Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Uncertainty Quantification for High-Dimensional Sparse Nonparametric Additive Models

Qi Gao, Randy C. S. Lai, Thomas C. M. Lee & Yao Li

To cite this article: Qi Gao, Randy C. S. Lai, Thomas C. M. Lee & Yao Li (2020) Uncertainty Quantification for High-Dimensional Sparse Nonparametric Additive Models, Technometrics, 62:4, 513-524, DOI: 10.1080/00401706.2019.1665591

To link to this article: https://doi.org/10.1080/00401706.2019.1665591

TECHNOMETRICS
2020, VOL. 62, NO. 4, 513–524
https://doi.org/10.1080/00401706.2019.1665591

Uncertainty Quantification for High-Dimensional Sparse Nonparametric Additive Models

Qi Gao[a], Randy C. S. Lai[b], Thomas C. M. Lee[a], and Yao Li[a]

[a]Department of Statistics, University of California at Davis, Davis, CA; [b]Department of Mathematics and Statistics, University of Maine, Orono, ME

ABSTRACT
Statistical inference in high-dimensional settings has recently attracted enormous attention within the literature. However, most published work focuses on the parametric linear regression problem. This article considers an important extension of this problem: statistical inference for high-dimensional sparse nonparametric additive models. To be more precise, this article develops a methodology for constructing a probability density function on the set of all candidate models. This methodology can also be applied to construct confidence intervals for various quantities of interest (such as noise variance) and confidence bands for the additive functions. This methodology is derived using a generalized fiducial inference framework. It is shown that results produced by the proposed methodology enjoy correct asymptotic frequentist properties. Empirical results obtained from numerical experimentation verify this theoretical claim. Lastly, the methodology is applied to a gene expression dataset and discovers new findings for which most existing methods based on parametric linear modeling failed to observe.

ARTICLE HISTORY
Received June 2018
Accepted September 2019

KEYWORDS
Confidence bands; Confidence intervals; Generalized fiducial inference; Large p small n; Variability estimation

1. Introduction

Nonparametric additive models, given their flexibility, have long been a popular tool for studying the effects of covariates in regression problems (e.g., Friedman and Stuetzle 1981; Stone 1985). Given a set of n independently and identically distributed observations $\{(Y_i, X_i)\}_{i=1}^n$, with Y_i being the ith response and $X_i = (X_{i1}, \ldots, X_{ip})^\top$ as the ith p-dimensional covariate, a nonparametric additive model is defined as

$$Y_i = \mu + \sum_{j=1}^{p} f_j(X_{ij}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where μ is an intercept term, the f_j's are unknown (and usually smooth) functions, and ε_i is an independent random error with mean zero and finite variance σ^2. Here this article allows the possibility that p is greater than n, which implies some of the functions f_j's are zero.

There is a rich literature on the estimation of the functions f_j's in (1) when $p < n$ is fixed. For example, Stone (1985) developed spline estimators that achieve the same optimal rate of convergence for general p as for $p = 1$ under some assumptions. Buja, Hastie, and Tibishirani (1989) proposed a backfitting algorithm to estimate the functions with linear smoothers and prove its convergence. For fixed p and under some mild regularity conditions, Horowitz, Klemela, and Mammen (2006) obtained oracle efficient estimators using a two-step procedure which are asymptotically normal with convergence rate $n^{-2/5}$ in probability.

In high-dimensional settings where $p > n$, much work has also been done in variable selection; that is, selecting (and estimating) the significant f_j's. Meier, Van De Geer, and Buhlmann (2009) proposed using a new sparsity-smoothness penalty for variable selection and provide oracle results which lead to asymptotic optimality of their estimator for high-dimensional sparse additive models. Ravikumar et al. (2009) derived a sparse backfitting algorithm for variable selection with a penalty based on the l_2 norm of the mean value of the nonparametric components. Their algorithm decouples smoothing and sparsity and is applicable to any nonparametric smoother. Huang, Horowitz, and Wei (2010) applied adaptive group Lasso to select significant f_j's and provide conditions for achieving selection consistency.

In recent years there has been a growing body of work in statistical inference for high-dimensional linear parametric models. For example, Bühlmann (2013), Javanmard and Montanari (2014), Van de Geer et al. (2014), and Zhang and Zhang (2014) studied hypothesis testing and confidence intervals for low-dimensional parameters in high-dimensional linear and generalized linear models. Their approaches are mostly based on "de-biasing" or "de-sparsifying" a regularized regression estimator such as Lasso. Chatterjee and Lahiri (2013) and Lopes (2014) examined properties of the residual bootstrap for high-dimensional regression. Lee et al. (2016) and Tibshirani et al. (2016) considered the exact post-selection inference for sequential regression procedures conditioning on the selected models. Finally, the empirical Bayes approach has also been adopted (see, e.g., Martin, Mess, and Walker 2017).

However, much less attention is given to statistical inference for nonparametric additive models, especially in high-dimensional settings. Fan and Jiang (2005) extended the

CONTACT Thomas C. M. Lee tcmlee@ucdavis.edu Department of Statistics, University of California at Davis, 4118 Mathematical Sciences Building, One Shields Avenue, Davis, CA 95616.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/TECH.

generalized likelihood ratio tests to additive models estimated by backfitting to determine if a specific additive component is significant or admits a certain parametric form. However, these authors did not consider the cases where $p > n$ and inferences for some parameters such as σ. More recently Lu, Kolar, and Liu (2015) proposed two types of confidence bands for the marginal influence function in a novel high-dimensional nonparametric model, termed ATLAS, which is a generalization of the sparse additive model, although no inference procedure is provided for other model components. Finally, various Bayesian methods have also been proposed, including Scheipl, Fahrmeir, and Kneib (2012) and Shang and Li (2014). However, none of these methods is designed to provide uncertainty quantification for high-dimensional nonparametric additive models.

The main goal of this article is to address the inference problem for high-dimensional nonparametric additive models. To be more specific, this article develops a method that quantifies the uncertainties in the estimated parameters and selected models. This method is based on the generalized fiducial inference (GFI) framework (Hannig et al. 2016), which has been shown to possess extremely good properties, both theoretical and empirical, in various inference problems. To the best of our knowledge, this is the first time that uncertainty quantification is being formally considered for high-dimensional additive models.

The remainder of this article proceeds as follows. In the next section, we first present a spline representation of nonparametric additive models upon which our inference will be based. In Section 3, we introduce the GFI framework and formally describe our proposed inference method for sparse and high-dimensional nonparametric additive models. Section 4 examines the theoretical properties of the proposed method while Section 5 illustrates its empirical properties via numerical experiments and a real data example. Lastly, concluding remarks are offered in Section 6 while proofs of theoretical results are delayed in Appendix A.

2. Spline Modeling of Additive Functions

The functions f_j's in nonparametric additive models are commonly modeled by splines f_{nj}'s in practice. A spline function is a piecewise polynomial function, usually cubic, that is connected together at knots. Here we state the standard conditions and definition for spline functions following (e.g., Stone 1985; Huang, Horowitz, and Wei 2010).

Suppose that $X_j \in \mathcal{X}_j$ where $\mathcal{X}_j = [a, b]$ for finite numbers $a < b$ and $E(Y^2) < \infty$. To ensure identifiability, we assume $Ef_j(X_j) = 0$ for $j = 1, \dots, p$. Let K be the number of knots for a partition of $[a, b]$ satisfying condition (A2) stated in Section 4. Let $\mathcal{S}_n$ be the collection of functions s on $[a, b]$ satisfying the following two conditions: (i) s is a polynomial of degree l (or less) on each subinterval, and, (ii) for two integers l and l' satisfying $l \geq 2$ and $0 \leq l' < l - 1$, s is l'-times continuously differentiable on $[a, b]$.

Then there exists a normalized B-spline basis $\{\varphi_k(\cdot), k = 1, \dots, h_n\}$, $h_n = K + l$ for $\mathcal{S}_n$, such that for any $f_{nj} \in \mathcal{S}_n$,

$$f_{nj}(x) = \sum_{k=1}^{h_n} \beta_{jk}\varphi_{jk}(x), \tag{2}$$

where β_{jk} is the coefficient of the basis function $\varphi_{jk}(x)$, $k = 1, \dots, h_n$. As shown in Lemma 1, f_j's can be well approximated by functions in $\mathcal{S}_n$ under certain smoothness conditions. Thus, in the rest of this article, for the purpose of expediting technical calculations, we shall assume that the spline representation is exact for the additive functions f_j's.

In matrix notation, Equation (1) can be rewritten in the following form

$$\boldsymbol{Y} = \mu\boldsymbol{1} + \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3}$$

where $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$, $\boldsymbol{Z}$ is a $n \times (h_n p)$ matrix with ith row equals to $(\varphi_{11}(X_{i1}), \varphi_{12}(X_{i1}), \dots, \varphi_{1h_n}(X_{i1}), \dots, \varphi_{p1}(X_{ip}), \varphi_{p2}(X_{ip}), \dots, \varphi_{ph_n}(X_{ip}))$, $\boldsymbol{\beta} = (\beta_{11}, \dots, \beta_{1h_n}, \dots, \beta_{p1}, \dots, \beta_{ph_n})^\top$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$. This linear representation of additive models provides us a proxy to apply the GFI methodology on high-dimensional regression models as described in Lai, Hannig, and Lee (2015).

3. Methodology

3.1. *Generalized Fiducial Inference*

The original idea of fiducial inference can be dated back to the 1930s. Fisher (1930) introduced fiducial inference as an alternative to Bayesian procedures with the goal to assign an appropriate statistical distribution on the parameters of a parametric family of distributions. One well-known criticism of the classical Bayesian procedures is the need of specifying prior distributions for the parameters. Fisher's proposal aims to avoid such an issue by considering a switching mechanism between the parameters and the observations, in a way very similar to the procedure of obtaining parameter estimates by maximizing the likelihood function. In spite of Fisher's continuous effort in establishing a formal inference framework via the fiducial argument, it has been overlooked for many years by the majority of the statistics community. Interested readers are referred to Hannig et al. (2016) where a detailed discussion about the history of fiducial inference and numerous related references can be found.

In recent years, there has been increasing interest in reformulating the somewhat abandoned fiducial concepts. These modern modifications include Dempster–Shafer theory (Dempster 2008), its relative inferential models (Martin, Zhang, and Liu 2010; Martin and Liu 2013, 2015) and confidence distribution (Xie and Singh 2013). One such modern formulation of Fisher's fiducial inference is the so-called generalized fiducial inference or GFI (Hannig 2009; Hannig et al. 2016). GFI has been applied successfully in many classical and modern problems, including wavelet regression (Hannig and Lee 2009), linear mixed models (Cisewski and Hannig 2012), and logistic regression model (Liu and Hannig 2016). In particular, Lai, Hannig, and Lee (2015) successfully apply GFI on ultrahigh-dimension regression models and show that the resulting GFI inference procedure has excellent theoretical and practical performance.

3.2. *A Recipe for Applying GFI*

The most significant idea behind the philosophy of GFI is a switching principle. It begins by realizing that any

n-dimensional observation $\boldsymbol{Y}$ can be viewed as an outcome of an equation:

$$\boldsymbol{Y} = \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{U}), \tag{4}$$

where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a p-dimensional fixed parameter vector which determines the distribution of $\boldsymbol{Y}$, $\boldsymbol{U}$ is a random variable whose distribution is known and does not depend on $\boldsymbol{\theta}$, and $\boldsymbol{G}$ is a parametric deterministic function relating $\boldsymbol{Y}$ and $\boldsymbol{\theta}$. Such a relationship is sometimes known as a "structural equation" in other areas of study. There may be more than one structural equation for any given distribution of a random vector $\boldsymbol{Y}$. If the elements of $\boldsymbol{Y}$ are independent, a naive choice of $\boldsymbol{G}$ would be the inverse distribution function for each element and $\boldsymbol{U}$ would be just an iid uniformly $(0, 1)$ random vector.

The switching principle states that, if $\boldsymbol{Y} = \boldsymbol{y}$ is observed, a distribution of $\boldsymbol{\theta}$ can be defined by inverting the relationship of $\boldsymbol{y}$ and $\boldsymbol{\theta}$ while continuing to believe that the same relation holds and the distribution of $\boldsymbol{U}$ remains unchanged. With this thinking, for any $\boldsymbol{y}$, one could define the set $\{\boldsymbol{\theta} : \boldsymbol{y} = \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{U}^*)\}$ as the inverse mapping of $\boldsymbol{G}$ and $\boldsymbol{U}^*$ is distributed identically as $\boldsymbol{U}$. This random set could be empty if there are no $\boldsymbol{\theta}$'s such that $\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{U}^*)$, or it could have more than one element if there is more than one $\boldsymbol{\theta}$ such that $\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{U}^*)$. The support of $\boldsymbol{U}^*$ could be renormalized to assure that there is at least one solution of the equation. For those values of $\boldsymbol{U}^*$ resulting in multiple solutions, Hannig (2009) suggested randomly picking an element from the random set $\{\boldsymbol{\theta} : \boldsymbol{y} = \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{U}^*)\}$.

This algorithm yields a random sample of $\boldsymbol{\theta}$ if $\boldsymbol{U}^*$ is repeatedly sampled. The resulting random sample of $\boldsymbol{\theta}$ is called a fiducial sample of $\boldsymbol{\theta}$, on which statistical inferences of $\boldsymbol{\theta}$ could be based. The density function of $\boldsymbol{\theta}$ is also implicitly defined via this algorithm and is denoted as $r(\boldsymbol{\theta}|\boldsymbol{y})$. The function $r(\boldsymbol{\theta}|\boldsymbol{y})$ is called the generalized fiducial density and Hannig et al. (2016) show that, under reasonable smoothness assumptions of the likelihood function of $\boldsymbol{Y}$, a version of the generalized fiducial density is given by

$$r(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}, \boldsymbol{\theta}) J(\boldsymbol{y}, \boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} f(\boldsymbol{y}, \boldsymbol{\theta}') J(\boldsymbol{y}, \boldsymbol{\theta}') d\boldsymbol{\theta}'}, \tag{5}$$

where

$$J(\boldsymbol{y}, \boldsymbol{\theta}) = D\left(\frac{d}{d\boldsymbol{\theta}} \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{u})|_{\boldsymbol{u}=\boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})}\right),$$

$D(A) = (\det A^{\top} A)^{1/2}$ and $\boldsymbol{u} = \boldsymbol{G}^{-1}(\boldsymbol{y}, \boldsymbol{\theta})$ is the value of $\boldsymbol{u}$ such that $\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{\theta}, \boldsymbol{u})$.

Although the generalized fiducial density in Equation (5) provides an explicit expression for the distribution of $\boldsymbol{\theta}$, it is not always possible to calculate its form analytically. For example, it is very often that $r(\boldsymbol{\theta}|\boldsymbol{y})$ is known only up to a normalizing constant, and in such cases one may need to use Monte Carlo techniques to simulate a fiducial sample. Besides conventional Monte Carlo techniques, Hannig, Lai, and Lee (2014) considered a nonintrusive method for models for which closed form densities are not available.

Model selection was introduced into the GFI paradigm by Hannig and Lee (2009) in the context of wavelet regression. The most significant challenge is to incorporate the uncertainty due to model selection into the problem setup. To facilitate the notation, denote now the structural equation of a particular model M as

$$\boldsymbol{Y} = \boldsymbol{G}(M, \boldsymbol{\theta}_M, \boldsymbol{U}), \quad M \in \mathcal{M}, \tag{6}$$

where $\mathcal{M}$ is a collection of models. Thus, for any given model, Equation (5) gives the corresponding generalized fiducial density for $\boldsymbol{\theta}$, which is now represented as $r(\boldsymbol{\theta}|\boldsymbol{y}, M)$. As stated by Hannig et al. (2016), similar to MLE, GFI tends to favor large models, therefore, additional penalty and assumptions about the model size are needed to account for the model complexity. These authors also argued for introducing penalty in the GFI framework which leads to the following marginal generalized fiducial probability $r(M)$ of model M:

$$r(M) = \frac{\int r(\boldsymbol{\theta}|\boldsymbol{y}, M) q^{|M|} d\boldsymbol{\theta}_M}{\sum_{M' \in \mathcal{M}} \int r(\boldsymbol{\theta}|\boldsymbol{y}, M') q^{|M'|} d\boldsymbol{\theta}_{M'}}, \tag{7}$$

where q is a constant determined by the penalty and $|M|$ is the number of parameters of the model M. Note that for brevity we suppress the dependence of $\boldsymbol{y}$ in the notation of $r(M)$. The value of q can be interpreted as the prior sparsity rate of the predictors under the Bayesian framework, or can be viewed as a solely penalty term as in the context of frequentists. In GFI, q can be thought as the probability of observing a structural equation for a specific predictor. For the $p < n$ scenario, one can choose q as $n^{-1/2}$ which results in the classical BIC penalty. However, for the more general and high-dimensional setting, the choice of q will need to be adjusted. One possibility is to set $q \propto p^{-1}$ which matches the extended Bayesian information criterion (EBIC) of Luo and Chen (2013) with $\gamma = 1$, where γ is a user-specified parameter for EBIC. Such a choice of q is justified by the theoretical results to be presented below. Throughout all our numerical work, we set $q = 0.2p^{-1}$.

3.3. GFI for Nonparametric Additive Models

This subsection applies the above results to nonparametric additive models and obtains the corresponding generalized fiducial probability. Without loss of generality, first assume that in (3) $\mu = 0$ and the random error $\boldsymbol{\varepsilon}$ is normally distributed with covariance $\text{diag}(\sigma^2, \ldots, \sigma^2)$. Let M denote any candidate model, M_0 be the true model and $\boldsymbol{H}$ be the projection matrix of Z; that is, $\boldsymbol{H} = Z(Z^T Z)^{-1} Z^T$. The residual sum of squares RSS is given by $\text{RSS} = \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}\|^2$. The structural equation (6) is now

$$\boldsymbol{Y} = \boldsymbol{G}(M, \boldsymbol{\theta}_M, \boldsymbol{U}) = Z\boldsymbol{\beta} + \sigma \boldsymbol{U}, \quad M \in \mathcal{M}. \tag{8}$$

It can be shown that for the parameters $\boldsymbol{\theta} = (\sigma, \boldsymbol{\beta})^{\top}$ in model (3) (with $\mu = 0$) (e.g., Lai, Hannig, and Lee 2015)

$$J(\boldsymbol{y}, \boldsymbol{\theta}) = \sigma^{-1} |\det(Z'Z)|^{1/2} \text{RSS}^{1/2}.$$

Therefore, the generalized fiducial density of $\boldsymbol{\theta}$ given any model M is

$$r(\boldsymbol{\theta}|\boldsymbol{y}, M) = \frac{\sigma^{-1}\left[\det(Z^{\top}Z)\right]^{1/2} \text{RSS}^{1/2} \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - Z\boldsymbol{\beta})^{\top}(\boldsymbol{y} - Z\boldsymbol{\beta})\right\}}{\int \sigma^{-1}\left[\det(Z^{\top}Z)\right]^{1/2} \text{RSS}^{1/2} \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - Z\boldsymbol{\beta})^{\top}(\boldsymbol{y} - Z\boldsymbol{\beta})\right\} d\boldsymbol{\theta}}. \tag{9}$$

Let p^* be the length of $\boldsymbol{\beta}$. The numerator of Equation (7) becomes

$$\begin{aligned}\int \sigma^{-1}\left[\det(\mathbf{Z}^\top \mathbf{Z})\right]^{1/2} \mathrm{RSS}^{1/2}\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \\ \times \exp\left\{-\frac{1}{2\sigma^2}(y-\mathbf{Z}\boldsymbol{\beta})^\top(y-\mathbf{Z}\boldsymbol{\beta})\right\} q^{p^*} d\theta \\ = (2\pi)^{(p^*-n)/2}\mathrm{RSS}^{1/2}\int \sigma^{p^*-n-1}\exp\left(-\frac{\mathrm{RSS}}{2\sigma^2}\right) q^{p^*} d\sigma \\ = (2\pi)^{(p^*-n)/2} 2^{(n-p^*-2)/2}\mathrm{RSS}^{(p^*-n+1)/2}\Gamma\left(\frac{n-p^*}{2}\right) q^{p^*}.\end{aligned} \tag{10}$$

Thus, the generalized fiducial probability $r(M)$ of any candidate model M is

$$r(M) \propto R(M) = (2\pi)^{(p^*-n)/2} 2^{(n-p^*-2)/2} \mathrm{RSS}^{(p^*-n+1)/2}\Gamma\left(\frac{n-p^*}{2}\right) q^{p^*}. \tag{11}$$

3.4. Generating Fiducial Samples

This subsection describes how to practically generate fiducial samples $(M, \sigma, \boldsymbol{\beta})$ for the current nonparametric additive modeling problem.

First, to reduce the "search space," we consider only candidate models from a subset $\mathcal{M}^*$ of $\mathcal{M}$. This subset $\mathcal{M}^*$ should contain only candidate models with nonnegligible values of $r(M)$. The way we obtain $\mathcal{M}^*$ is to apply group Lasso (Yuan and Lin 2006) to the spline representation in (2), in a manner described below. Notice that group Lasso is used here as it enforces that all β_{jk}'s with the same j to be zero or nonzero simultaneously.

Without the loss of generality, we assume that the first m_0 functions f_j's in (1) are nonzero. Let $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jh_n})^\top$ for $j = 1 \ldots, p$, then $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)^\top$. The group Lasso estimator $\hat{\boldsymbol{\beta}}$ is the minimizer of

$$L(\boldsymbol{\beta}) = \|Y - \mathbf{Z}\boldsymbol{\beta}\|_2^2 + \lambda\sum_{j=1}^{p}\|\boldsymbol{\beta}_j\|_2$$

subject to the constraint that

$$\sum_{i=1}^{n}\sum_{k=1}^{h_n}\beta_{ik}\varphi_k(Z_{ij}) = 0,$$

where λ is a penalty parameter. The constraint can be dropped if we initially center the response and the basis functions. Changing the values of λ will lead to a sequence of fitted models; that is, a solution path. Those fitted models that are on the solution path of group Lasso are taken as candidate models for $\mathcal{M}^*$. For the purpose of not missing any candidate models with nonnegligible $r(M)$ values, we repeat the group Lasso procedure to a number of bootstrapped data and take all the fitted models that lie on the solution paths as $\mathcal{M}^*$. In this way the size of $\mathcal{M}^*$ is substantially smaller than the size of $\mathcal{M}$, and we expect $\sum_{M\in\mathcal{M}^*} r(M)$ to be very close to 1.

For each $M \in \mathcal{M}^*$, we can compute

$$R(M) = (2\pi)^{(m-n)/2} 2^{(n-m-2)/2}\mathrm{RSS}^{(m-n+1)/2}\Gamma\left(\frac{n-m}{2}\right) \times q^m$$

with m as the number of nonzero functions in M. The generalized fiducial probability $r(M)$ can then be well approximated by

$$r(M) \approx \frac{R(M)}{\sum_{M^*\in\mathcal{M}^*} R(M^*)}. \tag{12}$$

For a given model M, σ and $\boldsymbol{\beta}$ can then be sampled from, respectively,

$$\mathrm{RSS}_M/\sigma^2 \sim \chi^2_{n-m} \tag{13}$$

and

$$\boldsymbol{\beta} \sim N(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}, \sigma^2(\mathbf{Z}_M^\top \mathbf{Z}_M)^{-1}), \tag{14}$$

where RSS_M is the residual sum of squares of the candidate model M, $\mathbf{Z}_M$ is the design matrix of M, and $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ is the MLE of $\boldsymbol{\beta}$ for M.

To summarize, we can generate a fiducial sample $(\tilde{M}, \tilde{\sigma}, \tilde{\boldsymbol{\beta}})$ by first drawing a model $\tilde{M}$ from (12), and then $\tilde{\sigma}$ and $\tilde{\boldsymbol{\beta}}$ from (13) and (14), respectively. Notice that in the above no computationally intensive technique like MCMC is required so the generation of a fiducial sample is relatively fast. Using a 2018 MacBook Pro with a dataset of $n = 400$ and $p = 600$, the proposed method typically takes around 50 sec to generate 10^5 fiducial samples.

Finally, we discuss the practical choice of K, the number of knots. It is widely known K will introduce bias if its value is too small, or it will inflate the variance if it is too large. Our experience is that, as long as K is larger than a certain value, the resulting estimates are very often similar (and acceptable), as the use of group Lasso will shrink those insignificant knots to zero. From a theoretical standpoint, the calculations of Lai, Huang, and Lee (2012) suggest that K should be of order $\log(n)$. So in practice we recommend choosing K as the smallest integer larger than $\log(n)$. Tables 1 and 2 suggest that the numerical results are relatively insensitive to the choice of K (as long as K is large enough).

3.5. Point Estimates, Confidence Intervals and Prediction Intervals

Repeating the above procedure multiple times will result in a fiducial sample for $(M, \sigma, \boldsymbol{\beta})$ which can be used for inference, in a manner similar to that for a Bayesian posterior sample. Instead of selecting one single model, the $r(M)$ approximated in (12) estimates how likely each candidate model would be the true model; this affects the models being selected in the fiducial sample. For σ, one can use the average or median of all $\tilde{\sigma}$'s as a point estimate, and the $\alpha/2$ and $1-\alpha/2$ percentiles to construct a $100(1-\alpha)\%$ confidence interval. Similarly, a confidence interval for $E(Y_i|\mathbf{x}_i)$ given the observation $(\mathbf{x}_i, Y_i)$ can be found by computing the percentiles from $z_s\tilde{\boldsymbol{\beta}}$, where z_s is the spline representation of $\mathbf{x}_i$. In addition, prediction intervals for Y can be obtained by taking the percentiles from $\mathbf{Z}\tilde{\boldsymbol{\beta}} + \tilde{\sigma}W$, where $W \sim N(\mathbf{0}, \mathbf{I_n})$.

However, constructing confidence bands for the f_j's is a less trivial task, as it is possible that any particular f_j would appear only in some but not all of the fiducial samples. To handle this issue, we use the following strategy. First from the fiducial

Table 1. Empirical coverage rates of confidence intervals for σ^2.

			90%	95%	99%
$(n, p, \sigma) = (200, 1000, 1)$	$l = 3, K = 6$	Proposed	86.40% (0.392)	92.90% (0.466)	95.10% (0.630)
		oracle	89.70% (0.374)	95.60% (0.447)	98.50% (0.595)
	$l = 3, K = 8$		86.60% (0.410)	91.20% (0.501)	94.82% (0.672)
			90.40% (0.378)	94.30% (0.454)	99.10% (0.609)
	$l = 4, K = 6$		86.80% (0.429)	89.60% (0.535)	94.50% (0.714)
			91.60% (0.376)	94.40% (0.451)	98.30% (0.599)
$(n, p, \sigma) = (200, 1000, 0.8)$	$l = 3, K = 6$	Proposed	89.80% (0.242)	94.40% (0.286)	98.60% (0.384)
		oracle	90.00% (0.243)	94.60% (0.288)	99.10% (0.384)
	$l = 3, K = 8$		88.89% (0.244)	93.00% (0.295)	98.20% (0.395)
			89.59% (0.242)	93.50% (0.292)	99.20% (0.389)
	$l = 4, K = 6$		90.40% (0.241)	93.70% (0.289)	99.10% (0.383)
			89.80% (0.242)	93.00% (0.290)	99.10% (0.386)
$(n, p, \sigma) = (250, 1500, 0.8)$	$l = 3, K = 6$	Proposed	89.50% (0.207)	94.80% (0.248)	98.30% (0.329)
		oracle	89.00% (0.208)	94.60% (0.207)	98.10% (0.331)
	$l = 3, K = 8$		90.90% (0.210)	94.29% (0.252)	98.80% (0.333)
			91.00% (0.210)	94.09% (0.253)	98.60% (0.334)
	$l = 4, K = 6$		88.70% (0.212)	92.69% (0.253)	98.68% (0.338)
			88.10% (0.214)	92.89% (0.255)	98.38% (0.340)

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

Table 2. Empirical coverage rates of confidence intervals for $E(Y_i|\mathbf{x}_i)$.

			90%	95%	99%
$(n, p, \sigma) = (200, 1000, 1)$	$l = 3, K = 6$	Proposed	87.35% (1.401)	93.19% (1.668)	98.05% (2.205)
		oracle	88.75% (1.385)	93.92% (1.648)	98.53% (2.167)
	$l = 3, K = 8$		86.46% (1.601)	91.39% (1.923)	97.05% (2.562)
			89.41% (1.524)	94.38% (1.817)	98.76% (2.396)
	$l = 4, K = 6$		87.70% (1.489)	93.36% (1.789)	98.17% (2.353)
			89.08% (1.452)	94.40% (1.731)	98.77% (2.272)
$(n, p, \sigma) = (200, 1000, 0.8)$	$l = 3, K = 6$	Proposed	89.08% (1.170)	93.63% (1.328)	98.50% (1.757)
		oracle	89.00% (1.167)	93.55% (1.323)	98.57% (1.741)
	$l = 3, K = 8$		89.14% (1.225)	94.38% (1.467)	98.72% (1.937)
			89.31% (1.220)	94.45% (1.457)	98.81% (1.916)
	$l = 4, K = 6$		88.86% (1.168)	94.13% (1.395)	98.72% (1.839)
			88.83% (1.165)	94.10% (1.389)	98.66% (1.825)
$(n, p, \sigma) = (250, 1500, 0.8)$	$l = 3, K = 6$	Proposed	88.17% (0.991)	93.62% (1.183)	98.47% (1.557)
		oracle	88.14% (0.989)	93.53% (1.179)	98.46% (0.983)
	$l = 3, K = 8$		89.33% (1.092)	94.38% (1.306)	98.79% (1.716)
			89.28% (1.090)	94.32% (1.302)	98.76% (1.707)
	$l = 4, K = 6$		87.48% (1.050)	93.04% (1.251)	98.33% (1.651)
			87.41% (1.048)	92.98% (1.247)	98.29% (1.643)

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

samples, we identify the model $\hat{M}$ that appears most. If $\hat{M}$ appears for more than 50% of the times (which has always been the case for all the simulated datasets and the real data example examined), we treat it as the selected model and declare all its nonzero $\hat{f}_j$'s as significant. For each of these selected function f_j's, we then form a confidence band by finding the corresponding percentiles from $Z_f\beta_f$ where Z_f and β_f are, respectively, the spline representation and the part of $\tilde{\beta}$ corresponding to this selected function.

4. Theoretical Properties

This section presents some asymptotic properties of the above generalized fiducial based method. We assume that p is diverging and the theoretical properties are established under the following conditions.

(A1) Let $\mathcal{H}$ be the class of functions h on $[a, b]$ which satisfies a Lipschitz condition of order α:

$$|h^{(k)}(s) - h^{(k)}(t)| \leq C|s - t|^{\alpha} \quad \text{for } s, t \in [a, b],$$

where k is a nonnegative integer and $\alpha \in (0, 1]$ so that $d = k + \alpha > 0.5$. Then $f_j \in \mathcal{H}$ for $1 \leq j \leq q$.

(A2) Let $a = \xi_0 < \xi_1 < \xi_2 < \cdots < \xi_K < \xi_{K+1} = b$ denote a partition of $[a, b]$ into $K + 1$ subintervals where the tth subinterval $I_t = [\xi_{t-1}, \xi_t)$ for $t = 1, \ldots, K$ and $I_{K+1} = [\xi_K, \xi_{K+1}]$. We assume that these knots are not overly sparse; that is, let $0 < \nu < 0.5$ and $K = n^{\nu}$ be a positive integer such that $\max_{1 \leq t \leq K+1} |\xi_t - \xi_{t-1}| = O(n^{-\nu})$.

(A3) There exists a constant c_0 such that $\min_{1 \leq j \leq q} \|f_j\|_2 \geq c_0$, where $\|f\|_2 = [\int_a^b f^2(x)dx]^{1/2}$ whenever the integral exists.

(A4) X has a continuous density and there exist constants C_1 and C_2 such that the density function g_j of X_j satisfies $0 < C_1 \leq g_j(X) \leq C_2 > \infty$.

(A5) Let m and m_0 be the number of nonzero functions selected for models M and M_0, respectively. Then $p^* = h_n m$ for model M. We consider only $M \in \mathcal{M}$ where $\mathcal{M} = \{M : m \leq km_0\}$ for a finite constant $k > 1$; that is, the model whose size is comparable to the true model.

(A6) Let $\Delta(M) = \|\mu - H_M\mu\|$ where $\mu = Z_{M_0}\beta_{M_0}$. We assume the following identifiability condition:

$$\lim_{n\to\infty} \min\left\{\frac{\Delta(M)}{h_n m_o \log p} : M_0 \notin \mathcal{M}, m \le k m_0\right\} = \infty.$$

This condition ensures that the true model can be differentiated from the other models.

(A7) There exists a variable screening procedure to reduce the size of $\mathcal{M}$ when p is too large in practice. Denote the class of candidate models resulting from the screening procedure by $\mathcal{M}^*$. Existence follows from

$$P(M_0 \in \mathcal{M}^*) \to 1 \quad \text{and} \quad \log(|\mathcal{M}_j^*|) = o(h_n j \log n), \quad (15)$$

where $\mathcal{M}_j^*$ denotes the set of all submodels in $\mathcal{M}^*$ of size j. These two limiting criteria ensure that the true model is contained in $\mathcal{M}^*$ and the size of the model space $\mathcal{M}^*$ is not too large.

The following theorem summarizes our main results and its proof can be found in Appendix A.

Theorem 1. Assume A1–A6 hold. As $n \to \infty$, $p \to \infty$, $h_n m_0 \log(p) = o(n)$, $\log(h_n m_0)/\log(p) \to \delta$ and $-\log(q)/\log(p) = \gamma$, we have

$$\max_{M \ne M_0, M \in \mathcal{M}} \frac{r(M)}{r(M_0)} \xrightarrow{P} 0, \quad (16)$$

for $1 + \delta < \gamma < C$ with C being a constant.

Moreover if A7 also holds, with the same γ we have

$$r(M_0) \xrightarrow{P} 1 \quad (17)$$

over the class $\mathcal{M}^*$.

Theorem 1 states that the true model M_0 has the highest generalized fiducial probability amongst all the candidate models under some regularity conditions, and if in addition Equation (15) holds, the true model will be selected with probability tending to 1. Note that Equation (16) does not imply (17) in general since we assume a diverging p. Here γ plays a role similar to that of the tuning parameter in EBIC of Luo and Chen (2013), which controls a penalty for the size of the class of submodels and it must fall within a specified range to ensure that the generalized fiducial distribution is consistent.

In practice, we use group Lasso on bootstrapped data to generate candidate models as discussed in Section 3.4. The resulting model space satisfies Equation (15), since group Lasso is selection consistent for some λ as shown in Nardi and Rinaldo (2008). Theorem 1 also implies that statistical inference based on the generalized fiducial density (5) will retain the exact asymptotic frequentist property as shown in Theorems 2 and 3 of Hannig et al. (2016), which ensure the consistency of our inferential procedure.

We close this section with the following two remarks. First, in general, nonparametric function estimators are biased. We handle this issue by imposing some restrictions on the true functions f_j's. Lemma 1 implies that these functions can be well approximated by polynomial splines adopted in the proposed method (i.e., the bias vanishes asymptotically). With these somewhat strong restrictions, we are able to obtain the above theoretical results. The second remark is that establishing theoretical results for the estimation of β_{jk}'s is more challenging. One reason is that the optimal number of β_{jk}'s can vary for different f_j's and we do not have any method or consistency results for estimating this number under the current setting. Another difficulty is that, in practice, some fitted models in the fiducial samples contain a certain set of β_{jk}'s while some other fitted models do not. This makes it very difficult to even construct confidence intervals for such β_{jk}'s, let alone conduct rigorous study of any theoretical property.

5. Empirical Properties

This subsection investigates the empirical properties of the proposed method via numerical experiments and a real data example.

5.1. Simulation Experiments

Following the simulation settings in Huang, Horowitz, and Wei (2010), we use the model

$$y_i = \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i, \quad i = 1, \ldots n, \quad \varepsilon_i \sim \text{iid}\, N(0, \sigma^2)$$

to generate simulated data, where

$$\begin{aligned}
f_1(x) &= 5x,\\
f_2(x) &= 3(2x-1)^2,\\
f_3(x) &= 4\sin(2\pi x)/\{2 - \sin(2\pi x)\},\\
f_4(x) &= 6\{0.1\sin(2\pi x) + 0.2\cos(2\pi x)\} + 0.3\sin^2(2\pi x)\\
&\quad +0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x),\\
f_j(x) &= 0 \quad \text{for} \quad 5 \le j \le p,
\end{aligned}$$

and the noise variance σ^2 is chosen such that the signal-to-noise ratio is greater than 1 for each nonzero function.

For each set of simulated data, we first use B-spline expansions to transform our data to representation (2). Then a set of candidate models $\mathcal{M}$ are generated by using group Lasso on the transformed data and 10 sets of bootstrapped data. For each M, we run a simple linear regression to obtain RSS_M and compute the fiducial probability $r(M)$ as shown in (11). Then we can draw samples of (M, σ^2, β) based on $r(M)$, (13) and (14) and construct confidence intervals or bands.

Figure 1 summarizes some results of applying the proposed method to a typical simulated dataset with $n = 200$, $p = 1000$ and $\sigma = 0.8$. For the B-spline expansion we use $l = 3$ and $K = 8$, and 10,000 samples of (M, σ^2, β) are generated. Using these samples a 95% confidence interval for σ is obtained, which is $(0.756, 0.947)$ and includes the true value 0.8. The left panel in Figure 1 depicts the histogram of the 10,000 samples of σ which can be seen to be approximately normally distributed. The right panel shows the 95% pointwise confidence band of $f_4(x)$, where the black line is the true function and the red lines are the two bounds. We use $f_4(x)$ here since it is the most complicated of the four nonzero functions. We can see that the confidence band covers the true function very well.

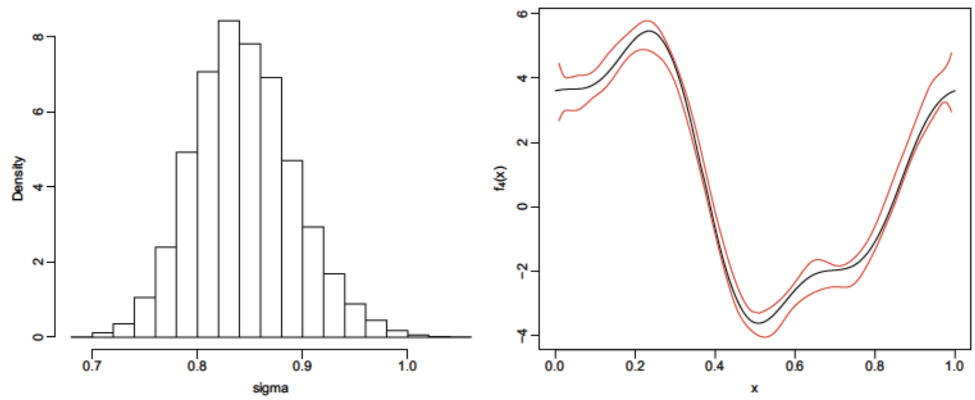

Figure 1. Left: Histogram of the fiducial samples of σ. Right: A 95% pointwise confidence band of f_4. The black line is the true function while the red lines show the band.

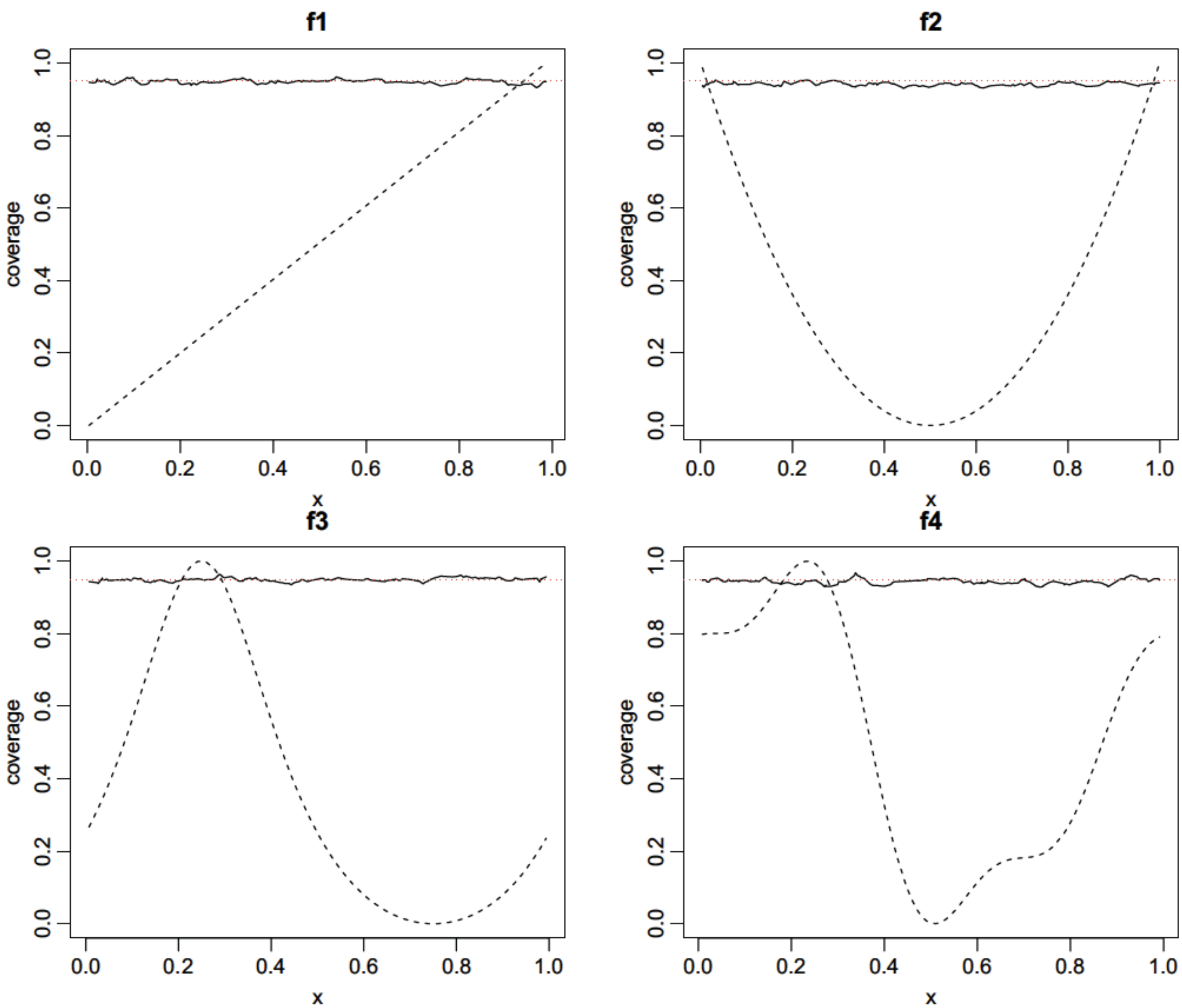

Figure 2. Empirical coverage rates for each nonzero function with experimental parameters $n = 200, p = 1000, \sigma = 0.8, \alpha = 5\%, l = 3$, and $K = 8$.

To test the coverage of these confidence intervals, we generate 1000 simulated datasets and apply the proposed method to compute the confidence intervals for σ^2 and the mean function $E(Y_i|\mathbf{x}_i)$ evaluated at n design points $\mathbf{x}_i$'s. We compare the performance of our method with the "oracle" method which uses the true model and classical theories in linear models based on the spline representation to derive confidence intervals. Different combinations of n, p, σ, l, K, and α are tested and the numerical results are summarized in Tables 1 and 2. The empirical coverage rates are reported together with the average widths of the intervals shown in parentheses.

To evaluate the performance visually, we also plot the empirical coverage rates of all four nonzero functions for one combination of experimental parameters; see Figure 2. In each panel the black dashed line depicts the true value of the function, the horizontal red dashed line is the target confidence level (95% in this case) while the black solid line represents the empirical coverage rates. One can see that these rates are very close to the target confidence level.

5.2. Comparison With an Existing Method

This subsection compares the performances of the proposed method with those from the kernel-sieve hybrid estimator developed by Lu, Kolar, and Liu (2015). As with the settings in Lu, Kolar, and Liu (2015), the same set of test functions in the previous subsection are used with $\sigma^2 = 1.5^2$, $p = 600$, and $n \in \{400, 500, 600\}$. For the kernel-sieve hybrid estimator, we follow the parameter selection used in Lu, Kolar, and Liu (2015). For the proposed method, we set $l = 3$ and $K = 8$.

The empirical coverage rates that the 95% confidence bands cover the true $f_4(x)$ on the first 100 data points are computed based on 500 repetitions. These coverage rates are summarized in Table 3. One can see that both methods produce very reasonable results, with those from the proposed method being closer to the nominal significance level. For visual evaluation, 95% confidence bands for typical datasets are displayed in Figures 3 and 4. These plots suggest that the proposed method produces tighter confidence bands.

Table 3. Empirical coverage rates of 95% confidence bands targeted for $f_4(x)$.

n	Generalized fiducial inference	Kernel-sieve hybrid estimator
400	94.9	91.6
500	94.8	93.2
600	95.3	93.6

Finally, we report the execution times for the methods. For a typical dataset of $n = 400$, the proposed method and the kernel-sieve hybrid estimator take about 52 sec and 903 sec, respectively, to finish. The code for the proposed method is written in *R* while the code for the hybrid estimator (kindly provided by one of its authors) is in Matlab. The machine is a 2018 MacBook Pro with a 2.3 GHz Intel Core i5 processor.

5.3. Real Data Example

This subsection presents a real data analysis on the riboflavin (vitamin B_2) production dataset which is available in Supplementary Section A.1 of Bühlmann, Kalisch, and Meier (2014). The response variable is the logarithm of the riboflavin production rate in *Bacillus subtilis* for $n = 71$ samples while there are $p = 4088$ covariates measuring the logarithm of the expression level of 4088 genes. Bühlmann, Kalisch, and Meier (2014) and Javanmard and Montanari (2014) use linear models to detect significant genes that potentially affect riboflavin production. Bühlmann, Kalisch, and Meier (2014) locate the gene YXLD-at while Javanmard and Montanari (2014) identify the two genes YXLD-at and YXLE-at as significant. Here, instead of using a simple linear model, we assume a nonparametric additive model and apply the GFI methodology to select significant genes.

Following Bühlmann, Kalisch, and Meier (2014), we first adopt a screening procedure and use only the 100 genes with the largest empirical variances. We then apply the proposed method with $K = 2$ and $l = 3$ to the screened dataset and obtain 10,000 fiducial samples for (M, σ, β). It turns out with 63.2% fiducial probability, YXLD-at and YBFG-at are jointly selected while with 28.4% fiducial probability, YXLD-at and XHLA-at are jointly selected. In other words the proposed method is capable of detecting YXLD-at which is considered significant

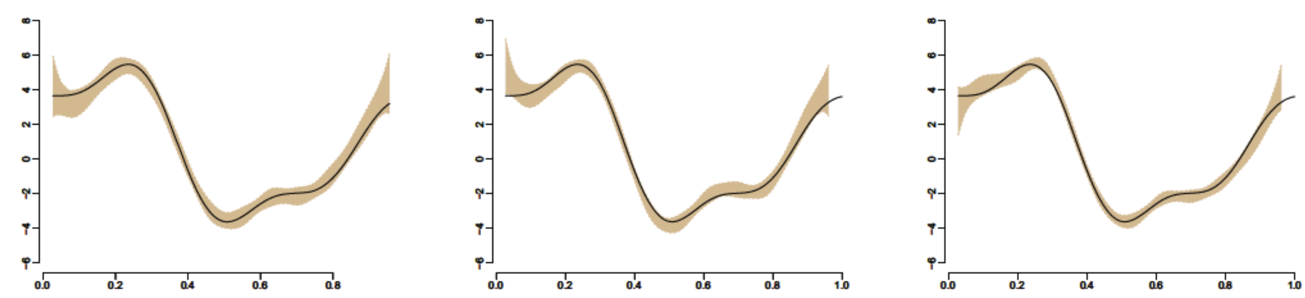

Figure 3. Plots of 95% confidence bands produced by the proposed method. These bands are targeting $f_4(x)$. From left to right, $n = 400, 500, 600$, respectively.

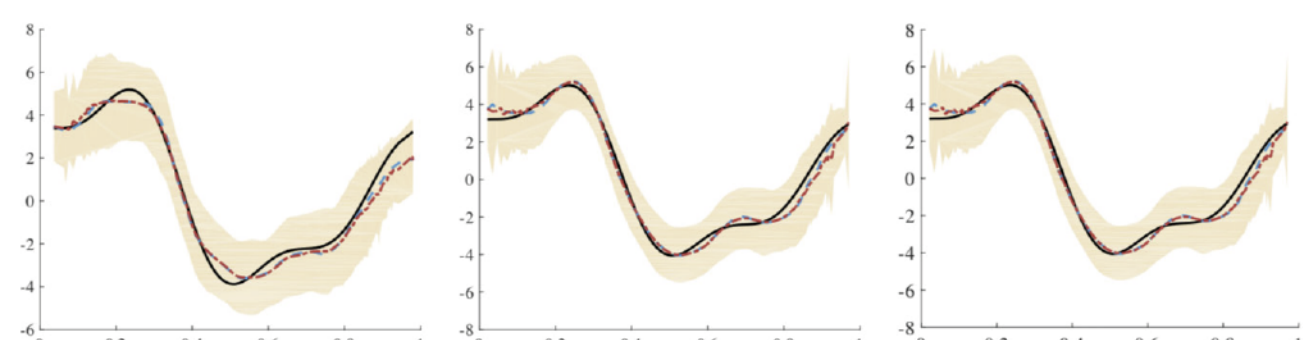

Figure 4. Similar to Figure 3 but for the kernel-sieve hybrid estimator.

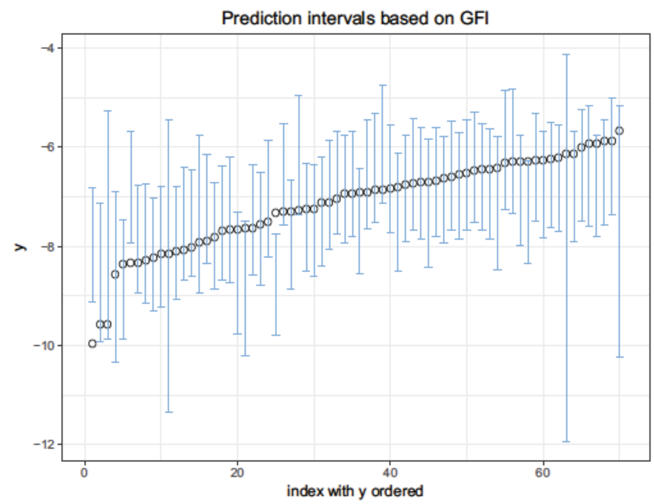

Figure 5. 95% prediction intervals, denoted as blue error bars, for the responses Y_i's, denoted as black circles. For clarity the Y_i's are sorted in ascending order.

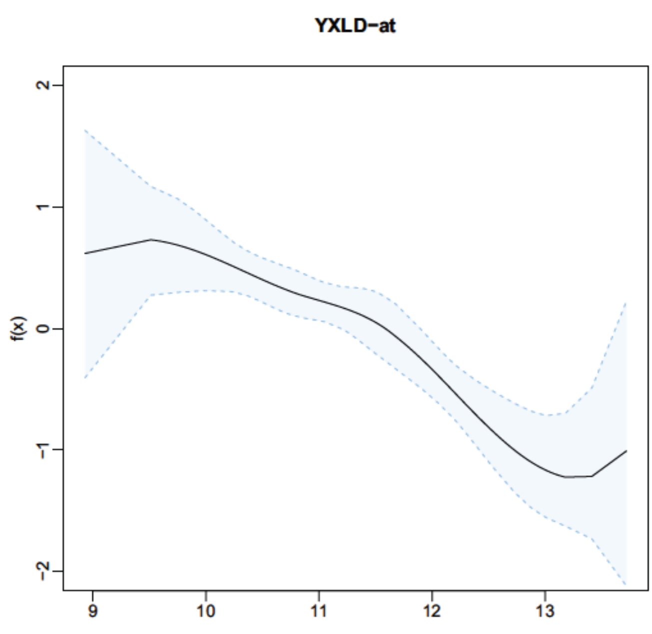

Figure 6. A 95% pointwise confidence band for YXLD_at. The black solid line is the median of the fiducial samples and the dashed blue lines represent the confidence band.

in most previous analyses of this dataset. Also, with the 10,000 fiducial samples we construct a 95% confidence interval for σ, which is (0.43, 0.62).

From the fiducial samples of $(M, \sigma, \boldsymbol{\beta})$, we also compute the leave-one-out 95% prediction intervals for the responses Y_i's and the results are displayed in Figure 5. Note that for clarity the Y_i's are sorted in ascending order. From the plot we can see that 68 out of 71 prediction intervals cover the value of Y_i's, which is around 95.8%. We also compute the 95% pointwise confidence band for YXLD-at which is shown in Figure 6. For the ith function, such a confidence band can be constructed by using the quantiles from $\mathbf{Z}_i\boldsymbol{\beta}_i$ where $\mathbf{Z}_i$ and $\boldsymbol{\beta}_i$ are, respectively, the design matrix and coefficients corresponding to the ith function after the B-spline expansion. In Figure 6, the black solid line is the median among all the samples as the true function is not available for real data, while the dashed lines represent the confidence band. This plot strongly suggests that this gene is indeed significant and the overall trend is more complicated than a simple straight line. We note that, although many previous methods based on high-dimensional linear regression have successfully identified this gene as significant, these methods fail to provide any flexible estimate for the trend, such as the one in Figure 6.

6. Conclusion

In this article, we adopted a GFI methodology to perform statistical inference on sparse high-dimensional nonparametric additive models. In particular, we developed a procedure to generate fiducial samples based on the generalized fiducial distribution of a set of candidate models obtained from group Lasso, and to construct various confidence intervals and prediction intervals by making use of these samples. The developed inferential procedure was shown to have an exact asymptotic frequentist property under some regularity conditions, which was confirmed by its promising performance in numerical simulations. We note that the current framework can in principle be extended to other more complicated and flexible models in high-dimensional settings, such as the generalized nonparametric additive models.

Appendix A: Technical Details

This appendix provides technical details, including the proof for Theorem 1. We begin with three lemmas.

A.1. Lemmas

Lemma 1. Let $\mathcal{F}$ be the class of functions f on $[a, b]$ which satisfies:

$$|f^{(k)}(s) - f^{(k)}(t)| \le C|s-t|^{\alpha} \quad \text{for } s, t \in [a, b],$$

where k is a nonnegative integer and $\alpha \in (0, 1]$ so that $d = k+\alpha > 0.5$. Let S_n^0 denote the space of centered polynomial splines. Suppose that $f \in \mathcal{F}$, $\mathrm{E}f(Z_j) = 0$ and $h_n = O(n^{1/(2d+1)})$, then there exists $f_n \in S_n^0$ satisfying

$$\|f_n - f\|_2 = O_p(h_n^{-d}) = O_p(n^{-d/(2d+1)}).$$

This lemma is proved in Huang, Horowitz, and Wei (2010) and it indicates that the f_j's can be well approximated by polynomial splines under certain smoothness assumptions. Therefore, the representation we consider in Equation (2) is exact.

Lemma 2. Let χ_j^2 denote a χ^2 random variable with degrees of freedom j. If $c \to \infty$ and $\frac{J}{c} \to 0$, then

$$P(\chi_j^2 > c) = \frac{1}{\Gamma(j/2)}(c/2)^{j/2-1}\exp(-c/2)(1+o(1))$$

uniformly for all $j \le J$.

The proof can be found in Luo and Chen (2013) by using integration by parts.

Lemma 3. Let χ_j^2 be a chi-square random variable with degrees of freedom j and $c_j = 2j[\log p + \log(j \log p)]$. If $p \to \infty$, then for any $J \le p$ and $h \ge 1$,

$$\sum_{j=1}^{J} \binom{p}{j} P(\chi_{hj}^2 > c_{hj}) \to 0.$$

Proof. Let $q_j = \sqrt{\frac{c_j}{(j\log p)^2}}$. By using $\binom{p}{j} \le p^j$ and Lemma 2,

$$\binom{p}{j} P(\chi^2_{hj} > c_{hj}) = \binom{p}{j} \frac{1}{2^{hj/2-1}\Gamma(hj/2)} c_{hj}^{hj/2-1} \exp(-c_{hj}/2)(1+o(1))$$
$$\le \frac{c_{hj}^{hj/2-1}}{(hj\log p)^{hj}}(1+o(1))$$
$$= \frac{q_{hj}^{hj}}{c_{hj}}(1+o(1))$$

uniformly over $j < hJ$ for any $J \le p$.

Since $q_j < 1$ for all j and $q_j \to 0$ when j is large enough, we have

$$\sum_{j=1}^{J} \binom{p}{j} P(\chi^2_{hj} > c_{hj}) \le \sum_{j=1}^{J} \frac{q_{hj}^{hj}}{c_{hj}}(1+o(1)) \to 0.$$

□

A.2. Proof of Theorem 1

Since

$$r(M) \propto (2\pi)^{(p^*-n)/2} 2^{(n-p^*-2)/2} \mathrm{RSS}^{(p^*-n+1)/2} \Gamma\left(\frac{n-p^*}{2}\right) \times q^{p^*},$$

we have

$$\frac{r(M)}{r(M_0)} = \exp\{-T_1 - T_2\},$$

where

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

and

$$T_2 = \frac{h_n(m_0-m)}{2}\log(\pi \mathrm{RSS}_{M_o}) + \log\left\{\Gamma(\frac{n-h_n m_0}{2})/\Gamma(\frac{n-h_n m}{2})\right\} + h_n(m_0-m)\log(q).$$

Case 1: $M_0 \notin M$.

Let $\mathcal{M}_j = \{M : |M| = j, M \in \mathcal{M}\}$. Recall $\boldsymbol{H}_M$ is the projection matrix for model M and $\boldsymbol{H}_{M_0}$ is the projection matrix for the true model M_0. Calculate

$$\mathrm{RSS}_{M_0} = (y - \boldsymbol{Z}_{M_0}\boldsymbol{\beta}_{M_0})^T(\boldsymbol{I} - \boldsymbol{H}_{M_0})(y - \boldsymbol{Z}_{M_0}\boldsymbol{\beta}_{M_0})$$
$$= \boldsymbol{\varepsilon}^T(\boldsymbol{I} - \boldsymbol{H}_{M_0})\boldsymbol{\varepsilon}$$
$$= \sum_{i=1}^{n-h_n m_0} Z_i^2 = (n - h_n m_0)(1+o_p(1)) = n(1+o_p(1)),$$

where Z_i's are iid standard normal variables.

Let $\Delta(M) = \|\boldsymbol{\mu} - \boldsymbol{H}_M\boldsymbol{\mu}\|$ with $\boldsymbol{\mu} = \boldsymbol{Z}_{M_0}\boldsymbol{\beta}_{M_0}$. Then

$$\mathrm{RSS}_M - \mathrm{RSS}_{M_0} = (\boldsymbol{\mu}+\boldsymbol{\varepsilon})^T(\boldsymbol{I}-\boldsymbol{H}_M)(\boldsymbol{\mu}+\boldsymbol{\varepsilon}) - \boldsymbol{\varepsilon}^T(\boldsymbol{I}-\boldsymbol{H}_{M_0})\boldsymbol{\varepsilon}$$
$$= \Delta(M) + 2\boldsymbol{\mu}^T(\boldsymbol{I}-\boldsymbol{H}_M)\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T\boldsymbol{H}_M\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T\boldsymbol{H}_{M_0}\boldsymbol{\varepsilon}, \quad \text{(A.1)}$$

where $\boldsymbol{\varepsilon}^T\boldsymbol{H}_{M_0}\boldsymbol{\varepsilon} = h_n m_0(1+o_p(1))$.

Express the second term in (A.1) as

$$\boldsymbol{\mu}^T(\boldsymbol{I}-\boldsymbol{H}_M)\boldsymbol{\varepsilon} = \sqrt{\Delta(M)}Z_M,$$

where $Z_M \sim N(0,1)$. Then for any $M \in \mathcal{M}$,

$$|\boldsymbol{\mu}^T(\boldsymbol{I}-\boldsymbol{H}_M)\boldsymbol{\varepsilon}| \le \sqrt{\Delta(M)}\max_{\mathcal{M}}|Z_M|.$$

Let $c_j = 2j\{\log p + \log(j\log p)\}$, according to Lemma 3 we have

$$P(\max_{\mathcal{M}}|Z_M| \ge \sqrt{c}) = P(\max_{M\in\mathcal{M}_j, 1\le j\le km_0}|Z_M| \ge \sqrt{c})$$
$$\le \sum_{j=1}^{km_0}\binom{p}{j}P(\chi_1^2 \ge c)$$
$$\le \sum_{j=1}^{km_0}\binom{p}{j}P(\chi_j^2 \ge c) \to 0.$$

Therefore, $|\boldsymbol{\mu}^T(\boldsymbol{I}-\boldsymbol{H}_M)\boldsymbol{\varepsilon}| = \sqrt{\Delta(M)O_p(km_o\log p)}$ uniformly over $\mathcal{M}$.

Similarly, for the third term in (A.1), as $\boldsymbol{\varepsilon}^T\boldsymbol{H}_M\boldsymbol{\varepsilon} = \chi^2_{h_n m}$ we have

$$P(\max_{\mathcal{M}}\boldsymbol{\varepsilon}^T\boldsymbol{H}_M\boldsymbol{\varepsilon} \ge c_{h_n j}) = P(\max_{M\in\mathcal{M}_j,1\le j\le km_0}\chi^2_{h_n j} \ge c_{h_n j})$$
$$\le \sum_{j=1}^{km_0}\binom{p}{j}P(\chi^2_{h_n j} \ge c_{h_n j}) \to 0.$$

Thus, we have

$$\max_{\mathcal{M}}\{\boldsymbol{\varepsilon}^T\boldsymbol{H}_M\boldsymbol{\varepsilon}\} = O_p(kh_n m_o\log p).$$

Assuming that $h_n m_o \log p = o(n)$, we have

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

and

$$T_1 = \frac{n-h_n m-1}{2}\log\left(1+\frac{\mathrm{RSS}_M - \mathrm{RSS}_{M_0}}{\mathrm{RSS}_{M_0}}\right)$$
$$= \frac{n(1+o_p(1))}{2}\log\left\{1+\frac{\Delta(M)(1+o_p(1))}{n}\right\}$$
$$= \frac{\Delta(M)(1+o_p(1))}{2}. \quad \text{(A.2)}$$

By Sterling's formula,

$$\log\left\{\Gamma\left(\frac{n-h_n m_0}{2}\right)/\Gamma\left(\frac{n-h_n m}{2}\right)\right\} = \frac{h_n(m-m_0)}{2}\log n(1+o(1)).$$

Therefore,

$$T_2 = \frac{h_n(m-m_0)}{2}\left\{\log n(o_p(1)) - \log(\pi q^2)\right\}$$
$$\ge -\frac{h_n m_0}{2}\left\{\log n(o_p(1)) - \log(\pi q^2)\right\}. \quad \text{(A.3)}$$

Case 2: $M_0 \in M$.

Let $\mathcal{M}^*$ be the collection of models that contain the true model; that is, $\mathcal{M}^* = \{M \in \mathcal{M}, M_0 \in M, M \ne M_0\}$. Moreover, let $\mathcal{M}_j^* = \{M, |M| = j, M_0 \in M\}$.

When $M_0 \in M$, $(\boldsymbol{I}-\boldsymbol{H}_M)\boldsymbol{Z}_{M_0} = 0$, therefore, $y^T(\boldsymbol{I}-\boldsymbol{H}_M)y = \boldsymbol{\varepsilon}^T(\boldsymbol{I}-\boldsymbol{H}_M)\boldsymbol{\varepsilon}$. Also

$$\mathrm{RSS}_M - \mathrm{RSS}_{M_0} = \boldsymbol{\varepsilon}^T(\boldsymbol{I}-\boldsymbol{H}_{M_0})\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T(\boldsymbol{I}-\boldsymbol{H}_M)\boldsymbol{\varepsilon}$$
$$= \boldsymbol{\varepsilon}^T(\boldsymbol{H}_M - \boldsymbol{H}_{M_0})\boldsymbol{\varepsilon}$$
$$= \chi^2_{h_n(m-m_0)}(M),$$

where $\chi^2_{h_n(m-m_0)}(M)$ follows chi-square distribution with degrees of freedom $h_n(m-m_0)$.

Let $c_j = 2j\{\log p + \log(j \log p)\}$. According to Lemma 3,

$$P\left(\max_{M\in\mathcal{M}_j^*, 1\le j\le km_0-m_0} \chi^2_{h_nj}(M) \ge c_{h_nj}\right) = \sum_{j=1}^{km_0-m_0} P(\max_{M\in\mathcal{M}_j^*} \chi^2_{h_nj}(M) \ge c_{h_nj}) = \sum_{j=1}^{km_0-m_0} \binom{p-m_0}{j} P(\chi^2_{h_nj}(M) \ge c_{h_nj}) = \sum_{j=1}^{km_0-m_0} \binom{p}{j} P(\chi^2_{h_nj}(M) \ge c_{h_nj}) \to 0.$$

Therefore, $\chi^2_{h_n(m-m_0)}(M) \le c_{h_n(m-m_0)}(1+o_p(1))$ and

$$T_1 = \frac{n-h_nm-1}{2}\log\left(\frac{\text{RSS}_M}{\text{RSS}_{M_0}}\right) = -\frac{n-h_nm-1}{2}\log\left\{1+\frac{\chi^2_{h_n(m-m_0)}(M)}{\text{RSS}_{M_0}-\chi^2_{h_n(m-m_0)}(M)}\right\} \ge -\frac{n-h_nm-1}{2}\left\{\frac{\chi^2_{h_n(m-m_0)}(M)}{\text{RSS}_{M_0}-\chi^2_{h_n(m-m_0)}(M)}\right\}.$$

Since $n^{-1}\text{RSS}_{M_0} \to \sigma^2$ as $n\to\infty$, we have $\text{RSS}_{M_0} = n(1+o(1))$,

$$T_1 \ge \frac{c_{h_n(m-m_0)}}{2}(1+o_p(1)) \ge -h_n(m-m_0)\left[1+\frac{\log\{h_n(km_0-m_0)\log p\}}{\log p}\right]\log p(1+o_p(1)) \ge -h_n(m-m_0)(1+\delta)\log p(1+o_p(1)) \tag{A.4}$$

uniformly over $\mathcal{M}^*$, and

$$T_2 = \frac{h_n(m-m_0)}{2}\{\log n(o_p(1)) - \log(\pi q^2)\} \tag{A.5}$$

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

Combing case 1 and case 2, we aim to show that

$$\max_{M\notin M_0, M\in\mathcal{M}} \frac{r(M)}{r(M_0)} = \max\{\max_{M_0\notin M} \exp(-T_1-T_2), \max_{M_0\in M}\exp(-T_1-T_2)\} \to 0.$$

By (A.2) and (A.3), for case 1,

$$T_1+T_2 \ge \frac{\Delta(M)(1+o_p(1))}{2} - \frac{h_nm_0}{2}\left\{\log n(o_p(1)) - \log(\pi q^2)\right\} = \frac{h_nm_0\log p}{2}\left\{\frac{\Delta(M)(1+o_p(1))}{h_nm_0\log p} - \frac{\log n o_p(1)}{\log p} + \frac{\log(\pi q^2)}{\log p}\right\}.$$

In order that

$$\min_{M_0\notin M} T_1+T_2 \to \infty,$$

we can choose q such that $-\log q = O(\log p)$; that is,

$$-\frac{\log q}{\log p} = O(1).$$

Similarly by (A.4) and (A.5), for case 2,

$$T_1+T_2 \ge \frac{h_n(m-m_0)\log p}{2}\left\{\frac{\log n(o_p(1))}{\log p} - \frac{\log(\pi q^2)}{\log p} - 2(1+\delta)(1+o_p(1))\right\}.$$

In order that

$$\min_{M_0\in M} T_1+T_2 \to \infty,$$

we have

$$-\frac{\log q}{\log p} > 1+\delta.$$

Therefore, for $1+\delta < \gamma = -\frac{\log q}{\log p} < C$ with C being a constant, we have

$$\max_{M\notin M_0, M\in\mathcal{M}} \frac{r(M)}{r(M_0)} \to 0.$$

Moreover, if condition (A7) holds, we have

$$\sum_{M\ne M_0, M\in\mathcal{M}^*} \frac{r(M)}{r(M_0)} \le \sum_{j=1}^{km_0}\sum_{\mathcal{M}^*} \frac{r(M)}{r(M_0)} \le km_0 \max_{M\ne M_0, M\in\mathcal{M}} |M_j^*| \frac{r(M)}{r(M_0)} \to 0.$$

This completes the proof for Theorem 1.

Acknowledgments

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

Funding

This work was partially supported by the National Science Foundation grants DMS-1512945, DMS-1513484, DMS-1811405, DMS-1811661, and DMS-1916125.

References

Bühlmann, P. (2013), "Statistical Significance in High-Dimensional Linear Models," *Bernoulli*, 19, 1212–1242. [513]

Bühlmann, P., Kalisch, M., and Meier, L. (2014), "High-Dimensional Statistics With a View Toward Applications in Biology," *Annual Review of Statistics and Its Application*, 1, 255–278. [520]

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models," *The Annals of Statistics*, 17, 453–555. [513]

Chatterjee, A., and Lahiri, S. N. (2013), "Rates of Convergence of the Adaptive Lasso Estimators to the Oracle Distribution and Higher Order Refinements by the Bootstrap," *The Annals of Statistics*, 41, 1232–1259. [513]

Cisewski, J., and Hannig, J. (2012), "Generalized Fiducial Inference for Normal Linear Mixed Models," *The Annals of Statistics*, 40, 2102–2127. [514]

Dempster, A. P. (2008), "The Dempster–Shafer Calculus for Statisticians," *International Journal of Approximate Reasoning*, 48, 365–377. [514]

Fan, J., and Jiang, J. (2005), "Nonparametric Inference for Addive Models," *Journal of the American Statistical Association*, 100, 890–907. [513]

Fisher, R. A. (1930), "Inverse Probability," in *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 26), Cambridge: Cambridge University Press, pp. 528–535. [514]

Friedman, J. H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823. [513]

Hannig, J. (2009), "On Generalized Fiducial Inference," *Statistica Sinica*, 19, 491–544. [514,515]

Hannig, J., Iyer, H., Lai, R. C. S., and Lee, T. C. M. (2016), "Generalized Fiducial Inference: A Review and New Results," *Journal of the American Statistical Association*, 111, 1346–1361. [514,515,518]

Hannig, J., Lai, R. C. S., and Lee, T. C. M. (2014), "Computational Issues of Generalized Fiducial Inference," *Computational Statistics & Data Analysis*, 71, 849–858. [515]

Hannig, J., and Lee, T. C. M. (2009), "Generalized Fiducial Inference for Wavelet Regression," *Biometrika*, 96, 847–860. [514,515]

Horowitz, J. L., Klemela, J., and Mammen, E. (2006), "Optimal Estimation in Additive Regression Models," *Bernoulli*, 12, 271–298. [513]

Huang, J., Horowitz, J. L., and Wei, F. (2010), "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics*, 38, 2282–2313. [513,514,518,521]

Javanmard, A., and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869–2909. [513,520]

Lai, R. C. S., Hannig, J., and Lee, T. C. M. (2015), "Generalized Fiducial Inference for Ultrahigh-Dimensional Regression," *Journal of the American Statistical Association*, 110, 760–772. [514,515]

Lai, R. C. S., Huang, H.-C., and Lee, T. C. M. (2012), "Fixed and Random Effects Selection in Nonparametric Additive Mixed Models," *Electronic Journal of Statistics*, 6, 810–842. [516]

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016)," Exact Post-Selection Inference, With Application to the Lasso," *The Annals of Statistics*, 44, 907–927. [513]

Liu, Y., and Hannig, J. (2016), "Generalized Fiducial Inference for Binary Logistic Item Response Models," *Psychometrika*, 81, 290–324. [514]

Lopes, M. (2014), "A Residual Bootstrap for High-Dimensional Regression With Near Low-Rank Designs," in *Advances in Neural Information Processing Systems*, pp. 3239–3247. [513]

Lu, J., Kolar, M., and Liu, H. (2015), "Kernel Meets Sieve: Post-Regularization Confidence Bands for High Dimensional Nonparametric Models With Local Sparsity," arXiv no. 1503.02978. [514,520]

Luo, S., and Chen, Z. (2013), "Extended BIC for Linear Regression Models With Diverging Number of Relavant Features and High or Ultra-High Feature Spaces," *Journal of Statistical Planning and Inference*, 143, 494–504. [515,518,521]

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. [514]

——— (2015), *Inferential Models: Reasoning With Uncertainty* (Vol. 145), Boca Raton, FL: CRC Press. [514]

Martin, R., Mess, R., and Walker, S. G. (2017), "Empirical Bayes Posterior Concentration in Sparse High-Dimensional Linear Models," *Bernoulli*, 23, 1822–1847. [513]

Martin, R., Zhang, J., and Liu, C. (2010), "Dempster-Shafer Theory and Statistical Inference With Weak Beliefs," *Statistical Science*, 25, 72–87. [514]

Meier, L., Van De Geer, S., and Buhlmann, P. (2009), "High Dimensional Additive Modeling," *The Annals of Statistics*, 37, 3779–3821. [513]

Nardi, Y., and Rinaldo, A. (2008), "On the Asymptotic Properties of the Group Lasso Estimator for Linear Models," *Electronic Journal of Statistics*, 2, 605–633. [518]

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009), "Sparse Additive Models," *Journal of the Royal Statistical Society*, Series B, 71, 1009–1030. [513]

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012), "Spike-and-Slab Priors for Function Selection in Structured Additive Regression Models," *Journal of the American Statistical Association*, 107, 1518–1532. [514]

Shang, Z., and Li, P. (2014), "High-Dimensional Bayesian Inference in Nonparametric Additive Models," *Electronic Journal of Statistics*, 8, 2804–2847. [514]

Stone, C. J. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 689–705. [513,514]

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016), "Exact Post-Selection Inference for Sequential Regression Procedures," *Journal of the American Statistical Association*, 111, 600–620. [513]

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [513]

Xie, M.-G., and Singh, K. (2013), "Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review," *International Statistical Review*, 81, 3–39. [514]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [516]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [513]