Statistica Sinica **28** (2018), 2409-2433
doi:https://doi.org/10.5705/ss.202017.0041

EMPIRICAL LIKELIHOOD RATIO TESTS FOR COEFFICIENTS IN HIGH-DIMENSIONAL HETEROSCEDASTIC LINEAR MODELS

Honglang Wang[1], Ping-Shou Zhong[2] and Yuehua Cui[2]

[1]*Indiana University-Purdue University Indianapolis*
and [2]*Michigan State University*

Abstract: This paper considers hypothesis testing problems for a low-dimensional coefficient vector in a high-dimensional linear model with heteroscedastic variance. Heteroscedasticity is a commonly observed phenomenon in many applications, including finance and genomic studies. Several statistical inference procedures have been proposed for low-dimensional coefficients in a high-dimensional linear model with homoscedastic variance, which are not applicable for models with heteroscedastic variance. The heterscedasticity issue has been rarely investigated and studied. We propose a simple inference procedure based on empirical likelihood to overcome the heteroscedasticity issue. The proposed method is able to make valid inference even when the conditional variance of random error is an unknown function of high-dimensional predictors. We apply our inference procedure to three recently proposed estimating equations and establish the asymptotic distributions of the proposed methods. Simulation studies and real data applications are conducted to demonstrate the proposed methods.

Key words and phrases: Empirical likelihood, heteroscedastic linear models, high-dimensional data, low-dimensional coefficients.

1. Introduction

In the last two decades, rapid progress has been made in high-dimensional statistics. In particular, high-dimensional linear regression models have received much attention. Many regularization methods have been proposed for simultaneous estimation and variable selection in linear models, including LASSO (Tibshirani (1996)), SCAD (Fan and Li (2001)), MCP (Zhang (2010)), among others. Most of this literature has focused on the estimation for coefficients in linear models with homoscedastic random errors. An excellent review can be found in Bühlmann and Van De Geer (2011).

The issue of heteroscedasticity is commonly seen in practice, but it has not received much attention in high-dimensional statistics literature. Wang, Wu and Li

(2012) analyzed the heteroscedasticity in a high-dimensional case using quantile regression. Daye, Chen and Li (2012) proposed a method that allows nonconstant error variances for high-dimensional estimation but with a parametric form of the variance function. More recently, Belloni, Chernozhukov and Wang (2014) came up with a self-tuning square root Lasso estimation method that solved the heteroscedasticity issue in high-dimensional regression analysis.

There is not much literature concerning statistical inference for regression coefficients in a high-dimensional model. Progress has been achieved for the inference about low-dimensional parameters in a high-dimensional model, including Zhang and Zhang (2014), Bühlmann (2013), Javanmard and Montanari (2013), van de Geer, Bühlmann and Ritov (2013), Lan et al. (2016), and Ning and Liu (2014). These procedures assume homoscedasticity for the error term, but this seldom holds in practice and there is rarely sufficient information to accurately specify a correct variance function. Moreover, the variances of these estimators are complex and difficult to estimate under the heteroscedasticity case. Incorrect variance models will lead to inferences that are not asymptotically valid (Belsley (2002)). Wagener and Dette (2012) generalized the asymptotic results of Knight and Fu (2000) for the case of a fixed dimension under heteroscedasitic errors, but there is little such work in the high-dimensional setting, aside from Dezeure, Bühlmann and Zhang (2016) who recently proposed bootstrap methods for inference under high-dimensional linear models with heteroscedastic errors.

This paper proposes to use Empirical Likelihood (EL) to test statistical hypotheses and construct confidence regions for low-dimensional components in high-dimensional liner models with heteroscedastic noise. EL (Owen (2001)) is a nonparametric approach for deriving estimations and confidence regions for unknown parameters (Owen (1990, 2001)). Professor Peter Hall made fundamental contributions to it. He showed that EL is Bartlett correctable (Hall (1990); DiCiccio, Hall and Romano (1991)) and produces confidence regions with natural shape and orientation (Hall and La Scala (1990)). As EL is a data-driven nonparametric method, it does not need distribution assumptions except for some moment conditions. EL-based methods have been used for statistical inferences with heteroscedasiticity in the low-dimensional case. Tsao and Wu (2006) conducted EL inference for a common mean in the presence of heteroscedasticity. Chen and Qin (2003) considered the EL-based point-wise confidence intervals for a nonparametric regression function with heteroscedastic errors. Lu (2009) and Zhou, Kim and Bathke (2012) discussed EL analysis for heteroscedastic partially linear models and heteroscedastic accelerated failure time models, respectively.

However, the EL-based method has not been used for the problem considered in this paper. A comprehensive overview of EL methods can be found in Owen (2001) and a survey of recent developments is in Chen and Van Keilegom (2009).

Different from existing methods, our proposed procedure does not need to estimate the variance explicitly due to the internal studentizing ability of EL. This makes our procedure attractive especially under the heteroscedasticity setting, even when the conditional variance of the error term is an unknown function of high-dimensional predictors. The proposed EL-based method is a general unified framework suitable for various estimating equations as long as they satisfy some conditions specified later.

The paper is organized as follows. In Section 3, we study the asymptotic normality of Wald-type statistic for the existing methods under heteroscedastic noise. In Section 4, we introduce a general EL-based method for the problems considered here. In addition, we provide explicit examples of the general EL-based method. Section 5 provides numerical results and Section 6 shows some real data analysis, followed by discussions in Section 7. We relegate technical proofs to the Appendix.

2. Basic Setup and Notations

Consider the following linear regression model,

$$\mathbb{Y} = \mathbb{X}\boldsymbol{\beta}^0 + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\mathbb{Y} = (Y_1, Y_2, \ldots, Y_n)^\intercal \in \mathbb{R}^n$ is the response vector, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\intercal \in \mathbb{R}^n$ is the vector of noise, and $\mathbb{X} = ((X_{ij})) \in \mathbb{R}^{n\times p}$ is the random design matrix with p columns $\{\mathbb{X}_j \in \mathbb{R}^{n\times 1}\}_{j=1}^{p}$ and n rows $\{\mathbf{X}_i^\intercal \in \mathbb{R}^{1\times p}\}_{i=1}^{n}$. The row vectors are assumed to be independent and identically distributed (IID) with $\mathrm{E}(\mathbf{X}_i) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{X}_i) = \boldsymbol{\Sigma} = ((\sigma_{jl}))_{1\leq j,l\leq p}$, and $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ is a vector of unknown true regression coefficients. The independent error terms satisfy $\mathrm{E}(\epsilon_i|\mathbf{X}_i) = 0$, and $\mathrm{Var}(\epsilon_i) = \sigma_i^2$. This is the usual heteroscedastic model (White (1980); Li and Yao (2015); Daye, Chen and Li (2012); Bai, Pan and Yin (2016); Dezeure, Bühlmann and Zhang (2016)). Let $\mathbf{Z}_i = \epsilon_i\mathbf{X}_i$ be a random vector. With these assumptions, $\mathbf{X}_i$ and ϵ_i are uncorrelated, $\mathrm{E}(\mathbf{Z}_i) = \mathbf{0}$. In addition, marginally we assume $\mathrm{Var}(\epsilon_i^2) = \kappa_i$. We denote the covariance matrix of $\mathbf{Z}_i$ by $\boldsymbol{\Theta}_i = ((\theta_{i;jk}))$.

In practice, among thousands of regressors, investigators may wish to test whether some target coefficients are significant or not. For example, one may want to know if treatment effects are significant after accounting for the effects of many other variables. This paper focuses on assessing the significance of a single coefficient. We test the following hypothesis for any given $j \in \{1, 2, \ldots, p\}$,

$$H_0 : \beta_j^0 = 0 \quad \text{vs.} \quad H_1 : \beta_j^0 \neq 0, \tag{2.2}$$

in (2.1) with $p \gg n$, assuming heteroscedastic errors.

The following notations are adopted throughout. For $\mathbf{v} = (v_1, v_2, \ldots, v_d)^\intercal \in \mathbb{R}^d$, let $\|\mathbf{v}\|_q = (\sum_{i=1}^d |v_i|^q)^{1/q}$ for $0 < q < \infty$, $\|\mathbf{v}\|_0 = |\text{supp}(\mathbf{v})|$ where $\text{supp}(\mathbf{v}) = \{j : v_j \neq 0\}$, $|A|$ is the cardinality of a set A, and $\|\mathbf{v}\|_\infty = \max_{1 \leq j \leq d} |v_i|$. We denote $\mathbf{I}_d$ as a $d \times d$ identity matrix. If the dimension is obvious from the context, we just omit the subscript d. For $\mathcal{S} \subseteq \{1, 2 \ldots, d\}$, let $\mathbf{v}_\mathcal{S} = \{v_j : j \in \mathcal{S}\}$ be a subvector of $\mathbf{v}$. For any $k \in \{1, 2, \ldots, d\}$, write $\mathbf{M}_{j\mathcal{S}} = \{M_{jl}, l \in \mathcal{S}\}$ for a row vector and $\mathbf{M}_{\mathcal{S}j} = \{M_{lj} : l \in \mathcal{S}\}$ for a column vector. Let $\backslash k = \{1, 2, \ldots, k-1, k+1, \ldots, d\}$, the $(d-1)$-dim vector with the $k-$th component removed. For a sequence of random variables X_n, we use $X_n \overset{d}{\to} X$ to denote the convergence in distribution, and $X_n \overset{p}{\to} a$ to denote convergence in probability. Let $s = \|\boldsymbol{\beta}^0\|_0$ be the number of non-zeros of $\boldsymbol{\beta}^0$. We assume sparsity with $s < n$.

3. Asymptotic Properties of Some Existing Methods Under Heteroscedasticity

To motivate our proposed method, we first study three existing methods and derive their asymptotic properties under the heteroscedastic linear model (2.1). These methods were only studied under the homogeneous linear models and we generalize these results to the heteroscedasticity case.

3.1. Low-dimensional projection method

In this subsection, we introduce the low-dimensional projection method proposed by Zhang and Zhang (2014). Under model (2.1) and the low-dimensional scenario with $p < n$, the ordinary least square (OLS) estimator for β_j^0 is,

$$\hat{\beta}_j = \frac{(\mathbb{X}_j^\perp)^\intercal \mathbb{Y}}{(\mathbb{X}_j^\perp)^\intercal \mathbb{X}_j} = \frac{(\mathcal{Q}_{\backslash j}\mathbb{X}_j)^\intercal \mathbb{Y}}{(\mathcal{Q}_{\backslash j}\mathbb{X}_j)^\intercal \mathbb{X}_j} = \frac{(\mathcal{Q}_{\backslash j}\mathbb{X}_j)^\intercal (\mathcal{Q}_{\backslash j}\mathbb{Y})}{(\mathcal{Q}_{\backslash j}\mathbb{X}_j)^\intercal (\mathcal{Q}_{\backslash j}\mathbb{X}_j)} = \frac{\mathbb{X}_j^\intercal \mathcal{Q}_{\backslash j}\mathbb{Y}}{\mathbb{X}_j^\intercal \mathcal{Q}_{\backslash j}\mathbb{X}_j}, \tag{3.1}$$

where $\mathbb{X}_j^\perp$ is the projection of $\mathbb{X}_j$ to the orthogonal complement of the column space spanned by $\{\mathbb{X}_{\backslash j}\}$, where, with $\mathcal{S} \subseteq \{1, 2, \ldots, p\}$ and $|\mathcal{S}| < n$, $\mathcal{Q}_\mathcal{S} = \mathbf{I} - \mathcal{P}_\mathcal{S} = \mathbf{I} - \mathbb{X}_\mathcal{S}(\mathbb{X}_\mathcal{S}^\intercal \mathbb{X}_\mathcal{S})^- \mathbb{X}_\mathcal{S}^\intercal \in \mathbb{R}^{n \times n}$, with $(\mathbb{X}_\mathcal{S}^\intercal \mathbb{X}_\mathcal{S})^-$ a generalized inverse of $\mathbb{X}_\mathcal{S}^\intercal \mathbb{X}_\mathcal{S}$.

In the high-dimensional linear model with $p > n$, the OLS estimator in (3.1) is no longer valid because $\mathcal{Q}_{\backslash j}\mathbb{Y}$ and $\mathcal{Q}_{\backslash j}\mathbb{X}_j$ are always 0. To resolve this in the high-dimensional case, Zhang and Zhang (2014) proposed a de-biased estimator: if $\mathbb{Z}_j$ be an $n \times 1$ projection vector, an estimate of β_j^0 is

$$\hat{\beta}_j^{(\mathrm{lin})} = \frac{\mathbb{Z}_j^\intercal \mathbb{Y}}{\mathbb{Z}_j^\intercal \mathbb{X}_j} = \beta_j^0 + \frac{\mathbb{Z}_j^\intercal \boldsymbol{\epsilon}}{\mathbb{Z}_j^\intercal \mathbb{X}_j} + \mathrm{Bias}\left(\hat{\beta}_j^{(\mathrm{lin})}\right), \tag{3.2}$$

where $\mathrm{Bias}(\hat{\beta}_j^{(\mathrm{lin})}) = \sum_{k \neq j} \mathbb{Z}_j^\intercal \mathbb{X}_k \beta_k^0 / \mathbb{Z}_j^\intercal \mathbb{X}_j$ is the bias term. The second term in (3.2) has mean zero and is of order $1/\sqrt{n}$. Because the bias term is not ignorable, $\hat{\beta}_j^{(\mathrm{lin})}$ is not directly useful for inference. To make it so, we need to reduce the order of the bias term $\mathrm{Bias}(\hat{\beta}_j^{(\mathrm{lin})})$ to $o_p(1/\sqrt{n})$. Here Zhang and Zhang (2014) proposed the de-biased estimator,

$$\hat{\beta}_j^{(\mathrm{de})} = \frac{\mathbb{Z}_j^\intercal \mathbb{Y} - \sum_{k \neq j} \mathbb{Z}_j^\intercal \mathbb{X}_k \hat{\beta}_k^{(0)}}{\mathbb{Z}_j^\intercal \mathbb{X}_j}, \tag{3.3}$$

where $\hat{\boldsymbol{\beta}}^{(0)}$ is some initial regularized estimator of $\boldsymbol{\beta}^0$ so that $\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^0\|_1 = o(a_n)$ for some $a_n \to 0$. Then the bias of $\hat{\beta}_j^{(\mathrm{de})}$ is controlled by

$$\left| \sum_{k \neq j} \frac{\mathbb{Z}_j^\intercal \mathbb{X}_k (\beta_k^0 - \hat{\beta}_k^{(0)})}{\mathbb{Z}_j^\intercal \mathbb{X}_j} \right| \leq \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^0\|_1 \max_{k \neq j} \left| \frac{\mathbb{Z}_j^\intercal \mathbb{X}_k}{\mathbb{Z}_j^\intercal \mathbb{X}_j} \right|.$$

To make the right hand side of this inequality of order $o_p(1/\sqrt{n})$, removing the bias using $\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^0$ is not enough, because $\|\hat{\boldsymbol{\beta}}^0 - \boldsymbol{\beta}^0\|_1$ is typically of order $O_p(s\sqrt{\log p/n})$ (Belloni, Chernozhukov and Wang (2014)). Therefore, we need to make $\max_{k \neq j} |\mathbb{Z}_j^\intercal \mathbb{X}_k|$ small enough. Ideally, if $\mathbb{Z}_j$ is orthogonal to all $\mathbb{X}_k, k \neq j$, then $\max_{k \neq j} |\mathbb{Z}_j^\intercal \mathbb{X}_k|$ is 0. However, this cannot hold if $p > n$. Therefore, a key problem is the selection of projection vector $\mathbb{Z}_j$.

In Zhang and Zhang (2014), van de Geer, Bühlmann and Ritov (2013), and Ning and Liu (2014), the linear sparse regularized regression procedure, say LASSO, is used to select the projection vector. Define $\eta_{ij} := X_{ij} - \mathbf{X}_{i,\backslash j}^\intercal \boldsymbol{\Sigma}_{\backslash j,\backslash j}^{-1} \boldsymbol{\Sigma}_{\backslash j,j}$, so

$$X_{ij} = \mathbf{X}_{i,\backslash j}^\intercal \mathbf{w}_j^0 + \eta_{ij}, \text{ with } \mathbf{w}_j^0 = \boldsymbol{\Sigma}_{\backslash j,\backslash j}^{-1} \boldsymbol{\Sigma}_{\backslash j,j}, \text{ for } i = 1, 2, \ldots, n.$$

This leads to a de-biased version of (3.3) with $\mathbb{Z}_j = \mathbb{X}_j - \mathbb{X}_{\backslash j} \hat{\mathbf{w}}_j$ and with $\hat{\mathbf{w}}_j$ as a regularized estimator of $\mathbf{w}_j^0$.

Under the homoscedastic case, the inference procedure can be built on the asymptotic normality of $\hat{\beta}_j^{(\mathrm{de})}$, which requires one to estimate the asymptotic variance $\sigma_\epsilon^2/(\sigma_{jj} - \boldsymbol{\Sigma}_{j,\backslash j} \boldsymbol{\Sigma}_{\backslash j,\backslash j}^{-1} \boldsymbol{\Sigma}_{\backslash j,j})$. Zhang and Zhang (2014) and Dezeure, Bühlmann and Zhang (2016) used $\hat{\sigma}_\epsilon^2 \|\mathbb{Z}_j\|_2^2 / |\mathbb{Z}_j^\intercal \mathbb{X}_j|^2$ with $\hat{\sigma}_\epsilon^2$ estimated from the scaled LASSO-LSE (Zhang and Zhang (2014)) or the method recommended in Reid, Tibshirani and Friedman (2016). Under heteroscedastic noise, we can also establish the asymptotic normality but with a much more complicated asymptotic variance than in the homoscedastic case. Take the asymptotic variance of

$\hat{\beta}_j^{(\mathrm{de})}$ as

$$\sigma^2_{n,\mathrm{lasso}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\theta_{i;jj} - 2\boldsymbol{\Sigma}_{j,\backslash j}\boldsymbol{\Sigma}^{-1}_{\backslash j,\backslash j}\boldsymbol{\Theta}_{i;j,\backslash j} + \boldsymbol{\Sigma}_{j,\backslash j}\boldsymbol{\Sigma}^{-1}_{\backslash j,\backslash j}\boldsymbol{\Theta}_{i;\backslash j,\backslash j}\boldsymbol{\Sigma}^{-1}_{\backslash j,\backslash j}\boldsymbol{\Sigma}_{\backslash j,j}}{(\sigma_{jj} - \boldsymbol{\Sigma}_{j,\backslash j}\boldsymbol{\Sigma}^{-1}_{\backslash j,\backslash j}\boldsymbol{\Sigma}_{\backslash j,j})^2}. \quad (3.4)$$

As a special case, if ϵ_i and $\mathbf{X}_i$ are independent and the error term is homoscedastic, then $\sigma^2_{n,\mathrm{lasso}}$ can be simplified to $\sigma^2_\epsilon/\{\sigma_{jj} - \boldsymbol{\Sigma}_{j,\backslash j}\boldsymbol{\Sigma}^{-1}_{\backslash j,\backslash j}\boldsymbol{\Sigma}_{\backslash j,j}\}$, agrees with the result obtained by Zhang and Zhang (2014).

Proposition 1. *Under* (2.1) *with heteroscedastic noise, if Assumption* 1 *in the Appendix holds, then*

$$\sqrt{n}(\hat{\beta}_j^{(de)} - \beta_j^0) \xrightarrow{d} N(0, \sigma^2_{lasso}), \quad (3.5)$$

where σ^2_{lasso} is the asymptotic variance and $\sigma^2_{lasso} = \lim_{n\to\infty}\sigma^2_{n,lasso}$.

The complex asymptotic variance (3.4) makes it hard to use the Wald-type inference procedure in practice since it is difficult to get a good estimate for the asymptotic variance. Then using the Wald type test procedure of Zhang and Zhang (2014) in the heteroscedastic case leads to invalid results, as will be seen in the simulation study in Section 5.

3.2. KFC projection

Lan et al. (2016) proposed another way to construct an asymptotically unbiased estimator. The idea is similar to the low-dimensional projection method proposed by Zhang and Zhang (2014). In the estimator considered in (3.1), one projects $\mathbb{X}_j$ to all the variables except the j-th variable. Lan et al. (2016) projects $\mathbb{X}_j$ onto the so-called KFC set $\mathcal{S} = \{l \neq j : |\sigma_{jl}| > c\}$ for some pre-specified threshold value $c > 0$, essentially the set of all key confounders associated with X_j. Assume $|\mathcal{S}| \leq m$ for some m depending on the sample size n. After excluding the covariates that are highly correlated with $\mathbb{X}_j$, an approximate estimate of β_j can be obtained by the marginal regression of the profiled response $\tilde{\mathbb{Y}} = \mathcal{Q}_{\mathcal{S}}\mathbb{Y}$ on the profiled covariates $\tilde{\mathbb{X}}_j = \mathcal{Q}_{\mathcal{S}}\mathbb{X}_j$, namely

$$\hat{\beta}_j^{(\mathrm{kfc})} = \frac{\tilde{\mathbb{X}}_j^{\intercal}\tilde{\mathbb{Y}}}{\tilde{\mathbb{X}}_j^{\intercal}\tilde{\mathbb{X}}_j} = \frac{\mathbb{X}_j^{\intercal}\mathcal{Q}_{\mathcal{S}}\mathbb{Y}}{\mathbb{X}_j^{\intercal}\mathcal{Q}_{\mathcal{S}}\mathbb{X}_j}. \quad (3.6)$$

Based on this, we propose the de-biased KFC estimator

$$\hat{\beta}_j^{(\mathrm{kfc\text{-}de})} = \frac{\mathbb{X}_j^{\intercal}\mathcal{Q}_{\mathcal{S}}\mathbb{Y} - \sum_{k\in\mathcal{S}^*}\mathbb{X}_j^{\intercal}\mathcal{Q}_{\mathcal{S}}\mathbb{X}_k\hat{\beta}_k}{\mathbb{X}_j^{\intercal}\mathcal{Q}_{\mathcal{S}}\mathbb{X}_j}, \quad (3.7)$$

where $\mathcal{S}^* = \mathcal{S}^{+c}$, the complement of $\mathcal{S}^+ := \{j\}\cup\mathcal{S}$, and $\hat{\boldsymbol{\beta}}_{\mathcal{S}^*}$ is an initial estimator. The key difference between $\hat{\beta}_j^{(\mathrm{kfc\text{-}de})}$ and $\hat{\beta}_j^{(\mathrm{de})}$ is the selection approach of the low-dimensional projection space spanned by the subsets of covariates. $\hat{\beta}_j^{(\mathrm{de})}$ is based on the lasso approach while $\hat{\beta}_j^{(\text{kfc-de})}$ is based on the screening approach to find it.

If we assume ϵ_i and $\mathbf{X}_i$ are independent, the simple asymptotic variance of $\hat{\beta}_j^{(\text{kfc-de})}$ is $\sigma_\epsilon^2/(\sigma_{jj} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}j})$, as discussed in Lan et al. (2016). Under (2.1) with heteroscedastic errors, Proposition 2 proves the asymptotic normality of the de-biased estimator $\hat{\beta}_j^{(\text{kfc-de})}$,

Proposition 2. *Under the Assumption* 3 *in the Appendix, we have*

$$\sqrt{n}(\hat{\beta}_j^{(\textit{kfc-de})} - \beta_j^0) \xrightarrow{d} N(0, \sigma_{\textit{kfc}}^2), \tag{3.8}$$

where the asymptotic variance is

$$\sigma_{\textit{kfc}}^2 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\theta_{i;jj} - 2\mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Theta}_{i;j\mathcal{S}} + \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Theta}_{i;\mathcal{S}\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}j}}{(\sigma_{jj} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}j})^2}. \tag{3.9}$$

If we assume independence between ϵ_i and $\mathbf{X}_i$ and homoscedasticity for the error terms, we have $\sigma_{\mathrm{kfc}}^2 = \lim_{n\to\infty} \sigma_\epsilon^2/(\sigma_{jj} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}j})$, whose consistent estimator is discussed in Lan et al. (2016). However, based on Proposition 2, we can see that the adjusted KFC estimator is not easy to be implemented under heteroscedastic linear models.

3.3. Inverse projection

In the last two subsections, the test statistics were constructed based on the asymptotically unbiased estimator for β_j^0. To conduct the hypothesis testing problem (2.2), Liu and Luo (2014) proposed an equivalent test based on the projection of X_{ij} onto $(Y_i, \mathbf{X}_{i,\setminus j}^{\intercal})^{\intercal}$,

$$X_{ij} = (Y_i, \mathbf{X}_{i,\setminus j}^{\intercal})\boldsymbol{\gamma}_j^0 + \eta_{ij,y}, \tag{3.10}$$

where $\eta_{ij,y}$ satisfies $\mathrm{E}\eta_{ij,y} = 0, \mathrm{Cov}(\eta_{ij,y}, (Y_i, \mathbf{X}_{i,\setminus j}^{\intercal})) = \mathbf{0}$. Under (2.1) with heteroscedastic noise, as long as $\mathrm{Cov}(\mathbf{X}_i, \epsilon) = \mathbf{0}$, we can still show that the vector $\boldsymbol{\gamma}_j^0$ satisfies $\boldsymbol{\gamma}_j^0 = -\sigma_{\eta_{j,y}}^2\big(-\beta_j^0/\sigma_\epsilon^2, \beta_j^0\boldsymbol{\beta}_{\setminus j}^{0\intercal}/\sigma_\epsilon^2 + \mathbf{\Omega}_{\setminus j,j}\big)^{\intercal}$, where $\sigma_{\eta_{j,y}}^2 = \mathrm{Var}(\eta_{ij,y}) = \{(\beta_j^0)^2 + w_{jj}\}^{-1}$ with $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = ((w_{jk}))$. Because $\mathrm{Cov}(\epsilon_i, \mathbf{X}_i) = \mathbf{0}$, with γ_{j1}^0 as the first element of $\boldsymbol{\gamma}_j^0$, we have

$$\mathrm{Cov}(\epsilon_i, \eta_{ij,y}) = \gamma_{j1}^0\mathrm{Cov}(\epsilon_i, -Y_i) = -\sigma_{\eta_{j,y}}^2\beta_j^0 := -\mathfrak{b}_j^0. \tag{3.11}$$

Hence to test (2.2) is equivalent to test $H_0 : \mathfrak{b}_j^0 = 0$ because $\sigma_{\eta_{j,y}}^2 > 0$. Based on Liu and Luo (2014), we can estimate $\mathfrak{b}_j^0$ using

$$\hat{\mathfrak{b}}_j = -\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \mathbf{X}_i^{\intercal}\hat{\boldsymbol\beta}\big)\big\{X_{ij} - (Y_i, \mathbf{X}_{i,\backslash j}^{\intercal})\hat{\boldsymbol\gamma}_j\big\}, \tag{3.12}$$

where $\hat{\boldsymbol\beta}$ and $\hat{\boldsymbol\gamma}_j$ are some initial regularized estimators of $\boldsymbol\beta^0$ and $\boldsymbol\gamma_j^0$.

Let

$$\begin{aligned}\sigma^2_{i;n,\mathrm{inv}} = {} & \theta_{i;jj} + (\gamma^0_{j1})^2\boldsymbol\beta^{0\intercal}\boldsymbol\Theta_i\boldsymbol\beta^0 + (\gamma^0_{j1})^2\kappa_i + \boldsymbol\gamma^{0\intercal}_{j,\backslash 1}\boldsymbol\Theta_{i;\backslash j,\backslash j}\boldsymbol\gamma^0_{j,\backslash 1} - 2\gamma^0_{j1}\boldsymbol\beta^{0\intercal}\boldsymbol\Theta_{i;\cdot,j}\\ & - 2\gamma^0_{j1}\varpi_{i;j} - 2\boldsymbol\gamma^{0\intercal}_{j,\backslash 1}\boldsymbol\Theta_{i;\backslash j,j} + 2(\gamma^0_{j1})^2\boldsymbol\beta^{0\intercal}\boldsymbol\varpi_i + 2\gamma^0_{j1}\boldsymbol\beta^{0\intercal}\boldsymbol\Theta_{i;\cdot,\backslash j}\boldsymbol\gamma^0_{j,\backslash 1}\\ & + 2\gamma^0_{j1}\boldsymbol\gamma^{0\intercal}_{j,\backslash 1}\boldsymbol\varpi_{i;\backslash j},\end{aligned}$$

where $\boldsymbol\varpi_i = \mathrm{Cov}(\epsilon_i^2, \mathbf{Z}_i)$ with $\mathbf{Z}_i = \epsilon_i\mathbf{X}_i$.

Proposition 3. *Under Assumption* 2 *in the Appendix, we have*

$$\sqrt{n}(\hat{\mathfrak{b}}_j - \mathfrak{b}_j^0) \xrightarrow{d} N(0, \sigma^2_{inv}), \tag{3.13}$$

where $\sigma^2_{inv} = \lim_{n\to\infty}(1/n)\sum_{i=1}^n \sigma^2_{i;n,inv}$.

4. EL-based Approaches

The key of our proposed method is the fact that all the estimators in Section 3 can be considered as the solution of estimating equations $\sum_{i=1}^n m_{ni}(\beta_j) = 0$. In addition, $m_{ni}(\beta_j^0)$ admits an asymptotic decomposition when it is evaluated at the true value β_j^0:

$$m_{ni}(\beta_j^0) := m_n(\mathbf{X}_i, Y_i, \beta_j^0, \hat{\boldsymbol\beta}_{\backslash j}, \hat{\boldsymbol\theta}) := W_{ni} + R_{ni}, \tag{4.1}$$

where the nuisance parameters $\boldsymbol\beta_{\backslash j}$ and the other nuisance parameters denoted as $\boldsymbol\theta$ are replaced by their estimators $\hat{\boldsymbol\beta}_{\backslash j}$ and $\hat{\boldsymbol\theta}$. Moreover, the $\{W_{ni}\}_{i=1}^n$ are independent random variables, and the $\{R_{ni}\}_{i=1}^n$ satisfy the following.

(C0) $\mathrm{P}\big(\min_{1\le i\le n} m_{ni} < 0 < \max_{1\le i\le n} m_{ni}\big) \to 1$;

(C1) W_{ni}'s are independent with mean 0 and finite variance $\sigma^2_{i;n}$ such that $s_n^2/n \to \sigma_w^2$, where $s_n^2 = \sum_{i=1}^n \sigma^2_{i;n}$;

(C2) $n^{-1/2}\sum_{i=1}^n R_{ni} = o_p(1)$ and $\max_{1\le i\le n}|R_{ni}| = o_p(n^{1/2})$.

Condition (C0) implies that 0 is inside of the convex hull of "data points" m_{ni}'s, which ensures EL can be appropriately defined and computed. Condition (C1) and (C2), respectively, impose some conditions on the leading order term W_{ni} and small order term R_{ni} in the decomposition of $m_{ni}(\beta_j^0)$ so that the Wilks' theorem can be established for the EL ratio statistic based on m_{ni}'s. In particular, the condition (C2) implies that the errors due to the plug-in estimators of nuisance parameters $\hat{\boldsymbol\beta}_{\backslash j}, \hat{\boldsymbol\theta}$ are ignorable.

According to Owen (2001), with estimating equations, we can construct an EL statistic to make inference. Define the EL ratio function for the target parameter β_j as

$$\mathrm{EL}_n(\beta_j) = \max\left\{\prod_{i=1}^{n} np_i : p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i m_{ni}(\beta_j) = 0\right\}. \tag{4.2}$$

Under this unified framework, we have the following.

Theorem 1. *If* (C0)-(C2) *hold, then* $-2\log EL_n(\beta_j^0) \xrightarrow{d} \chi_1^2$.

Based on Theorem 1, an asymptotic α level test is given by rejecting H_0 if $-2\log \mathrm{EL}_n(\beta_j^0) > \chi^2_{1,\alpha}$ where $\chi^2_{1,\alpha}$ is the upper α quantile of χ^2_1. We can also construct a $(1-\alpha)100\%$ confidence interval for β_j as $\mathrm{CI}_\alpha = \{\beta_j : -2\log \mathrm{EL}_n(\beta_j) < \chi^2_{1,\alpha}\}$. Based on Propositions 1, 2, and 3, we see that the Wald-type inference procedure is hard to implement due to the complex asymptotic variance. Since the asymptotic distribution is chi-square, we do not need to estimate any additional parameters, such as the asymptotic variance. This is a great advantage, especially under the heteroscedastic linear regression models.

To apply Theorem 1 in practice, we need to find estimating equations $m_{ni}(\beta_j^0)$ for β_j^0 that admit decompositions that satisfy the conditions in Theorem 1. The following subsections outline three EL methods based on the estimators proposed in Sections 3.1, 3.2, and 3.3.

4.1. EL method based on low dimensional projection

The de-biased estimator (3.3) can be regarded as the solution to the estimating equation

$$\sum_{i=1}^{n} m_{ni}^{(\mathrm{lasso})}(\beta_j) := \sum_{i=1}^{n} \big(X_{ij} - \mathbf{X}_{i,\setminus j}^{\mathsf{T}}\hat{\mathbf{w}}_j\big)\big(Y_i - X_{ij}\beta_j - \mathbf{X}_{i,\setminus j}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{\setminus j}\big) = 0. \tag{4.3}$$

Here $\hat{\beta}^0_{\setminus j}$ is the estimation of a $p-1$ dimensional vector with all its elements from the initial estimator $\hat{\beta}$ except the j-th. The corresponding population counterpart of (4.3) is $\eta_{ij}\epsilon_i = \big\{X_{ij} - \mathrm{E}(X_{ij}|\mathbf{X}_{i,\setminus j})\big\}\big(Y_i - \mathbf{X}_i^{\mathsf{T}}\boldsymbol{\beta}^0\big)$. Simple algebra implies that $m_{ni}^{(\mathrm{lasso})}(\beta_j)$ has the decomposition

$$\begin{aligned} &m_{ni}^{(\mathrm{lasso})}(\beta_j^0)\\ &= \underbrace{\epsilon_i\eta_{ij}}_{W_{ni}^{(\mathrm{lasso})}} + \underbrace{\eta_{ij}(\boldsymbol{\beta}^0_{\setminus j} - \hat{\boldsymbol{\beta}}_{\setminus j})^{\mathsf{T}}\mathbf{X}_{i,\setminus j} + (\mathbf{w}_j^0 - \hat{\mathbf{w}}_j)^{\mathsf{T}}\mathbf{X}_{i,\setminus j}\big(Y_i - X_{ij}\beta_j^0 - \mathbf{X}_{i\setminus j}\hat{\boldsymbol{\beta}}_{\setminus j}\big)}_{R_{ni}^{(\mathrm{lasso})}}. \end{aligned}$$

For a fully understanding of the effect of heteroscedasticity, we study the asymptotics of $m_{ni}^{(\mathrm{lasso})}(\beta_j^0)$ in the following.

Proposition 4. *Under model* (2.1), $W_{ni}^{(lasso)}$ *has mean 0 and variance*

$$E\{(W_{ni}^{(lasso)})^2\} = \theta_{i;jj} - 2\mathbf{\Sigma}_{j,\backslash j}\mathbf{\Sigma}_{\backslash j,\backslash j}^{-1}\mathbf{\Theta}_{i;j,\backslash j} + \mathbf{\Sigma}_{j,\backslash j}\mathbf{\Sigma}_{\backslash j,\backslash j}^{-1}\mathbf{\Theta}_{i;\backslash j,\backslash j}\mathbf{\Sigma}_{\backslash j,\backslash j}^{-1}\mathbf{\Sigma}_{\backslash j,j}. \quad (4.4)$$

Here $\theta_{i;jj}$, $\mathbf{\Theta}_{i;j,\backslash j}$ *and* $\mathbf{\Theta}_{i;\backslash j,\backslash j}$ *are from the covariance matrix* $\mathbf{\Theta}_i = ((\theta_{i;jk}))$ *of* $\mathbf{Z}_i = \epsilon_i \mathbf{X}_i$. *Furthermore, if* ϵ_i *and* $\mathbf{X}_i$ *are independent and the error term is homoscedastic, then* $E\{(W_{ni}^{(lasso)})^2\} = \sigma_\epsilon^2(\sigma_{jj} - \mathbf{\Sigma}_{j,\backslash j}\mathbf{\Sigma}_{\backslash j,\backslash j}^{-1}\mathbf{\Sigma}_{\backslash j,j})$.

The comparison of the variances in Proposition 4 shows the difference between our heteroscedastic case and the homoscedastic case.

Let $\mathrm{EL}_n^{(\mathrm{lasso})}(\beta_j)$ be the EL-ratio test statistic defined by (4.2) using $m_{ni}^{(\mathrm{lasso})}(\beta_j)$ to replace $m_{ni}(\beta_j)$. The following Theorem demonstrates that the EL ratio test statistic $\mathrm{EL}_n^{(\mathrm{lasso})}(\beta_j)$ constructed based on the estimating equations (4.3) is asymptotically chi-square distributed.

Theorem 2. *Under some regularity conditions for the initial estimators as in Assumption* 1 *in the Appendix, assume that* $\mathbf{X}_i$ *and* ϵ_i *are both sub-Gaussian. As long as* $s\log p/\sqrt{n} = o(1)$, *the conditions* (C0)-(C2) *are satisfied. If* $\sigma_{n,lasso}^2 \to \sigma_{lasso}^2$ *for some* $\sigma_{lasso}^2 < \infty$, *then* $-2\log EL_n^{(lasso)}(\beta_j^0) \xrightarrow{d} \chi_1^2$.

Remark 1. Assumption 1 is needed to control the order of the remainder term $R_{ni}^{(\mathrm{lasso})}$ so that it satisfies the condition (C2). By applying appropriate inequalities, the order of remainder term is dominated by the orders of estimation errors of the initial estimators, and some quantities related to ϵ_i and $\mathbf{X}_i$, that can be, respectively, controlled by choosing appropriate initial regularized estimators (such as LASSO, SCAD and MCP) for $\boldsymbol{\beta}^0$ and $\mathbf{w}_j^0$, and the sub-Gaussian assumptions for ϵ_i and $\mathbf{X}_i$. For details, refer to the proof of Theorem 2.

Under the homoscedastic noise case, Zhang and Zhang (2014) and van de Geer, Bühlmann and Ritov (2013) used the Wald-type test statistic for testing H_0 based on the de-biased estimator $\hat{\beta}_j^{(\mathrm{de})}$. Ning and Liu (2014) considered the Score test statistic for testing H_0 based on the estimating equation (4.3). The Score test statistic and the Wald type test statistics are asymptotically equivalent. There still exist some differences between the two methods, as pointed out by Ning and Liu (2014). Our method constructs likelihood ratio tests based on the same estimating equation, thus it enjoys the nice properties of likelihood-based methods. Since we are using empirical likelihood, it not only enjoys the Wilk's phenomenon, but has other nice properties: the shape of the confidence interval is data driven, and our procedure is more robust to the distribution assumption for the error term since it only requires moment assumptions. Our method can be easily implemented under heteroscedasticity linear models due to the self studentization property of EL. Refer to the empirical studies in the simulation section for performance comparisons of our method with the Wald type test and the Score test.

4.2. EL method based on KFC method

The de-biased KFC estimator can be also represented as the solution to the estimating equation based on the population subject $\eta_{ij,\mathcal{S}}\epsilon_i := \big\{X_{ij} - \mathrm{E}(X_{ij}|\mathbf{X}_{i\mathcal{S}})\big\} \big(Y_i - \mathbf{X}_i^\intercal \boldsymbol{\beta}^0\big)$,

$$\sum_{i=1}^{n} m_{ni}^{(\mathrm{kfc})}(\beta_j) := \sum_{i=1}^{n} (\tilde{Y}_i - \tilde{X}_{ij}\beta_j - \tilde{\mathbf{X}}_{i\mathcal{S}^*}^\intercal \hat{\boldsymbol{\beta}}_{\mathcal{S}^*})\tilde{X}_{ij} = 0, \tag{4.5}$$

where $m_n^{(\mathrm{kfc})}(\beta_j^0)$ can be decomposed as, asymptotically,

$$\begin{aligned} m_{ni}^{(\mathrm{kfc})}(\beta_j^0) = {} & \epsilon_i \eta_{ij,\mathcal{S}} + \big(\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}} - X_{ij}\big)\mathbf{X}_{i\mathcal{S}}^\intercal(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\boldsymbol{\epsilon} \\ & + \big\{\epsilon_i - \mathbf{X}_{i\mathcal{S}}^\intercal(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\boldsymbol{\epsilon}\big\}\big\{\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}} - \mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbf{X}_{i\mathcal{S}}\big\} \\ & + \big\{X_{ij} - \mathbb{X}_j^\intercal\mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbf{X}_{i\mathcal{S}}\big\}\big\{\mathbf{X}_{i\mathcal{S}^*}^\intercal - \mathbf{X}_{i\mathcal{S}}^\intercal(\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\intercal\mathbb{X}_{\mathcal{S}^*}\big\}(\boldsymbol{\beta}_{\mathcal{S}^*}^0 \\ & - \hat{\boldsymbol{\beta}}_{\mathcal{S}^*}). \end{aligned}$$

We denote the first term as $W_{ni}^{(\mathrm{kfc})}$ and the others by $R_{ni}^{(\mathrm{kfc})}$. For simplicity, we assume the normality of $\mathbf{X}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ for the KFC projection section. Now, the $W_{ni}^{(\mathrm{kfc})} = \{\epsilon_i(X_{ij} - \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{X}_{i\mathcal{S}})\}_{i=1}^n$ are independent with $\mathrm{E}W_{ni}^{(\mathrm{kfc})} = 0$, and as in Proposition 4, it follows that $\mathrm{E}\{(W_{ni}^{(\mathrm{kfc})})^2\} = \theta_{i;jj} - 2\boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\boldsymbol{\Theta}_{i;j\mathcal{S}} + \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\boldsymbol{\Theta}_{i;\mathcal{S}\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}j}$. If we assume independence between ϵ_i and $\mathbf{X}_i$ and homoscedasticity for the error terms, we have $\mathrm{E}\{(W_{ni}^{(\mathrm{kfc})})^2\} = \sigma_\epsilon^2(\sigma_{jj} - \boldsymbol{\Sigma}_{j\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}j})$.

Let $\mathrm{EL}_n^{(\mathrm{kfc})}(\beta_j)$ be the empirical likelihood ratio test statistic defined by (4.2) with $m_{ni}^{(\mathrm{kfc})}(\beta_j)$ replaced by $m_{ni}(\beta_j)$.

Theorem 3. *Under Assumption* 3 *in the Appendix, the conditions* (C0)-(C2) *hold. If $\sigma_{n,kfc}^2 \to \sigma_{kfc}^2$ for some $\sigma_{kfc}^2 < \infty$, then $-2\log EL_n^{(kfc)}(\beta_j^0) \xrightarrow{d} \chi_1^2$.*

Remark 2. For the remainder term $R_{ni}^{(\mathrm{kfc})}$ to satisfy (C2), we need to control the error due to the initial estimators. We assume ϵ and $\mathbf{X}$ are sub-Gaussianian. In addition, for the KFC method, we need to control the partial correlation between X_j and any covariates that are not in the KFC set.

One of the key steps in our procedure is the selection of the KFC set. We propose the following. Based on the normality assumption of the predictors, we have the conditional distribution result for any give subset $\mathcal{S}$,

$$\rho_{jk}(\mathcal{S}) := \mathrm{Corr}(X_{ij}, X_{ik}|\mathbf{X}_{i\mathcal{S}}) = \sigma_{jk} - \boldsymbol{\Sigma}_{\mathcal{S}j}^\intercal\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\boldsymbol{\Sigma}_{\mathcal{S}k}.$$

The sample partial correlation can be evaluated as, $\hat{\rho}_{jk}(\mathcal{S}) = \tilde{\mathbb{X}}_j^\intercal \tilde{\mathbb{X}}_k / n$. For testing whether a partial correlation is zero or not, we can apply Fisher's z-transformation

$$\hat{F}_{jk} = \frac{1}{2} \log \left\{ \frac{1 + \hat{\rho}_{jk}(\mathcal{S})}{1 - \hat{\rho}_{jk}(\mathcal{S})} \right\}.$$

Classical decision theory then yields the following rule when using the significance level α: reject the null hypothesis $H_0 : \rho_{jk}(\mathcal{S}) = 0$ against the two-sided alternative $H_a : \rho_{jk}(\mathcal{S}) \neq 0$ if

$$\sqrt{n - |\mathcal{S}| - 3} |\hat{F}_{jk}| > z_{\alpha/2}.$$

We can then select the smallest size of $\mathcal{S}$ such that

$$\max_{k \in \mathcal{S}^*} \sqrt{n - |\mathcal{S}| - 3} |\hat{F}_{jk}| < z_{\alpha/2}.$$

To make this KFC set selection more stable, we adopt the stability selection proposed by Meinshausen and Bühlmann (2010) and Shah and Samworth (2013). According to Shah and Samworth (2013), we split the data into half B times and select the final KFC set with variables showing in at least 50% of those $2B$ KFC sets.

4.3. EL method based on the inverse method

We have $\hat{\mathfrak{b}}_j$ as the solution to the estimating equation

$$\sum_{i=1}^n m_{ni}^{(\mathrm{inv})}(\mathfrak{b}_j) := \sum_{i=1}^n \left(Y_i - \mathbf{X}_i^\intercal \hat{\boldsymbol{\beta}}\right) \left\{ X_{ij} - (Y_i, \mathbf{X}_{i,\setminus j}^\intercal) \hat{\boldsymbol{\gamma}}_j \right\} + n \mathfrak{b}_j = 0. \tag{4.6}$$

Simple algebra then immediately yields a decomposition of $m_{ni}^{(\mathrm{inv})}(\mathfrak{b}_j)$,

$$m_{ni}^{(\mathrm{inv})}(\mathfrak{b}_j^0) \\ = \underbrace{(\epsilon_i \eta_{ij,y} + \mathfrak{b}_j^0)}_{W_{ni}^{(\mathrm{inv})}} + \underbrace{\epsilon_i (Y_i, \mathbf{X}_{i,\setminus j}^\intercal)(\boldsymbol{\gamma}_j^0 - \hat{\boldsymbol{\gamma}}_j) + \mathbf{X}_i^\intercal (\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}) \left\{ X_{ij} - (Y_i, \mathbf{X}_{i,\setminus j}^\intercal) \hat{\boldsymbol{\gamma}}_j \right\}}_{R_{ni}^{(\mathrm{inv})}}.$$

Proposition 5. *Under model* (2.1)*, we have that $W_{ni}^{(inv)}$ has mean* 0 *and $E\{(W_{ni}^{(inv)})^2\} = \sigma_{i;n,inv}^2$. If ϵ_i and $\mathbf{X}_i$ are independent, then $E\{(W_{ni}^{(inv)})^2\} = Var(\epsilon_i)\ Var(\eta_{ij,y}) + (\gamma_{j1}^0)^2 \{Var(\epsilon_i^2) - Var^2(\epsilon_i)\}$. Under homoscedasticity and normality for ϵ_i, we have $E\{(W_{ni}^{(inv)})^2\} = \sigma_\epsilon^2 \sigma_{\eta_j,y}^2 + (\beta_j^0)^2 \sigma_{\eta_j,y}^4$.*

Let $\mathrm{EL}_n^{(\mathrm{inv})}(\beta_j)$ be the empirical likelihood ratio test statistic defined by (4.2) with $m_{ni}^{(\mathrm{inv})}(\beta_j)$ replaced by $m_{ni}(\beta_j)$.

Theorem 4. *Under conditions for the initial estimators as in Assumption* 2 *in the Appendix, and with $(\mathbf{X}_i^\intercal, \epsilon_i)^\intercal$ sub-Gaussian, if $s \log p / \sqrt{n} = o(1)$, the condi-*

tions (C0)-(C2) *are satisfied. If* $(1/n)\sum_{i=1}^{n}\sigma_{i;n,inv}^2 \to \sigma_{inv}^2$ *for some* $\sigma_{inv}^2 < \infty$, *then* $-2\log EL_n^{(inv)}(\mathfrak{b}_j^0) \xrightarrow{d} \chi_1^2$.

5. Simulation Studies

In this section, we report on simulation studies to investigate the finite sample performance of the proposed EL ratio tests, and we compare this performance with methods from in the existing literature.

We generated random samples according to model (2.1). The covariates were generated from a multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$. We considered three covariance matrices for $\boldsymbol{\Sigma} = ((\sigma_{jk}))$: a banded matrix with $\sigma_{jk} = \rho^{|j-k|}\mathbb{1}(|j-k| < 2)$, a Toeplitz matrix with $\sigma_{jk} = \rho^{|j-k|}$, and a block diagonal matrix with $\boldsymbol{\Sigma} = \mathbf{I}_{[p/3]} \otimes \mathbf{B}(\rho)$ where $\mathbf{B}(\rho)$ is a 3×3 matrix with the (i,j) component $\rho^{|i-j|}$. We set $\rho = 0.2$ and 0.5 in our simulation.

We also considered five scenarios for the error distribution: standard normal $\mathrm{N}(0,1)$, mixture normal distribution $0.7\mathrm{N}(0,1) + 0.3\mathrm{N}(0,5^2)$, t distribution with degrees of freedom 3, and heteroscedastic distributions $0.7X_1Z$ and $X_1Z\sum_{j=2}^{p}X_{j-1}X_j/(p-1)$ where $Z \sim \mathrm{N}(0,1)$ independent of $\mathbf{X}$. For the heteroscedastic distributions, ϵ is not independent of $\mathbf{X}$. For the first heteroscedastic case the conditional variance only depends on a low-dimensional covariate (the first component of the covariates $\mathbf{X}$), for the second, it depends on the the entire vector of covariates. Our goal is to test if the first coefficient is zero or not.

$$H_0 : \beta_1^0 = 0, \quad \text{v.s.} \quad H_1 : \beta_1^0 \neq 0.$$

The first component of the true coefficients $\boldsymbol{\beta}_1^0$ was set to $0, 0.1, 0.2, 0.3, 0.4$ and 0.5. Here 0 was used to evaluate the empirical size and the non-zero values were used to evaluate the power of the proposed methods. In addition, we set $\beta_4^0 = 1.5, \beta_7^0 = 2$ and all others to be 0. We chose $p = 100, 200, 500$ and $n = 200, 400$. The number of simulation replicates was 500.

We considered three EL-based methods proposed in Section 4, “EL-LASSO”, “EL-KFC”, and “EL-INV”, respectively, corresponds to the proposed method introduced in Section 4.1, 4.2, and 4.3. We compared them with two existing methods: the Wald type test proposed in Zhang and Zhang (2014) and and van de Geer, Bühlmann and Ritov (2013) (denoted by “Wald”) and the Score type test (denoted by “Score”) proposed in Ning and Liu (2014), with Lasso estimation for $\hat{\mathbf{w}}_1$. For initial estimators such as $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}_1$ and $\hat{\mathbf{w}}_1$, we applied the scaled Lasso of Sun and Zhang (2012), that has the advantage of being tuning insensitive. For the “EL-KFC”, to stabilize the KFC set selection, we used the stability selection procedure through sub-sampling proposed by Meinshausen and Bühlmann (2010) and Shah and Samworth (2013). According to the latter, we split the data into half 10 times, and selected the final KFC set with variables showing in at least 10 of those 20 KFC sets.

For the scenarios with normally distributed random errors, we observed that all the procedures were able to control type I error around the nominal level of 5%. The proposed EL-based approach with different estimating equations had very similar power. In general, the EL-based tests had better power performance than the existing methods, especially in the low sample size situations. Refer to the Supplemental Material for the simulation results in these cases.

Our main interest was to evaluate the performance of the proposed methods and some existing methods under the heteroscedastic linear regression model. Table 1 summarizes the results for the scenario with $\mathbf{X}$ generated as multivariate normal with the Toeplitz covariance matrix ($\rho = 0.2$) and the heteroscedastic error distribution $0.7X_1\mathrm{N}(0,1)$. Under this case, the EL-based inference procedures, "EL- KFC", "EL- INV" and "EL-LASSO", were asymptotically valid because they can control the type I errors reasonably well. For the 'Wald" and "Score" methods, type I errors were largely inflated, not surprising as these procedures were designed for linear models with homogeneous variance. In Table 2, we summarize the empirical size and power under another scenario with heteroscedastic error whose conditional error variance depends on high dimensional covariates generated according to $X_1 \sum_{j=2}^{p} X_{j-1}X_j\mathrm{N}(0,1)/(p-1)$. Although the error variance depends on a high-dimensional covariates, our methods were still able to control the type I error well under the null hypothesis. The "Wald" and "Score" methods had size distortion under the heteroscedastic error distribution.

6. An Empirical Study

We applied the proposed methods to study the association between gene expression and copy number alternation using a data set collected at multiple cancer centers (Feng, Fu and Sun (2010)). The data set contains gene expression and copy number alternation measured through primary breast tumor specimens in a few recent breast cancer cohort studies. In cells with cancer, mutations can cause a gene to be either deleted or duplicated on a chromosome, which leads to loss or gain of DNA copies of a gene. Comparative Genomic Hybridization (CGH) is a technique for measuring DNA copy numbers of genes of interest on the genome. The CGH array experiments return log_2 ratio between the number

Table 1. Empirical size and power of the proposed EL-based test procedures and two existing procedures under the heteroscedastic error case. In this table, covariates are generated by a multivariate normal distribution with covariance given by a Toeplitz matrix with $\rho = 0.2$, and the random error are generated according to $0.7X_1\mathrm{N}(0,1)$.

			β_1^0					
Method	p	n	0	0.1	0.2	0.3	0.4	0.5
EL-KFC	100	200	0.062	0.244	0.624	0.924	0.986	1.000
		400	0.040	0.366	0.916	0.998	1.000	1.000
	200	200	0.070	0.230	0.652	0.920	0.990	1.000
		400	0.076	0.350	0.890	0.990	1.000	1.000
	500	200	0.060	0.254	0.636	0.900	0.986	0.996
		400	0.058	0.402	0.902	0.992	1.000	1.000
EL-INV	100	200	0.058	0.230	0.620	0.910	0.986	1.000
		400	0.040	0.356	0.918	0.998	1.000	1.000
	200	200	0.058	0.222	0.652	0.910	0.988	1.000
		400	0.066	0.342	0.880	0.990	1.000	1.000
	500	200	0.060	0.236	0.624	0.898	0.980	0.996
		400	0.050	0.402	0.902	0.992	1.000	1.000
EL-LASSO	100	200	0.056	0.244	0.634	0.922	0.988	1.000
		400	0.046	0.376	0.926	1.000	1.000	1.000
	200	200	0.062	0.232	0.668	0.926	0.990	1.000
		400	0.072	0.356	0.890	0.988	1.000	1.000
	500	200	0.068	0.250	0.640	0.912	0.986	0.996
		400	0.052	0.412	0.902	0.992	1.000	1.000
Wald	100	200	0.256	0.496	0.860	0.986	1.000	1.000
		400	0.210	0.706	0.986	1.000	1.000	1.000
	200	200	0.234	0.464	0.848	0.980	1.000	1.000
		400	0.236	0.680	0.968	1.000	1.000	1.000
	500	200	0.208	0.516	0.874	0.978	1.000	1.000
		400	0.234	0.736	0.986	1.000	1.000	1.000
Score	100	200	0.256	0.490	0.860	0.986	1.000	1.000
		400	0.218	0.700	0.986	1.000	1.000	1.000
	200	200	0.234	0.470	0.846	0.980	1.000	1.000
		400	0.234	0.672	0.968	1.000	1.000	1.000
	500	200	0.204	0.518	0.870	0.978	1.000	1.000
		400	0.230	0.728	0.984	1.000	1.000	1.000

of DNA copies of a gene in the tumor cells and that in the reference cells. A positive (negative) measurement suggests a possible copy number gain (loss). After proper normalization, `cghFLasso` (Tibshirani and Wang (2008)) was used to estimate the underlying DNA copy numbers based on array outputs. Then the copy number alteration intervals (CNAIs), defined as basic CNA units (genome regions) in which all genes tend to be duplicated or deleted simultaneously, were

Table 2. Empirical size and power of the proposed EL-based test procedures and two existing procedures under the heteroscedastic error case. In this table, covariates are generated by a multivariate normal distribution with covariance given by a Toeplitz matrix with $\rho = 0.2$, and the random error are generated according to $X_1 \sum_{j=2}^{p} X_{j-1}X_j\mathrm{N}(0,1)/(p-1)$.

			β_1^0					
Method	p	n	0	0.1	0.2	0.3	0.4	0.5
EL-KFC	100	200	0.066	0.886	0.998	1.000	1.000	1.000
		400	0.048	0.988	1.000	1.000	1.000	1.000
	200	200	0.076	0.932	1.000	1.000	1.000	1.000
		400	0.068	0.988	1.000	1.000	1.000	1.000
	500	200	0.060	0.942	1.000	1.000	1.000	1.000
		400	0.054	1.000	1.000	1.000	1.000	1.000
EL-INV	100	200	0.062	0.872	0.998	1.000	1.000	1.000
		400	0.038	0.988	1.000	1.000	1.000	1.000
	200	200	0.074	0.936	1.000	1.000	1.000	1.000
		400	0.064	0.988	1.000	1.000	1.000	1.000
	500	200	0.056	0.938	1.000	1.000	1.000	1.000
		400	0.042	1.000	1.000	1.000	1.000	1.000
EL-LASSO	100	200	0.066	0.876	0.998	1.000	1.000	1.000
		400	0.046	0.988	1.000	1.000	1.000	1.000
	200	200	0.078	0.934	1.000	1.000	1.000	1.000
		400	0.064	0.988	1.000	1.000	1.000	1.000
	500	200	0.064	0.944	1.000	1.000	1.000	1.000
		400	0.046	1.000	1.000	1.000	1.000	1.000
Wald	100	200	0.222	0.982	1.000	1.000	1.000	1.000
		400	0.214	1.000	1.000	1.000	1.000	1.000
	200	200	0.244	0.990	1.000	1.000	1.000	1.000
		400	0.214	0.998	1.000	1.000	1.000	1.000
	500	200	0.260	0.990	1.000	1.000	1.000	1.000
		400	0.240	1.000	1.000	1.000	1.000	1.000
Score	100	200	0.226	0.984	1.000	1.000	1.000	1.000
		400	0.208	1.000	1.000	1.000	1.000	1.000
	200	200	0.236	0.990	1.000	1.000	1.000	1.000
		400	0.206	0.998	1.000	1.000	1.000	1.000
	500	200	0.260	0.990	1.000	1.000	1.000	1.000
		400	0.232	1.000	1.000	1.000	1.000	1.000

estimated by a clustering method based on the DNA copy numbers estimation. The gene expression data were collected by microarray expression experiments.

We had 172 specimens with both cDNA expression microarray and CGH array measurements. For each CNAI, the mean value of the estimated copy numbers of the genes falling into this CNAI was calculated. This resulted in a 172 (samples) by 384 (CNAIs) numeric matrix. We focused on a set of 654 breast cancer related genes based on seven published breast cancer gene lists. This resulted in a 172 (samples) by 654 (genes) numeric matrix. Please refer to Peng et al. (2010) for more details about the data preprocessing.

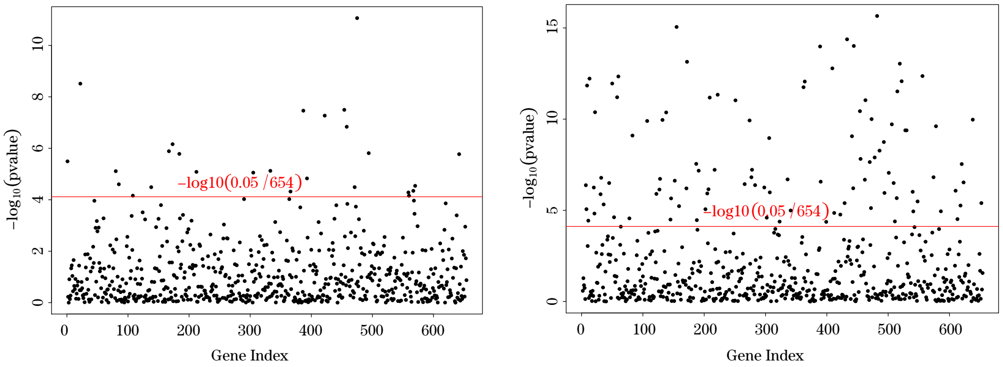
We studied the association between gene expression and DNA copy numbers through a high-dimensional linear regression model. Each of the 654 gene expression was used as response variable, and the DNA copy numbers were used as predictors. We focused on the genes with heteroscedastic error variance and first conducted a test to identify them.

We tested for the presence of heteroscedasticity for each of the 654 genes using test procedures proposed by Li and Yao (2015): the approximate likelihood-ratio test (ALRT) and the coefficient-of-variation test (CVT). They were constructed using the residuals obtained as $\mathbb{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}^0$, where $\hat{\boldsymbol{\beta}}^0$ is the ordinary least squares (OLS) estimate of $\boldsymbol{\beta}^0$. Although both the dimension of covariates and the sample size are allowed to grow to infinity simultaneously in their proposed test procedures, the covariates dimension needs to be less than the sample size, their proposed procedures were not directly applicable.

In order to apply them, for each of the 654 gene expressions we first selected variables by feature screening via distance correlation learning approach proposed by Li, Zhong and Zhu (2012), implemented in package *grpss*. This procedure has decent performance under heteroscedastic setting (Li, Zhong and Zhu (2012)). The p-values obtained by the two test procedures are summarized, respectively, in Figure 1 (a) and (b). To adjust for multiplicity, we applied the Bonferroni method to control the family-wise error rate. After the Bonferroni correction, 33 genes were declared to have significant heteroscedasticity based on the ALRT procedure, and 155 genes had significant heteroscedasticity based on the CVT procedure. Thus, heteroscedasticity exists for many genes in this data set.

We selected the top four genes with significant heteroscedasticity from the ALRT procedure among the common genes selected by both of ALRT and CVT for further analysis. The reason we chose ALRT is due to its robustness seen in Li and Yao (2015). The four selected genes are the 279-th gene named "SEMA3C" on Chr7, the 433-th gene named "POLR2F" on Chr22, the 493-th gene named "C18orf21" on Chr18, and the 610-th gene called "FOXA1" on Chr14.

We applied the proposed EL-based approaches to the four selected genes, and compared them with the "Wald" and "Score" tests described in the simulation studies. The results are showed in Figure 2. For each test procedure (EL-bassed approaches, "Wald" and "Score"), we obtained a sequence of p-values $\{p_j\}_{j=1}^{p}$,

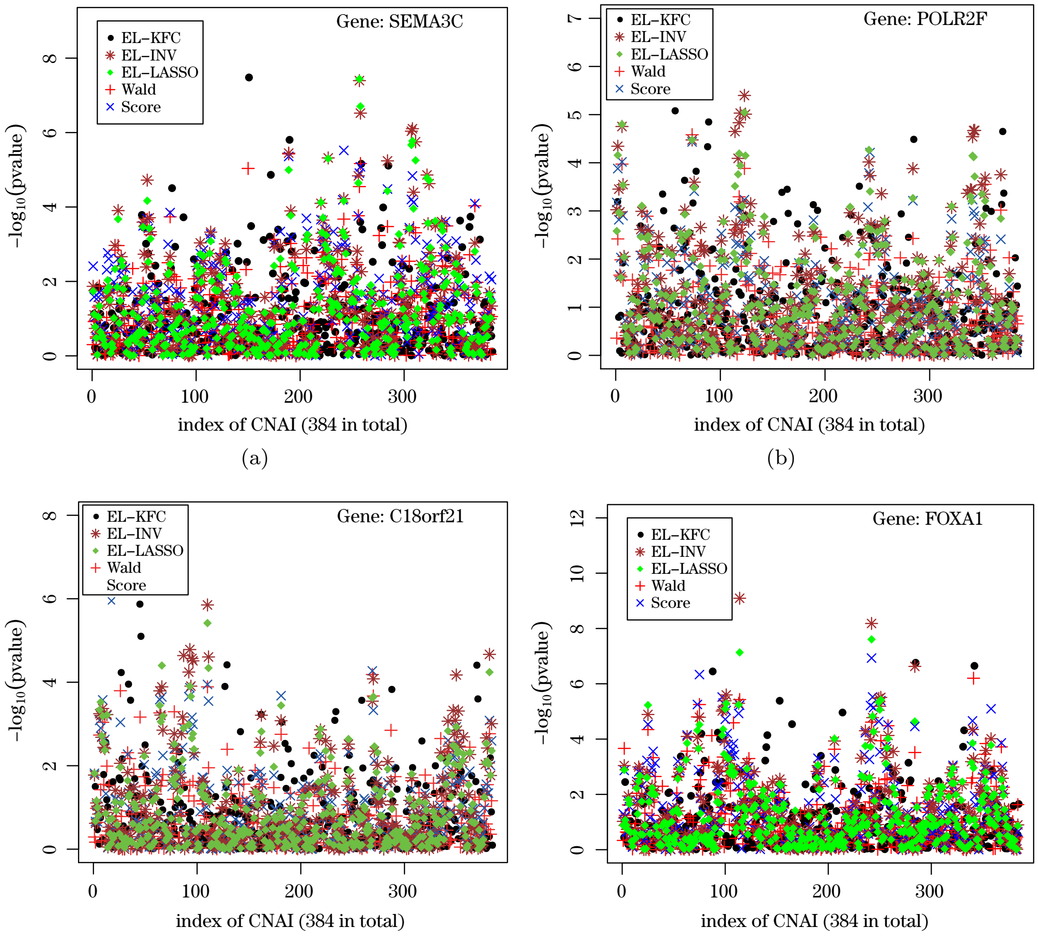


(a) ALRT (b) CVT

Figure 1. **p values for testing heteroscedasticity.** From ALRT, we got 33 genes with significant heteroscedasticity; from CVT, we got 155 genes with significant heteroscedasticity. The horizontal red line represents the Bonferoni threshold.

where p_j is the p-value for testing $H_{0j} : \beta_j^0 = 0$ vs $H_{1j} : \beta_j^0 \neq 0$ for $j = 1, \ldots, p$. Then we ordered p-values in an increasing order, $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(j)} \leq \cdots \leq p_{(p)}$, and applied the Benjamini-Hochberg (BH) to identify the significant hypotheses. Rejecting the null hypotheses $H_{0j} : \beta_j^0 = 0$ means here that the j-th CNAI are significantly associated with the gene expression.

As shown in Figure 2(d), for the gene "FOXA1" on chromosome 14, the 114-th and 258-th CNAIs were significant using the EL-based test procedures and the existing "Wald" and "Score" test procedures. For the gene "C18orf21" on chromosome 18, the 161-th CNAI was detected by the EL based methods as illustrated in Figure 2(c), but not detected by the "Score" and "Wald" tests. The 161-th CNAI corresponds to Cytoband 8p22. In the studies conducted by Tsuneizumi et al. (2002) and Voeghtly et al. (2012), it was found that the allelic loss in Cytoband 8p22 is closely related to the risk of breast cancer. Specifically, patients with tumors lost an allele at 8p22 had significantly higher risks of mortality than those with tumors retaining both alleles at those loci. In another study on the Human Protein Atlas (`http://www.proteinatlas.org/ENSG00000141428-C18orf21/cancer`), it was found that several cases of breast cancers exhibited moderate nuclear/nucleolar positivity of the gene "C18orf21". Finding the significant association between the expression of gene "C18orf21" and the CNA in Cytoband 8p22 can improve our understanding of the relationship between the discoveries in these studies. More importantly, it provided us some insight about the underlying disease mechanism of breast cancer. This



(a) (b)

(c) (d)

Figure 2. P-value Manhattan plots for top four genes showed significant heteroscedasticity.

shows the advantage of the EL-based proposed methods, and the necessarily of considering heteroscedasticity in this data set.

7. Discussion

We have studied inference problem for low-dimensional parameters in a high-dimensional heteroscedastic linear model. The asymptotic normalities of the existing estimators were established under the heteroscedastic linear model. but they are difficult to implement in practice due to the complicated asymptotic variance. To address the issue, we have proposed three EL-based approaches that avoid the explicit estimation of the variance. The key advantage of our EL-based methods is that they can allow for heteroscedastic error noise. In them, the conditional variance of random error is allowed to depend on the high-dimensional covariates so one can test statistical hypothesis and construct confidence intervals that data driven shapes. We do not require independence between the error term and the covariates, only that the error term and the covariates to be uncorrelated. The method we proposed provides a unified framework for testing low-dimensional coefficients in high-dimensional linear models when the estimating equations can be established and satisfy the conditions of Theorem 1. The procedure are simple to apply, where not needing to derive the asymptotic variances for estimators based on different estimating equations.

Supplementary Materials

In the supplemental file, we provide proofs to all the theoretical results presented in the paper, and some additional simulation results.

Acknowledgment

We are grateful to the guest Editor, an associate editor and two anonymous referees for constructive and insightful comments, which led to an improved manuscript. Zhong's research was partially supported by NSF DMS-1462156.

Appendix

Technical Assumptions

For a symmetric matrix $\mathbf{M} = ((M_{jk}))$, $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ are the minimal and maximal eigenvalues of $\mathbf{M}$. For any matrix $\mathbf{M} = ((M_{jk}))$, let $\|\mathbf{M}\|_{\max} = \max_{j,k} |M_{jk}|$, $\|\mathbf{M}\|_1 = \max_k \sum_j |M_{jk}|$, $\|\mathbf{M}\|_2 = \sqrt{\lambda_{\max}(\mathbf{M}^{\intercal}\mathbf{M})}$, and $\|\mathbf{M}\|_\infty = \max_j \sum_k |M_{jk}|$.

Assumption A1.

(1) The initial estimator $\hat{\boldsymbol{\beta}}$ satisfies $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 = O_p(s\sqrt{\log p/n})$.

(2) The initial estimators $\hat{\mathbf{w}}_j$ satisfy $\max_{1\le j\le p} \|\hat{\mathbf{w}}_j - \mathbf{w}_j^0\|_1 = O_p(a_n)$, where $a_n = o(1/\sqrt{\log p})$.

(3) The prediction errors satisfy $\|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/n = O_p(s\log p/n)$ and $\max_{1\le j\le p}$ $\|\mathbb{X}_{\setminus j}(\hat{\mathbf{w}}_j - \mathbf{w}_j^0)\|_2^2/n = O_p(b_n)$, where $\mathbb{X}_{\setminus j}$ is the design matrix $\mathbb{X}$ with the j-th column deleted and $b_n = o(1/\sqrt{n})$.

(4) $\mathbf{X}_i$ *and* ϵ_i *are all sub-Gaussian.*

(5) $s\log p/\sqrt{n} = o(1)$.

Remark 3. With (4), we have $X_{ik}\epsilon_i$ sub-exponential with $\mathrm{E}(\epsilon_i X_{ik}) = 0$. By the Bernstein inequality (Vershynin (2010)) and union bound inequality, we have

$$\mathrm{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i\right\|_\infty \geq t\right) \leq C_1 p \exp\left(-C\min\left(\frac{t^2}{C_2}, \frac{t}{C_3}\right)n\right).$$

By taking $t = C'\sqrt{\log p/n}$ for some positive constant C' such that $CC^{'2} > C_2$, we have

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i\right\|_\infty = O_p\left(\sqrt{\frac{\log p}{n}}\right). \tag{A.1}$$

With $\eta_{ij} = X_{ij} - \mathrm{E}(X_{ij}|\mathbf{X}_{i,\setminus j})$, η_{ij} sub-gaussian, for any $k \neq j$, we have $\mathrm{E}(X_{ik}\eta_{ij}) = \mathrm{E}[X_{ik}\{X_{ij} - \mathrm{E}(X_{ij}|\mathbf{X}_{i,\setminus j})\}] = \mathrm{E}\{X_{ik}X_{ij} - \mathrm{E}(X_{ik}X_{ij}|\mathbf{X}_{i,\setminus j})\} = 0$. Similarly, we have, for any $t > 0$ and $1 \leq j \neq k \leq p$,

$$\mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_{ik}\eta_{ij}\right| \geq t\right) \leq C_1 p \exp\left(-C\min\left(\frac{t^2}{C_2}, \frac{t}{C_3}\right)n\right),$$

which leads to

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\eta_{ij}\mathbf{X}_{i,\setminus j}\right\|_\infty = O_p\left(\sqrt{\frac{\log p}{n}}\right). \tag{A.2}$$

For the properties of the initial estimators in (1), (2) and (3) under the heteroscedasitic noise case, we can use the $\sqrt{\text{Lasso}}$ estimator as in Belloni, Chernozhukov and Wang (2014). According to Theorem 7 in Belloni, Chernozhukov and Wang (2014), we have that the $\sqrt{\text{Lasso}}$ estimators under certain conditions have these properties satisfied.

Assumption A2.

(1) *With the same assumption as in the Lasso projection case, initial estimator* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 = O_p(s\sqrt{\log p/n})$.

(2) *With the same assumption as in the Lasso projection case for the initial estimators* $\hat{\boldsymbol{\gamma}}_j$, $\max_{1\leq j\leq p}\|\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^0\|_1 = O_p(a_n)$, *where* $a_n = o(1/\sqrt{\log p})$.

(3) *With the same assumption as in the Lasso projection case for the prediction errors,* $\|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/n = O_p(s\log p/n)$ *and* $\max_{1\leq j\leq p}\|(\mathbb{Y}, \mathbb{X}_{\setminus j})(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^0)\|_2^2/n = O_p(b_n)$, *and* $b_n = o(1/\sqrt{n})$.

(4) $(\mathbf{X}_i^\intercal, \epsilon_i)^\intercal$ *is sub-Gaussian.*

(5) $s\log p/\sqrt{n} = o(1)$.

Remark 4. For the condition (2), if we assume $a = \max_{1\le j\le p} s_j$ with $s_j = \|\boldsymbol{\gamma}_j^0\|_0$, the $\sqrt{\text{Lasso}}$ estimators for $\boldsymbol{\gamma}_j^0$ satisfy this condition with $a_n = a\sqrt{\log p/n}$. For the condition (3), due to $\text{Cov}(\boldsymbol{\beta}^{0\intercal}\mathbf{X}_i, \epsilon_i) = \text{E}(\epsilon_i\boldsymbol{\beta}^{0\intercal}\mathbf{X}_i) = 0$, we have $\epsilon_i\boldsymbol{\beta}^{0\intercal}\mathbf{X}_i$ sub-exponential and, by the Bernstein inequality, we have for any $t > 0$,

$$\text{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^\intercal\boldsymbol{\beta}^0\epsilon_i\right| \ge t\right) \le 2\exp\left(-C_1 n\min\left(\frac{t^2}{C_2^2}, \frac{t}{C_2}\right)\right).$$

This leads to

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^\intercal\boldsymbol{\beta}^0\epsilon_i = O_p\left(\sqrt{\frac{\log p}{n}}\right), \tag{A.3}$$

as long as $\log p/n \to 0$. With the same argument, we have

$$\frac{1}{n}\sum_{i=1}^{n}X_{ik}\eta_{ij,y} = O_p\left(\sqrt{\frac{\log p}{n}}\right), \tag{A.4}$$

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i, \mathbf{X}_{i,\backslash j}^\intercal)\boldsymbol{\gamma}_j^0\eta_{ij,y} = O_p\left(\sqrt{\frac{\log p}{n}}\right). \tag{A.5}$$

Assumption A3.

(1) *For the eigenvalues of $\mathbf{\Sigma}$, there exist some constants $\lambda_{\min}$ and $\lambda_{\max}$ such that $0 < \lambda_{\min} < \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) < \lambda_{\max} < \infty$.*

(2) *$\mathbf{X}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$ and ϵ_i are sub-Gaussian.*

(3) *The initial estimator $\hat{\boldsymbol{\beta}}$ satisfies $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 = O_p(s\sqrt{\log p/n})$.*

(4) *$s\sqrt{(\log p)^2 m^3/n} = o(1)$ and $s\sqrt{(\log p)^3 m^2/n^2} = o(1)$ where m is the upper bound of the size of KFC set $|\mathcal{S}|$.*

(5) $s\sqrt{\log p}\sup_{\mathcal{S}:|\mathcal{S}|\le m}\max_{k\in\mathcal{S}^*}\left|\sigma_{jk} - \mathbf{\Sigma}_{j\mathcal{S}}\mathbf{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{S}k}\right| = o(1)$.

Remark 5. Condition (1) is a mild condition that assures the asymptotic identifiability of the model (Fan and Lv (2008); Wang (2009, 2012)). Condition (2) is a common condition used for simplification of theoretical proofs in high-dimensional setups; see for example, Wang (2009) and Zhang and Zhang (2014). Condition (4) is for controlling the size of the KFC set $|\mathcal{S}|$, and Condition (5) controls the partial correlation between the target covariate X_{ij} and $\mathbf{X}_{i\mathbf{S}^*}$.

References

Bai, Z., Pan, G. and Yin, Y. (2016). Homoscedasticity tests for both low and high-dimensional fixed design regressions. *arXiv preprint arXiv:1603.03830*.

Belloni, A., Chernozhukov, V., Wang, L., et al. (2014). Pivotal estimation via square-root Lasso in nonparametric regression. *The Annals of Statistics* **42**, 757–788.

Belsley, D. A. (2002). An investigation of an unbiased correction for heteroskedasticity and the effects of misspecifying the skedastic function. *Journal of Economic dynamics and Control* **26**, 1379–1396.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19**, 1212–1242.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, New York.

Chen, S.-X. and Qin, Y.-S. (2003). Coverage accuracy of confidence intervals in nonparametric regression. *Acta Mathematicae Applicatae Sinica (English Series)* **19**, 387–396.

Chen, S. X. and Van Keilegom, I. (2009). A review on empirical likelihood methods for regression. *Test* **18**, 415–447.

Daye, Z. J., Chen, J. and Li, H. (2012). High-dimensional heteroscedastic regression with an application to eqtl data analysis. *Biometrics* **68**, 316–326.

Dezeure, R., Bühlmann, P. and Zhang, C.-H. (2016). High-dimensional simultaneous inference with the bootstrap. *arXiv preprint arXiv:1606.03940*.

DiCiccio, T., Hall, P. and Romano, J. (1991). Empirical likelihood is bartlett-correctable. *The Annals of Statistics* **19**, 1053–1061.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96**, 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849–911.

Feng, J., Fu, W. and Sun, F. (2010). *Frontiers in Computational and Systems Biology*. Vol. 15. Springer Science & Business Media, New York.

Hall, P. (1990). Pseudo-likelihood theory for empirical likelihood. *The Annals of Statistics* **18**, 121–140.

Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *International Statistical Review/Revue Internationale de Statistique* **58**, 109–127.

Javanmard, A. and Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1306.3171*.

Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics* **28**, 1356–1378.

Lan, W., Zhong, P.-S., Li, R., Wang, H. and Tsai, C.-L. (2016). Testing a single regression coefficient in high dimensional linear models. *Journal of Econometrics* **195**, 154–168.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107**, 1129–1139.

Li, Z. and Yao, J. (2015). Testing for heteroscedasticity in high-dimensional regressions. *arXiv preprint arXiv:1510.00097*.

Liu, W. and Luo, S. (2014). Hypothesis testing for high-dimensional regression models. manuscript.

Lu, X. (2009). Empirical likelihood for heteroscedastic partially linear models. *Journal of Multivariate Analysis* **100**, 387–396.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 417–473.

Ning, Y. and Liu, H. (2014). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *arXiv preprint arXiv:1412.8765.*

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics* **18**, 90–120.

Owen, A. B. (2001). *Empirical Likelihood.* CRC Press, Boca Raton, FL.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R. and Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics* **4**, 53–77.

Reid, S., Tibshirani, R. and Friedman, J. (2016). A study of error variance estimation in Lasso regression. *Statistica Sinica* **26**, 35–67.

Shah, R. D. and Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 55–80.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99**, 879–898.

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

Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for cgh data using the fused Lasso. *Biostatistics* **9**, 18–29.

Tsao, M. and Wu, C. (2006). Empirical likelihood inference for a common mean in the presence of heteroscedasticity. *Canadian Journal of Statistics* **34**, 45–59.

Tsuneizumi, M., Emi, M., Hirano, A., Utada, Y., Tsumagari, K., Takahashi, K., Kasumi, F., Akiyama, F., Sakamoto, G., Kazui, T., et al. (2002). Association of allelic loss at 8p22 with poor prognosis among breast cancer cases treated with high-dose adjuvant chemotherapy. *Cancer Letters* **180**, 75–82.

van de Geer, S., Bühlmann, P. and Ritov, Y. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv preprint arXiv:1303.0518.*

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027.*

Voeghtly, L. M., Mamula, K., Campbell, J. L., Shriver, C. D. and Ellsworth, R. E. (2012). Molecular alterations associated with breast cancer mortality. *PloS one* **7**, e46814.

Wagener, J. and Dette, H. (2012). Bridge estimators and the adaptive Lasso under heteroscedasticity. *Mathematical Methods of Statistics* **21**, 109–126.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104**, 1512–1524.

Wang, H. (2012). Factor profiled sure independence screening. *Biometrika* **99**, 15–28.

Wang, L., Wu, Y. and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association* **107**, 214–222.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–838.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942.

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 (Statistical Methodology)* **76**, 217–242.

Zhou, M., Kim, M.-O. and Bathke, A. C. (2012). Empirical likelihood analysis for the heteroscedastic accelerated failure time model. *Statistica Sinica* **22**, 295–316.

Department of Mathematical Sciences, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA.

E-mail: hlwang@iupui.edu

Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA.

E-mail: pszhong@stt.msu.edu

Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA.

E-mail: cui@stt.msu.edu

(Received January 2017; accepted July 2017)