Biometrika (2020), **???**, ?, *pp.* 1–21
Printed in Great Britain Advance Access publication on ?? ?? 20??

Large-scale model selection in misspecified generalized linear models

BY EMRE DEMIRKAYA
Department of Business Analytics and Statistics, University of Tennessee, Knoxville, Tennessee, 37996, U.S.A.
edemirka@utk.edu

YANG FENG
Department of Biostatistics, School of Global Public Health, New York University, 715 Broadway, New York, New York, 10003, U.S.A.
yang.feng@nyu.edu

PALLAVI BASU
Indian School of Business, Gachibowli, Hyderabad Telangana, India
plvibasu.work@gmail.com

AND JINCHI LV
Data Sciences and Operations Department, University of Southern California, Los Angeles, California, 90089, U.S.A.
jinchilv@marshall.usc.edu

SUMMARY

Model selection is crucial to high-dimensional learning and inference for contemporary big data applications in pinpointing the best set of covariates among a sequence of candidate interpretable models. Most existing work assumes implicitly that the models are correctly specified or have fixed dimensionality. Yet both features of model misspecification and high dimensionality are prevalent in practice. In this paper, we exploit the framework of model selection principles under the misspecified generalized linear models presented in Lv and Liu (2014) and investigate the asymptotic expansion of the posterior model probability in the setting of high-dimensional misspecified models. With a natural choice of prior probabilities that encourages interpretability and incorporates the Kullback–Leibler divergence, we suggest the high-dimensional generalized Bayesian information criterion with prior probability for large-scale model selection with misspecification. Our new information criterion characterizes the impacts of both model misspecification and high dimensionality on model selection. We further establish the consistency of covariance contrast matrix estimation and the model selection consistency of the new information criterion in ultra-high dimensions under some mild regularity conditions. The numerical studies demonstrate that our new method enjoys improved model selection consistency compared to its main competitors.

Some key words: Model misspecification; High dimensionality; Big data; Model selection; Bayesian principle; Kullback–Leibler divergence; GBIC$_p$; GIC; Robustness

1. INTRODUCTION

With rapid advances of modern technology, big data of unprecedented size, such as genetic and proteomic data, fMRI and functional data, and panel data in economics and finance, are frequently encountered in many contemporary applications. In these applications, the dimensionality p can be comparable to or even much larger than the sample size n. A key assumption that often makes large-scale learning and inference feasible is the sparsity of signals, meaning that only a small fraction of covariates contribute to the response when p is large compared to n. High-dimensional modeling with dimensionality reduction and feature selection plays an important role in these problems, e.g., Fan & Lv (2010); Bühlmann & van de Geer (2011); Fan & Lv (2018). A sparse modeling procedure typically produces a sequence of interpretable candidate models, each involving a possibly different subset of covariates. An important question is how to compare different models in high dimensions when models are possibly misspecified.

The problem of model selection has been studied extensively by many researchers in the past several decades. Among others, well-known model selection criteria include the Akaike information criterion (AIC) (Akaike, 1973, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978), where the former is based on the Kullback–Leibler divergence principle of model selection and the latter is originated from the Bayesian principle of model selection. A great deal of work has been devoted to understanding and extending these model selection criteria to different model settings; see, for example, Bozdogan (1987); Foster & George (1994); Konishi & Kitagawa (1996); Ing (2007); Chen & Chen (2008); Chen & Chan (2011); Liu & Yang (2011); Ninomiya & Kawano (2016); Eguchi (2017); Hsu et al. (2019). Fong & Holmes (2020) studied the links between the cross-validation and Bayesian model selection. The connections between the Akaike information criterion and cross-validation have been investigated in Stone (1977); Hall (1990); Peng et al. (2013) for various contexts. In particular, Fan & Tang (2013) showed that classical information criteria such as Akaike information criterion and Bayesian information criterion can no longer be consistent for model selection in ultra-high dimensions and proposed the generalized information criterion (GIC) for tuning parameter selection in high-dimensional penalized likelihood, for the scenario of correctly specified models. See also Barber & Candès (2015); Bühlmann & van de Geer (2015); Candès et al. (2018); Shah & Bühlmann (2018); Fan et al. (2020, 2019) for some recent work on high-dimensional inference for feature selection.

Most existing work on model selection and feature selection usually make an implicit assumption that the model under study is correctly specified or of fixed dimensions. Given the practical importance of model misspecification, White (1982) laid out a general theory of maximum likelihood estimation in misspecified models for the case of fixed dimensionality and independent and identically distributed observations. Cule et al. (2010) also studied the maximum likelihood estimation of a multi-dimensional log-concave density when the model is misspecified. Recently, Lv & Liu (2014) investigated the problem of model selection with model misspecification and originated asymptotic expansions of both Kullback–Leibler divergence and Bayesian principles in misspecified generalized linear models, leading to the generalized Akaike information criterion (GAIC) and generalized Bayesian information criterion (GBIC), for the case of fixed dimensionality. A specific form of prior probabilities motivated by the Kullback–Leibler divergence principle led to the generalized Bayesian information criterion with prior probability (GBIC$_p$). Yet both features of model misspecification and high dimensionality are prevalent in contemporary big data applications. Thus an important question is how to characterize the impacts of both model misspecification and high dimensionality on model selection. We intend to provide some partial answers to this question in the paper.

Let us first gain some insights into the challenges of the aforementioned problem by considering a motivating example. Assume that the response Y depends on the covariate vector $(X_1, \ldots, X_p)^T$ through the functional form $Y = f(X_1) + f(X_2 - X_3) + f(X_4 - X_5) + \varepsilon$, where $f(x) = x^3/(x^2 + 1)$ and the remaining setting is the same as in Section 4.2. Consider sample size $n = 200$ and vary dimensionality p from 100 to 3200. Without any prior knowledge of the true model structure, we take the linear regression model

$$y = Z\beta + \varepsilon \tag{1}$$

as the working model and apply some information criteria to hopefully recover the oracle working model, where y is an n-dimensional response vector, Z is an $n \times p$ design matrix, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a p-dimensional regression coefficient vector, and ε is an n-dimensional error vector. Following Candès et al. (2018) , we define the oracle working model $\mathfrak{M}_0$ as the Markov blanket for Y, that is, $\mathfrak{M}_0$ is the smallest subset of indices such that Y is independent of $X_{\mathfrak{M}_0^c}$ conditional on $X_{\mathfrak{M}_0}$; see Lauritzen (1996) and Pearl (2014). In this example, the oracle working model consists of the first five covariates. When $p = 100$, the traditional Akaike information criterion and Bayesian information criterion, which ignore model misspecification, tend to select a model with size larger than five. In contrast, GBIC_p in Lv & Liu (2014) selects the oracle working model around 60% of the time. However, when p is increased to 3200, these methods fail to select such a model with significant probability and the prediction performance of the selected models deteriorates. This motivates us to study the problem of model selection in high-dimensional misspecified models. In contrast, our new method can recover the oracle working model with significant probability in this challenging scenario.

The main contributions of our paper are threefold. First, we provide the asymptotic expansion of the posterior model probability in high-dimensional misspecified generalized linear models, which involves delicate and challenging technical analysis. Motivated by the asymptotic expansion and a natural choice of prior probabilities that encourages interpretability and incorporates Kullback–Leibler divergence, we suggest the high-dimensional generalized Bayesian information criterion with prior probability (HGBIC_p) for large-scale model selection with misspecification. Second, our work provides rigorous theoretical justification of the covariance contrast matrix estimator that incorporates the effect of model misspecification and is crucial for practical implementation. Such an estimator is shown to be consistent in the general setting of high-dimensional misspecified models. Third, we establish the model selection consistency of our new information criterion HGBIC_p in ultra-high dimensions under some mild regularity conditions. In particular, our work provides important extensions to the studies in Lv & Liu (2014) and Fan & Tang (2013) to the cases of high dimensionality and model misspecification, respectively. The aforementioned contributions make our work distinct from other studies on model misspecification including Bühlmann & van de Geer (2015); Hsu et al. (2019); Shah & Bühlmann (2018). Since Lv & Liu (2014) is closely related to our paper, we want to reiterate the main differences between these two works. First, the study in Lv & Liu (2014) has focused on fixed dimensionality. Hence, our model selection criterion differs from that in Lv & Liu (2014) in how it penalizes the model complexity as discussed in Section 2.2. Although both criteria rely on the estimation of the covariance contrast matrix, the consistency result of the covariance contrast matrix estimator in Lv & Liu (2014) does not allow model misspecification. We establish the consistency of the estimator for the covariance contrast matrix even under model misspecification in Section 3.3. Finally, in light of the new consistency result, we further provide a model selection consistency theorem for our model selection criterion, which result was missing in Lv & Liu (2014).

2. LARGE-SCALE MODEL SELECTION WITH MISSPECIFICATION

2.1. *Model misspecification*

The main focus of this paper is investigating ultra-high dimensional model selection with model misspecification in which the dimensionality p can grow nonpolynomially with sample size n. Let Z be the $n \times p$ design matrix with all available covariates. We denote by $\mathfrak{M}$ an arbitrary subset with size d of all p available covariates and $X = (x_1, \ldots, x_n)^T$ the corresponding $n \times d$ fixed design matrix given by the covariates in model $\mathfrak{M}$. Assume that conditional on the covariates in model $\mathfrak{M}$, the response vector $Y = (Y_1, \ldots, Y_n)^T$ has independent components and each Y_i follows distribution $G_{n,i}$ with density $g_{n,i}$, with all the distributions $G_{n,i}$ unknown to us in practice. Denote by $g_n = \prod_{i=1}^n g_{n,i}$ the product density and G_n the corresponding *true* distribution of the response vector Y.

Since the collection of true distributions $\{G_{n,i}\}_{1 \le i \le n}$ is unknown to practitioners, one often chooses a family of working models to fit the data. One class of popular working models is the family of the generalized linear models McCullagh & Nelder (1989) with a canonical link and natural parameter vector $\theta = (\theta_1, \ldots, \theta_n)^T$ with $\theta_i = x_i^T \beta$, where x_i is a d-dimensional covariate vector and $\beta = (\beta_1, \ldots, \beta_d)^T$ is a d-dimensional regression coefficient vector. Let $\tau > 0$ be the dispersion parameter. Then under the working model of the generalized linear models, the conditional density of response y_i given the covariates in model $\mathfrak{M}$ is assumed to take the form

$$f_{n,i}(y_i) = \exp\{y_i \theta_i - b(\theta_i) + c(y_i, \tau)\}, \tag{2}$$

where $b(\cdot)$ and $c(\cdot, \cdot)$ are some known functions with $b(\cdot)$ twice continuously differentiable and $b''(\cdot)$ bounded away from 0 and ∞. F_n denotes the corresponding distribution of the n-dimensional response vector $y = (y_1, \ldots, y_n)^T$ with the product density $f_n = \prod_{i=1}^n f_{n,i}$ assuming the independence of components. Since the generalized linear model is chosen by the user, the *working* distribution F_n can be generally different from the true unknown distribution G_n.

For the generalized linear models in (2) with natural parameter vector θ, let us define two vector-valued functions $b(\theta) = \{b(\theta_1), \ldots, b(\theta_n)\}^T$ and $\mu(\theta) = \{b'(\theta_1), \ldots, b'(\theta_n)\}^T$, and a matrix-valued function $\Sigma(\theta) = \mathrm{diag}\{b''(\theta_1), \ldots, b''(\theta_n)\}$. The basic properties of the generalized linear models give the mean vector $E(y) = \mu(\theta)$ and the covariance matrix $\mathrm{cov}(y) = \Sigma(\theta)$ with $\theta = X\beta$. The product density of the response vector y can be written as

$$f_n(y; \beta, \tau) = \prod_{i=1}^n f_{n,i}(y_i) = \exp\{y^T X\beta - 1^T b(X\beta) + \sum_{i=1}^n c(y_i, \tau)\}, \tag{3}$$

where 1 represents the n-dimensional vector with all components being one. Since the generalized linear models is only our working model, (3) results in the quasi-log-likelihood function White (1982)

$$\ell_n(y; \beta, \tau) = \log f_n(y; \beta, \tau) = y^T X\beta - 1^T b(X\beta) + \sum_{i=1}^n c(y_i, \tau). \tag{4}$$

Hereafter we treat the dispersion parameter τ as a known parameter and focus on our main parameter of interest β. Whenever there is no confusion, we will slightly abuse the notation and drop the functional dependence on τ.

The quasi-maximum likelihood estimator for the parameter vector β in our working model of the generalized linear models (2) is defined as $\widehat{\beta}_n = \arg\max_{\beta \in \mathbb{R}^d} \ell_n(y, \beta)$, which is the solution to the score equation

$$\Psi_n(\beta) = \partial \ell_n(y, \beta)/\partial \beta = X^T\{y - \mu(X\beta)\} = 0. \quad (5)$$

For the linear regression model with $\mu(X\beta) = X\beta$, such a score equation becomes the familiar normal equation $X^T y = X^T X\beta$. Such a vector β is called quasi-maximum likelihood estimator when the model is misspecified. Hereafter, we call β maximum likelihood estimator for simplicity since we do not know whether the model is misspecified or not in practice. The Kullback–Leibler divergence (Kullback & Leibler, 1951) of our working model F_n from the true model G_n is defined as $I\{g_n; f_n(\cdot, \beta)\} = E\{\log g_n(Y)\} - E\{\ell_n(Y, \beta)\}$ with the response vector Y following the true distribution G_n. As in Lv & Liu (2014), we consider the best working model that is closest to the true model under the Kullback–Leibler divergence. Such a model has parameter vector $\beta_{n,0} = \arg\min_{\beta \in \mathbb{R}^d} I\{g_n; f_n(\cdot, \beta)\}$, which solves the equation

$$X^T\{E(Y) - \mu(X\beta)\} = 0. \quad (6)$$

We see that equation (6) is simply the population version of the score equation given in (5).

Following Lv & Liu (2014), we introduce two matrices, Fisher information in outer product form and in Hessian form. These matrices play a key role in model selection with model misspecification. Under the true distribution G_n, we have $\operatorname{cov}\left(X^T Y\right) = X^T \operatorname{cov}(Y) X$. Computing the score equation at $\beta_{n,0}$, Fisher information matrix in outer product form is defined by

$$B_n = \operatorname{cov}\{\Psi_n(\beta_{n,0})\} = \operatorname{cov}\left(X^T Y\right) = X^T \operatorname{cov}(Y) X \quad (7)$$

with $\operatorname{cov}(Y) = \operatorname{diag}\{\operatorname{var}(Y_1), \ldots, \operatorname{var}(Y_n)\}$ by the independence assumption and under the true model. Under the working model F_n, it holds that $\operatorname{cov}\left(X^T Y\right) = X^T \Sigma(X\beta) X$. The Fisher information matrix in Hessian form is defined by

$$A_n(\beta) = \frac{\partial^2 I\{g_n; f_n(\cdot, \beta)\}}{\partial \beta^2} = -E\left\{\frac{\partial^2 \ell_n(Y, \beta)}{\partial \beta^2}\right\} = X^T \Sigma(X\beta) X, \quad (8)$$

and denote by $A_n = A_n(\beta_{n,0})$. Hence we see that matrices A_n and B_n are the covariance matrices of $X^T Y$ under the best working model $F_n(\beta_{n,0})$ and the true model G_n, respectively. To account for the effect of model misspecification, we define the covariance contrast matrix $H_n = A_n^{-1} B_n$ as revealed in Lv & Liu (2014). Observe that A_n and B_n coincide when the best working model and the true model are the same. In this case, H_n is an identity matrix of size d.

2.2. *High-dimensional generalized Bayesian information criterion with prior probability*

Given a set of competing models $\{\mathfrak{M}_m : m = 1, \ldots, M\}$, a popular model selection procedure using Bayesian principle of model selection is to first put nonzero prior probability $\alpha_{\mathfrak{M}_m}$ on each model $\mathfrak{M}_m$, and then choose a prior distribution $\mu_{\mathfrak{M}_m}$ for the parameter vector in the corresponding model. We use $d_m = |\mathfrak{M}_m|$ to denote the dimensionality of candidate model $\mathfrak{M}_m$ and suppress the subscript m for conciseness whenever there is no confusion. Assume that the density function of $\mu_{\mathfrak{M}_m}$ is bounded in $\mathbb{R}^{\mathfrak{M}_m} = \mathbb{R}^{d_m}$ and locally bounded away from zero in a shrinking neighborhood of $\beta_{n,0}$. The Bayesian principle of model selection is to choose the most probable model *a posteriori*; that is, choose the model $\mathfrak{M}_{m_0}$ such that $m_0 = \arg\max_{m \in \{1,\ldots,M\}} S(y, \mathfrak{M}_m; F_n)$, where

$$S(y, \mathfrak{M}_m; F_n) = \log \int \alpha_{\mathfrak{M}_m} \exp\{\ell_n(y, \beta)\} d\mu_{\mathfrak{M}_m}(\beta) \quad (9)$$

with the log-likelihood $\ell_n(y, \beta)$ as defined in (4) and the integral over $\mathbb{R}^{d_m}$.

The choice of prior probabilities $\alpha_{\mathfrak{M}_m}$ is important in high dimensions. Lv & Liu (2014) suggested the use of prior probability $\alpha_{\mathfrak{M}_m} \propto e^{-D_m}$ for each candidate model $\mathfrak{M}_m$, where the quantity D_m is defined as $D_m = E[I\{g_n; f_n(\cdot, \widehat{\beta}_{n,m})\} - I\{g_n; f_n(\cdot, \beta_{n,m,0})\}]$ with the subscript m indicating a particular candidate model. The motivation is that the further the maximum likelihood estimator $\widehat{\beta}_{n,m}$ is away from the best misspecified generalized linear models $F_n(\cdot, \beta_{n,m,0})$, the lower prior probability we assign to that model. In the high-dimensional setting when dimensionality p can be much larger than sample size n, it is sensible to also take into account the complexity of the space of all possible sparse models with the same size as $\mathfrak{M}_m$. Such an observation motivates us to consider a new prior probability of the form

$$\alpha_{\mathfrak{M}_m} \propto p^{-d} e^{-D_m} \tag{10}$$

with $d = |\mathfrak{M}_m|$. The complexity factor p^{-d} is motivated by the asymptotic expansion of $\{p!/(p-d)!\}^{-1}$. In fact, an application of Stirling's formula yields $\log\{p!/(p-d)!\}^{-1} \approx -d \log p = \log(p^{-d})$ up to an additive term of order $o(d)$ when $d = o(p)$. The factor of $[p!/\{(p-d)!d!\}]^{-1}$ was also exploited in Chen & Chen (2008) who showed that using the term $[p!/\{(p-d)!d!\}]^{-\gamma}$ with some constant $0 < \gamma \leq 1$, the extended Bayesian information criterion can be model selection consistent for the scenario of correctly specified models with $p = O(n^{\kappa})$ for some positive constant κ satisfying $1 - (2\kappa)^{-1} < \gamma$. A different way of integrating the number of candidate models into the prior was considered in Szulc (2012) when the model under study is correctly specified. Moreover, we add the term $d!$ to reflect a stronger prior on model sparsity. See also Fan & Tang (2013) for the characterization of model selection in ultra-high dimensions with correctly specified models.

A similar normalization term can also be found in some fully Bayesian methods; see, e.g., Castillo et al. (2015) for more details. However, the fully Bayesian methods need to specify the distribution of parameter β, whereas our method only puts some prior probabilities on the candidate models $\mathfrak{M}_m$, and the distribution $\mu_{\mathfrak{M}_m}(\beta)$ of parameter β given model $\mathfrak{M}_m$ does not need to be specified. Furthermore, fully Bayesian approaches require posterior computation, which may limit their use in high dimensions; see, e.g., George (2000).

The asymptotic expansion of the posterior model probability in Theorem 1 to be presented in Section 3.2 motivates us to introduce the high-dimensional generalized Bayesian information criterion with prior probability (HGBIC_p) for large-scale model selection with misspecification.

DEFINITION 1. *We define* $\text{HGBIC}_p = \text{HGBIC}_p(y, \mathfrak{M}_m; F_n)$ *of model* $\mathfrak{M}_m$ *as*

$$\text{HGBIC}_p = -2\ell_n(y, \widehat{\beta}_n) + 2(\log p^*)|\mathfrak{M}_m| + \text{tr}(\widehat{H}_n) - \log|\widehat{H}_n|, \tag{11}$$

where $\widehat{H}_n$ *is a consistent estimator of* H_n *and* $p^* = pn^{1/2}$. *Here, consistency is in terms of trace and log determinant of the matrix.*

In correctly specified models, $H_n = A_n^{-1} B_n = I_d$ and so the term $\text{tr}(\widehat{H}_n) - \log|\widehat{H}_n|$ in (11) is asymptotically close to $|\mathfrak{M}_m|$ when $\widehat{H}_n$ is a consistent estimator of H_n. Thus compared to the Bayesian information criterion with factor $\log n$, the HGBIC_p contains a larger factor of order $\log p$ when dimensionality p grows nonpolynomially with sample size n. This leads to a heavier penalty on model complexity, similarly to that in Fan & Tang (2013).

As shown in Lv & Liu (2014), the HGBIC_p defined in (11) can also be viewed as a sum of three terms: the goodness of fit, model complexity, and model misspecification; see Lv & Liu (2014) for more details. Furthermore, HGBIC_p is also related to Takeuchi's information criterion $\text{TIC} = -2\ell_n(y, \widehat{\beta}_n) + 2\text{tr}(\widehat{H}_n)$ in Takeuchi (1976), which contains similar model misspecification term $\text{tr}(\widehat{H}_n)$, but lacks any model complexity term.

Our new information criterion HGBIC_p provides an important extension of the original model selection criterion $\text{GBIC}_p = -2\ell_n(y, \widehat{\beta}_n) + (\log n)|\mathfrak{M}_m| + \text{tr}(\widehat{H}_n) - \log|\widehat{H}_n|$ in Lv & Liu (2014), which was proposed for the scenario of model misspecification with fixed dimensionality, by explicitly taking into account the high dimensionality of the whole feature space. Moreover, in view of (11) and the definition of p^*, HGBIC_p has an additional model complexity term $2(\log p)|\mathfrak{M}_m|$.

3. Asymptotic properties of HGBIC_p

3.1. *Technical assumptions*

We list the technical assumptions required to prove the main results and the asymptotic properties of the maximum likelihood estimator with diverging dimensionality. Denote by Z the full design matrix of size $n \times p$ whose (i, j)th entry is x_{ij}. For any subset $\mathfrak{M}_m$ of $\{1, \ldots, p\}$, $Z_{\mathfrak{M}_m}$ denotes the submatrix of Z formed by columns whose indices are in $\mathfrak{M}_m$. When there is no confusion, we drop the subscript and use $X = Z_{\mathfrak{M}_m}$ for fixed $\mathfrak{M}$. For theoretical reasons, we restrict the parameter space to $\mathcal{B}_0$ which is a sufficiently large convex and compact set of $\mathbb{R}^p$. We consider parameters with bounded support. Namely, we define $\mathcal{B}(\mathfrak{M}_m) = \{\beta \in \mathcal{B}_0 : \text{supp}(\beta) = \mathfrak{M}_m\}$ and $\mathcal{B} = \cup_{|\mathfrak{M}_m| \leq K} \mathcal{B}(\mathfrak{M}_m)$ where the maximum support size K is taken to be $o(n)$. Moreover, we assume that $c_0 \leq b''(Z\beta) \leq c_0^{-1}$ for any $\beta \in \mathcal{B}$ where c_0 is some positive constant.

We use the following notation. For matrices, $\|\cdot\|_2$, $\|\cdot\|_\infty$, and $\|\cdot\|_F$ denote the matrix operator norm, entrywise maximum norm, and matrix Frobenius norm, respectively. For vectors, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ denote the vector L_2-norm and maximum norm, and $(v)_i$ represents the ith component of vector v. Denote by $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ the smallest and largest eigenvalues of a given matrix, respectively.

Assumption 1. There exists some positive constant c_1 such that for each $i = 1, \ldots, n$, $pr(|W_i| > t) \leq c_1 \exp(-c_1^{-1}t)$ for any $t > 0$, where $W = (W_1, \ldots, W_n)^T = Y - E(Y)$. The variances of Y_i are bounded below uniformly in i and n.

Assumption 2. Let u_1 and u_2 be some positive constants and $\widetilde{m}_n = O(n^{u_1})$ a diverging sequence. We have the following bounds $\max\{\|E(Y)\|_\infty, \sup_{\beta \in \mathcal{B}} \|\mu(Z\beta)\|_\infty\} \leq \widetilde{m}_n$, and $\sum_{i=1}^n \left(\frac{[E(Y_i) - \{\mu(X\beta_{n,0})\}_i]^2}{\text{var}(Y_i)}\right)^2 = O(n^{u_2})$. For simplicity, we also assume that $\widetilde{m}_n$ diverges faster than $\log n$.

Assumption 3. Let $K = o(n)$ be a positive integer. There exist positive constants c_2 and u_3 such that, for any $\mathfrak{M}_m \subset \{1, \ldots, p\}$ such that $|\mathfrak{M}_m| \leq K$, $c_2 \leq \lambda_{\min}(n^{-1} Z_{\mathfrak{M}_m}^T Z_{\mathfrak{M}_m}) \leq \lambda_{\max}(n^{-1} Z_{\mathfrak{M}_m}^T Z_{\mathfrak{M}_m}) \leq c_2^{-1}$, and $\|Z\|_\infty = O(n^{u_3})$. For simplicity, we assume that columns of Z are normalized: $\sum_{i=1}^n x_{ij}^2 = n$ for all $j = 1, \ldots, p$.

Assumption 1 is a standard tail assumption on the response variable Y. This assumption ensures that the sub-exponential norm of the response is bounded. Assumptions 2 and 3 have their counterparts in Fan & Tang (2013). However, Assumption 2 is modified to deal with model misspecification. More specifically, the means of the true distribution and fitted model, as well as their relations, are assumed in Assumption 2. The first part simultaneously controls the tail behavior of the response and fitted model. The second part ensures that the mean of the fitted distribution does not deviate from the true mean too significantly. We would like to point out that such an assumption does not limit the generality of model misspecification since the misspecification considered in the paper is due to the distributional mismatch between the working model and the underlying true model. Even in the misspecified scenario, the fitted mean vector from the working model can approximate the true mean vector under certain regularity conditions. Assumption 3 is on the design matrix X. The first part is important for the consistency of the maximum likelihood estimator $\widehat{\beta}_n$ and the uniqueness of the population parameter. Assumptions 2 and 3 also provide bounds for the eigenvalues of $A_n(\beta)$ and B_n. See Fan & Tang (2013) for further discussions on these assumptions.

For the following assumptions, we define a neighborhood around $\beta_{n,0}$. Let $\delta_n = \widetilde{m}_n(\log p)^{1/2} = O\{n^{u_1}(\log p)^{1/2}\}$. We define the neighborhood $N_n(\delta_n) = \{\beta \in \mathbb{R}^d : \|(n^{-1}B_n)^{1/2}(\beta - \beta_{n,0})\|_2 \leq (n/d)^{-1/2}\delta_n\}$. We assume that $(n/d)^{-1/2}\delta_n$ converges to zero so that $N_n(\delta_n)$ is an asymptotically shrinking neighborhood of $\beta_{n,0}$.

Assumption 4. Assume that the prior density relative to the Lebesgue measure μ_0 on $\mathbb{R}^d$, $\pi\{h(\beta)\} = d\mu_{\mathfrak{M}_m}/d\mu_0\{h(\beta)\}$ satisfies $\inf_{\beta \in N_n(2\delta_n)} \pi\{h(\beta)\} \geq c_3$ and $\sup_{\beta \in \mathbb{R}^d} \pi\{h(\beta)\} \leq c_3^{-1}$, where c_3 is a positive constant, and $h(\beta) = (n^{-1}B_n)^{1/2}\beta$.

Assumption 5. Let $V_n(\beta) = B_n^{-1/2}A_n(\beta)B_n^{-1/2}$, $V_n = V_n(\beta_{n,0}) = B_n^{-1/2}A_nB_n^{-1/2}$, and $\widetilde{V}_n(\beta_1, \ldots, \beta_d) = B_n^{-1/2}\widetilde{A}_n(\beta_1, \ldots, \beta_d)B_n^{-1/2}$, where $\widetilde{A}_n(\beta_1, \ldots, \beta_d)$ is the matrix whose jth row is the corresponding row of $A_n(\beta_j)$ for each $j = 1, \ldots, d$. There exists some sequence $\rho_n(\delta_n)$ such that $\rho_n(\delta_n)\delta_n^2 d$ converges to zero, $\max_{\beta_1,\ldots,\beta_d \in N_n(\delta_n)} \|\widetilde{V}_n(\beta_1, \ldots, \beta_d) - V_n\|_2 \leq \rho_n(\delta_n)$, and $\max_{\beta \in N_n(2\delta_n)} \max[|\lambda_{\min}\{V_n(\beta) - V_n\}|, |\lambda_{\max}\{V_n(\beta) - V_n\}|] \leq \rho_n(\delta_n)$.

Similar versions of Assumptions 4 and 5 were imposed in Lv & Liu (2014). Under Assumption 4, the prior density is bounded above globally and bounded below in a neighborhood of $\beta_{n,0}$. This assumption is used in Theorem 1 for the asymptotic expansion of the posterior model probability. Assumption 5 is on the continuity of the matrix-valued function V_n and $\widetilde{V}_n$ in a shrinking neighborhood $N_n(2\delta_n)$ of $\beta_{n,0}$. The first and second parts control the expansions of expected log-likelihood and score functions, respectively. Assumption 5 ensures that the remainders are negligible in approximating $S(y, \mathfrak{M}_m; F_n)$. Some detailed discussion on Assumption 5 is provided in Section D of the Supplementary Material. See also Lv & Liu (2014) for more discussions on these assumptions.

3.2. *Asymptotic expansion of the Bayesian principle of model selection*

We now provide the asymptotic expansion of the posterior model probability with the prior introduced in Section 2.2. As mentioned earlier, the Bayesian principle chooses the model that maximizes $S(y, \mathfrak{M}_m; F_n)$ given in (9). To ease the presentation, for any $\beta \in \mathbb{R}^d$, we define a quantity

$$\ell_n^*(y, \beta) = \ell_n(y, \beta) - \ell_n(y, \widehat{\beta}_n), \tag{12}$$

which is the deviation of the quasi-log-likelihood from its maximum. Then from (9) and (12), we have

$$S(y, \mathfrak{M}_m; F_n) = \ell_n(y, \widehat{\beta}_n) + \log E_{\mu_{\mathfrak{M}_m}}\{U_n(\beta)^n\} + \log \alpha_{\mathfrak{M}_m}, \tag{13}$$

where $U_n(\beta) = \exp\{n^{-1}\ell_n^*(y, \beta)\}$. With the choice of the prior probability in (10), it is clear that

$$\log \alpha_{\mathfrak{M}_m} = -D_m - d\log p. \tag{14}$$

Aided by (13) and (14), some delicate technical analysis unveils the following expansion of $S(y, \mathfrak{M}_m; F_n)$.

THEOREM 1. *Assume that Assumptions* 1–5 *hold and let* $\alpha_{\mathfrak{M}_m} = Cp^{-d}e^{-D_m}$ *with* $C > 0$ *some normalization constant. If* $(n/d)^{-1/2}\delta_n = o(1)$*, then we have with probability tending to one,*

$$S(y, \mathfrak{M}_m; F_n) = \ell_n(y, \widehat{\beta}_n) - (\log p^*)|\mathfrak{M}_m| - \frac{1}{2}\mathrm{tr}(H_n) + \frac{1}{2}\log|H_n| + \log(Cc_4) + o(\widetilde{\mu}_n), \tag{15}$$

where $H_n = A_n^{-1}B_n$*,* $p^* = pn^{1/2}$*,* $\widetilde{\mu}_n = max\{\mathrm{tr}(A_n^{-1}B_n), 1\}$*,* $c_3 \leq c_4 \leq c_3^{-1}$*, and* c_3 *is the positive constant given in Assumption* 4.

Theorem 1 lays the foundation for investigating high-dimensional model selection with model misspecification. Based on the asymptotic expansion in (15), our new information criterion HGBIC$_p$ in (11) is defined by replacing the covariance contrast matrix H_n with a consistent estimator $\widehat{H}_n$. The HGBIC$_p$ naturally characterizes the impacts of both model misspecification and high dimensionality on model selection. A natural question is how to ensure a consistent estimator for H_n. We address such a question in the next section.

3.3. *Consistency of covariance contrast matrix estimation*

For practical implementation of HGBIC$_p$, it is of vital importance to provide a consistent estimator for the covariance contrast matrix H_n. To this end, we consider the plug-in estimator $\widehat{H}_n = \widehat{A}_n^{-1}\widehat{B}_n$ with $\widehat{A}_n$ and $\widehat{B}_n$ defined as follows. Since the maximum likelihood estimator $\widehat{\beta}_n$ provides a consistent estimator of $\beta_{n,0}$ in the best misspecified generalized linear models $F_n(\cdot, \beta_{n,0})$, a natural estimate of matrix A_n is

$$\widehat{A}_n = A_n(\widehat{\beta}_n) = X^T\Sigma(X\widehat{\beta}_n)X.$$

When the model is correctly specified, the following simple estimator

$$\widehat{B}_n = X^T\mathrm{diag}\left[\left\{y - \mu(X\widehat{\beta}_n)\right\} \circ \left\{y - \mu(X\widehat{\beta}_n)\right\}\right] X$$

with $\circ$ denoting the componentwise product gives an asymptotically unbiased estimator of the matrix B_n.

THEOREM 2. *Assume that Assumptions* 1–3 *hold,* $n^{-1}A_n(\beta)$ *is Lipschitz in operator norm in the neighborhood* $N_n(\delta_n)$*,* $d = O(n^{\kappa_1})$*, and* $\log p = O(n^{\kappa_2})$ *with constants satisfying* $0 < \kappa_1 < 1/4$*,* $0 < u_3 < 1/4 - \kappa_1$*,* $0 < u_2 < 1 - 4\kappa_1 - 4u_3$*,* $0 < u_1 < 1/2 - 2\kappa_1 - u_3$*, and* $0 < \kappa_2 < 1 - 4\kappa_1 - 2u_1 - 2u_3$*. Then the plug-in estimator* $\widehat{H}_n = \widehat{A}_n^{-1}\widehat{B}_n$ *satisfies that* $\mathrm{tr}(\widehat{H}_n) = \mathrm{tr}(H_n) + o_P(1)$ *and* $\log|\widehat{H}_n| = \log|H_n| + o_P(1)$ *with significant probability* $1 - O(n^{-\delta} + p^{1-8c_2\gamma_n^2})$*, where* δ *is some positive constant and* γ_n *is a slowly diverging sequence such that* $\gamma_n\widetilde{m}_n(K^{1/2}n^{-1}\log p)^{1/2} \to 0$.

Theorem 2 improves the result in Lv & Liu (2014) in two important aspects. First, the consistency of the covariance contrast matrix estimator was justified in Lv & Liu (2014) only for the scenario of correctly specified models. Our new result shows that the simple plug-in estimator $\widehat{H}_n$ still enjoys consistency in the general setting of model misspecification. Second, the result in Theorem 2 holds for the case of high dimensionality. These theoretical guarantees are crucial to the practical implementation of the new information criterion HGBIC$_p$. Our numerical studies in Section 4 reveal that such an estimate works well in a variety of model misspecification settings.

3.4. *Model selection consistency of* HGBIC$_p$

We further investigate the model selection consistency property of information criterion HGBIC$_p$. Assume that there are $M = o(n^\delta)$ sparse candidate models $\mathfrak{M}_1, \ldots, \mathfrak{M}_M$, where δ is some sufficiently large positive constant. At first glance, such an assumption may seem slightly restrictive since it rules out an exhaustive search over all $p!/\{(p-d)!d!\}$ possible candidate models. However, our goal in the paper is to provide practitioners with some tools for comparing a set of candidate models that are available to them. In fact, the set of sparse models under model comparison in practice can be often smaller, e.g., polynomial instead of exponential in sample size, even under the ultra-high dimensional setting. One example is that people may apply different algorithms each of which can lead to a possibly different model. Another example is the use of a certain regularization method with a sequence of sparse models generated by a path algorithm, which will be demonstrated in our numerical studies. For each candidate model $\mathfrak{M}_m$, we have the HGBIC$_p$ criterion as defined in (11)

$$\text{HGBIC}_p(\mathfrak{M}_m) = -2\ell_n(y, \widehat{\beta}_{n,m}) + 2(\log p^*)|\mathfrak{M}_m| + \text{tr}(\widehat{H}_{n,m}) - \log|\widehat{H}_{n,m}|, \qquad (16)$$

where $\widehat{H}_{n,m}$ is a consistent estimator of $H_{n,m}$ and $p^* = pn^{1/2}$. Assume that there exists an oracle working model in the sequence $\{\mathfrak{M}_m : m = 1, \ldots, M\}$ that has support identical to the set of all important features in the true model. Without loss of generality, suppose that $\mathfrak{M}_1$ is such oracle working model.

THEOREM 3. *Assume that all the assumptions of Theorems* 1–2 *hold and the population version of* HGBIC$_p$ *criterion in* (16) *is minimized at* $\mathfrak{M}_1$ *such that for some positive sequence* Δ_n *slowly converging to zero,*

$$\min_{m>1} \left\{\text{HGBIC}_p^*(\mathfrak{M}_m) - \text{HGBIC}_p^*(\mathfrak{M}_1)\right\} > \Delta_n \qquad (17)$$

with $\text{HGBIC}_p^*(\mathfrak{M}_m) = -2\ell_n(y, \beta_{n,m,0}) + 2(\log p^*)|\mathfrak{M}_m| + \text{tr}(H_{n,m}) - \log|H_{n,m}|$. *Then it holds that*

$$\min_{m>1} \{\text{HGBIC}_p(\mathfrak{M}_m) - \text{HGBIC}_p(\mathfrak{M}_1)\} > \Delta_n/2$$

for large enough n *with asymptotic probability one.*

Theorem 3 formally establishes the model selection consistency property of the new information criterion HGBIC$_p$ for large-scale model selection with misspecification in that the oracle working model can be selected among a large sequence of candidate sparse models with significant probability. Such a desired property is an important consequence of results in Theorems 1 and 2. Furthermore, assumption (17) is intrinsically necessary for this kind of theorem. For any model selection criteria, when the models are indistinguishable at the population level, the criteria cannot differentiate them in the sample version. Theorem 3 ensures that the gap in the population version is preserved in the sample version giving a slight leeway.

4. NUMERICAL STUDIES

4.1. *Setup*

We now investigate the finite-sample performance of the information criterion HGBIC$_p$ in comparison to the information criteria such as the Akaike information criterion (AIC), Bayesian information criterion (BIC), extended Bayesian information criterion (EBIC) (Chen & Chen, 2008), generalized information criterion (GIC) (Fan & Tang, 2013), generalized Akaike information criterion (equivalently Takeuchi information criterion), generalized Bayesian information criterion

(GBIC), and GBIC_p, in high-dimensional misspecified models via three simulation examples: a multiple index model (Section 4.2), a logistic regression model with interaction effects (Section B in the Supplementary Material), and a Poisson regression model with interaction effects (Section C in the Supplementary Material). For each candidate model $\mathfrak{M}_m$, the EBIC and GIC criteria are defined as

$$\text{EBIC}(\mathfrak{M}_m) = -2\ell_n(y, \widehat{\beta}_{n,m}) + (\log n)|\mathfrak{M}_m| + \log \binom{p}{|\mathfrak{M}_m|}, \tag{18}$$

$$\text{GIC}(\mathfrak{M}_m) = -2\ell_n(y, \widehat{\beta}_{n,m}) + (\log n)(\log \log p)|\mathfrak{M}_m|. \tag{19}$$

4.2. *Multiple index model*

The first model we consider is the following multiple index model

$$Y = f(\beta_1 X_1) + f(\beta_2 X_2 + \beta_3 X_3) + f(\beta_4 X_4 + \beta_5 X_5) + \varepsilon, \tag{20}$$

where the response depends on the covariates X_j's only through the first five ones in a nonlinear fashion and $f(x) = x^3/(x^2+1)$. Here the rows of the $n \times p$ design matrix Z are sampled as independent copies from $N(0, I_p)$, and the n-dimensional error vector $\varepsilon \sim N(0, \sigma^2 I_n)$. We set the true parameter vector $\beta_0 = (1, -1, 1, 1, -1, 0, \ldots, 0)^T$ and $\sigma = 1$. We vary the dimensionality p from 100 to 3200 while keeping the sample size n fixed at 200. We would like to investigate the behavior of different information criteria when the dimensionality increases. Although the data was generated from model (20), we fit the linear regression model (1). This is a typical example of model misspecification. Since the first five variables are independent of the other variables, the oracle working model is $M_0 = \text{supp}(\beta_0) = \{1, \ldots, 5\}$. Due to the high dimensionality, it is computationally prohibitive to implement the best subset selection. Thus we first applied Lasso followed by least-squares refitting to build a sequence of sparse models and then selected the final model using a model selection criterion. In practice, one can apply any preferred variable selection procedure to obtain a sequence of candidate interpretable models.

We report the consistent selection probability (the proportion of simulations where selected model $\widehat{M} = M_0$), the sure screening probability Fan & Lv (2008); Fan & Fan (2008) (the proportion of simulations where selected mode $\widehat{M} \supset M_0$), and the prediction error $E(Y - z^T\widehat{\beta})^2$ with $\widehat{\beta}$ an estimate and (z, Y) an independent observation for $z = (X_1, \ldots, X_p)^T$. To evaluate the prediction performance of different criteria, we calculated the average prediction error on an independent test sample of size 10,000. The results for prediction error and model selection performance are summarized in Table 1. In addition, we calculate the average number of false positives for each method in Table 2.

From Table 1, we observe that as the dimensionality p increases, the consistent selection probability tends to decrease for all criteria except the newly suggested HGBIC_p, which maintains at least 95% consistent selection probability throughout all dimensionalities considered. Generally speaking, generalized Akaike information criterion improved over Akaike information criterion, and GBIC, GBIC_p performed better than BICin terms of both prediction and variable selection. The high-dimensional information criteria EBICand GICoutperformed the traditional Akaike information criterion and BIC. In particular, the model selected by our new information criterion HGBIC_p delivered the best performance with the smallest prediction error and highest consistent selection probability across all settings.

An interesting observation is the comparison among GBIC_p, GIC, and HGBIC_p in terms of model selection consistency property. While GBIC_p is comparable to HGBIC_p when the dimensionality is not large (e.g., $p = 100$), the difference between these two methods increases as the dimensionality increases. In the case when $p = 3200$, HGBIC_p has 95% of success for consistent

Table 1. *Average results over 100 repetitions for Example* 4.2 *with all entries multiplied by 100.*

p	AIC	BIC	EBIC	GIC	GAIC	GBIC	GBIC$_p$	HGBIC$_p$	Oracle
Consistent selection probability with sure screening probability in parentheses									
100	0(100)	29(100)	70(100)	66(100)	0(100)	33(100)	57(100)	100(100)	100(100)
200	0(100)	6(100)	57(100)	59(100)	0(100)	9(100)	32(100)	99(100)	100(100)
400	0(100)	1(100)	57(100)	68(100)	0(100)	3(100)	13(100)	99(100)	100(100)
800	0(100)	0(100)	51(100)	64(100)	0(100)	0(100)	10(100)	98(100)	100(100)
1600	0(100)	0(100)	39(100)	59(100)	0(100)	0(100)	9(100)	98(100)	100(100)
3200	0(100)	0(100)	43(100)	64(100)	0(100)	0(100)	4(100)	95(99)	100(100)
Mean prediction error with standard error in parentheses									
100	151(2)	126(2)	122(1)	122(1)	137(2)	126(2)	123(1)	119(1)	119(1)
200	166(2)	131(2)	121(1)	121(1)	139(2)	130(2)	124(1)	117(1)	117(1)
400	181(3)	140(2)	124(1)	123(1)	146(2)	139(2)	129(2)	120(1)	119(1)
800	187(2)	149(2)	127(1)	125(1)	151(2)	147(2)	136(2)	121(1)	121(1)
1600	185(2)	154(2)	128(2)	124(1)	152(2)	152(2)	137(2)	119(1)	119(1)
3200	178(2)	151(2)	123(1)	120(1)	146(2)	150(2)	134(2)	117(1)	116(1)

selection, while GBIC$_p$ has a success rate of only 4%. This confirms the necessity of including the $\log p^*$ factor with $p^* = pn^{1/2}$ in the model selection criterion to take into account the high dimensionality, which is in line with the results in Fan & Tang (2013) for the case of correctly specified models. On the other hand, due to the lack of consideration of model misspecification, GICis still outperformed by the newly proposed HGBIC$_p$ throughout all dimensionalities considered.

Table 2. *Average false positives over 100 repetitions for Example* 4.2.

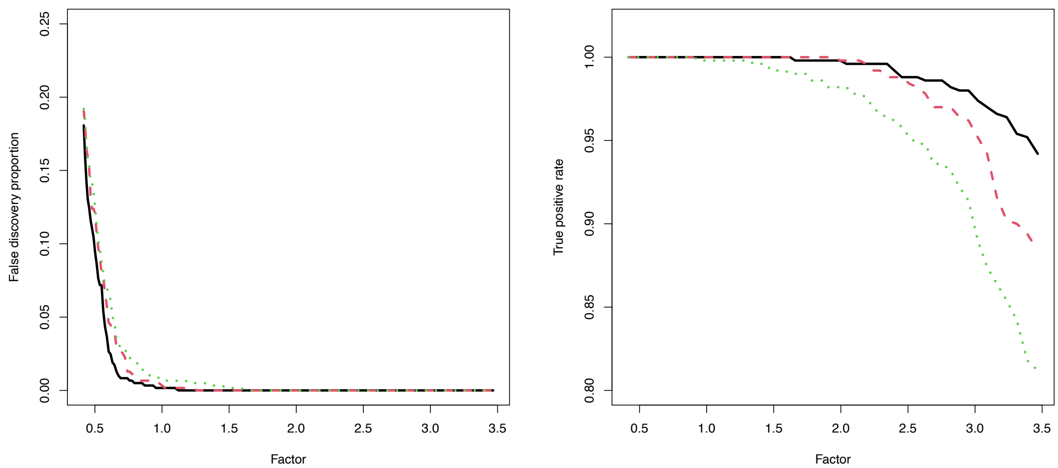
p	AIC	BIC	EBIC	GIC	GAIC	GBIC	GBIC$_p$	HGBIC$_p$
100	15.35	1.84	0.49	0.58	7.05	1.75	0.86	0.00
200	24.30	3.53	0.76	0.70	7.43	3.07	1.39	0.01
400	31.46	5.58	0.73	0.53	8.32	5.11	1.98	0.01
800	34.12	7.21	0.87	0.60	8.26	6.20	2.58	0.02
1600	34.41	8.74	1.23	0.56	7.65	7.58	3.12	0.02
3200	33.41	8.64	0.93	0.48	7.25	8.28	3.26	0.04

We further study a family of model selection criteria induced by the HGBIC$_p$ and characterized as follows

$$\text{HGBIC}_{p,\zeta}(\mathfrak{M}_m) = -2\ell_n(y, \widehat{\beta}_{n,m}) + \zeta \left\{ 2(\log p^*)|\mathfrak{M}_m| + \mathrm{tr}(\widehat{H}_{n,m}) - \log|\widehat{H}_{n,m}| \right\}, \quad (21)$$

where ζ is a positive factor controlling the penalty level on both model misspecification and high dimensionality. HGBIC$_{p,\zeta}$ with $\zeta = 1$ reduces to our original HGBIC$_p$. Here we examine the impact of the factor ζ on the false discovery proportion and the true positive rate for the selected model $\widehat{M}$ compared to the oracle working model M_0. In Figure 1, we observe that as ζ increases, the average false discovery proportion drops sharply as it gets close to 1. In addition, we have the desired model selection consistency property (with the false discovery proportion close to 0 and true positive rate close to 1 when $\zeta \in [1, 1.5]$. This figure demonstrates the robustness of the introduced HGBIC$_{p,\zeta}$ criteria.

Fig. 1. The average false discovery proportion (left panel) and the true positive rate (right panel) as the factor ζ varies for Example 4.2 when $p = 200$ (black solid), $p = 800$ (red dashed), and $p = 3200$ (green dot-dash).



Acknowledgement

The authors sincerely thank the editor, associate editor, and referees for comments that significantly improved the paper. This work was supported by the U.S. National Science Foundation, a grant from the Simons Foundation, and Adobe Data Science Research Award. Demirkaya and Feng contribute equally to this work.

Supplementary material

Supplementary material available online contains additional numerical studies, examples to compute HGBIC$_p$, all the proofs of main results, and additional technical details.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In Second International Symposium on Information Theory (eds. B. N. Petrov and F. Csaki), Akademiai Kiado, Budapest* , 267–281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Barber, R. & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43**, 2055–2085.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52**, 345–370.

Bühlmann, P. & van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Bühlmann, P. & van de Geer, S. (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics* **9**, 1449–1473.

Candès, E. J., Fan, Y., Janson, L. & Lv, J. (2018). Panning for gold: 'model–X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B* **80**, 551–577.

Castillo, I., Schmidt-Hieber, J., Van der Vaart, A. et al. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics* **43**, 1986–2018.

CHEN, J. & CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–771.

CHEN, K. & CHAN, K.-S. (2011). Subset ARMA selection via the adaptive lasso. *Statistics and Its Interface* **4**, 197–205.

CULE, M., SAMWORTH, R. & STEWART, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density (with discussion). *J. Roy. Statist. Soc. Ser. B* **72**, 545–607.

EGUCHI, S. (2017). Model comparison for generalized linear models with dependent observations. *Econometrics and Statistics* **5**, 171–188.

ERDŐS, L., YAU, H.-T. & YIN, J. (2012). Bulk universality for generalized wigner matrices. *Probability Theory and Related Fields* **154**, 341–407.

FAN, J. & FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics* **36**, 2605–2637.

FAN, J. & LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B* **70**, 849–911.

FAN, J. & LV, J. (2010). A selective overview of variable selection in high dimensional feature space (invited review article). *Statistica Sinica* **20**, 101–148.

FAN, J. & LV, J. (2018). Sure independence screening (invited review article). *Wiley StatsRef: Statistics Reference Online* .

FAN, Y., DEMIRKAYA, E., LI, G. & LV, J. (2020). RANK: large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association* **115**, 362–379.

FAN, Y., DEMIRKAYA, E. & LV, J. (2019). Nonuniformity of p-values can occur early in diverging dimensions. *Journal of Machine Learning Research* **20**, 1–33.

FAN, Y. & TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society Series B* **75**, 531–552.

FONG, E. & HOLMES, C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika* **107**, 489–496.

FOSTER, D. & GEORGE, E. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics* , 1947–1975.

GEORGE, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association* **95**, 1304–1308.

HALL, P. (1990). Akaike's information criterion and kullback-leibler loss for histogram density estimation. *Probability Theory and Related Fields* **85**, 449–467.

HORN, R. A. & JOHNSON, C. R. (1985). *Matrix Analysis*. Cambridge University Press.

HSU, H.-L., ING, C.-K. & TONG, H. (2019). On model selection from a finite family of possibly misspecified time series models. *The Annals of Statistics* **47**, 1061–1087.

ING, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *The Annals of Statistics* **35**, 1238–1277.

KONISHI, S. & KITAGAWA, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875–890.

KULLBACK, S. & LEIBLER, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86.

LAURITZEN, S. L. (1996). *Graphical models*, vol. 17. Clarendon Press.

LIU, W. & YANG, Y. (2011). Parametric or nonparametric? a parametricness index for model selection. *The Annals of Statistics* , 2074–2102.

LV, J. & LIU, J. S. (2014). Model selection principles in misspecified models. *Journal of the Royal Statistical Society Series B* **76**, 141–167.

MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

NINOMIYA, Y. & KAWANO, S. (2016). AIC for the Lasso in generalized linear models. *Electronic Journal of Statistics* **10**, 2537–2560.

PEARL, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.

PENG, H., YAN, H. & ZHANG, W. (2013). The connection between cross-validation and Akaike information criterion in a semiparametric family. *Journal of Nonparametric Statistics* **25**, 475–485.

SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

SHAH, R. D. & BÜHLMANN, P. (2018). Goodness-of-fit tests for high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 113–135.

STONE, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society Series B* , 44–47.

SZULC, P. (2012). Weak consistency of modified versions of Bayesian information criterion in a sparse linear regression. *Probability and Mathematical Statistics* **32**, 47–55.

TAKEUCHI, K. (1976). The distribution of information statistics and the criterion of goodness of fit of models. *Mathematical Science* **153**, 12–18.

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.