stats

MDPI

Article

A High Dimensional Omnibus Regression Test

Ahlam M. Abid [1], Paul A. Quaye [2] and David J. Olive [3,*]

[1] Department of Basic Sciences, Saudi Electronic University, Madinah 42351, Saudi Arabia; a.abid@seu.edu.sa
[2] Department of Statistical Sciences, Wake Forest University, Winston-Salem, NC 27109, USA; quayep@wfu.edu
[3] School of Mathematical & Statistical Sciences, Southern Illinois University, Carbondale, IL 62901, USA
* Correspondence: dolive@siu.edu

Abstract

Consider regression models where the response variable Y only depends on the $p \times 1$ vector of predictors $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ through the sufficient predictor $SP = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}$. Let the covariance vector $\mathrm{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Assume the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are independent and identically distributed random vectors for $i = 1, \ldots, n$. Then for many such regression models, $\boldsymbol{\beta} = \boldsymbol{0}$ if and only if $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$ where $\boldsymbol{0}$ is the $p \times 1$ vector of zeroes. The test of $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \boldsymbol{0}$ is equivalent to the high dimensional one sample test $H_0 : \boldsymbol{\mu} = \boldsymbol{0}$ versus $H_A : \boldsymbol{\mu} \neq \boldsymbol{0}$ applied to $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ where $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \mu_Y)$ and the expected values $E(\boldsymbol{x}) = \boldsymbol{\mu}_{\boldsymbol{x}}$ and $E(Y) = \mu_Y$. Since $\boldsymbol{\mu}_{\boldsymbol{x}}$ and μ_Y are unknown, the test of $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{\beta} \neq \boldsymbol{0}$ is implemented by applying the one sample test to $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y})$ for $i = 1, \ldots, n$. This test has milder regularity conditions than its few competitors. For the multiple linear regression one component partial least squares and marginal maximum likelihood estimators, the test can be adapted to test $H_0 : (\beta_{i_1}, \ldots, \beta_{i_k})^T = \boldsymbol{0}$ versus $H_1 : (\beta_{i_1}, \ldots, \beta_{i_k})^T \neq \boldsymbol{0}$ where $1 \leq k \leq p$.

Keywords: generalized linear models; multiple linear regression; one sample test; partial least squares; U-statistics

Academic Editor: Wei Zhu

Received: 13 September 2025
Revised: 27 October 2025
Accepted: 31 October 2025
Published: 5 November 2025

Citation: Abid, A.M.; Quaye, P.A.; Olive, D.J. A High Dimensional Omnibus Regression Test. *Stats* **2025**, *8*, 107. https://doi.org/10.3390/stats8040107

1. Introduction

This section reviews regression models where the response variable Y depends on the $p \times 1$ vector of predictors $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ only through the sufficient predictor $SP = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}$. Then there are n cases $(Y_i, \boldsymbol{x}_i^T)^T$. For the regression models, the conditioning and subscripts, such as i, will often be suppressed. This paper gives a high dimensional test for $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{\beta} \neq \boldsymbol{0}$ where $\boldsymbol{0} = (0, \ldots, 0)^T$ is the $p \times 1$ vector of zeroes.

A useful multiple linear regression (MLR) model is

$$Y_i = \alpha + x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + e_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \quad (1)$$

for $i = 1, \ldots, n$. Assume that the e_i are independent and identically distributed (iid) with expected value $E(e_i) = 0$ and variance $V(e_i) = \sigma^2$. In matrix form, this model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\phi} + \boldsymbol{e}, \quad (2)$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times (p+1)$ matrix with ith row $(1, \boldsymbol{x}_i^T)$, $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta}^T)^T$ is a $(p+1) \times 1$ vector, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Also $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

For a multiple linear regression model with heterogeneity, assume model (1) holds with $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{e}) = \boldsymbol{\Sigma}_{\boldsymbol{e}} = diag(\sigma_i^2) = diag(\sigma_1^2, \ldots, \sigma_n^2)$ is an $n \times n$ positive

definite matrix. Under regularity conditions, the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\phi}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ can be shown to be a consistent estimator of $\boldsymbol{\phi}$.

For estimation with ordinary least squares, let the covariance matrix of $\boldsymbol{x}$ be $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T]$ and the $p \times 1$ vector $\boldsymbol{\eta} = \text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = E[(\boldsymbol{x} - E(\boldsymbol{x})(Y - E(Y))] = (\text{Cov}(x_1, Y), \ldots, \text{Cov}(x_p, Y))^T$. Let

$$\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y})$$

and

$$\tilde{\boldsymbol{\eta}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{n-1}{n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}.$$

For a multiple linear regression model with iid cases, $\hat{\boldsymbol{\beta}}_{OLS}$ is a consistent estimator of $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ under mild regularity conditions, while $\hat{\alpha}_{OLS}$ is a consistent estimator of $E(Y) - \boldsymbol{\beta}_{OLS}^T E(\boldsymbol{x})$.

Ref. [1] showed that the one component partial least squares (OPLS) estimator $\hat{\boldsymbol{\beta}}_{OPLS} = \hat{\lambda}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ estimates $\lambda\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\beta}_{OPLS}$ where

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x}Y}}{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}} \text{ and } \hat{\lambda} = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}}{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}} \tag{3}$$

for $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \boldsymbol{0}$. If $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$, then $\boldsymbol{\beta}_{OPLS} = \boldsymbol{0}$. Also see [2–4]. Ref. [5] derived the large sample theory for $\hat{\boldsymbol{\eta}}_{OPLS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ and OPLS under milder regularity conditions than those in the previous literature, where $\boldsymbol{\eta}_{OPLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Ref. [6] showed that for iid cases $(\boldsymbol{x}_i, Y_i)$, these results still hold for multiple linear regression models with heterogeneity.

The marginal maximum likelihood estimator (MMLE or marginal least squares estimator) is due to [7,8]. This estimator computes the marginal regression of Y on x_i, such as Poisson regression, resulting in the estimator $(\hat{\alpha}_{i,M}, \hat{\beta}_{i,M})$ for $i = 1, \ldots, p$. Then $\hat{\boldsymbol{\beta}}_{MMLE} = (\hat{\beta}_{1,M}, \ldots, \hat{\beta}_{p,M})^T$.

For multiple linear regression, the marginal estimators are the simple linear regression estimators. Hence

$$\hat{\boldsymbol{\beta}}_{MMLE} = [diag(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})]^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}. \tag{4}$$

If the $\boldsymbol{t}_i$ are the predictors that are scaled or standardized to have unit sample variances, then

$$\hat{\boldsymbol{\beta}}_{MMLE} = \hat{\boldsymbol{\beta}}_{MMLE}(\boldsymbol{t}, Y) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t},Y} = \hat{\boldsymbol{\eta}}_{OPLS}(\boldsymbol{t}, Y) \tag{5}$$

where $(\boldsymbol{t}, Y)$ denotes that Y was regressed on $\boldsymbol{t}$. Ref. [6] derived large sample theory for the MMLE for multiple linear regression models, including models with heterogeneity.

For Poisson regression and related models, the response variable Y is a nonnegative count variable. A useful *Poisson regression (PR) model* is $Y \sim Poisson\left(e^{SP}\right)$. This model has $E(Y|SP) = V(Y|SP) = \exp(SP)$. The *quasi-Poisson regression model* has $E(Y|SP) = \exp(SP)$ and $V(Y|SP) = \phi\exp(SP)$ where the dispersion parameter $\phi > 0$. Note that this model and the Poisson regression model have the same conditional mean function, and the conditional variance functions are the same if $\phi = 1$.

Some notation is needed for the negative binomial regression model. If Y has a (generalized) negative binomial distribution, $Y \sim NB(\mu, \kappa)$, then the probability mass function (pmf) of Y is

$$P(Y = y) = \frac{\Gamma(y + \kappa)}{\Gamma(\kappa)\Gamma(y + 1)}\left(\frac{\kappa}{\mu + \kappa}\right)^{\kappa}\left(1 - \frac{\kappa}{\mu + \kappa}\right)^{y}$$

for $y = 0, 1, 2, \ldots$ where $\mu > 0$ and $\kappa > 0$. Then $E(Y) = \mu$ and $\text{V}(Y) = \mu + \mu^2/\kappa$.

The *negative binomial regression model* states that $Y_1, \ldots, Y_n$ are independent random variables with

$$Y|SP \sim NB(\exp(SP), \kappa).$$

This model has $E(Y|SP) = \exp(SP)$ and

$$V(Y|SP) = \exp(SP)\left(1 + \frac{\exp(SP)}{\kappa}\right) = \exp(SP) + \tau \exp(2\ SP).$$

Following Ref. [9] (p. 560), as $\tau \equiv 1/\kappa \to 0$, it can be shown that the negative binomial regression model converges to the Poisson regression model.

Let the log transformation $Z_i = \log(Y_i)$ if $Y_i > 0$ and $Z_i = \log(0.5)$ if $Y_i = 0$. This transformation often results in a linear model with heterogeneity:

$$Z_i = \alpha_Z + \boldsymbol{x}_i^T \boldsymbol{\beta}_Z + e_i \tag{6}$$

where the e_i are independent with expected value $E(e_i) = 0$ and variance $V(e_i) = \sigma_i^2$. For Poisson regression, the minimum chi-square estimator is the weighted least squares estimator from the regression of Z_i on $\boldsymbol{x}_i$ with weights $w_i = e^{Z_i}$. See [9] (pp. 611–612).

If the regression model for Y depends on $\boldsymbol{x}$ only through $\alpha + \boldsymbol{x}^T\boldsymbol{\beta}$, and if the predictors $\boldsymbol{x}_i$ are iid from a large class of elliptically contoured distributions, then [10,11] showed that, under regularity conditions, $\boldsymbol{\beta}_{OLS} = c\boldsymbol{\beta}$. Hence $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = c\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}$. Thus $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = d\boldsymbol{\beta}$ if $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \tau^2 \boldsymbol{I}_p$ where $\tau^2 > 0$ and $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. If $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$ in this case, then $\beta_i = 0$ implies that $Cov(x_i, Y) = 0$. The constant c is typically nonzero unless the model has a lot of symmetry about the distribution of $\alpha + \boldsymbol{x}^T\boldsymbol{\beta}$. Simulation with $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ can be difficult if the population values of c and d are unknown. Results from [12] (p. 89) suggest that for Poisson regression model, a rough approximation is $\hat{\boldsymbol{\beta}}_{PR} \approx \hat{\boldsymbol{\beta}}_{OLS}/\overline{Y}$. Results from [13] suggest that for binary logistic regression, a rough approximation is $\hat{\boldsymbol{\beta}}_{LR} \approx \hat{\boldsymbol{\beta}}_{OLS}/MSE$ where MSE is the mean square error from the OLS regression.

Ref. [14] has an interesting result for the multiple linear regression model (1). Assume that the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid with $E(Y) = \mu_Y$, $E(\boldsymbol{x}) = \boldsymbol{\mu}_{\boldsymbol{x}}$ and nonsingular $Cov(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$. Let $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Then testing $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ versus $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ is equivalent to testing $H_0 : \boldsymbol{\mu} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{0}$ with $\boldsymbol{\mu} = E(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{x}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ where $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \mu_Y - (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})^T\boldsymbol{\beta}_0)$, and a one sample test can be applied to $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y} - (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T\boldsymbol{\beta}_0)$.

Ref. [14] notes that there are only a few high dimensional analogs of the low dimensional multiple linear regression F-test for H_0 versus H_1. See [15–18]. The assumptions on the predictors in these four papers are very strong.

This paper uses the above test for $\boldsymbol{\beta}_0 = \boldsymbol{0}$, which is equivalent to a test for $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$. The resulting test is not limited to OLS for multiple linear regression with iid errors. As shown below and in the following paragraph, the test can be used for multiple linear regression when heterogeneity is present, and the test can also be used for many regression models that depend on the predictors only through $\boldsymbol{x}_i^T\boldsymbol{\beta}$. Suppose $\boldsymbol{\beta}_D = \boldsymbol{D}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ where $\boldsymbol{D}$ is a $p \times p$ positive definite matrix. Then $\boldsymbol{\beta}_D = \boldsymbol{0}$ if and only if $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$. Then $\boldsymbol{D}^{-1} = \lambda \boldsymbol{I}$ for OPLS, $\boldsymbol{D}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ for OLS, and $\boldsymbol{D}^{-1} = [diag(\boldsymbol{\Sigma}_{\boldsymbol{x}})]^{-1}$ for the MMLE. The k-component partial least squares estimator can be found by regressing Y on a constant and on $W_i = \hat{\boldsymbol{\eta}}_i^T\boldsymbol{x}$ for $i = 1, \ldots, k$ where $\hat{\boldsymbol{\eta}}_i = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{i-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ for $i = 1, \ldots, k$. See [19]. Hence $\boldsymbol{\beta}_{kPLS} = \boldsymbol{0}$ if $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$. Thus if the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid, then using $\boldsymbol{\beta}_0 = \boldsymbol{0}$ gives tests for $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$, $H_0 : \boldsymbol{\beta}_{MMLE} = \boldsymbol{0}$, $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$, $H_0 : \boldsymbol{\beta}_{OPLS} = \boldsymbol{0}$, and $H_0 : \boldsymbol{\beta}_{kPLS} = \boldsymbol{0}$. For multiple linear regression with heterogeneity, $\hat{\boldsymbol{\beta}}_{OLS}$ is still a consistent estimator of $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Hence the test can be used when the constant variance assumption is violated.

Under iid cases with $\boldsymbol{\beta} = \boldsymbol{0}$, if the response variables Y_i depend on the $\boldsymbol{x}_i$ only through $\boldsymbol{x}_i^T\boldsymbol{\beta}$, then $Y_i \sim Y_i|\alpha \sim Y_i|(\alpha + \boldsymbol{x}_i^T\boldsymbol{\beta})$. Hence the Y_i are iid and do not depend on $\boldsymbol{x}$, and

thus satisfy a multiple linear regression model with $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta} = \mathbf{0}$. For a parametric regression, such as a generalized linear model, assume $Y_i \sim D(\tau(\alpha + \boldsymbol{x}_i^T\boldsymbol{\beta}), \boldsymbol{\theta})$ where D is the parametric distribution and τ is a real valued function. For example, D could be the negative binomial distribution with $\tau(SP) = e^{SP}$ and $\theta = \kappa$. If $\boldsymbol{\beta} = \mathbf{0}$, then the iid $Y_i \sim D(\tau(\alpha), \boldsymbol{\theta})$. Typically, if $\boldsymbol{\beta} \neq \mathbf{0}$, then $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \mathbf{0}$, and the test can have good power. An exception is when there is a lot of symmetry which rarely occurs with real data. For example, suppose $Y = m(SP) + e$ where the iid errors $e_i \sim N(0, \sigma_1^2)$ are independent of the predictors, $SP \sim N(0, \sigma_2^2)$, and the function m is symmetric about 0, e.g., $m(SP) = (SP)^2$. Then $\boldsymbol{\beta}_{OLS} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \mathbf{0}$ even if $\boldsymbol{\beta} \neq \mathbf{0}$.

If $\boldsymbol{\beta}_0 = \mathbf{0}$, then $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \mu_Y)$, and $E(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Then apply a high dimensional one sample test on the $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y})$. Note that the sample mean $\overline{\boldsymbol{v}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$.

Section 2.1 reviews and derives some results for the one sample test that will be used. Section 2.2 reviews some two sample tests. Section 2.3 gives theory for the test given in the above paragraph.

2. Materials and Methods

2.1. A High Dimensional One Sample Test

This section reviews and derives some results for the one sample test that will be used. Suppose $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are iid random vectors with $E(\boldsymbol{x}) = \boldsymbol{\mu}$ and covariance matrix $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$. Then the test $H_0 : \boldsymbol{\mu} = \mathbf{0}$ versus $H_1 : \boldsymbol{\mu} \neq \mathbf{0}$ is equivalent to the test $H_0 : \boldsymbol{\mu}^T\boldsymbol{\mu} = 0$ versus $H_1 : \boldsymbol{\mu}^T\boldsymbol{\mu} \neq 0$. Let $\boldsymbol{S} = \hat{\boldsymbol{\Sigma}}$. A U-statistic for estimating $\boldsymbol{\mu}^T\boldsymbol{\mu}$ is

$$T_n = T_n(\boldsymbol{x}) = \frac{1}{n(n-1)} \sum_{i \neq j} \boldsymbol{x}_i^T \boldsymbol{x}_j = \frac{n\overline{\boldsymbol{x}}^T\overline{\boldsymbol{x}} - tr(\boldsymbol{S})}{n} \tag{7}$$

where tr() is the trace function. See, for example, [20].

To see that the last equality holds, note that

$$T_n = \frac{1}{n(n-1)} \left[\sum_i \sum_j \boldsymbol{x}_i^T \boldsymbol{x}_j - \sum_i \boldsymbol{x}_i^T \boldsymbol{x}_i \right] = \frac{n^2 \overline{\boldsymbol{x}}^T \overline{\boldsymbol{x}} - \sum_i \boldsymbol{x}_i^T \boldsymbol{x}_i}{n(n-1)}.$$

Now

$$\boldsymbol{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = \frac{1}{n-1} \left[\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^T - n\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^T \right].$$

Thus

$$tr(\boldsymbol{S}) = \frac{1}{n-1} \left[\sum_i tr(\boldsymbol{x}_i \boldsymbol{x}_i^T) - ntr(\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^T) \right] = \frac{1}{n-1} \left[\sum_i \boldsymbol{x}_i^T \boldsymbol{x}_i - n\overline{\boldsymbol{x}}^T\overline{\boldsymbol{x}} \right].$$

Thus

$$n\overline{\boldsymbol{x}}^T\overline{\boldsymbol{x}} - tr(\boldsymbol{S}) = n\overline{\boldsymbol{x}}^T\overline{\boldsymbol{x}} + \frac{n}{n-1}\overline{\boldsymbol{x}}^T\overline{\boldsymbol{x}} - \frac{1}{n-1}\sum_i \boldsymbol{x}_i^T\boldsymbol{x}_i = \frac{n^2\overline{\boldsymbol{x}}^T\overline{\boldsymbol{x}} - \sum_i \boldsymbol{x}_i^T\boldsymbol{x}_i}{n-1}.$$

Next, we derive a simple test. Let the variance $V(\boldsymbol{x}_i^T\boldsymbol{x}_j) = V(W) = V(W_{ij}) = \sigma_W^2$ for $i \neq j$. Let $m = floor(n/2) = \lfloor n/2 \rfloor$ be the integer part of $n/2$. So floor(100/2) = floor(101/2) = 50. Let the iid random variables $W_1 = \boldsymbol{x}_1^T\boldsymbol{x}_2, W_2 = \boldsymbol{x}_3^T\boldsymbol{x}_4, , \ldots, W_m = \boldsymbol{x}_{2m-1}^T\boldsymbol{x}_{2m}$. Note that $E(W_i) = \boldsymbol{\mu}^T\boldsymbol{\mu}$ and $V(W_i) = \sigma_W^2$. Let S_W^2 be the sample variance of the W_i:

$$S_W^2 = \frac{1}{m-1} \sum_{i=1}^{m} (W_i - \overline{W})^2.$$

The following new theorem follows from the univariate central limit theorem.

Theorem 1. *Assume $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are iid, $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$, and the variance $V(\boldsymbol{x}_i^T \boldsymbol{x}_j) = \sigma_W^2$ for $i \neq j$. Let $W_1, \ldots, W_m$ be defined as above. Then*

(a) $\sqrt{m}(\overline{W} - \boldsymbol{\mu}^T\boldsymbol{\mu}) \xrightarrow{D} N(0, \sigma_W^2)$.

$$(b)\ \frac{\sqrt{m}(\overline{W} - \boldsymbol{\mu}^T\boldsymbol{\mu})}{S_W} \xrightarrow{D} N(0,1)$$

as $n \to \infty$.

The following theorem derives the variance $V(T_n)$ under simpler regularity conditions than those in the literature, and the new proof of the theorem is also simpler.

Theorem 2. *Assume $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are iid, $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$, and the variance $V(\boldsymbol{x}_i^T \boldsymbol{x}_j) = \sigma_W^2$ for $i \neq j$. Let $W_{ij} = \boldsymbol{x}_i^T \boldsymbol{x}_j$ for $i \neq j$. Let $\theta = Cov(W_{ij}, W_{id}) = \boldsymbol{\mu}^T \boldsymbol{\Sigma} \boldsymbol{\mu}$ where $j \neq d$, $i < j$, and $i < d$. Then*

$$(a)\ V(T_n) = \frac{2\sigma_W^2}{n(n-1)} + \frac{4(n-2)\theta}{n(n-1)}.$$

(b) If $H_0 : \boldsymbol{\mu} = \mathbf{0}$ is true, then $\theta = 0$ and

$$V_0 = V(T_n) = \frac{2\sigma_W^2}{n(n-1)}.$$

Proof. (a) To find the variance $V(T_n)$ with T_n from Equation (7), let $W_{ij} = \boldsymbol{x}_i^T \boldsymbol{x}_j = W_{ji}$, and note that

$$T_n = \frac{2}{n(n-1)} H_n \ \ where \ \ H_n = \sum_{i<} \sum_{j} \boldsymbol{x}_i^T \boldsymbol{x}_j = \sum_{i<j} \boldsymbol{x}_i^T \boldsymbol{x}_j.$$

Then $V(H_n) = Cov(H_n, H_n) =$

$$Cov\left(\sum_{i<} \sum_{j} W_{ij}, \sum_{k<} \sum_{d} W_{kd} \right) = \sum_{i<} \sum_{j} \sum_{k<} \sum_{d} Cov(W_{ij}, W_{kd}). \tag{8}$$

Let $V(W_{ij}) = \sigma_W^2$ for $i \neq j$. The covariances are of 3 types. First, if $(ij) = (kd)$ with $i < j$, then $Cov(W_{ij}, W_{kd}) = V(W_{ij}) = \sigma_W^2$. Second, if i, j, k, d are distinct with $i < j$ and $k < d$, then W_{ij} and W_{kd} are independent with $Cov(W_{ij}, W_{kd}) = 0$. Third, there are terms where exactly three of the four subscripts are distinct, which have $Cov(W_{ij}, W_{id}) = \theta$ where $j \neq d$, $i < j$, and $i < d$ or $Cov(W_{ij}, W_{kj}) = \theta$ where $i \neq k$, $i < j$, and $k < j$. These covariance terms are all equal to the same number θ since $W_{ij} = W_{ji}$. The number of ways to get three distinct subscripts is

$$a - b - c = \binom{n}{2}^2 - \binom{n}{2}\binom{n-2}{2} - \binom{n}{2} = n(n-1)(n-2)$$

since a is the number of terms on the right hand side of (8), b is the number of terms where i, j, k, d are distinct with $i < j$ and $k < d$, and c is the number of terms where $(ij) = (kd)$ with $i < j$. [Note that $n(n-1)$ terms have i and j distinct. Half of these terms have $i < j$ and half have $i > j$. Similarly, $n(n-1)(n-2)(n-3)$ terms have $ijkd$ distinct, and half of the $n(n-1)$ terms have $i < j$, while half of the $(n-2)(n-3)$ terms have $k < d$.] Thus

$$V(H_n) = 0.5n(n-1)\sigma_W^2 + n(n-1)(n-2)\theta.$$

This calculation was adapted from [21] (pp. 336–337). Thus

$$V(T_n) = \frac{4}{[n(n-1)]^2} V(H_n) = \frac{2\sigma_W^2}{n(n-1)} + \frac{4(n-2)\theta}{n(n-1)}.$$

(b) Now $\theta = Cov(\boldsymbol{x}_i^T\boldsymbol{x}_j, \boldsymbol{x}_i^T\boldsymbol{x}_k)$ where $\boldsymbol{x}_i, \boldsymbol{x}_j$, and $\boldsymbol{x}_k$ are iid. Hence $\theta =$

$$Cov(\sum_d x_{id}x_{jd}, \sum_t x_{it}x_{kt}) = \sum_d\sum_t Cov(x_{id}x_{jd}, x_{it}x_{kt}) =$$

$$\sum_d\sum_t [E(x_{id}x_{jd}x_{it}x_{kt}) - E(x_{id}x_{jd})E(x_{it}x_{kt})] =$$

$$\sum_d\sum_t [E(x_{id}x_{it})E(x_{jd})E(x_{kt}) - E(x_{id})E(x_{jd})E(x_{it})E(x_{kt})] =$$

$$\sum_d\sum_t [E(x_{jd})E(x_{kt})(E(x_{id}x_{it}) - E(x_{id})E(x_{it}))] =$$

$$\sum_d\sum_t [E(x_{jd})E(x_{kt})\ Cov(x_{id}, x_{it})] = \boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu}.$$

Under H_0, $\boldsymbol{\mu} = \mathbf{0}$ and thus $\theta = 0$. □

Note that T_n is the sample mean of the $0.5n(n-1)$ distinct, identically distributed $W_{ij} = \boldsymbol{x}_i^T\boldsymbol{x}_j$ for $i < j$. When $\boldsymbol{\mu} = \mathbf{0}$, Theorem 2 proves that the W_{ij} are uncorrelated. Hence when H_0 is true, $V(T_n)$ satisfies (Theorem 2b). Ref. [14] (p. 2024) showed that $V(W_{ij}) = V(\boldsymbol{x}_i^T\boldsymbol{x}_j) = \sigma_W^2 = tr(\boldsymbol{\Sigma}^2) + 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu}$. Plugging this value into (Theorem 2a) gives the [22] result

$$V(T_n) = \frac{2}{n(n-1)}tr(\boldsymbol{\Sigma}^2) + \frac{4\boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu}}{n}.$$

Note that $\theta = \boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu}$ can be consistently estimated as follows. Let $g = floor(n/3)$. Let $W_1 = \boldsymbol{x}_1^T\boldsymbol{x}_2$, $Z_1 = \boldsymbol{x}_1^T\boldsymbol{x}_3$, $W_2 = \boldsymbol{x}_4^T\boldsymbol{x}_5$, $Z_2 = \boldsymbol{x}_4^T\boldsymbol{x}_6$, …, $W_g = \boldsymbol{x}_{3g-2}^T\boldsymbol{x}_{3g-1}$, $Z_g = \boldsymbol{x}_{3g-2}^T\boldsymbol{x}_{3g}$. Then $\hat{\theta}$ is the sample covariance of the (W_i, Z_i) where $i = 1, \ldots, g$. Note that a consistent estimator of $tr(\boldsymbol{\Sigma}^2)$ is $S_W^2 - 2\hat{\theta}$.

Let $\hat{V}(T_n)$ and $\hat{V}_0(T_n)$ be consistent estimators of $V(T_n)$ and $V_0(T_n)$, respectively. Then ref. [22–25], and others proved that under mild regularity conditions when H_0 is true,

$$T_n/\sqrt{\hat{V}(T_n)} = T_n/\sqrt{\hat{V}_0(T_n)} \xrightarrow{D} N(0,1).$$

Under regularity conditions when H_0 is true, ref. [25] proved that $T_n/\sqrt{\hat{V}_0(T_n)} \xrightarrow{D} t_k$ as $p \to \infty$ for fixed $n \geq 3$ where $k = 0.5n(n-1) - 1$.

A consistent estimator of $V_0(T_n)$ needs a consistent estimator of $\sigma_W^2 = 0.5n(n-1)$ $V_0(T_n)$. Let $s_n^2 = \hat{V}_0(T_n)$. Then one estimator is $0.5n(n-1)s_n^2 = S_W^2$ from Theorem 1. An estimator nearly the same as the one used by [25] is

$$0.5n(n-1)s_n^2 = \hat{\sigma}_W^2 = \frac{1}{n(n-1)}\sum\sum_{i\neq j}(\boldsymbol{x}_i^T\boldsymbol{x}_j - T_n)^2 = \frac{1}{n(n-1)}\sum\sum_{i\neq j}(W_{ij} - T_n)^2.$$

Note that σ_W can be proportional to p since σ_W is the standard deviation of a sum of p random variables. Thus to have good asymptotic power against all alternatives, likely need $p/n \to 0$ as $n, p \to \infty$. When $\boldsymbol{\mu} \neq \mathbf{0}$, $T_n/\sqrt{\hat{V}_0(T_n)}$ tends to have more power than $T_n/\sqrt{\hat{V}(T_n)}$ since $V_0(T_n) < V(T_n)$. Suppose $\boldsymbol{\mu} = \delta\mathbf{1}$ where the constant $\delta > 0$ and $\mathbf{1}$ is the $p \times 1$ vector of ones. Then $\boldsymbol{\mu}^T\boldsymbol{\mu} = \delta^2 p$, and the test using $\hat{V}_0(T_n)$ may have good power for $T_n/\sqrt{\hat{V}_0(T_n)} > 1.96 \approx 2$ or for

$$\frac{\delta^2\ p}{\sqrt{\frac{2\sigma_W^2}{n(n-1)}}} > 2 \ \ or \ \ \delta^2 > \frac{2\sqrt{2}\,\sigma_W}{n\ p}.$$

For computing $\hat{V}_0(T_n)$, a question is whether to use an estimator of σ_W^2 or of $\tau^2 = tr(\boldsymbol{\Sigma}^2)$. Let the ijth element of $\boldsymbol{\Sigma}$ be σ_{ij} with $\boldsymbol{\Sigma} = (\sigma_{ij})$. Let $\|\boldsymbol{\Sigma}\|_F$ be the Frobenius norm of $\boldsymbol{\Sigma}$, and $\|\boldsymbol{a}\|$ be the Euclidean norm of vector $\boldsymbol{a}$. Let $vec(\boldsymbol{\Sigma})$ be the vector formed by stacking the columns of $\boldsymbol{\Sigma}$ into a vector. Then $\tau^2 = tr(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}) = \|\boldsymbol{\Sigma}\|_F^2 = \sum_{i=1}^p \sum_{j=1}^p \sigma_{ij}^2 = \|vec(\boldsymbol{\Sigma})\|^2$. There is a level-power tradeoff. Using $\hat{\sigma}_W^2$ is good for controlling the level = P(type I) error when H_0 is true. Since $\sigma_W^2 = \tau^2 + 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu} = \tau^2 + 2\theta$, the parameter τ^2 can be much smaller than σ_W^2, and using a good estimator of τ^2 may result in better power.

In high dimensions, it is often very difficult to estimate a $k \times 1$ vector $\boldsymbol{\theta}$ when $k > n$. This result is a form of "the curse of dimensionality." If a $\sqrt{n}$ consistent estimator of $\boldsymbol{\theta}$ is available, then the squared norm

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = \sum_{i=1}^{k} (\hat{\theta}_i - \theta_i)^2 \propto k/n.$$

Hence estimators $\hat{\tau}^2$ that use many parameters, such as plug in estimators $\hat{\boldsymbol{\Sigma}}$, are likely to be poor. The two parameter estimator $\hat{\tau}^2 = \hat{\sigma}_W^2 - 2\hat{\theta}$ likely has more variability than $\hat{\sigma}_W^2$ when H_0 is true, and better estimators of θ are needed. In simulations, $\hat{\tau}_1^2 = \hat{\sigma}_W^2 - 2\hat{\theta}$ was often negative. Let $\hat{\tau}_2^2 = \hat{\tau}_1^2$ if $\hat{\tau}_1^2 > 0$ and $\hat{\tau}_2^2 = \hat{\sigma}_W^2$, otherwise. In limited simulations, this estimator did about as well as $\hat{\tau}_3^2 = \hat{\sigma}_W^2$. Obtaining an estimator that clearly outperforms $\hat{\sigma}_W^2$ would improve the omnibus test, but is beyond the scope of this paper.

We also considered replacing $\boldsymbol{x}_i$ by $\boldsymbol{z}_i = ss(\boldsymbol{x}_i)$ where the spatial sign function $ss(\boldsymbol{x}_i) = \boldsymbol{0}$ if $\boldsymbol{x}_i = \boldsymbol{0}$, and $ss(\boldsymbol{x}_i) = \boldsymbol{x}_i/\|\boldsymbol{x}_i\|$ otherwise. This function projects the nonzero $\boldsymbol{x}_i$ onto the unit p-dimensional hypersphere centered at $\boldsymbol{0}$. Let $T_n(\boldsymbol{w})$ denote the statistic T_n computed from an iid sample $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$. Since the $\boldsymbol{z}_i$ are iid if the $\boldsymbol{x}_i$ are iid, use $T_n(\boldsymbol{z})$ to test $H_0 : \boldsymbol{\mu}_{\boldsymbol{z}} = \boldsymbol{0}$ versus $H_A : \boldsymbol{\mu}_{\boldsymbol{z}} \neq \boldsymbol{0}$ where $\boldsymbol{\mu}_{\boldsymbol{z}} = E(\boldsymbol{z}_i)$. In general, $\boldsymbol{\mu}_{\boldsymbol{z}} \neq \boldsymbol{\mu} = \boldsymbol{\mu}_{\boldsymbol{x}} = E(\boldsymbol{x}_i)$, but $\boldsymbol{\mu}_{\boldsymbol{z}} = \boldsymbol{\mu} = \boldsymbol{0}$ can occur if the $\boldsymbol{x}_i$ have a lot of symmetry about $\boldsymbol{0}$. In particular, $\boldsymbol{\mu}_{\boldsymbol{z}} = \boldsymbol{\mu} = \boldsymbol{0}$ if the $\boldsymbol{x}_i$ are iid from an elliptically contoured distribution with $E(\boldsymbol{x}_i) = \boldsymbol{\mu} = \boldsymbol{0}$. The test based on the statistic $T_n(\boldsymbol{z})$ can be useful if the first or second moments of the $\boldsymbol{x}_i$ do not exist, for example if the $\boldsymbol{x}_i$ are iid from a multivariate Cauchy distribution. These results may be useful for understanding papers such as [26].

The nonparametric bootstrap draws a bootstrap data set $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*$ with replacement from the $\boldsymbol{x}_i$ and computes T_1^* by applying T_n on the bootstrap data set. This process is repeated B times to get a bootstrap sample $T_1^*, \ldots, T_B^*$. For the statistic T_n, the nonparametric bootstrap fails in high dimensions because terms like $\boldsymbol{x}_j^T\boldsymbol{x}_j$ need to be avoided, and the nonparametric bootstrap has replicates: the proportion of cases in the bootstrap sample that are not replicates is about $1 - e^1 \approx 2/3 \approx 7/11$. The m out of n bootstrap draws a sample of size m without replacement from the n cases. Using $m = floor(2n/3)$ worked well in simulations. Sampling without replacement is also known as subsampling and the delete d jackknife.

2.2. Three High Dimensional Two Sample Tests

If $(\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i})$ come in correlated pairs, a high dimensional analog of the paired t test applies the one sample test on $\boldsymbol{z}_i = \boldsymbol{x}_{1i} - \boldsymbol{x}_{2i}$.

Now suppose there are two independent random samples $\boldsymbol{x}_{1,1}, \ldots, \boldsymbol{x}_{1,n_1}$ and $\boldsymbol{x}_{2,1}, \ldots, \boldsymbol{x}_{2,n_2}$ from two populations or groups, and that it is desired to test $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ where $E(\boldsymbol{x}_i) = \boldsymbol{\mu}_i$ are $p \times 1$ vectors. Let $n = n_1 + n_2$. Let $\boldsymbol{S}_i$ be the sample covariance matrix of $\boldsymbol{x}_i$ and let $Cov(\boldsymbol{x}_i) = \boldsymbol{\Sigma}_i$ for $i = 1, 2$.

A simple test takes $m = \min(n_1, n_2)$ and $\boldsymbol{z}_i = \boldsymbol{x}_{1i} - \boldsymbol{x}_{2i}$ for $i = 1, \ldots, m$. Then apply the one sample test from Theorem 2 to the $\boldsymbol{z}_i$. This paired test might work well in high dimensions because of the superior power of the Theorem 2 test, but in low dimensions, it is known that there are better tests.

Let $\boldsymbol{x}_1$ be the $\boldsymbol{x}_i$ that has $n_1 \leq n_2$. Then let

$$\boldsymbol{y}_i = \boldsymbol{x}_{1i} - \sqrt{\frac{n_1}{n_2}}\boldsymbol{x}_{2i} + \frac{1}{\sqrt{n_1 n_2}}\sum_{j=1}^{n_1}\boldsymbol{x}_{2j} - \bar{\boldsymbol{x}}_2 = \boldsymbol{x}_{1i} - \sqrt{\frac{n_1}{n_2}}\boldsymbol{x}_{2i} + \boldsymbol{a}_{n_1,n_2} - \bar{\boldsymbol{x}}_2$$

for $i = 1, \ldots, n_1$. Note that $\boldsymbol{y}_i = \boldsymbol{z}_i = \boldsymbol{x}_{1i} - \boldsymbol{x}_{2i}$ if $n_1 = n_2$. Ref. [27] (pp. 177–178) proved that $\overline{\boldsymbol{y}} = \bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2$, that $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ are uncorrelated for $i \neq j$, that $E(\boldsymbol{y}_i) = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, and that $Cov(\boldsymbol{y}_i) = Cov(\boldsymbol{x}_1) + (n_1/n_2)Cov(\boldsymbol{x}_2)$ for $i = 1, \ldots, n_1$. Ref. [25] showed that $T_n(\boldsymbol{y})/\sqrt{\hat{V}_0(\boldsymbol{y})} \xrightarrow{D} N(0,1)$ where the $\boldsymbol{y}$ denotes that the one sample test was computed using the $\boldsymbol{y}_i$.

Note that $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ holds if and only if $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 = \boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2^T\boldsymbol{\mu}_2 - 2\boldsymbol{\mu}_1^T\boldsymbol{\mu}_2 = 0$. These terms can be estimated by $T_n = T_n(\boldsymbol{x}, \boldsymbol{y}) = T_1 + T_2 - 2T_3$ where T_1 and T_2 are the one sample test statistic applied to samples 1 and 2 and $n_1 n_2 T_3 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\boldsymbol{x}_{1i}^T\boldsymbol{x}_{2j}$.

Let $X_{ij} = \boldsymbol{x}_{1i}^T\boldsymbol{x}_{1j} = X_{ji}$ and $Y_{ij} = \boldsymbol{x}_{2i}^T\boldsymbol{x}_{2j} = Y_{ji}$ where $i \neq j$. Let $Z_{ij} = \boldsymbol{x}_{1i}^T\boldsymbol{x}_{2j} = Z_{ji}$. Let $\sigma_X^2 = V(X_{ij})$, $\sigma_Y^2 = V(Y_{ij})$, and $\sigma_Z^2 = V(Z_{ij})$. Let $V_0(T_n)$ be the variance of T_n when H_0 is true. Assume s_n^2 is a consistent estimator of $V_0(T_n)$. Under $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and additional regularity conditions, ref. [22] showed that $V_0(T_n) =$

$$\frac{2}{n_1(n_1-1)}tr(\boldsymbol{\Sigma}_1^2) + \frac{2}{n_2(n_2-1)}tr(\boldsymbol{\Sigma}_2^2) + \frac{4}{n_1 n_2}tr(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)$$

and that

$$\frac{T_n}{s_n} \xrightarrow{D} N(0,1).$$

Let $\theta_1 = Cov(X_{ij}, X_{it}) = \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1\boldsymbol{\mu}_1$ where $j \neq t, i < j$, and $i < t, \theta_2 = Cov(Y_{ij}, Y_{it}) = \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2\boldsymbol{\mu}_2$ where $j \neq t$, $i < j$, and $i < t$, $\theta_3 = Cov(Z_{ij}, Z_{it}) = \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_1\boldsymbol{\mu}_2$ where $j \neq t$, and $\theta_4 = Cov(Z_{ij}, Z_{kj}) = \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_2\boldsymbol{\mu}_1$ where $i \neq k$.

Ref. [22] showed that

$$V(T_3) = \frac{tr(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)}{n_1 n_2} + \frac{\theta_3}{n_1} + \frac{\theta_4}{n_2}.$$

Ref. [28], using arguments similar to Theorem 2, showed

$$V(T_3) = \frac{\sigma_Z^2}{n_1 n_2} + \frac{\theta_3(n_2-1)}{n_1 n_2} + \frac{\theta_4(n_1-1)}{n_1 n_2}.$$

Thus $\sigma_Z^2 = tr(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2) + \theta_3 + \theta_4$ and $tr(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2) = \sigma_Z^2 - \theta_3 - \theta_4$. Hence

$$V_0(T_n) = \frac{2(\sigma_X^2 - 2\theta_1)}{n_1(n_1-1)} + \frac{2(\sigma_Y^2 - 2\theta_2)}{n_2(n_2-1)} + \frac{4(\sigma_Z^2 - \theta_3 - \theta_4)}{n_1 n_2}.$$

If $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{0}$, then the $\theta_i = 0$, and the formula with the $\theta_i = 0$ worked well in simulations. Note that σ_X^2, σ_Y^2, and the θ_i can be estimated as in Section 2.1. Let $m = \min(n_1, n_2)$, and $Z_i = \boldsymbol{x}_{1i}^T\boldsymbol{x}_{2i}$ for $i = 1, \ldots, m$. Let S_Z^2 be the sample variance of the Z_i. Another estimator of σ_Z^2 is

$$\hat{\sigma}_Z^2 = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\boldsymbol{x}_{1i}^T\boldsymbol{x}_{2j} - T_3(\boldsymbol{x}, \boldsymbol{y}))^2.$$

2.3. Theory for Testing $H_0 : \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$

Consider tests of the form $H_0 : \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \boldsymbol{0}$. The omnibus test uses $\boldsymbol{A} = \boldsymbol{I}_p$ and tests $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \boldsymbol{0}$.

Let $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \mu_Y)$ and $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y})$ for $i = 1, \ldots, n$. Then $T_n(\boldsymbol{w})/s_n(\boldsymbol{w}) \xrightarrow{D} N(0,1)$ under mild regularity conditions by Section 2.1 where $\boldsymbol{w}$ indicates that the test was applied to the $\boldsymbol{w}_i$. Ref. [14] showed that $T_n(\boldsymbol{v})/s_n(\boldsymbol{v}) \xrightarrow{D} N(0,1)$ and used $p = 1.5n$ for multiple linear regression in their simulations.

Let $O = \{i_1, \ldots, i_k\}$, $\boldsymbol{x}_O = (x_{i_1}, \ldots, x_{i_k})^T$, and $\boldsymbol{x}_{i,O} = (x_{i,i_1}, \ldots, x_{i,i_k})^T$. Then testing $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}_O Y} = \boldsymbol{0}$ uses the one sample test on the $\boldsymbol{v}_{i,O} = (\boldsymbol{x}_{i,O} - \overline{\boldsymbol{x}}_O)(Y_i - \overline{Y})$. This test is equivalent to testing $H_0 : \boldsymbol{\beta}_{OPLS,O} = \boldsymbol{0}$ and $H_0 : \boldsymbol{\beta}_{MMLE,O} = \boldsymbol{0}$. Note that data splitting could be used to select O. For multiple linear regression and the MMLE and OPLS estimators, these tests are high dimensional analogs for the OLS partial F tests for testing whether a reduced model is good. If $I \cup O = \{1, \ldots, p\}$, then I corresponds to the predictors in the reduced model while O corresponds to the predictors out of the reduced model.

In low dimensions, important tests for regression include (a) $H_0 : \beta_i = 0$ (the Wald tests for MLR), (b) $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ (the Anova F test for MLR), and (c) $H_0 : (\beta_{i_1}, \ldots, \beta_{i_k})^T = \boldsymbol{0}$ (the partial F test for MLR). The above paragraph shows how to do these high dimensional tests for the multiple linear regression OPLS and MMLE estimators, with or without heterogeneity. Data splitting is not needed if O is known. Note that (a) corresponds to testing $H_0 : \mathrm{Cov}(x_i, Y) = 0$ while (c) corresponds to testing $H_0 : (\mathrm{Cov}(x_{i_1}, Y), \ldots, \mathrm{Cov}(x_{i_k}, Y))^T = \boldsymbol{0}$.

The next subsection reviews competitors for the above tests when k is small compared to n.

2.4. Theory for Certain $\boldsymbol{A}$

This subsection reviews some large sample theory for $\hat{\boldsymbol{\eta}}_{OPLS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ and OPLS for the multiple linear regression model, including some high dimensional tests for low dimensional quantities such as $H_0 : \beta_i = 0$ or $H_0 : \beta_i - \beta_j = 0$. These tests depended on iid cases, but not on linearity or the constant variance assumption. Hence the tests are useful for multiple linear regression with heterogeneity.

The following [5] theorem gives the large sample theory for $\hat{\boldsymbol{\eta}} = \widehat{\mathrm{Cov}}(\boldsymbol{x}, Y)$. Ref. [6] gave alternative proofs. This theory needs $\boldsymbol{\eta} = \boldsymbol{\eta}_{OPLS} = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ to exist for $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$ to be a consistent estimator of $\boldsymbol{\eta}$. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ and let $\boldsymbol{w}_i$ and $\boldsymbol{z}_i$ be defined below where

$$\mathrm{Cov}(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{w}} = E[(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})^T (Y_i - \boldsymbol{\mu}_Y)^2)] - \boldsymbol{\Sigma}_{\boldsymbol{x}Y} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T.$$

Then the low order moments are needed for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$ to be a consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{w}}$.

Theorem 3. *Assume the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid. Assume $E(x_{ij}^k \, Y_i^m)$ exist for $j = 1, \ldots, p$ and $k, m = 0, 1, 2$. Let $\boldsymbol{\mu}_{\boldsymbol{x}} = E(\boldsymbol{x})$ and $\mu_Y = E(Y)$. Let $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \boldsymbol{\mu}_Y)$ with sample mean $\overline{\boldsymbol{w}}_n$. Let $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$. Then (a)*

$$\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}), \; \sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}), \tag{9}$$

$$\text{and } \sqrt{n}(\tilde{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}).$$

(b) Let $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)(Y_i - \overline{Y}_n)$. Then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}} + O_P(n^{-1/2})$. Hence $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}} + O_P(n^{-1/2})$. (c) Let $\boldsymbol{A}$ be a $k \times p$ full rank constant matrix with $k \leq p$, assume $H_0 : \boldsymbol{A}\boldsymbol{\beta}_{OPLS} = \boldsymbol{0}$ is true, and assume $\hat{\lambda} \xrightarrow{P} \lambda \neq 0$. Then

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) \xrightarrow{D} N_k(\boldsymbol{0}, \lambda^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{A}^T). \tag{10}$$

For the following theorem, consider a subset of k distinct elements from $\tilde{\boldsymbol{\Sigma}}$ or from $\hat{\boldsymbol{\Sigma}}$. Stack the elements into a vector, and let each vector have the same ordering. For example, the largest subset of distinct elements corresponds to

$$vech(\tilde{\boldsymbol{\Sigma}}) = (\tilde{\sigma}_{11}, \ldots, \tilde{\sigma}_{1p}, \tilde{\sigma}_{22}, \ldots, \tilde{\sigma}_{2p}, \ldots, \tilde{\sigma}_{p-1,p-1}, \tilde{\sigma}_{p-1,p}, \tilde{\sigma}_{pp})^T = [\tilde{\sigma}_{jk}].$$

For random variables $x_1, \ldots, x_p$, use notation such as $\overline{x}_j$ = the sample mean of the x_j, $\mu_j = E(x_j)$, and $\sigma_{jk} = Cov(x_j, x_k)$. Let

$$n\ vech(\tilde{\mathbf{\Sigma}}) = [n\ \tilde{\sigma}_{jk}] = \sum_{i=1}^{n}[(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)].$$

For general vectors of elements, the ordering of the vectors will all be the same and be denoted by vectors such as $\hat{\boldsymbol{c}} = [\hat{\sigma}_{jk}]$, $\tilde{\boldsymbol{c}} = [\tilde{\sigma}_{jk}]$, $\boldsymbol{c} = [\sigma_{jk}]$, $\boldsymbol{v}_i = [(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)]$, and $\boldsymbol{w}_i = [(x_{ij} - \mu_j)(x_{ik} - \mu_k)]$. Let $\overline{\boldsymbol{w}}_n = \sum_{i=1}^{n} \boldsymbol{w}_i/n$ be the sample mean of the $\boldsymbol{w}_i$. Assuming that $Cov(\boldsymbol{w}_i) = \mathbf{\Sigma}_{\boldsymbol{w}}$ exists, then $E(\boldsymbol{w}_i) = E(\overline{\boldsymbol{w}}_n) = \boldsymbol{c}$.

The following [6] theorem provides large sample theory for $\hat{\boldsymbol{c}}$ and $\tilde{\boldsymbol{c}}$. We use $Cov(\boldsymbol{w}_i) = \mathbf{\Sigma}_{\boldsymbol{d}}$ to avoid confusion with the $\mathbf{\Sigma}_{\boldsymbol{w}}$ used in Theorem 3. Note that $\boldsymbol{x}_i$ are dummy variables and could be replaced by $\boldsymbol{u}_i = (Y_{i1}, \ldots, Y_{im}, x_{i1}, \ldots, x_{ip})^T$ to get information about m response variables $Y_1, \ldots, Y_m$. Testing $H_0 : (\mathbf{\Sigma}_{\boldsymbol{x}Y_1}^T, \mathbf{\Sigma}_{\boldsymbol{x}Y_2}^T)^T = \mathbf{0}$ could likely be done applying the one sample test to $\boldsymbol{z}_1 = ((\boldsymbol{x}_1 - \overline{\boldsymbol{x}})^T(Y_{1,1} - \overline{Y}_1), (\boldsymbol{x}_2 - \overline{\boldsymbol{x}})^T(Y_{2,2} - \overline{Y}_2))^T, \ldots, \boldsymbol{z}_m = ((\boldsymbol{x}_{n-1} - \overline{\boldsymbol{x}})^T(Y_{1,n-1} - \overline{Y}_1), (\boldsymbol{x}_n - \overline{\boldsymbol{x}})^T(Y_{2,n} - \overline{Y}_2))^T$ assuming $n = 2m$ and iid cases.

Theorem 4. *Assume the cases $\boldsymbol{x}_i$ are iid and that $Cov(\boldsymbol{w}_i) = \mathbf{\Sigma}_{\boldsymbol{d}}$ exists. Using the above notation with $\boldsymbol{c}$ a $k \times 1$ vector,*

(i) $\sqrt{n}(\tilde{\boldsymbol{c}} - \boldsymbol{c}) \xrightarrow{D} N_k(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{d}})$.

(ii) $\sqrt{n}(\hat{\boldsymbol{c}} - \boldsymbol{c}) \xrightarrow{D} N_k(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{d}})$.

(iii) $\hat{\mathbf{\Sigma}}_{\boldsymbol{d}} = \hat{\mathbf{\Sigma}}_{\boldsymbol{v}} + O_P(n^{-1/2})$ *and* $\tilde{\mathbf{\Sigma}}_{\boldsymbol{d}} = \tilde{\mathbf{\Sigma}}_{\boldsymbol{v}} + O_P(n^{-1/2})$.

2.5. Testing

As noted by [5], the following simple testing method reduces a possibly high dimensional problem to a low dimensional problem. Testing $H_0 : \boldsymbol{A}\boldsymbol{\beta}_{OPLS} = \mathbf{0}$ versus $H_1 : \boldsymbol{A}\boldsymbol{\beta}_{OPLS} \neq \mathbf{0}$ is equivalent to testing $H_0 : \boldsymbol{A}\boldsymbol{\eta} = \mathbf{0}$ versus $H_1 : \boldsymbol{A}\boldsymbol{\eta} \neq \mathbf{0}$ where $\boldsymbol{A}$ is a $k \times p$ constant matrix. Let $\mathrm{Cov}(\hat{\mathbf{\Sigma}}_{\boldsymbol{x}Y}) = \mathrm{Cov}(\hat{\boldsymbol{\eta}}) = \mathbf{\Sigma}_{\boldsymbol{w}}$ be the asymptotic covariance matrix of $\hat{\boldsymbol{\eta}} = \hat{\mathbf{\Sigma}}_{\boldsymbol{x}Y}$. In high dimensions where $n < 5p$, we can't get a good nonsingular estimator of $\mathrm{Cov}(\hat{\mathbf{\Sigma}}_{\boldsymbol{x}Y})$, but we can get good nonsingular estimators of $\mathrm{Cov}(\hat{\mathbf{\Sigma}}_{\boldsymbol{u}Y}) = \mathrm{Cov}((\hat{\eta}_{i_1}, \ldots, \hat{\eta}_{i_k})^T)$ with $\boldsymbol{u} = \boldsymbol{x}_I = (x_{i_1}, \ldots, x_{i_k})^T$ where $n \geq Jk$ with $J \geq 10$. Here $I = \{i_1, \ldots, i_k\}$ denotes predictors that are in the model. (Values of J much larger than 10 may be needed if some of the k predictors and/or Y are skewed.) Simply apply Theorem 3 to the predictors $\boldsymbol{u}$ used in the hypothesis test, and thus use the sample covariance matrix of the vectors $(\boldsymbol{u}_i - \overline{\boldsymbol{u}})(Y_i - \overline{Y})$. Hence we can test hypotheses like $H_0 : \beta_i - \beta_j = 0$. In particular, testing $H_0 : \beta_i = 0$ is equivalent to testing $H_0 : \eta_i = \sigma_{x_i,Y} = 0$ where $\sigma_{x_i,Y} = \mathrm{Cov}(x_i, Y)$.

2.6. High Dimensional Outlier Detection

High dimensional outlier detection is important. This subsection follows [29] closely. See [29,30] for examples and simulations. Let $\boldsymbol{W}$ be a data matrix, where the rows $\boldsymbol{w}_i$ correspond to cases. For example, $\boldsymbol{w}_i = \boldsymbol{x}_i$ or $\boldsymbol{w}_i = \boldsymbol{z}_i = (Y_i, x_{i1}, \ldots, x_{ip})^T$. One of the simplest outlier detection methods uses the Euclidean distances of the $\boldsymbol{w}_i$ from the coordinatewise median $D_i = D_i(\mathrm{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median MED_j: the coordinatewise median computed from the "half set" of cases $\boldsymbol{w}_i$ with $D_i^2 \leq \mathrm{MED}(D_i^2(\mathrm{MED}_{j-1}, \boldsymbol{I}_p))$ where $\mathrm{MED}_0 = \mathrm{MED}(\boldsymbol{W})$. We often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\mathrm{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \mathrm{MED}(D_1, \ldots, D_n) + k\mathrm{MAD}(D_1, \ldots, D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise. Using $k \geq 0$ insures that at least half of the cases get weight 1. This weighting corresponds to the weighting that would be used in a one sided metrically trimmed mean (Huber type skipped mean) of the distances. Here, the sample median

absolute deviation is $\text{MAD}(n) = \text{MAD}(D_1, \ldots, D_n) = \text{MED}(|D_i - \text{MED}(n)|,\ i = 1, \ldots, n)$ where $\text{MED}(n) = \text{MED}(D_1, \ldots, D_n)$ is the sample median of $D_1, \ldots, D_n$.

Let the *covmb2 set B* of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set B. If $\boldsymbol{w}_i = \boldsymbol{x}_i$, then

$$T = \frac{\sum_{i=1}^{n} W_i \boldsymbol{x}_i}{\sum_{i=1}^{n} W_i} \text{ and } \boldsymbol{C} = \frac{\sum_{i=1}^{n} W_i (\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^{n} W_i \ - \ 1}.$$

This estimator was built for speed, applications, and outlier resistance.

Another method to get an outlier resistant estimator $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ is to use the following identity. If X and Y are random variables, then

$$\text{Cov}(\text{X}, \text{Y}) = [\text{Var}(\text{X} + \text{Y}) - \text{Var}(\text{X} - \text{Y})]/4.$$

Then replace $\text{Var}(W)$ by $[\hat{\sigma}(W)]^2$ where $\hat{\sigma}(W)$ is a robust estimator of scale or standard deviation and $W = X + Y$ or $W = X - Y$. We used $\hat{\sigma}(W) = 1.483MAD(W)$ where $MAD(W) = MAD(n) = MAD(W_1, \ldots, W_n)$. Hence

$$\widehat{\text{Cov}}(X, Y) = ([1.483MAD(X + Y)]^2 - [1.483MAD(X - Y)]^2)/4. \tag{11}$$

The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers.

The function `rcovxy` makes the classical and three robust estimators of $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$, and makes a scatterplot matrix of the four estimated sufficient predictors $\hat{\boldsymbol{\eta}}^T \boldsymbol{x}$ and Y. Only two robust estimators are made if $n \leq 2.5p$.

3. Results

Example 1. *The [31] data was collected from $n = 26$ districts in Prussia in 1843. Let Y = the number of women married to civilians in the district with a constant and predictors x_1 = the population of the district in 1843, x_2 = the number of married civilian men in the district, x_3 = the number of married men in the military in the district, and x_4 = the number of women married to husbands in the military in the district. Sometimes the person conducting the survey would not count a spouse if the spouse was not at home. Hence Y and x_2 are highly correlated but not equal. Similarly, x_3 and x_4 are highly correlated but not equal. We expect $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS} \approx (0, 1, 0, 0)^T$. Then $\hat{\boldsymbol{\beta}}_{OLS} = (0.00035, 0.9995, -0.2328, 0.1531)^T$, $\hat{\boldsymbol{\beta}}_{MMLE} = (0.1782, 1.0010, 48.5630, 51.5513)^T$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = (9285758004, 1674298902, 9855702, 9653811)^T$, and $\hat{\boldsymbol{\beta}}_{OPLS} = (0.1727, 0.0311, 0.0002, 0.0002)^T$. Let the omnibus test statistic $Z(\boldsymbol{v}) = T_n(\boldsymbol{v})/\sqrt{\hat{V}_0(T_n(\boldsymbol{v}))}$ applied to the $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y})$. Then $Z(\boldsymbol{v}) = 9.3281$ and the hypotheses $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$, $H_0 : \boldsymbol{\beta}_{OLS} = \boldsymbol{0}$, $H_0 : \boldsymbol{\beta}_{OPLS} = \boldsymbol{0}$ and $H_0 : \boldsymbol{\beta}_{MMLE} = \boldsymbol{0}$ are all rejected. The classical F-test also rejects H_0 with p-value=0.*

Example 2. *The [32] pottery data has $n = 36$ pottery shards of Roman earthware produced between second century B.C. and fourth century A.D. Often the pottery was stamped by the manufacturer. A chemical analysis was done for $p = 20$ chemicals (variables), the types of pottery were 1-Arretine, 2-not-Arretine, 3-North Italian, 4-Central Italian, 5-questionable origin. Let the binary response variable $Y = 1$ for type 1 and 0 for types 2–5. The omnibus test had $Z = 2.146$ for a two sided p-value of 0.0319 and the more correct right tailed p-value of 0.016. The chi-square logistic regression test for $\boldsymbol{\beta} = \boldsymbol{0}$ had p-value=0.0002, but the GLM did not converge.*

3.1. One Sample Tests

In the simulations, we examined five one sample tests. The first "test" used the m out of n bootstrap to compute $T_1^*, \ldots, T_B^*$ with $B = 100$. We used the shorth bootstrap confidence interval described in [30] (ch. 2). This "test" has not been proven to have level α. The second test computed the usual t confidence interval

$$[\overline{W} - t_{1-\alpha/2,m-1} S_W/\sqrt{m}, \overline{W} + t_{1-\alpha/2,m-1} S_W/\sqrt{m}]$$

for $\boldsymbol{\mu}^T\boldsymbol{\mu}$ based on the W_i from Theorem 1. The third and fourth tests used Theorem 2 (b) and $T_n/s_n \xrightarrow{D} N(0,1)$ if s_n^2 is a consistent estimator of $V(T_n)$ when H_0 is true. The third test used $0.5n(n-1)s_n^2 = \hat{\sigma}_W^2$, while the fourth test used $0.5n(n-1)s_n^2 = S_W^2$ based on Theorem 1. These two tests computed intervals ("confidence intervals for 0")

$$[T_n - t_{1-\alpha/2,m-1}\, s_n, T_n + t_{1-\alpha/2,m-1}\, s_n].$$

The tests 2–4 use the same cutoff $t_{1-\alpha/2,m-1}$ so that the average interval lengths are more comparable. The fifth test used the Theorem 2 test applied to the spatial sign vectors with S_W^2.

The simulation used four distribution types where $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{y} + \delta\mathbf{1}$ with $E(\boldsymbol{x}) = \delta\mathbf{1}$ where $\mathbf{1}$ is the $p \times 1$ vector of ones. Type 1 used $\boldsymbol{y} \sim N_p(\mathbf{0}, \boldsymbol{I})$, type 2 used a mixture distribution $\boldsymbol{y} \sim 0.6N_p(\mathbf{0}, \boldsymbol{I}) + 0.4N_p(\mathbf{0}, 25\boldsymbol{I})$, type 3 for a multivariate t_4 distribution, and type 4 for a multivariate lognormal distribution where $\boldsymbol{y} = (y_1, \ldots, y_p)$ with $w_i = \exp(Z)$ where $Z \sim N(0,1)$ and $y_i = w_i - E(w_i)$ where $E(w_i) = \exp(0.5)$. The covariance matrix type depended on the matrix $\boldsymbol{A}$. Type 1 used $\boldsymbol{A} = \boldsymbol{I}_p$, type 2 used $\boldsymbol{A} = diag(\sqrt{1}, \ldots, \sqrt{p})$, and type 3 used $\boldsymbol{A} = \psi\mathbf{1}\mathbf{1}^T + (1-\psi)\boldsymbol{I}_p$ giving $\text{cor}(x_{ij}, x_{ik}) = \rho$ for $j \neq k$ where $\rho = 0$ if $\psi = 0$, $\rho \to 1/(c+1)$ as $p \to \infty$ if $\psi = 1/\sqrt{cp}$ where $c > 0$, and $\rho \to 1$ as $p \to \infty$ if $\psi \in (0,1)$ is a constant. We used $\delta = 0$ and $\delta > 0$ chosen so at least one test had good power. The simulation used 5000 runs, the 4 $\boldsymbol{x}$ distributions, and the 3 matrices $\boldsymbol{A}$. For the third $\boldsymbol{A}$, we used $\psi = 1/\sqrt{p}$.

Tables 1 and 2 summarize some simulation results. There are two lines for each simulation scenario. The first line gives the simulated power = proportion of times $H_0 : \boldsymbol{\mu} = \mathbf{0}$ was rejected. The second line gives the average length of the confidence interval for 0 where H_0 is rejected if 0 is not in the confidence interval. When $\delta = 0$, observed coverage between 0.04 and 0.06 suggests coverage = power = level is close to the nominal value 0.05. For larger δ, want the coverage near 1 for good power. See [28] for more simulations.

The bootstrap test corresponds to the boot column, the tests using $(\overline{w}, S_W)$, $(T_n, \hat{\sigma}_W)$, and (T_n, S_W) correspond to the next three columns. The last column corresponds to the spatial sign test. This test tends to have much shorter lengths because of the transformation of the data. The test using $(\overline{w}, S_W)$ has simple large sample theory, but low power compared to the other methods. This test's length is approximately $\sqrt{n-1}$ times the length of that corresponding to (T_n, S_W) where $\sqrt{99} \approx 10$ in the tables. The bootstrap test was sometimes conservative with observed coverage < 0.04 when $\delta = 0$. For xtype = 4 and $\delta = 0$, H_0 was not true for the spatial test. Hence the coverage for the spatial test was sometimes higher than 0.06 for this scenario. For $\delta = 0$, the test with $(T_n, \hat{\sigma}_W)$ sometimes had coverage less than 0.04, while the test with (T_n, S_W) sometimes had coverage greater than 0.06. In the simulations, the spatial test often performed well, but typically $E(\boldsymbol{z}_i) = \boldsymbol{\mu}_{\boldsymbol{z}} \neq \boldsymbol{\mu}_{\boldsymbol{x}} = E(\boldsymbol{x}_i)$, which makes the spatial test harder to use. For testing $H_0 : \boldsymbol{\mu}_{\boldsymbol{x}} = \mathbf{0}$, the test with $(T_n, \hat{\sigma}_W)$ appeared to perform better than the three competitors.

Table 1. One sample tests, covtyp = 1, cov = observed type I error for $\delta = 0$ and power for $\delta > 0$. Boldface for good performance not including spatial.

n	p	psi/xtype	δ	Boot	$(\overline{w}, S_W)$	$(T_n, \hat{\sigma}_W)$	(T_n, S_W)	Spatial
100	100	0	0	0.0230	0.0580	0.0400	0.0452	0.0444
	len	1		0.6732	5.6520	0.5711	0.5681	0.0057
100	100	0	0.075	0.8160	0.0688	**0.9216**	0.9176	0.9166
	len	1		0.8081	5.7018	0.5741	0.5731	0.0057
100	100	0	0	0.0236	0.0436	**0.0466**	0.0776	0.0478
	len	2		7.0590	58.2593	6.0094	5.8553	0.0057
100	100	0	0.15	0.1938	0.0506	0.3128	0.3490	0.9988
	len	2		7.5830	58.1417	6.0204	5.8435	0.0057
100	100	0	0	0.0222	0.0466	0.0450	0.0680	0.0468
	len	3		1.3031	10.6946	**1.1140**	1.0749	0.0057
100	100	0	0.1	0.7536	0.0544	0.8720	0.8714	0.9956
	len	3		1.5563	10.8976	1.1260	1.0953	0.0057
100	100	0	0	0.0206	0.0556	0.0372	0.0656	0.0906
	len	4		3.1105	25.4558	2.6543	2.5584	0.0057
100	100	0	0.17	0.9024	0.0546	**0.9622**	0.9496	0.7668
	len	4		3.7816	25.5420	2.6708	2.5671	0.0057
100	1000	0	0	0.0236	0.0482	0.0448	**0.0506**	0.0506
	len	1		2.1403	17.8302	1.8059	1.7920	0.0018
100	1000	0	0.0415	0.872	0.068	**0.9438**	0.9398	0.9388
	len	1		2.2771	17.9004	1.8089	1.7991	0.0018
100	1000	0	0	0.0236	0.0448	**0.0458**	0.0712	0.0558
	len	2		22.4434	185.1105	19.0973	18.6043	0.0018
100	1000	0	0.075	0.142	0.0480	0.2222	0.2616	0.9978
	len	2		22.8203	182.6556	18.9772	18.3576	0.0018
100	1000	0	0	0.0214	0.0432	**0.0436**	0.0650	0.0450
	len	3		4.1649	34.1708	**3.5444**	3.4343	0.0018
100	1000	0	0.05	0.6458	0.0558	0.7642	0.7770	0.9908
	len	3		4.3708	34.0483	3.5586	3.4220	0.0018
100	1000	0	0	0.0192	0.0544	0.0378	**0.0518**	0.0484
	len	4		9.9417	82.3953	8.4267	8.2810	0.0018
100	1000	0	0.087	0.8430	0.0576	**0.9282**	0.9242	0.8774
	len	4		10.5664	82.8816	8.4523	8.3299	0.0018

Table 2. One sample tests, covtyp = 2, p = 10,000, cov = observed type I error for $\delta = 0$ and power for $\delta > 0$. Boldface for good performance not including spatial.

n	p	psi/xtype	δ	boot	$(\overline{w}, S_W)$	$(T_n, \hat{\sigma}_W)$	(T_n, S_W)	Spatial
100	10,000	0	0	0.0272	0.0536	0.045	0.0502	0.0496
	len	1		39,006.52	326,271.5	32,976.41	32,791.52	0.0007
100	10,000	0	1.69	0.8482	0.0582	0.93	0.9294	0.9286
	len	1		39,690.34	327,648.8	32,994.63	32,929.94	0.0007
100	10,000	0	0	0.0244	0.0442	**0.0486**	0.0876	0.0526
	len	2		408,860	3,330,506	347,476	334,728.5	0.0007
100	10,000	0	3	0.1126	0.0488	0.1778	0.2148	0.9952
	len	2		411,196.1	3,349,674	347,862.6	336,654.9	0.0007
100	10,000	0	0	0.0206	0.044	**0.0436**	0.0632	0.051
	len	3		75,976.41	624,134.1	64,858.9	62,727.84	0.0007
100	10,000	0	2.5	0.8918	0.0608	0.9462	0.9454	1
	len	3		77,389.1	625,801.9	64,740.62	62,895.46	0.0007
100	10,000	0	0	0.0236	0.0534	0.038	**0.0444**	0.0454
	len	4		181,871.7	1,517,807	154,052.2	152,545.3	0.0007
100	10,000	0	3.80	0.8952	0.0578	0.9558	0.9522	0.948
	len	4		185,192.4	1,518,189	154,094.1	152,583.7	0.0007

3.2. Two Sample Tests

In the simulations, we examined three two sample tests. The first "test" used the m out of n bootstrap where $m_i = 2n_i/3$ to bootstrap the [22] test that estimates $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 = \boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2^T\boldsymbol{\mu}_2 - 2\boldsymbol{\mu}_1^T\boldsymbol{\mu}_2$. The second test was the "paired test" with $m = \min(n_1, n_2)$ and $\boldsymbol{z}_i = \boldsymbol{x}_{1i} - \boldsymbol{x}_{2i}$ for $i = 1, \ldots, m$. Then apply the one sample test from Theorem 2 to the $\boldsymbol{z}_i$. The third test was the [25] Li test. Both of these tests used S_W^2 applied to the $\boldsymbol{z}_i$ or the $\boldsymbol{y}_i$.

The simulation used four distribution types where $\boldsymbol{x}_1 = \boldsymbol{A}_1\boldsymbol{y}_1 + \delta\mathbf{1}$ and $\boldsymbol{x}_2 = \boldsymbol{A}_2\boldsymbol{y}_2$ where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ had the same distribution, with $E(\boldsymbol{x}_1) = \delta\mathbf{1}$ and $E(\boldsymbol{x}_2) = \mathbf{0}$. Type 1 used $\boldsymbol{y} \sim N_p(\mathbf{0}, \boldsymbol{I})$, type 2 used a mixture distribution $\boldsymbol{y} \sim 0.6N_p(\mathbf{0}, \boldsymbol{I}) + 0.4N_p(\mathbf{0}, 25\boldsymbol{I})$, type 3 for a multivariate t_4 distribution, and type 4 for a multivariate lognormal distribution where $\boldsymbol{y} = (y_1, \ldots, y_p)$ with $w_i = \exp(Z)$ where $Z \sim N(0, 1)$ and $y_i = w_i - E(w_i)$ where $E(w_i) = \exp(0.5)$. The covariance matrix type depended on the matrix $\boldsymbol{A}$.

For the covariance types, $Cov(\boldsymbol{x}_1) = \boldsymbol{I}, Cov(\boldsymbol{x}_2) = \sigma^2 Cov(\boldsymbol{x}_1)$ for covtyp = 1. $Cov(\boldsymbol{x}_1) = diag(1, 2, \ldots, p), Cov(\boldsymbol{x}_2) = \sigma^2 Cov(\boldsymbol{x}_1)$ for covtyp = 2. $Cov(\boldsymbol{x}_1) = \boldsymbol{I}$, $Cov(\boldsymbol{x}_2) = \sigma^2 diag$ (1,2,...,p) for covtyp = 3. Table 3 shows some results. Two lines were used for each simulation scenario, with coverages on the first line and lengths on the second line. When $n_1 = n_2$, the paired test and Li test gave the same results. When n_1/n_2 was not near 1, the Li test had better power and shorter length. Increasing δ could greatly increase the length for the bootstrap test, but the coverage would be 1. Improving the one sample test would improve the Li test, but the Li test performed well in simulations.

Table 3. Two sample tests, covtyp = 1, cov = observed type I error for $\delta = 0$ and power for $\delta > 0$. Boldface for better performance.

(n_1, n_2, σ, p)	xtype	covtype	δ	Boot	Pair	Li
(100, 100, 1, 100)	1	1	0	0.0246	0.0494	0.0494
len	1	1	0	1.3426	1.1389	1.1389
(100, 100, 1, 100)	1	1	0.1	0.7224	0.8586	0.8586
len	1	1	0.1	1.5789	1.1417	1.1417
(100, 200, 1, 100)	1	1	0	0.0256	0.0456	0.0462
len	1	1	0	1.0019	1.1360	**0.8535**
(100, 200, 1, 100)	1	1	0.1	0.9166	0.8602	**0.9612**
len	1	1	0.1	1.2396	1.1432	0.8609

3.3. Theorem 3 Tests

We illustrate Theorem 3 and Section 2.5 for Poisson regression and negative binomial regression. This simulation is similar to that done by [6] for multiple linear regression with and without heterogeneity. Let $\boldsymbol{x} \sim N_{p-1}(\mathbf{0}, \boldsymbol{I})$ be the $(p-1) \times 1$ vector of nontrivial predictors. Let $SP_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} = 1 + 1x_{i,1} + \cdots + 1x_{i,k}$ for $i = 1, \ldots, n$. Hence $\alpha = 1$ and $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta}^T)^T = (1, .., 1, 0, \ldots, 0)^T$ with $k+1$ ones and $p-k-1$ zeros. Here $\boldsymbol{\beta}$ is the Poisson regression parameter vector $\boldsymbol{\beta}_{PR}$ or the negative binomial regression parameter vector $\boldsymbol{\beta}_{NBR}$. Let $Z_i = \log(Y_i)$ if $Y_i > 0$ and $Z_i = \log(0.5)$ if $Y_i = 0$. Then a multiple linear regression model with heterogeneity is $Z_i = \alpha_Z + \boldsymbol{x}_i^T\boldsymbol{\beta}_Z + e_i$ where the e_i are independent with expected value $E(e_i) = 0$ and variance $V(e_i) = \sigma_i^2$. Since the cases $(\boldsymbol{x}_i, Y_i)$ are iid, the OLS estimator $\boldsymbol{\beta}_{OLS} = c_o\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Z} = \boldsymbol{\Sigma}_{\boldsymbol{x}Z}$ because $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{I}_{p-1}$. Thus $\boldsymbol{\Sigma}_{\boldsymbol{x}Z} = (c_o, \ldots, c_o, 0, \ldots, 0)^T$ with the first k values equal to c_o and $p-k-1$ zeros.

Let $\boldsymbol{\eta}_{OPLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}Z} = (\eta_1, \ldots, \eta_{p-1})^T$. Then the Theorem 3 large sample $100(1-\delta)$ confidence interval (CI) is $\hat{\eta}_i \pm t_{n-1,1-\delta/2}SE(\hat{\eta}_i)$ could be computed for each η_i. If 0 is not in the confidence interval, then $H_0 : \eta_i = 0$ and $H_0 : \beta_{iE} = 0$ are both rejected for estimators E = OPLS and MMLE for the multiple linear regression model with Z. In the simulations with $n = 50$, $p = 4$, and $\psi > 0$, the maximum observed undercoverage was

about $0.05 = 5\%$. Hence the program has the option to replace the cutoff $t_{n-1,1-\delta/2}$ by $t_{n-1,up}$ where $up = min(1 - \delta/2 + 0.05, 1 - \delta/2 + 2.5/n)$ if $\delta/2 > 0.1$,

$$up = min(1 - \delta/4, 1 - \delta/2 + 12.5\delta/n)$$

if $\delta/2 \leq 0.1$. If $up < 1 - \delta/2 + 0.001$, then use $up = 1 - \delta/2$. This correction factor was used in the simulations for the nominal 95% CIs, where the correction factor uses a cutoff that is between $t_{n-1,0.975}$ and the cutoff $t_{n-1,0.9875}$ that would be used for a 97.5% CI. The nominal coverage was 0.95 with $\delta = 0.05$. Observed coverage between 0.94 and 0.96 suggests coverage is close to the nominal value. Ref. [33] noted that weighted least squares tests tend to reject H_0 too often (liberal tests with undercoverage).

To summarize the $p - 1$ confidence intervals, the average length of the $p - 1$ confidence intervals over 5000 runs was computed. Then the minimum, mean, and maximum of the average lengths was computed. The proportion of times each confidence interval contained zero was computed. These proportions were the observed coverages of the $p - 1$ confidence intervals. Then the minimum observed coverage was found. The percentage of the observed coverages that were ≥ 0.9, 0.92, 0.93, 0.94, and 0.96 were also recorded. The test $H_0 : (\eta_i, \eta_j)^T = (0, 0)^T$ was also done where H_0 was true. The coverage of the test was recorded and a correction factor was not used. Negative binomial regression and Poisson regression were used, where $\kappa = \infty$ indicates that Poisson regression was used.

Table 4 illustrates Theorem 3(a) where $k = 1$ and Table 4 replaces Y with Z. For Table 4, confidence intervals were made for $\eta_i = Cov(x_i, Z)$ for $i = 1, \ldots, 99$ and the coverage was the percentage of the 5000 CIs that contained 0. Here $\eta_1 \neq 0$, but $\eta_i = 0$ for $i = 2, \ldots, 99$. The first two lines of Table 4 correspond to Poisson regression. The confidence interval for η_1 never contained 0, hence the minimum coverage was 0 with observed power $= 1 - 0 = 1$. The proportion of CIs that had coverage ≥ 0.94 was 0.9898 (98/99 CIs). Hence this was also the proportion of CIs with coverage $\geq 0.90, 0.92$ and 0.93. The proportion of CIs that had coverage ≥ 0.96 was 0.8081 (80/99 CIs). The typical coverage was near 0.965, hence the correction factor was slightly too large. The test $H_0 : (\eta_{98}, \eta_{99})^T = (0, 0)^T$ did not use a correction factor, and coverage was 0.9438. The minimum average CI length was 0.4166, the sample mean of the average CI lengths was 0.4187, and the maximum average length was 0.4875, corresponding to η_1. The second two lines and below for Table 4 were for the negative binomial regression with kappa $= \kappa = 0.5, 1, 10, 100$. For $\kappa = 1000$ and 10,000, the simulations were very similar to those for $\kappa = \infty$. Using Y instead of Z gave similar results with longer lengths.

Table 4. Cov(x,Z), n = 100, p = 100, k = 1, want cov > 0.94 except for mincov and cov96.

κ	mincov	cov90	cov92	cov93	cov94	cov96	testcov
∞	0.0000	0.9899	0.9899	0.9899	0.9899	0.8081	0.9438
len	0.4166	0.4187	0.4875				
0.5	0.0062	0.9899	0.9899	0.9899	0.9899	0.7576	0.9440
len	0.5050	0.5084	0.5686				
1	0.0000	0.9899	0.9899	0.9899	0.9899	0.7475	0.9410
len	0.4809	0.4834	0.5421				
10	0.0000	0.9899	0.9899	0.9899	0.9899	0.6970	0.9412
len	0.4258	0.4279	0.4929				
100	0.0000	0.9899	0.9899	0.9899	0.9899	0.6566	0.9464
len	0.4174	0.4195	0.4882				

3.4. Omnibus Test

Multiple Linear Regression

For this simulation, the $\boldsymbol{x}$ were generated as in Section 3.1 with $\boldsymbol{\mu} = \mathbf{0}$, and then $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$ where $\boldsymbol{\beta} = \delta\mathbf{1}$. Hence $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is true when $\delta = 0$. The one sample test was applied on the $\boldsymbol{v}_i$ using S_W and $\hat{\sigma}_W$. The zero mean iid errors e_i were iid from five distributions: (i) N(0,1), (ii) t_3, (iii) EXP(1) $-$ 1, (iv) uniform$(-1, 1)$, and (v) 0.9 N(0,1) + 0.1 N(0,100). Only distribution (iii) is not symmetric. With 5000 runs, would like the coverage to be between 0.04 and 0.06 when $\delta = 0$. In Table 5, the coverage was a bit high when S_W was used (second to last column) instead of $\hat{\sigma}_W$ (fourth column). Power near 0.95 was good for $\delta = 0.0035$.

Table 5. Omnibus test for multiple linear regression, cov = observed type I error for $\delta = 0$ and power for $\delta > 0$. Boldface for good performance.

(n, p)	(xtype,etype,ψ)	δ	cov	len	S_W cov	len
(100, 100)	(1, 1, 0)	0	**0.0574**	6.3950	0.0710	5.9478
(100, 100)	(1, 1, 0)	0.0035	0.9524	9.0769	0.9542	8.3259
(100, 100)	(1, 1, 0.1)	0	**0.0524**	6.3914	0.0684	6.0284
(100, 100)	(1, 1, 0.1)	0.0035	0.9592	9.1093	0.9550	8.3757
(200, 100)	(1, 1, 0)	0	0.0484	3.2456	0.0586	**3.1284**
(500, 100)	(1, 1, 0)	0	0.0488	1.3147	0.0548	**1.2821**
(100, 100)	(3, 2, 0)	0	**0.0518**	30.055	0.1394	24.370
(100, 500)	(3, 2, 0)	0	**0.0466**	657.47	0.1320	524.95
(50, 500)	(2, 4, 0)	0	**0.0572**	940.29	0.1372	799.28
(50, 500)	(2, 4, 0)	0.0001	0.9400	2587.0	**0.9600**	1547.5
(10, 500)	(2, 4, 0)	0	**0.0258**	4417.8	0.1546	3734.8

Poisson Regression

For this simulation, the $\boldsymbol{x}_i$ were generated in a manner similar to Section 3.1 when the $\boldsymbol{x}_i$ were from a multivariate normal distribution. Let $\boldsymbol{\beta} = \delta(1, \ldots, 1, 0, \ldots, 0)^T$ where there were k 1's and $p - k$ 0's. Then the $\boldsymbol{x}_i$ were scaled such that $SP \sim N(1, 1)$ when $\delta = 1$. In general, $SP \sim N(1, \delta^2)$ for $\delta \geq 0$. Hence the population Poisson regression was fairly strong for $\delta = 1$ and rather weak for $\delta = 0.25$. Table 6 shows that using $\hat{\sigma}_W$ controlled the nominal level 0.05 better than using S_W. As p got larger, the power performance could decrease. See line 8 of Table 6.

Table 6. Omnibus test for Poisson regression, cov = observed type I error for $\delta = 0$ and power for $\delta > 0$. Boldface for good performance.

(n, p)	(k,ψ)	δ	cov	len	S_W cov	len
(100, 100)	(100, 0)	0	**0.0448**	0.0151	0.0878	0.0144
(100, 100)	(100, 0)	0.7	**0.9318**	0.0616	0.9184	0.0548
(100, 100)	(100, 0.1)	0	**0.0568**	0.0015	0.0672	0.0014
(100, 100)	(100, 0.1)	0.25	**0.9758**	0.0024	0.9742	0.0021
(200, 100)	(100, 0)	0	**0.0438**	0.0075	0.0652	0.0073
(500, 100)	(100, 0)	0	**0.0484**	0.0030	0.0606	0.0030
(100, 200)	(100, 0)	0	**0.0478**	0.0215	0.0852	0.0205
(100, 500)	(100, 0)	2	0.3196	61.606	**0.6008**	38.942
(100, 500)	(100, 0)	0	**0.0502**	0.0344	0.0852	0.0330
(50, 500)	(100, 0)	0	**0.0370**	0.0741	0.0906	0.0702
(30, 500)	(100, 0)	0	**0.0232**	0.1413	0.0896	0.1302

Sample *R* code for the above two tables is shown below.

```
source (‘‘http://parker.ad.siu.edu/Olive/slpack.txt’’)
```

```
mlrcovxysim (n=100,p=500,nruns=5000,xtype=3,etype=2,delta=0)
prcovxysim (n=500,p=100,k=100,nruns=5000,psi=0,delta=0)
```

4. Discussion

The omnibus test is resistant to model misspecification. For example, (a) the constant variance multiple linear regression model could be assumed when there is heterogeneity, and (b) for count data, a multiple linear regression model, or a negative binomial regression model, or a quasi-Poisson regression model may fit the data much better than the count model actually chosen. The test can also be used in low dimensions when the MLE fails to converge.

Based on the simulations and the theory, (a) the omnibus test and one sample test will not have good power against all alternatives unless $\sigma_W/n \to 0$ as $n, p \to \infty$. (b) The omnibus test and one sample test tended to have simulated observed level near the nominal level (control the type I error) if $\hat{\sigma}_W$ was used, but the omnibus test could be conservative if n was small: $n = 10$ for multiple linear regression and $n = 30$ for Poisson regression in the simulations. Sometimes $\hat{\sigma}_W$ exploded if p was large or if H_0 was false. (c) The omnibus test and one sample test have little outlier resistance. Thus it is important to check for outliers before performing the tests. (d) Both tests worked fairly well in simulations for $n \geq 50$ and $p \leq 10n$, and Ref. [14] used $p = 1.5n$ in their simulations for multiple linear regression.

Right tail tests should be used for $\boldsymbol{\mu}^T\boldsymbol{\mu}$ since they have more power, but two tail tests are easier to explain and compare. Ref. [14] used the statistic

$$t_n = \frac{2}{n(n-1)} \sum_{i<j} [\boldsymbol{v}_i^T \boldsymbol{v}_j + k_n(\boldsymbol{a}^T \boldsymbol{v}_i)(\boldsymbol{v}_j^T \boldsymbol{a})]$$

with $\boldsymbol{a} = \mathbf{1}/\sqrt{p}$ and $k_n = (p/ln(p))^{1/2}$. This statistic can also be used for an omnibus test when $\boldsymbol{\beta}_0 = \mathbf{0}$. The extra term was used to increase power and is likely a good idea, but better formulas for $\hat{V}_0(t_n)$ may be needed.

Ref. [28] has many references for high dimensional one and two sample tests. For classification with two groups, let $\boldsymbol{\Sigma}$ be the pooled covariance matrix. Then $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{0}$ if and only if $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$, which can be tested with a two sample test. For the importance of $\boldsymbol{\beta}$ in discriminant analysis, see, for example, [34].

Let the "fail to reject region" be the compliment of the rejection region. Often the fail to reject region is a confidence region for the parameter or parameter vector of interest, where a confidence interval is a special case of a confidence region. In high dimensions, the length or volume of the fail to reject region does not necessarily converge to 0 as $n, p \to \infty$, and the volume could diverge to ∞ if $p/n \to \infty$. For the one sample test, the fail to reject region using V_0 has much more power than using a confidence interval for $\boldsymbol{\mu}^T\boldsymbol{\mu}$.

Simulations were done in *R*. See [35]. The collection of [30] *R* functions *slpack*, available from (http://parker.ad.siu.edu/Olive/slpack.txt, accessed on 3 November 2025). has some useful functions for the inference. The function *hdomni* does the omnibus test. The relevant *R* code is shown below.

```
hdomni(x,y,alpha=0.05)
k <- n*(n-1)
xx <- scale(x,scale=F) #centered but not scaled
v <- xx*c(y-mean(y))
a <- apply(v,2,sum)
Thd <- (t(a)%*%a - sum(v^2))/k       #1 by 1 matrix
Thd <- as.double(Thd) #so the test statistic Thd=Tn is a scalar
sscp <- v%*%t(v)
```

```
ss <- sscp - Thd
ss <- ss^2
vw1 <- (sum(ss) - sum(diag(ss)))/k
Vohat <- 2*vw1/k
Z <- Thd/sqrt(Vohat)
pval <- 2*pnorm(-abs(Z)) #two tail pvalue
rpval=1-pnorm(Z) #right tail pvalue
```

The function *hdhot1sim3* was used to simulate the five one sample tests, and was used for Tables 1 and 2. The function *hdhot1sim4* added the test using $\hat{\tau}_2$. The function *hdhot2sim* simulates the two sample test which applies the fast paired test on the $\boldsymbol{z}_i = \boldsymbol{x}_{i1} - \boldsymbol{x}_{i2}$ for $i = 1, \ldots, m$, the [25] test, and the two sample [22] test based on subsampling with $m_i = floor(2n_i/3)$ for i = 1, 2. See Table 3. Proofs for Theorems 3 and 4 were not given, but are available from preprints of the corresponding published papers from (http://parker.ad.siu.edu/Olive/preprints.htm, accessed on 3 November 2025).

For Table 4, the function `nbinroplssimz` was used to create negative binomial regression data sets for finite κ, while the function `PRoplssimz` was used to create the Poisson regression data sets corresponding to $\kappa = \infty$. The functions without the z do not use the $Z = log(Y)$ transformation.

For the omnibus test, the function `mlrcovxysim` was used for multiple linear regression, while the function `prcovxysim` was used for Poisson regression.

The spatial sign vectors have a some outlier resistance. If the predictor variables are all continuous, the *covmb2* and *ddplot5* functions are useful for detecting outliers in high dimensions. See [30] (section 1.4.3). Ref. [36] gave estimators for the variance of U-statistics.

Author Contributions: Conceptualization, A.M.A., P.A.Q and D.J.O.; methodology, A.M.A., P.A.Q. and D.J.O.; writing-original draft preparation, D.J.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data sets are available from (http://parker.ad.siu.edu/Olive/sldata.txt, accessed on 28 October 2025).

Acknowledgments: The authors thank the editors and referees for their work.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CI	confidence interval
iid	independent and identically distributed
MDPI	Multidisciplinary Digital Publishing Institute
MLR	Multiple Linear Regression
MMLE	marginal maximum likelihood estimator
OLS	ordinary least squares
OPLS	one component partial least squares
SP	sufficient predictor

References

1. Cook, R.D.; Helland, I.S.; Su, Z. Envelopes and partial least squares regression. *J. Roy. Stat. Soc. B* **2013**, *75*, 851–877. [CrossRef]
2. Basa, J.; Cook, R.D.; Forzani, L.; Marcos, M. Asymptotic distribution of one-component partial least squares regression estimators in high dimensions. *Can. J. Stat.* **2024**, *52*, 118–130. [CrossRef]

3. Cook, R.D.; Forzani, L. *Partial Least Squares Regression: And Related Dimension Reduction Methods*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2024.
4. Wold, H. Soft modelling by latent variables: The non-linear partial least squares (NIPALS) approach. *J. Appl. Prob.* **1975**, *12*, 117–142. [CrossRef]
5. Olive, D.J.; Zhang, L. One component partial least squares, high dimensional regression, data splitting, and the multitude of models. *Commun. Stat. Theory Methods* **2025**, *54*, 130–145. [CrossRef]
6. Olive, D.J.; Alshammari, A.A.; Pathiranage, K.G.; Hettige, L.A.W. Testing with the one component partial least squares and the marginal maximum likelihood estimators. *Commun. Stat. Theory Methods* **2025**. [CrossRef]
7. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Stat. Soc. B* **2008**, *70*, 849–911. [CrossRef]
8. Fan, J.; Song, R. Sure independence screening in generalized linear models with np-Dimensionality. *Ann. Stat.* **2010**, *38*, 3217–3841. [CrossRef]
9. Agresti, A. *Categorical Data Analysis*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2002.
10. Li, K.C.; Duan, N. Regression analysis under link violation. *Ann. Stat.* **1989**, *17*, 1009–1052. [CrossRef]
11. Chen, C.H.; Li, K.C. Can SIR be as popular as multiple linear regression? *Stat. Sinica* **1998**, *8*, 289–316.
12. Cameron, A.C.; Trivedi, P.K. *Regression Analysis of Count Data*; Cambridge University Press: Cambridge, UK, 1998.
13. Haggstrom, G.W. Logistic regression and discriminant analysis by ordinary least squares. *J. Bus. Econ. Stat.* **1983**, *1*, 229–238. [CrossRef]
14. Zhao, A.; Li, C.; Li, R.; Zhang, Z. Testing high-dimensional regression coefficients in linear models. *Ann. Stat.* **2024**, *52*, 2034–2058. [CrossRef]
15. Cui, H.; Guo, W.; Zhong, W. Test for high-dimensional regression coefficients using refitted cross-validation variance estimation. *Ann. Stat.* **2018**, *46*, 958–988. [CrossRef]
16. Goeman, J.J.; van de Geer, S.A.; van Houwelingen, H.C. Testing against a high dimensional alternative. *J. R. Stat. Soc. B* **2006**, *68*, 477–493. [CrossRef]
17. Lan, W.; Wang, H.; and Tsai, C.-L. Testing covariates in high-dimensional regression. *Ann. Inst. Statist. Math.* **2014**, *66*, 279–301. [CrossRef]
18. Zhong, P.-S.; Chen, S.X. Tests for high dimensional regression coefficients with factorial designs. *J. Amer. Stat. Assoc.* **2011**, *106*, 260–274. [CrossRef]
19. Helland, I.S. Partial least squares regression and statistical models. *Scand. J. Stat.* **1990**, *17*, 97–114.
20. Park, J.; Ayyala, D.N. A test for the mean vector in large dimension and small samples. *J. Stat. Plan. Inf.* **2013**, *143*, 929–943. [CrossRef]
21. Lehmann, E.L. *Nonparametrics: Statistical Methods Based on Ranks*; Holden-Day: San Francisco, CA, USA, 1975.
22. Chen, S.X.; Qin, Y.L. A two sample test for high-dimensional data with applications to gene-set testing. *Ann. Stat.* **2010**, *38*, 808–835. [CrossRef]
23. Srivastava, M.S.; Du, M. A test for the mean vector with fewer observations than the dimension. *J. Mult. Anal.* **2008**, *99*, 386–402. [CrossRef]
24. Bai, Z.D.; Saranadasa, H. Effects of high dimension: By an example of a two sample problem. *Stat. Sinica* **1996**, *6*, 311–329.
25. Li, J. Finite sample *t*-tests for high-dimensional means. *J. Mult. Anal.* **2023**, *196*, 105183. [CrossRef] [PubMed]
26. Wang, L.; Peng, B.; and Li, R. A high-dimensional nonparametric multivariate test for mean vector. *J. Am. Stat. Assoc.* **2015**, *110*, 1658–1669. [CrossRef]
27. Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*, 2nd ed.; Wiley: New York, NY, USA, 1984.
28. Abid, A.M. Some Simple High Dimensional One and Two Sample Tests. Ph.D. Thesis, Southern Illinois University, Carbondale, IL, USA, 2025. Available online: http://parker.ad.siu.edu/Olive/sAhlam.pdf (accessed on 28 October 2025).
29. Olive, D.J. Some useful techniques for high dimensional statistics. *Stats* **2025**, *8*, 60. [CrossRef]
30. Olive, D.J. Prediction and Statistical Learning, Online Course Notes, 2025. Available online: http://parker.ad.siu.edu/Olive/slearnbk.htm (accessed on 28 October 2025).
31. Hebbler, B. Statistics of Prussia. *J. Roy. Stat. Soc. A* **1847**, *10*, 154–186. [CrossRef]
32. Wisseman, S.U.; Hopke, P.K.; Schindler-Kaudelka, E. Multielemental and multivariate analysis of Italian terra sigillata in the world heritage museum, university of Illinois at Urbana-Champaign. *Archeomaterials* **1987**, *1*, 101–107.
33. Pötscher, B.M.; Preinerstorfer, D. How reliable are bootstrap-based heteroskedasticity robust tests? *Econ. Theory* **2023**, *39*, 789–847. [CrossRef]
34. Wang, Y.; Wu, Z.; Wang, C. High dimensional discriminant analysis under weak sparsity. *Commun. Stat. Theory Methods* **2025**, *54*, 2657–2674. [CrossRef]

35. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2024.
36. Xu, T.; Zhu, R.; Shao, X. On variance estimation of random forests with infinite-order U-statistics. *Electr. J. Stat.* **2024**, *18*, 2135–2207. [CrossRef]