960 (2.251)	0.983 (2.958)
		Oracle	0.902 (0.409)	0.953 (0.488)	0.991 (0.641)
	$b = 2/\sqrt{8}$	Proposed	0.907 (0.435)	0.955 (0.522)	0.991 (0.697)
	$\rho = 0$	RCV	0.951 (1.573)	0.980 (1.874)	0.997 (2.463)
		Oracle	0.903 (0.409)	0.951 (0.487)	0.990 (0.640)
	$b = 3/\sqrt{8}$	Proposed	0.900 (0.429)	0.951 (0.515)	0.990 (0.687)
	$\rho = 0$	RCV	0.957 (1.187)	0.983 (1.415)	0.998 (1.860)
$(n, p, d) = (500, 50{,}000, 8)$		Oracle	0.898 (0.409)	0.949 (0.488)	0.989 (0.641)
	$b = 1/\sqrt{8}$	Proposed	0.829 (0.501)	0.892 (0.601)	0.958 (0.803)
	$\rho = 0.5$	RCV	0.945 (1.713)	0.978 (2.041)	0.996 (2.682)
		Oracle	0.905 (0.408)	0.951 (0.486)	0.992 (0.639)
	$b = 2/\sqrt{8}$	Proposed	0.907 (0.430)	0.956 (0.517)	0.993 (0.693)
	$\rho = 0.5$	RCV	0.951 (0.708)	0.979 (0.844)	0.997 (1.109)
		Oracle	0.900 (0.408)	0.951 (0.487)	0.992 (0.640)
	$b = 3/\sqrt{8}$	Proposed	0.903 (0.421)	0.953 (0.505)	0.991 (0.675)
	$\rho = 0.5$	RCV	0.900 (0.417)	0.949 (0.497)	0.990 (0.653)
		Oracle	0.898 (0.409)	0.949 (0.487)	0.990 (0.640)

NOTE: The numbers in the parentheses are the averaged widths of the corresponding confidence intervals.

oracle estimates. The RCV estimates sometimes have very large bias.

We also obtained two sets of 90%, 95%, and 99% confidence intervals for σ^2 from each simulated dataset. The first set was computed using the proposed generalized fiducial method, and the second was calculated by applying classical theory to the true model. Again, the last method cannot be used in practice, and is used for benchmark comparisons, that is, the oracle method. The empirical coverage rates of these confidence intervals are summarized in Table 2. It can be seen that the generalized fiducial confidence intervals are nearly as good as the oracle confidence intervals.

Finally, for each simulated dataset we applied three methods to compute the confidence intervals for the regression coefficients β_j's and the mean function $E(Y_i|\boldsymbol{x}_i)$ evaluated at 50 randomly selected design points $\boldsymbol{x}_i$'s. The three methods are the proposed generalized fiducial method, the RCV method of Fan, Guo, and Hao (2012), and the oracle method that uses the true model. As before the empirical coverage rates of these confidence intervals are calculated and they are reported in Tables 3 and 4. Note that only the confidence intervals for β_1 are reported, as the confidence intervals for other β_j's have similar coverage rates. Overall one can see that the generalized fiducial method gave quite reliable results, except for a few experimental settings where the confidence intervals were over-liberal.

In an attempt to produce a single summary statistic for comparing the empirical coverage rates of the confidence intervals produced by different methods, the following calculation has been done. For all the 90% generalized fiducial confidence intervals for β_1, we counted the number of times that their empirical coverage rates are within the range $(1-\alpha) \pm 1.96\sqrt{\alpha(1-\alpha)/N_{\text{sim}}}$, where $\alpha = 0.10$ and $N_{\text{sim}} = 1000$ is the number of repetitions performed for each experimental setting. Similar calculations were then performed for the 95% and 99% (i.e., $\alpha = 0.05$ and $\alpha = 0.01$) confidence intervals. We observed that for the proposed generalized fiducial method, out of the 54 empirical coverage rates, 33 of them are within their corresponding target ranges. We have also done the same calculations for the RCV and the oracle methods, and the numbers of their empirical coverage rates that are inside their target ranges are, respectively, 17 and 50. Finally, we repeated the same calculations for the empirical coverage rates for $E(Y_i|\boldsymbol{x}_i)$, and the corresponding numbers for the proposed, RCV, and oracle methods are, respectively, 44, 23, and 54. Of course, these numbers are not perfect for judging the relative merits of the different methods, but they seem to suggest that the proposed generalized fiducial method provides improvement over the RCV method.

4.2 Real Data Example: Housing Price Appreciation

This section analyzes a dataset that contains 119 months of housing price appreciation (HPA) of the national house price index (HPI) for 381 core-based statistical areas (CBSAs) in the united states. Here, HPA is defined as the percentage of monthly change in log-HPI for each of the 381 CBSAs. The goal of the analysis is to predict future HPA values for these CBSAs using existing data. This dataset was recorded from 1996 to 2005, and has been studied, for example, by Fan, Guo, and Hao (2012).

Of course, house prices depend on geographical locations and various macroeconomic factors. As argued by Fan, Guo, and Hao (2012), effects from macroeconomic factors can be well summarized by the national HPA. Let $X_{t,j}$ be the HPA of the jth CBSA in month t, and $X_{t,\text{N}}$ be the national HPA of month t. Then for any $k = 1, \ldots, 381$, a reasonable model for a 1 year ahead HPA prediction for the kth CBSA is

$$X_{t,k} = \sum_{j=1}^{381} \beta_j^{(k)} X_{t-1,j} + \beta_{\text{N}}^{(k)} X_{t-1,\text{N}} + \epsilon_{t-1},$$

where $\beta_j^{(k)}$'s and $\beta_{\text{N}}^{(k)}$ are model parameters and ϵ_{t-1} is an independent random error. Given the national HPA $X_{t-1,\text{N}}$, it is reasonable to assume that areas that are far away would have minimal influence on the local house prices, therefore one can assume the $\beta_j^{(k)}$'s are sparse. Note that for any given k, we have "$p > n$," as $p = 382$ and $n = 119$.

For illustrative purposes, we apply the proposed generalized fiducial procedure to the above model for one of the CBSAs: San Francisco–San Mateo–Redwood. Two fitted models with nonnegligible fiducial probabilities are returned: with probability 0.335 the housing appreciation of this area depends on itself and it is nearby CBSA San Jose–San Francisco–Oakland, while with probability about 0.663, it depends only on the CBSA San Jose–San Francisco–Oakland.

We also obtained estimate for the noise standard deviation σ, which can be interpreted as a measure of prediction accuracy when forecasting the housing appreciation. Our point estimate for σ is 0.56 with a 95% confidence as (0.48, 0.65). Our point estimate agrees with those reported in Fan, Guo, and Hao (2012), although no confidence intervals are reported there.

5. CONCLUSION

In this article, we studied the issue of uncertainty quantification in the ultrahigh-dimensional regression problem. We applied the generalized fiducial inference methodology to develop an inferential procedure for this problem. Our theoretical results show that estimates obtained by this procedure are consistent, while confidence intervals constructed by this procedure are asymptotically correct in the frequentist sense. Numerical results from simulation experiments confirm with these theoretical findings. To the best of our knowledge, there are very few published articles that are devoted to quantify uncertainties in the ultrahigh-dimensional regression problem, and hence the current article is one of the first to provide a systematic treatment to this problem. It also opens the possibility for using fiducial and related methods for conducting statistical inference for other "large p small n" problems, such as classification and covariance matrix estimation.

APPENDIX A: DERIVATION OF (6)

This appendix derives the generalized fiducial density (6). We first derive a simplified version of the Jacobian (4). Then we derive the fiducial density in our the ultrahigh-dimensional model obtaining penalty (7). At the end of this appendix, we discuss relationship of our solution with the objective Bayesian model selection.

A.1 Derivation of a Simpler General Jacobian Formula

First observe that the term $J(\boldsymbol{y}, \boldsymbol{\theta})$ in (4) can be further simplified. The product of Jacobian matrices in each of the summands of (4) simplifies to a matrix containing the m-columns of the $n \times d$ matrix $\{\frac{d}{d\boldsymbol{y}}\boldsymbol{G}^{-1}(\boldsymbol{y}, \boldsymbol{\theta})\}^{-1}\frac{d}{d\boldsymbol{\theta}}\boldsymbol{G}^{-1}(\boldsymbol{y}, \boldsymbol{\theta})$ and the $n - m$ columns of the identity matrix with columns $i_1, \ldots, i_m$ removed. Thus, we have

$$J(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{\substack{\boldsymbol{i} = (i_1, \ldots, i_m) \\ 1 \le i_1 < \cdots < i_m \le n}} \left|\det\left[\left\{\frac{\boldsymbol{d}}{\boldsymbol{d}\boldsymbol{y}}\boldsymbol{G}^{-1}(\boldsymbol{y}, \boldsymbol{\theta})\right\}^{-1} \times \frac{\boldsymbol{d}}{\boldsymbol{d}\boldsymbol{\theta}}\boldsymbol{G}^{-1}(\boldsymbol{y}, \boldsymbol{\theta})\right]_{\boldsymbol{i}}\right|,$$

where for any $n \times m$ matrix $\boldsymbol{A}$, the sub-matrix $(\boldsymbol{A})_{\boldsymbol{i}}$ is the $d \times d$ matrix containing the rows $i_1, \ldots, i_m$ of $\boldsymbol{A}$. Finally, using the implicit function theorem we obtain our final expression

$$J(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{\substack{\boldsymbol{i} = (i_1, \ldots, i_m) \\ 1 \le i_1 < \cdots < i_m \le n}} \left|\det\left[\frac{\boldsymbol{d}}{\boldsymbol{d}\boldsymbol{\theta}}\boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})\Big|_{\boldsymbol{u}=\boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})}\right]_{\boldsymbol{i}}\right|. \quad \text{(A.1)}$$

A.2 Rederivation of Penalized Fiducial Distribution

The most natural naive data-generating equation for this model is

$$\boldsymbol{y} = \boldsymbol{G}(M, \boldsymbol{\beta}_M, \sigma^2, \boldsymbol{Z}) = \boldsymbol{X}_M\boldsymbol{\beta}_M + \sigma\boldsymbol{Z},$$

where $\boldsymbol{y}$ is the observations, M is the model considered (collection of parameters that can be nonzero), $\boldsymbol{X}_M$ is the design matrix for model M, $\boldsymbol{\beta}_M \in \mathbb{R}^{|M|}$ and $\sigma > 0$ are parameters, and $\boldsymbol{Z}$ is a vector of iid standard normal random variables.

Hannig (2013) proposed to define a generalized fiducial distribution using discretization. The proposal can be expressed as a weak limit, as $\epsilon \to 0$, of the conditional distributions

$$\arg\min_{M,\boldsymbol{\beta}_M,\sigma^2} \|\boldsymbol{y} - \boldsymbol{G}(M, \boldsymbol{\beta}_M, \sigma^2, \boldsymbol{Z})\|_\infty \mid \{\min_{M,\boldsymbol{\beta}_M,\sigma^2} \|\boldsymbol{y} - \boldsymbol{G}(M, \boldsymbol{\beta}_M, \sigma^2, \boldsymbol{Z})\|_\infty < \epsilon\},$$

where $\|\cdot\|_\infty$ is the l_∞ norm.

Let the collection of models $\mathcal{M}'$ under consideration satisfies that for any two models $M_1 \neq M_2 \in \mathcal{M}'$, we have

$$P\left(\bigcap_{i=1,2}\{\min_{\boldsymbol{\beta}_{M_i},\sigma^2} \|\boldsymbol{y} - \boldsymbol{G}(M_i, \boldsymbol{\beta}_{M_i}, \sigma^2, \boldsymbol{Z})\|_\infty < \epsilon\}\right) = o\left(\max_{i=1,2} P(\min_{\boldsymbol{\beta}_{M_i},\sigma^2} \|\boldsymbol{y} - \boldsymbol{G}(M_i, \boldsymbol{\beta}_{M_i}, \sigma^2, \boldsymbol{Z})\|_\infty < \epsilon)\right), \quad \epsilon \to 0.$$

Then the marginal fiducial distribution for each model $M \in \mathcal{M}'$ is the limit, as $\epsilon \to 0$, of the conditional probabilities

$$r(M) = \lim_{\epsilon\to 0} \frac{P(\min_{\boldsymbol{\beta}_M,\sigma^2} \|\boldsymbol{y} - \boldsymbol{G}(M, \boldsymbol{\beta}_M, \sigma^2, \boldsymbol{Z})\|_\infty < \epsilon)}{\sum_{M'\in\mathcal{M}'} P(\min_{\boldsymbol{\beta}_{M'},\sigma^2} \|\boldsymbol{y} - \boldsymbol{G}(M', \boldsymbol{\beta}_{M'}, \sigma^2, \boldsymbol{Z})\|_\infty < \epsilon)}. \quad \text{(A.2)}$$

Fix a model of size $m = |M| < n - 1$. Expression (A.1) can be simplified as

$$J_M(\boldsymbol{y}, \boldsymbol{\theta}) = \sigma^{-2} \sum_{\substack{\boldsymbol{i} = (i_0, \ldots, i_{m+1}) \\ 1 \le i_0 < \cdots < i_{m+1} \le n}} \left|\det(\boldsymbol{y}, \boldsymbol{X}_M)_{\boldsymbol{i}}\right|.$$

Consequently, following the calculations in the proof of Theorem 3.1 of Hannig (2013) we have

$$P(\min_{\boldsymbol{\beta}_M,\sigma^2} \|\boldsymbol{y} - \boldsymbol{G}(M, \boldsymbol{\beta}_M, \sigma^2, \boldsymbol{Z})\|_\infty < \epsilon) = \epsilon^{n-m-1} \sum_{\substack{\boldsymbol{i} = (i_0, \ldots, i_{m+1}) \\ 1 \le i_0 < \cdots < i_{m+1} \le n}} \left|\det(\boldsymbol{y}, \boldsymbol{X}_M)_{\boldsymbol{i}}\right| \Gamma\left(\frac{n-m}{2}\right) \times (\pi \mathrm{RSS}_M)^{-\frac{n-m}{2}} |\det\left(\boldsymbol{X}_M^T\boldsymbol{X}_M\right)|^{-\frac{1}{2}} + o(\epsilon^{n-m-1}), \quad \text{(A.3)}$$

where RSS_M denotes the residual sum of squares of model M when the parameters are estimated using maximum likelihood.

Equation (A.3) illuminates two major shortcomings of the naive data-generating equation. First, all models that do not have the full dimension have fiducial probability zero. Second, the formula for Jacobian requires to compute a sum of $\binom{n}{m+1}$ terms, which is very computationally expensive.

Hannig and Lee (2009) proposed to solve the first issue by including additional data-generating equations. However, their proposal is not directly applicable to the ultrahigh-dimensional situation and needs to be modified. Our data-generating equation will be selected to solve both issues at the same time.

We propose to use the following modified data-generating equation $\bar{\boldsymbol{G}}(M, \boldsymbol{\beta}_M, \sigma^2, \boldsymbol{Z}, \boldsymbol{B}, \boldsymbol{P})$. The data part is

$$\boldsymbol{v}_M = \left[\left(\boldsymbol{X}_M^T\boldsymbol{X}_M\right)^{-1/2}\boldsymbol{X}_M^T\boldsymbol{y}; (\mathrm{RSS}_M)^{1/2}; \left\{\boldsymbol{i} - \boldsymbol{X}_M\left(\boldsymbol{X}_M^T\boldsymbol{X}_M\right)^{-1}\boldsymbol{X}_M^T\right\}\boldsymbol{y}/\mathrm{RSS}_M\right],$$

where $\boldsymbol{y} = \boldsymbol{X}_M\boldsymbol{\beta}_M + \sigma\boldsymbol{Z}$. The additional data-generating equations are

$$b_m = B_m, \quad p_k = P_k,\ k = 1, \ldots, m,$$

where $m = |M|$ is the dimension of M, B_m is a Bernoulli$(1 - r_m)$ random variable that will be used to penalize for the number of models that have the same size m; and P_k are iid continuous random variables with $f_P(0) = q$ independent of B_m that will be used to penalize for the size of models. Notice that the number of additional equation is the same as the number of unknown parameters in the model. Since we never actually observe the outcomes of the extra data-generating equation, we will select their values as $b_m = p_k = 0$.

For this data-generating equation, the Jacobian (A.1) simplifies to

$$\bar{J}_M(\boldsymbol{y}, \boldsymbol{\theta}) = \sigma^{-2}\left|\det(\boldsymbol{X}_M'\boldsymbol{X}_M)\right|^{\frac{1}{2}}\mathrm{RSS}_M^{\frac{1}{2}}. \quad \text{(A.4)}$$

Consequently,

$$P(\min_{\boldsymbol{\beta}_M,\sigma^2} \|(\boldsymbol{v}_M, b_m, \boldsymbol{p}) - \bar{\boldsymbol{G}}(M, \boldsymbol{\beta}_M, \sigma^2, \boldsymbol{Z}, B_m, \boldsymbol{P})\|_\infty < \epsilon) = \epsilon^{n-1}\Gamma\left(\frac{n-m}{2}\right)(\pi\mathrm{RSS}_M)^{-\frac{n-m-1}{2}} r_m q^m + o(\epsilon^{n-1}). \quad \text{(A.5)}$$

Following recommendation of Hannig and Lee (2009), we select $q = n^{-1/2}$. Additionally, we select $r_m = \binom{p}{m}^{-\gamma}$. The second choice is to penalize for the fact that there are a large number of models that all have the same size. The most natural choice is $\gamma = 1$ for which r_m is the probability of randomly selecting a model M from all models of size m. However, to match the EBIC penalty of Chen and Chen (2008), we allow for other choices of γ.

Equation (6) now follows from (A.2) and (A.5) provided the following identifiability condition holds. For any size m, the residual vectors $\boldsymbol{i} - \boldsymbol{X}_M(\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^T\}\boldsymbol{y}/\mathrm{RSS}_M$ are distinct for all the models $M \in \mathcal{M}'$ of size m. This assumption is implied eventually almost surely by the identifiability assumption (11).

A.3 Comparison to Objective Bayesian Model Selection

By inspecting the form of the generalized fiducial density (3), we see that the Jacobian term $J(\boldsymbol{y}, \boldsymbol{\theta})$ plays a role analogous to a

"data-dependent" prior. Additionally, the Jacobian $\bar{J}_M(\boldsymbol{y}, \boldsymbol{\theta})$ displayed in (A.4) has the form $C_M(\boldsymbol{y})\sigma^{-2}$, where $C_M(\boldsymbol{y})$ is a specific constant depending only on the observed data. Therefore, this particular Jacobian can be viewed as an improper Bayesian prior $\pi(\boldsymbol{\beta}_M, \sigma^2) \propto \sigma^{-2}$. As discussed in Berger and Pericchi (2001), one of the issues with the use of improper priors in Bayesian model selection is that a selection of a constant C_B in the prior $C_B\sigma^{-2}$ is arbitrary. This is not a problem when a posterior with respect to one model is considered because the arbitrary constant cancels. However, it becomes a problem in model selection as the arbitrary constants C_B influence the result making the use of improper prior for model selection difficult. Thus, a contribution of fiducial inference could be viewed as suggesting the choice of a particular constant $|\det(\boldsymbol{X}'_M\boldsymbol{X}_M)|^{\frac{1}{2}}\mathrm{RSS}_M^{\frac{1}{2}}$ for each of the model.

APPENDIX B: PROOF OF THEOREM 3.1

B.1 Lemmas

First, we present three lemmas, where detailed proofs can be found in Luo and Chen (2013). Lemma B.1 is proved by applying Stirling's formula. Lemma B.2 is proved by integration by parts and Lemma B.3 is proved by applying Lemma B.2.

Lemma B.1. If $\log j/\log p \to \delta$ as $p \to \infty$, then

$$\log\binom{p}{j} = j\log p(1-\delta)(1+o(1)).$$

Lemma B.2. Let χ_j^2 be a chi-square random variable with degrees of freedom j. If $c \to \infty$ and $J/c \to 0$, then

$$P\left(\chi_j^2 > c\right) = \frac{1}{\Gamma(j/2)}(c/2)^{k/2-1}e^{-c/2}(1+o(1)),$$

uniformly over $j \le J$.

Lemma B.3. Let χ_j^2 be a chi-square random variable with degrees of freedom j. Let $c_j = 2j\{\log p + \log(j\log p)\}$. If $p \to \infty$, then for any $J \le p$,

$$\sum_{j=1}^{J}\binom{p}{j}P\left(\chi_j^2 > c_j\right) \to 0.$$

B.2 Proof of Theorem 3.1

This appendix presents the proof of Theorem 3.1. Some of the arguments are similar to those in Luo and Chen (2013).

Denote $\mathcal{M}$ as the collection of models for which (11) holds, that is, $\mathcal{M} = \{M : |M| \le k|M_0|\}$ for some fixed k. We first prove that $\max_{\mathcal{M}} r_\gamma(M)/r_\gamma(M_0) \xrightarrow{P} 0$. WLOG, assume that $\sigma^2 = 1$. Let $m = |M|$ and $m_0 = |M_0|$ whenever there is no ambiguity. Notice that $m_0 = o(n)$ and $m = o(n)$. Rewrite

$$R_\gamma(M)/R_\gamma(M_0) = \exp\{-T_1 - T_2\}$$

where

$$T_1 = \frac{n-m-1}{2}\log\left(\frac{\mathrm{RSS}_M}{\mathrm{RSS}_{M_0}}\right),$$

$$\begin{aligned} T_2 = {} & \frac{m-m_0}{2}\log n + \frac{m-m_0}{2}\log(\pi\mathrm{RSS}_{M_0}) \\ & + \log\left\{\Gamma\left(\frac{n-m_0}{2}\right)/\Gamma\left(\frac{n-m}{2}\right)\right\} \\ & - \gamma\log\binom{p}{m_0} + \gamma\log\binom{p}{m}. \end{aligned}$$

We are going to show that the followings hold uniformly for all M:

$$\begin{cases} T_1 = \dfrac{\Delta_M(1+o_p(1))}{2} & \text{if } M_0 \not\subset M, \\ T_2 \ge -\dfrac{3}{2}m_0\log n - \gamma m_0\log p & \text{if } M_0 \not\subset M, \end{cases} \tag{B.1}$$

$$\begin{cases} T_1 \ge -(m-m_0)(1+\delta)\log p(1+o_p(1)) & \text{if } M_0 \subset M, \\ T_2 = \dfrac{3}{2}(m-m_0)\log n(1+o_p(1)) & \\ \quad +\gamma(1-\delta)(m-m_0)\log p(1+o(1)) & \text{if } M_0 \subset M. \end{cases} \tag{B.2}$$

Case 1: $M_0 \not\subset M$.
Let $\mathcal{M}_j = \{M : |M| = j, M \in \mathcal{M}\}$.
First note that $\mathrm{RSS}_{M_0} = (n-m_0)\left(1+o_p(1)\right) = n\left(1+o_p(1)\right)$,

$$\begin{aligned} \mathrm{RSS}_M - \mathrm{RSS}_{M_0} = {} & \Delta(M) + 2\boldsymbol{\mu}^T(\boldsymbol{I}-\boldsymbol{H}_M)\boldsymbol{\epsilon} \\ & + \boldsymbol{\epsilon}^T\boldsymbol{H}_M\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T\boldsymbol{H}_{M_0}\boldsymbol{\epsilon} \end{aligned} \tag{B.3}$$

and $\boldsymbol{\epsilon}^T\boldsymbol{H}_{M_0}\boldsymbol{\epsilon} = m_0(1+o_p(1))$.

Consider the second term in (B.3) and denote $Z_M = \boldsymbol{\mu}^T(\boldsymbol{I}-\boldsymbol{H}_M)\boldsymbol{\epsilon}/\sqrt{\Delta_M}$, we have

$$\boldsymbol{\mu}^T(\boldsymbol{I}-\boldsymbol{H}_M)\boldsymbol{\epsilon} = \sqrt{\Delta_M}Z_M$$

and $Z_M \sim N(0,1)$. Let $c_j = 2j\{\log p + \log(j\log p)\}$. For simplicity, denote $c_{|M|}$ by c_m. Then, by Lemma B.3,

$$\begin{aligned} P\left(\max_{\mathcal{M}}|Z_M/\sqrt{c_m}| > 1\right) &\le \sum_{j=1}^{km_0}\sum_{\mathcal{M}_j}P\left(Z_M^2 > c_j\right) \\ &= \sum_{j=1}^{km_0}\binom{p}{j}P\left(\chi_1^2 > c_j\right) \\ &\le \sum_{j=1}^{km_0}\binom{p}{j}P\left(\chi_j^2 > c_j\right) \to 0. \end{aligned}$$

Therefore, $|\boldsymbol{\mu}^T(I-\boldsymbol{H}_M)\boldsymbol{\epsilon}| \le \sqrt{\Delta_M}|Z_M| \le \sqrt{\Delta_M}\sqrt{c_m}(1+o_p(1))$ uniformly over $\mathcal{M}$. Since $c_m = O(m_0\log p)$, and by the identifiability condition (11), $m_0\log p = o(\Delta_M)$ uniformly over $\mathcal{M}$ s.t. $M_0 \not\subset M$,

$$|\boldsymbol{\mu}^T(I-\boldsymbol{H}_M)\boldsymbol{\epsilon}| = o_p(\Delta_M).$$

Now consider the third term in (B.3), by Lemma B.3 again,

$$\begin{aligned} P\left(\max_{\mathcal{M}}\boldsymbol{\epsilon}^T\boldsymbol{H}_M\boldsymbol{\epsilon}/c_m > 1\right) &\le \sum_{j=1}^{km_0}\sum_{\mathcal{M}_j}P\left(\boldsymbol{\epsilon}^T\boldsymbol{H}_M\boldsymbol{\epsilon} > c_j\right) \\ &= \sum_{j=1}^{km_0}\binom{p}{j}P\left(\chi_j^2 > c_j\right) \to 0. \end{aligned}$$

So $\boldsymbol{\epsilon}^T\boldsymbol{H}_M\boldsymbol{\epsilon} \le c_m(1+o_p(1))$ and

$$\boldsymbol{\epsilon}^T\boldsymbol{H}_M\boldsymbol{\epsilon} = o_p(\Delta_M)$$

uniformly over $\mathcal{M}$ s.t. $M_0 \not\subset M$.

Therefore,

$$\mathrm{RSS}_M - \mathrm{RSS}_{M_0} = \Delta(M)(1+o_p(1)),$$

and

$$\begin{aligned} \log\left(\frac{\mathrm{RSS}_M}{\mathrm{RSS}_{M_0}}\right) &= \log\left(1+\frac{\mathrm{RSS}_M - \mathrm{RSS}_{M_0}}{\mathrm{RSS}_{M_0}}\right) \\ &= \log\left\{1+\frac{\Delta(M)}{n}(1+o_p(1))\right\} \end{aligned}$$

uniformly for all $M \in \mathcal{M}$ s.t. $M_0 \not\subset M$. Therefore,

$$T_1 = \frac{n(1+o(1))}{2}\log\left\{1+\frac{\Delta(M)}{n}(1+o_p(1))\right\} = \frac{\Delta(M)(1+o_p(1))}{2}$$

uniformly for all $M \in \mathcal{M}$ s.t. $M_0 \not\subset M$.

Moreover,

$$\begin{aligned}&\frac{m-m_0}{2}\log(\pi \mathrm{RSS}_{M_0}) + \log\left\{\Gamma\left(\frac{n-m_0}{2}\right)/\Gamma\left(\frac{n-m}{2}\right)\right\}\\ &= \frac{m-m_0}{2}\log n(1+o_p(1)) + \frac{m-m_0}{2}\log n(1+o(1))\\ &= (m-m_0)\log n(1+o_p(1)).\end{aligned}$$

Finally,

$$\begin{aligned}T_2 &= \frac{3}{2}(m-m_0)\log n(1+o_p(1)) - \gamma\log\binom{p}{m_0} + \gamma\log\binom{p}{m}\\ &\geq -\frac{3}{2}m_0\log n(1+o_p(1)) - \gamma m_0\log p.\end{aligned}$$

Case 2: $M_0 \subset M$.

Let $\mathcal{M}^* = \{M \in \mathcal{M},\ M_0 \subset M,\ M \neq M_0\}$ and $\mathcal{M}_j^* = \{M,\ |M| = j,\ M_0 \subset M\}$.

First notice that $\mathrm{RSS}_{M_0} - \mathrm{RSS}_M = \chi^2_{m-m_0}(M)$, where $\chi^2_{m-m_0}(M)$ is a chi-square random variable depending on M with degrees of freedom $m - m_0$.

Recall $c_j = 2j\{\log p + \log(j\log p)\}$, by Lemma B.3 again,

$$\begin{aligned}&P\left(\max_{1\leq j\leq km_0-m_0}\max_{M\in\mathcal{M}_j^*}\chi_j^2(M)/c_j \geq 1\right)\\ &\leq \sum_{j=1}^{km_0-m_0} P\left(\max_{M\in\mathcal{M}_j^*}\chi_j^2(M) \geq c_j\right)\\ &= \sum_{j=1}^{km_0-m_0}\binom{p-m_0}{j}P\left(\chi_j^2(M)\geq c_j\right)\\ &\leq \sum_{j=1}^{km_0-m_0}\binom{p}{j}P\left(\chi_j^2(M)\geq c_j\right)\to 0.\end{aligned}$$

It implies that

$$\chi^2_{m-m_0}(M) \leq c_{m-m_0}(1+o_p(1)).$$

Note that $c_{m-m_0} = o(n)$ uniformly, therefore

$$\begin{aligned}&\frac{n-m-1}{2}\log\left(\frac{RSS_M}{RSS_{M_0}}\right)\\ &= -\frac{n-m-1}{2}\log\left(1+\frac{\chi^2_{m-m_0}(M)}{\mathrm{RSS}_{M_0}-\chi^2_{m-m_0}(M)}\right)\\ &\geq -\frac{n-m-1}{2}\left(\frac{\chi^2_{m-m_0}(M)}{\mathrm{RSS}_{M_0}-\chi^2_{m-m_0}(M)}\right)\\ &\geq -\frac{c_{m-m_0}}{2}(1+o_p(1)) \geq -(m-m_0)\\ &\times\left[1+\frac{\log\{(km_0-m_0)\log p\}}{\log p}\right]\log p(1+o_p(1))\\ &\geq -(m-m_0)(1+\delta)\log p(1+o_p(1))\end{aligned}$$

uniformly over $\mathcal{M}^*$.

Therefore, we show that

$$T_1 \geq -(m-m_0)(1+\delta)\log p(1+o_p(1))$$

uniformly over $\mathcal{M}^*$.

By Lemma B.1, for $m_0 < m < km_0$, $\log\binom{p}{m} = (1-\delta)m\log p(1+o(1))$ uniformly over $\mathcal{M}^*$.

Therefore,

$$\begin{aligned}T_2 = &\frac{3}{2}(m-m_0)\log n(1+o_p(1))\\ &+\gamma(1-\delta)(m-m_0)\log p(1+o(1))\end{aligned}$$

uniformly over $\mathcal{M}^*$. Finally,

$$\begin{aligned}\max_{M\neq M_0, M\in\mathcal{M}} R_\gamma(M)/R_\gamma(M_0) = \max\Bigg\{&\max_{M_0\not\subset M}\exp(-T_1-T_2),\\ &\max_{M_0\subset M}\exp(-T_1-T_2)\Bigg\}.\end{aligned}$$

By (B.1),

$\max_{M_0\not\subset M}\exp(-T_1-T_2) \xrightarrow{P} 0$ since

$$\min_{M_0\not\subset M} T_1 + T_2 \to \infty$$

and by (B.2), $\max_{M_0\subset M}\exp(-T_1-T_2)\to 0$ if $\gamma > \frac{1+\delta}{1-\delta} - \frac{3\eta}{2(1-\delta)}$. It proves that

$$\max_{M\neq M_0, M\in\mathcal{M}} R_\gamma(M)/R_\gamma(M_0) \xrightarrow{P} 0.$$

Moreover, if (14) holds and denote m'_j as the number of models in $\mathcal{M}'_j$,

$$\begin{aligned}\sum_{M\neq M_0, M\in\mathcal{M}'} R_\gamma(M)/R_\gamma(M_0) &\leq \sum_{j=1}^{km_0}\sum_{\mathcal{M}'_j} R_\gamma(M)/R_\gamma(M_0)\\ &\leq \sum_{j=1}^{km_0}\max_{M\neq M_0, M\in\mathcal{M}'_j} m'_j R_\gamma(M)\\ &\quad /R_\gamma(M_0) \xrightarrow{P} 0.\end{aligned}$$

Equivalently,

$$R_\gamma(M_0)/\sum_{M\in\mathcal{M}'} R_\gamma(M) \xrightarrow{P} 1.$$

[Received April 2013. Revised March 2014.]

REFERENCES

Barbieri, M. M., and Berger, J. O. (2004), "Optimal Predictive Model Selection," *The Annals of Statistics*, 32, 870–897. [763]

Berger, J. O., Bernardo, J. M., and Sun, D. (2009), "The Formal Definition of Reference Priors," *The Annals of Statistics*, 37, 905–938. [761]

Berger, J. O., and Pericchi, L. R. (2001), "Objective Bayesian Methods for Model Selection: Introduction and Comparison," in *Model Selection of IMS Lecture Notes Monograph Series* (Vol. 38), ed. P. Lahiri, Beachwood, OH: Institute of Mathematical Statistics, pp. 135–207. [770]

Bühlmann, P., Kalisch, M., and Maathuis, M. H. (2010), "Variable Selection in High-Dimensional Linear Models: Partially Faithful Distributions and the PC-Simple Algorithm," *Biometrika*, 97, 261–278. [760]

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criteria for Model Selection With Large Model Spaces," *Biometrika*, 95, 759–771. [760,762,769]

Cho, H., and Fryzlewicz, P. (2011), "High Dimensional Variable Selection via Tilting," *Journal of the Royal Statistical Society,* Series B, 74, 593–622. [760]

Cisewski, J., and Hannig, J. (2012), "Generalized Fiducial Inference for Normal Linear Mixed Models," *The Annals of Statistics*, 40, 2102–2127. [761]

Dempster, A. P. (2008), "The Dempster-Shafer Calculus for Statisticians," *International Journal of Approximate Reasoning*, 48, 365–377. [761]

Dong, Y. (2007), "Inference After Model Selection," Ph.D. dissertation, University of Minnesota. [764]

Edlefsen, P. T., Liu, C., and Dempster, A. P. (2009), "Estimating Limits From Poisson Counting Data Using Dempster–Shafer Analysis," *Annals of Applied Statistics*, 3, 764–790. [761]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [762]

Fan, J., Feng, Y., and Song, R. (2011), "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models," *Journal of the American Statistical Association*, 106, 544–557. [760]

Fan, J., Guo, S., and Hao, N. (2012), "Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression," *Journal of the Royal Statistical Society,* Series B, 74, 37–65. [760,765,768]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [760,764]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society,* Series B, 70, 849–911. [760,762,763]

——— (2010), "A Selective Overview of Variable Selection in High Dimensional Feature Space," *Statistica Sinica*, 20, 101–148. [760]

——— (2011), "Non-Concave Penalized Likelihood With Np-Dimensionality," *IEEE Transactions on Information Theory*, 57, 5467–5484. [760]

Fan, J., and Peng, H. (2004), "Nonconcave Penalized Likelihood With a Diverging Number of Parameters," *The Annals of Statistics*, 32, 928–961. [764]

Fisher, R. A.. (1930), "Inverse Probability," *Proceedings of the Cambridge Philosophical Society*, xxvi, 528–535. [760,761]

Hannig, J. (2009), "On Generalized Fiducial Inference," *Statistica Sinica*, 19, 491–544. [761,763]

——— (2013), "Generalized Fiducial Inference via Discretization," *Statistica Sinica*, 23, 489–514. [761,763,769]

Hannig, J., Iyer, H. K., and Patterson, P. (2006), "Fiducial Generalized Confidence Intervals," *Journal of American Statistical Association*, 101, 254–269. [761]

Hannig, J., and Lee, T. C. M. (2009), "Generalized Fiducial Inference for Wavelet Regression," *Biometrika*, 96, 847–860. [761,762,769]

Huang, J., Ma, S., and Zhang, C. (2008), "Adaptive Lasso for Sparse High-Dimensional Regression Models," *Statistica Sinica*, 18, 1603–1618. [764]

Leeb, H., and Pötscher, B. M. (2008), "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?," *Econometric Theory*, 24, 338–376. [764]

Lidong, E., Hannig, J., and Iyer, H. K. (2008), "Fiducial Intervals for Variance Components in an Unbalanced Two-Component Normal Mixed Linear Model," *Journal of American Statistical Association*, 103, 854–865. [761]

Luo, S., and Chen, Z. (2013), "Extended BIC for Linear Regression Models With Diverging Number of Relevant Features and High or Ultra-High Feature Spaces," *Journal of Statistical Planning and Inference*, 143, 494–504. [763,770]

Martin, R., and Liu, C. (2013), "Inferential Models: A Framework for Prior-Free Posterior Probabilistic Inference," *Journal of the American Statistical Association*, 108, 301–313. [761]

Martin, R., Zhang, J., and Liu, C. (2010), "Dempster-Shafer Theory and Statistical Inference With Weak Beliefs," *Statistical Science*, 25, 72–87. [761]

McNally, R. J., Iyer, H. K., and Mathew, T. (2003), "Tests for Individual and Population Bioequivalence Based on Generalized P-Values," *Statistics in Medicine*, 22, 31–53. [761]

Meier, L., Van De Geer, S., and Bühlmann, P. (2009), "High-Dimensional Additive Modeling," *The Annals of Statistics*, 37, 3779–3821. [760]

Meinshausen, N., and Buhlmann, P. (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34, 1436–1462. [764]

Pötscher, B. M., and Leeb, H. (2009), "On the Distribution of Penalized Maximum Likelihood Estimators: The LASSO, SCAD, and Thresholding," *Journal of Multivariate Analysis*, 100, 2065–2082. [764]

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009), "Sparse Additive Models," *Journal of the Royal Statistical Society,* Series B, 71, 1009–1030. [760]

Rissanen, J. (1989), *Stochastic Complexity in Statistical Inquiry*, Singapore: World Scientific. [762]

——— (2007), *Information and Complexity in Statistical Modeling*, New York: Springer. [762]

Salome, D.. (1998), "Statistical Inference via Fiducial Methods," Ph.D. dissertation, University of Groningen. [761]

Shen, X., Huang, H.-C., and Ye, J. (2004), "Inference After Model Selection," *Journal of the American Statistical Association*, 99, 751–762. [764]

Singh, K., Xie, M., and Strawderman, W. E. (2005), "Combining Information From Independent Sources Through Confidence Distributions," *The Annals of Statistics*, 33, 159–183. [761]

Sonderegger, D., and Hannig, J. (2014), "Fiducial Theory for Free-Knot Splines," in *Contemporary Developments in Statistical Theory, a Festschrift in honor of Professor Hira L. Koul*, eds. S. Lahiri, A. Schick, A. SenGupta, and T. N. Sriram, New York: Springer, pp. 155–189. [763]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [760,764]

Wang, H. (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association*, 104, 1512–1524. [760]

Wang, J. C.-M., and Iyer, H. K. (2005), "Propagation of Uncertainties in Measurements Using Generalized Inference," *Metrologia*, 42, 145–153. [761]

Weerahandi, S. (1993), "Generalized Confidence Intervals," *Journal of the American Statistical Association*, 88, 899–905. [761]

——— (1995), *Exact Statistical Methods for Data Analysis, Springer Series in Statistics*, New York: Springer-Verlag. [761]

Xie, M., and Singh, K. (2013), "Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review," *International Statistical Review*, 81, 3–39. [761]

Xie, M., Singh, K., and Strawderman, W. E. (2011), "Confidence Distributions and a Unified Framework for Meta-Analysis," *Journal of the American Statistical Association*, 106, 320–333. [761]

Zhang, C., and Huang, J. (2008), "The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression," *The Annals of Statistics*, 36, 1567–1594. [763,764]

Zhang, J., and Liu, C. (2011), "Dempster-Shafer Inference With Weak Beliefs," *Statistica Sinica*, 21, 475–494. [761]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *The Journal of Machine Learning Research*, 7, 2541–2563. [764]