Test effects of high-dimensional covariates via aggregating cumulative covariances

LIPING ZHU

Institute of Stat and Big Data, Renmin University of China

DEC-14-2023

Section 1: The Problem

Introduction

- Let us start from the linear regression.

$$Y = \alpha + \mathbf{x}^\intercal \beta + \varepsilon,$$

- A hypothesis test:

$$\begin{gathered} H_0 : \beta = 0, \text{ or equivalently, } \beta_1 = \cdots = \beta_p = 0, \\ \text{versus} \\ H_1 : \beta \neq 0. \end{gathered}$$

Introduction

- The cardiomyopathy microarray data ($n = 30$): Redfern et al. (2000) [1] and Segal et al. (2003)[2].
- The overexpression of G protein-coupled receptor Ro1 in hearts of adult mice would lead to a lethal dilated cardiomyopathy.
- Are the gene expressions ($p = 6,319$) really predictive for the gene expression level of Ro1?
- This amounts to testing $H_0 : \beta = 0$, versus $H_1 : \beta \neq 0$ in the linear model $Y = \alpha + \mathbf{x}^\intercal \beta + \varepsilon$.

[1] Redfern, C.H., et al. (2000). Conditional expression of a gi-coupled receptor causes ventricular conduction de-lay and a lethal cardiomyopathy. *Proceedings of the National Academy of Sciences*, 97(9), 4826-4831.

[2] Segal, M.R., Dahlquist, K.D., and Conklin, B.R. (2003). Regression approaches for microarray data analysis. *Journal of Computational Biology*, 10(6), 961-980.

Three challenges: (1) High dimensions

$p = 6,319$ and $n = 30$.

- If p is small relative to n, the classical F-test can be used to infer the overall significance of linear regression coefficients.
- Zhong and Chen (2011) [3] showed that the power of F-test is adversely impacted by an increasing ratio p/n even when $p < n - 1$.
- In “large p, small n" situations, F-test is no longer applicable.

[3] Zhong, P. S. and Chen, S. X. (2011). Tests for high-dimensional regression coefficients with factorial designs. *Journal of the American Statistical Association*, 106(493), 260-274.

Introduction

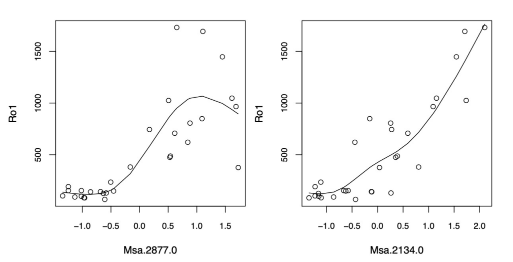
- To accommodate the high-dimensionality issue, Zhong and Chen (2011) modified the F-test and suggested using

$$\sum_{s=1}^{p} \operatorname{cov}^2(Y, X_s).$$

- The test statistic they use is

$$\mathrm{ZC}_{n,p} = \{4(n)_4\}^{-1} \sum_{s=1}^{p} \sum_{(i,j,k,l)}^{n} (Y_i - Y_j)(Y_k - Y_l)(X_{is} - X_{js})(X_{ks} - X_{ls}).$$

Three challenges: (2) Nolinear dependence



Introduction

- To take nonlinear dependence into account, Zhang, Yao and Shao(2018)[4] considered tesing

$$H_0 : E(Y \mid X_s) = E(Y) \text{ almost surely, for all } 1 \le s \le p.$$

- Zhang, Yao and Shao(2018) suggested using the summation of martingale difference divergence, which can be used to measure arbitrarily nonlinear mean dependence,

$$\sum_{s=1}^{p} \mathrm{MDD}(Y \mid X_s)^2,$$

where
$\mathrm{MDD}(Y \mid X)^2 = -E[\{Y - E(Y)\}\{Y' - E(Y')\}|X - X'|]$.

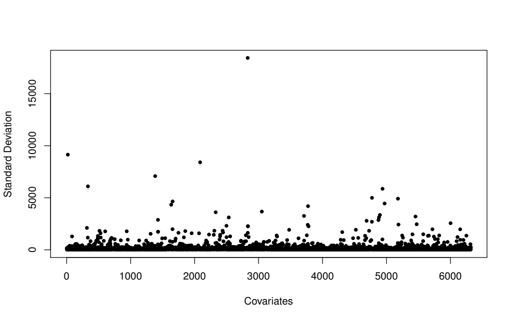
[4] Zhang, X., Yao, S., and Shao, X. (2018). Conditional mean and quantile dependence testing in high dimension. *The Annals of Statistics*, 46(1), 219-246.

Introduction

The test statistic of Zhang, Yao and Shao(2018) is built upon

$$\begin{aligned}\mathrm{ZYS}_{n,p} &= \{4(n)_4\}^{-1}\sum_{s=1}^{p}\sum_{(i,j,k,l)}^{n}(Y_i - Y_j)(Y_k - Y_l)(|X_{is} - X_{ls}| \\ &\quad +|X_{js} - X_{ks}| - |X_{is} - X_{ks}| - |X_{js} - X_{ls}|).\end{aligned}$$

Three challenges: (3) Covariate heteroscedasticity



The standard deviation of each gene expression level, which ranges from 17.34 to 18,437.96.

Introduction

- The heterogeneous variances of covariates can seriously affect the performances of the testing procedures, whose test statistics are not invariant after scale transformations of the covariates.
- Standardizing each covariate by its corresponding variance before applying the tests can achieve the scale-invariance.
- Since the sample variance is only root-n consistent, such a standardization causes asymptotically nonnegligible bias-terms when p diverges to infinity much faster than n.

INTRODUCTION

- **Our Ambition: Address the above three challenges simultaneously.**

Section 2: Conditional Mean Independence Test

Conditional Mean Independence Test

Equivalence

$E(Y \mid X) = E(Y) \Leftrightarrow \mathrm{cov}\{Y,\ I(X < x)\} = 0 \quad \text{for all } x.$

- The cumulative covariance $\mathrm{CCov}(Y \mid X)$ is defined as
$$\mathrm{CCov}(Y \mid X) = E[\mathrm{cov}^2\{Y, I(X < \widetilde{X}) \mid \widetilde{X}\}],;$$
where $(\widetilde{X}, \widetilde{Y})$ is an independent copy of (X, Y).
- It is equivalent to Pearson correlation if X and Y are jointly normal, and $\mathrm{CCov}(Y \mid X)$ is zero if and only if $E(Y \mid X) = E(Y)$.

Conditional Mean Independence Test

- To test $H_0 : E(Y \mid X_s) = E(Y)$ for all $1 \leq s \leq p$, it is natural to use the summation of all marginal cumulative covariances,

$$\sum_{s=1}^{p} \mathrm{CCov}(Y \mid X_s),$$

- A straightforward estimate is given by

$$W_{n,p} = n^{-3} \sum_{s=1}^{p} \sum_{j=1}^{n} \left[\sum_{i=1}^{n} \left(Y_i - \overline{Y} \right) \{ I(X_{is} < X_{js}) - F_{n,s}(X_{js}) \} \right]^2 ,$$

where
$\overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i,$ and $F_{n,s}(X_{js}) = n^{-1} \sum_{i=1}^{n} I(X_{is} < X_{js}).$

Conditional Mean Independence Test

- It can be verified that $T_{n,p}$ is unbiased, which is given by

$$T_{n,p} = \{4(n)_5\}^{-1} \sum_{s=1}^{p} \sum_{(i,j,k,l,r)}^{n} (Y_i - Y_j)(Y_k - Y_l) \\ \times \psi(X_{is}, X_{js}, X_{rs}) \psi(X_{ks}, X_{ls}, X_{rs}),$$

where $(n)_m = n(n-1)\dots(n-m+1)$ for $1 \leq m \leq n$, and $\psi(X_1, X_2, X_3) = I(X_1 < X_3) - I(X_2 < X_3)$.

Conditional Mean Independence Test

- A fast algorithm:

$$
\begin{aligned}
T_{n,p} &= \{(n)_5\}^{-1}\Bigg[(n-2)(n-3)\sum_{s=1}^{p}\sum_{j=2}^{n}\Big(\sum_{i=1}^{j-1}\dot{Y}_{(i)s}\Big)^2 \\
&+ 2\sum_{s=1}^{p}\sum_{j=2}^{n}\Big\{(nj-2n-2j+2)\dot{Y}_{(j)s}\sum_{i=1}^{j-1}\dot{Y}_{(i)s}\Big\} \\
&- \sum_{s=1}^{p}\sum_{j=2}^{n}\Big\{(n^2-2nj-n+4j-4)\sum_{i=1}^{j-1}\dot{Y}^2_{(i)s}\Big\} \\
&- \{n(n^2-3n+8)/3\}\sum_{s=1}^{p}\sum_{i=1}^{n}\dot{Y}^2_{(i)s} + 2\sum_{s=1}^{p}\sum_{i=1}^{n}(i-1)^2\dot{Y}^2_{(i)s}\Bigg].
\end{aligned}
$$

Conditional Mean Independence Test

Theorem 1: Asymptotic null distribution

Under the null hypothesis and certain regularity conditions, as $n, p \to \infty$,

$$\{n(n-1)/2\}^{1/2}\, T_{n,p}/S \xrightarrow{D} N(0,1).$$

Conditional Mean Independence Test

Next we provide an estimate for S^2.

$$S^2_{n,p} = \{4c_n n(n-1)\}^{-1} \sum_{i \neq j}^{n} K_0(\dot{Y}_i, \dot{Y}_j)^2 \Big[\sum_{s=1}^{p} K_1\{F_{n,s}(X_{is}), F_{n,s}(X_{js})\} \Big]^2,$$

where $K_0(Y_1, Y_2) = \{Y_1 - E(Y)\}\{Y_2 - E(Y)\}$,
$K_1\{F_{n,s}(X_{1s}), F_{n,s}(X_{2s})\} =$
$F^2_{n,s}(X_{1s}) + F^2_{n,s}(X_{2s}) - 2\max\{F_{n,s}(X_{1s}), F_{n,s}(X_{2s})\} + 2/3$, and
$F_{n,s}(\cdot)$ is the empirical cumulative distribution function of X_s.

Conditional Mean Independence Test

Theorem 2: Ratio consistency

Under certain regularity conditions, as $n, p \to \infty$,

$$S_{n,p}^2 / S^2 \xrightarrow{P} 1.$$

Therefore, under H_0,

$$\{n(n-1)/2\}^{1/2} \, T_{n,p} / S_{n,p} \xrightarrow{D} N(0,1).$$

Section 3: Asymptotic Relative Efficiency

Asymptotic Relative Efficiency

- The modified F-statistic under linear model assumption: Zhong and Chen (2011)

$$\mathrm{ZC}_{n,p} = \{4(n)_4\}^{-1} \sum_{s=1}^{p} \sum_{(i,j,k,l)}^{n} (Y_i - Y_j)(Y_k - Y_l)(X_{is} - X_{js})(X_{ks} - X_{ls}).$$

- The martingale difference divergence without model assumptions: Zhang, Yao and Shao (2018)

$$\begin{aligned} \mathrm{ZYS}_{n,p} &= \{4(n)_4\}^{-1} \sum_{s=1}^{p} \sum_{(i,j,k,l)}^{n} (Y_i - Y_j)(Y_k - Y_l)(|X_{is} - X_{ls}| \\ &\quad + |X_{js} - X_{ks}| - |X_{is} - X_{ks}| - |X_{js} - X_{ls}|). \end{aligned}$$

Asymptotic Relative Efficiency

- We study the asymptotic powers of these three tests under high-dimensional linear models, and anticipate that similar conclusions can be drawn from nonlinear models.
- Let us consider the model

$$Y = \mathbf{x}^{\intercal}\boldsymbol{\beta} + \varepsilon,$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\intercal}$, $\mathbf{x} =(X_1, \ldots, X_p)^{\intercal} \sim N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathrm{diag}(d_1, \ldots, d_p)$, $E(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = \sigma^2$.

Asymptotic Relative Efficiency

- All three test statistics are asymptotically standard normal.
- The power of the three tests under the local alternatives is

$$\Psi_{n,p} = \{1 + o(1)\}\Phi\left(-z_\alpha + \mathrm{SNR}\right),$$

  - $\mathrm{SNR}_{\mathrm{NEW}} = \left\{15n(n-1)/(4\pi^2\sigma^4 p)\right\}^{1/2} \sum_{s=1}^{p} d_s \beta_s^2.$
  - $\mathrm{SNR}_{\mathrm{ZC}} = \left\{n(n-1)/(2\sigma^4)\right\}^{1/2} \sum_{s=1}^{p} d_s^2 \beta_s^2 \Big(\sum_{s=1}^{p} d_s^2 \Big)^{-1/2}.$
  - $\mathrm{SNR}_{\mathrm{ZYS}}$ $= \left[n(n-1)/\{8\sigma^4(1-\sqrt{3}+\pi/3)\}\right]^{1/2} \sum_{s=1}^{p} d_s^{3/2} \beta_s^2 \Big(\sum_{s=1}^{p} d_s \Big)^{-1/2}.$

Asymptotic Relative Efficiency

In the homoscedastic case, $d_1 = \ldots = d_p$.

- $\mathrm{ARE}(\mathrm{NEW}, \mathrm{ZC}) \approx 0.872$,
- $\mathrm{ARE}(\mathrm{NEW}, \mathrm{ZYS}) \approx 0.979$,
- $\mathrm{ARE}(\mathrm{ZYS}, \mathrm{ZC}) \approx 0.891$.

Asymptotic Relative Efficiency

- For simplicity, we assume that all non-zero coefficients β_s have the same magnitude, that is, $\beta_s = \kappa I(1 \leq s \leq q), s = 1, \ldots, p,$, $q \in \{1, \ldots, p\}$ is fixed.
- We further assume the condition

$$p = o\Big\{ \min\big(\sum_{s=1}^{p} d_s^2, \sum_{s=1}^{p} d_s\big) \Big\}.$$

- Consider an explicit scenario: There is a parameter $\delta > 0$ not depending on the dimension p such that

$$d_s \asymp s^{\delta}, \text{for } s = 1, \ldots, p.$$

Asymptotic Relative Efficiency

- In the ultrahigh dimension setting $\log p \asymp n^{\theta}$, the signal to noise ratios of three tests are

$$\begin{aligned} \mathrm{SNR}_{\mathrm{ZC}} &\asymp (\log p)^{1/\theta} p^{-(1+2\delta)/2}, \\ \mathrm{SNR}_{\mathrm{ZYS}} &\asymp (\log p)^{1/\theta} p^{-(1+\delta)/2}, \text{ and} \\ \mathrm{SNR}_{\mathrm{NEW}} &\asymp (\log p)^{1/\theta} p^{-1/2}. \end{aligned}$$

- The explicit order of asymptotic relative efficiency:

$$\begin{aligned} \mathrm{ARE}(\mathrm{NEW}, \mathrm{ZC}) &\asymp p^{\delta}, \\ \mathrm{ARE}(\mathrm{NEW}, \mathrm{ZYS}) &\asymp p^{\delta/2}, \text{ and} \\ \mathrm{ARE}(\mathrm{ZYS}, \mathrm{ZC}) &\asymp p^{\delta/2}. \end{aligned}$$

- $\mathrm{ARE}(\mathrm{NEW}, \mathrm{ZC}) \to \infty$ and $\mathrm{ARE}(\mathrm{NEW}, \mathrm{ZYS}) \to \infty$, as $p \to \infty$. The asymptotic powers of three tests arranged in a descending order: our proposed test, ZYS test and ZC test.

Section 4: Numerical Studies

Numerical Studies

We consider three models:

$$\begin{aligned} Y_i &= \mathbf{x}_i^\intercal \boldsymbol{\beta}_1 + \varepsilon_i, && (1) \\ Y_i &= 3(\mathbf{x}_i^\intercal \boldsymbol{\beta}_3) + \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_4/2) + \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_2 - 1)\varepsilon_i, && (2) \\ Y_i &= (\mathbf{x}_i^\intercal \boldsymbol{\beta}_5)\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_2/\sqrt{2}) + \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}_5/\sqrt{2q}) + \varepsilon_i, && (3) \end{aligned}$$

where $\mathbf{x}_i = (X_{i1}, \dots, X_{ip})^\intercal$ is generated from the MA model

$$X_{is} = s^{\delta/2}\{\rho_1 Z_{is} + \rho_2 Z_{i(s+1)} + \cdots + \rho_T Z_{i(s+T-1)}\},$$

for $\delta \geq 0$, $T = 8$ and $s = 1, \dots, p$.

- $(Z_{i1}, \dots, Z_{i(p+T-1)})^\intercal$ is drawn from a $(p+T-1)$-dimensional standard normal distribution.
- $\{\rho_k\}_{k=1}^T$ are generated independently from the uniform distribution on $[0, 1]$ and are kept fixed once generated.

Numerical Studies

Table: Empirical sizes and powers for linear model at significance level 5%, where δ controls the degree of heteroscedasticity.

			Normal error			Gamma error		
(n, p)	Hypothesis	δ	ZC	ZYS	NEW	ZC	ZYS	NEW
$(120, 1116)$	H_0	0.00	0.047	0.048	0.044	0.052	0.049	0.052
		0.25	0.043	0.041	0.044	0.057	0.052	0.052
		0.50	0.042	0.044	0.044	0.058	0.053	0.052
		0.75	0.039	0.046	0.044	0.064	0.057	0.052
		1.00	0.039	0.047	0.044	0.063	0.057	0.052
	Non-sparse H_1	0.00	0.849	0.814	0.797	0.842	0.811	0.794
		0.25	0.731	0.918	0.981	0.749	0.909	0.979
		0.50	0.466	0.912	1.000	0.471	0.918	0.999
		0.75	0.246	0.830	1.000	0.251	0.836	1.000
		1.00	0.139	0.655	1.000	0.150	0.661	1.000
	Sparse H_1	0.00	0.670	0.612	0.593	0.643	0.620	0.602
		0.25	0.231	0.452	0.796	0.242	0.452	0.796
		0.50	0.101	0.303	0.933	0.110	0.304	0.941
		0.75	0.063	0.190	0.982	0.079	0.201	0.989
		1.00	0.056	0.133	0.998	0.063	0.131	1.000

Real data analysis revisited

- The cardiomyopathy microarray data contains 6,319 gene expression levels from 30 mice.
- We aim to test whether these genes are really predictive to the expression level of Ro1.

Real data analysis

- We divide the whole dataset into two subsets with $n_1 = 16$ and $n_2 = 14$.
- On the first subset, we screen out unimportant genes by marginally testing the conditional mean independence between the expression levels of each gene and R01.
- The Benjamini-Hochberg procedure is applied to control the false discovery rate at 0.001.
- We randomly pick 6, 7, 8, 9 and 10 samples from the second subset of data and test the overall effects of selected genes.

Real data analysis

Table: The empirical powers for ZC, ZYS tests and our proposed test.

random samples	ZC	ZYS	NEW
6	0.363	0.105	0.463
7	0.432	0.328	0.731
8	0.569	0.632	0.912
9	0.692	0.845	0.980
10	0.831	0.975	1.000

References

- Redfern, C. H. et al. (2000).
  Conditional expression of a Gi-coupled receptor causes ventricular conduction delay and a lethal cardiomyopathy.
  Proceedings of the National Academy of Sciences, 97(9):4826–4831.

- Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003).
  Regression approaches for microarray data analysis.
  Journal of Computational Biology, 10(6):961–980.

- Zhang, X., Yao, S., and Shao, X. (2018).
  Conditional mean and quantile dependence testing in high dimension.
  The Annals of Statistics, 46(1):219–246.

- Zhong, P.-S. and Chen, S. X. (2011).
  Tests for high-dimensional regression coefficients with factorial designs.
  Journal of the American Statistical Association, 106(493):260–274.

THANKS!

QUESTIONS?