Ultra-high-dimensional Longitudinal Quantile Regression and Application in Blood Pressure Analysis

Tianhai Zu

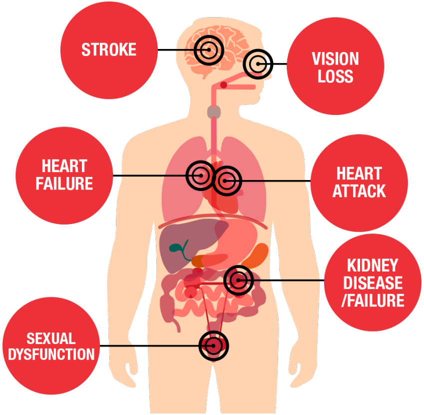
Operations, Business Analytics, and Information Systems

Carl H. Lindner College of Business

email: zuti@mail.uc.edu

Coauthors: Heng Lian, Brittany Green, Yan Yu

High Blood Pressure



- **High blood pressure** (also referred to as hypertension) is when your blood pressure, the force of blood flowing through your blood vessels, is consistently too high.
- **7.8 million** deaths were found to be related to systolic blood pressure above 140mmHg[1].
- Lately, high blood pressure has even been associated with **COVID 19** complications[2].

[1] Forouzanfar et al., “Global burden of hypertension and systolic blood pressure of at least 110 to 115 mm Hg, 1990-2015”

[2] Caillon et al., “High systolic blood pressure at hospital admission is an important risk factor in models predicting outcome of COVID-19 patients”

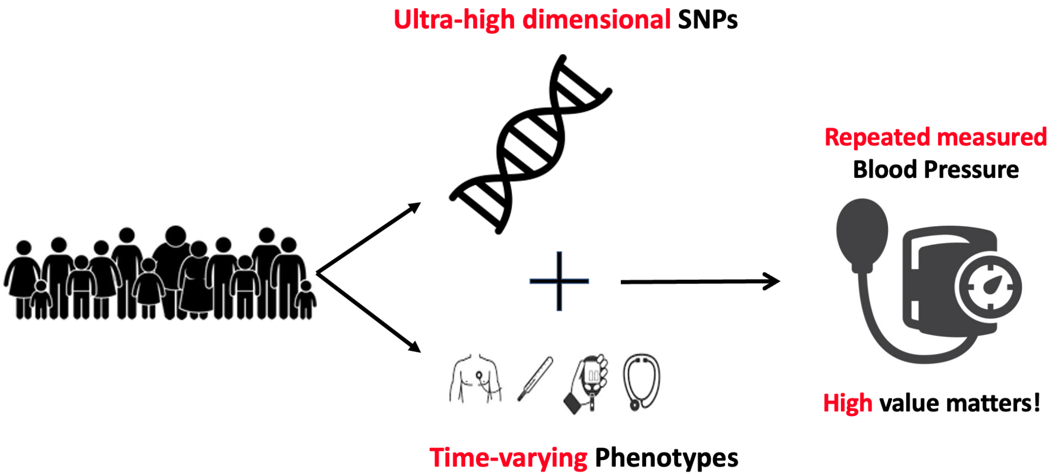
Framingham Heart Study Data

National Heart, Lung, and Blood Institute

- NIH granted.
- Collected about every 4 years.
- Huge and complex.

FRAMINGHAM HEART STUDY
Three Generations of Dedication
1948 1971 2002
A Project of the National Heart, Lung, and Blood Institute and Boston University

Research Background



Research Goal

Over half million SNPs, along with time-varying phenotypes, we aim to find the **important** genetic risk factors for **high** blood pressure.

Our Approach

We developed a novel penalized quantile generalized estimating equation method that can do simultaneous variable selection and modelling on ultra-high dimensional data.

- **Simultaneous variable selection and modelling.**
- **Quantile model.**
- **Longitudinal nature, correlation within subject.**

The first, even for cross-sectional data.

Traditional Approach

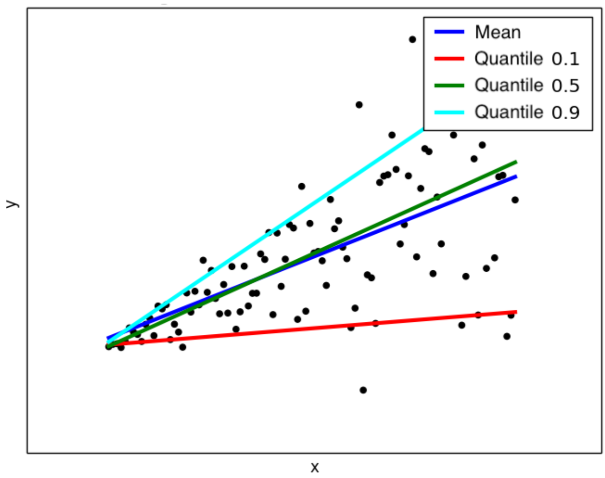
Traditional genome-wide association studies (GWAS)

- One at a time, marginal correlation.
- Mean relationship.
- High threshold of significance levels (10^{-8}).
- High false positive rate.

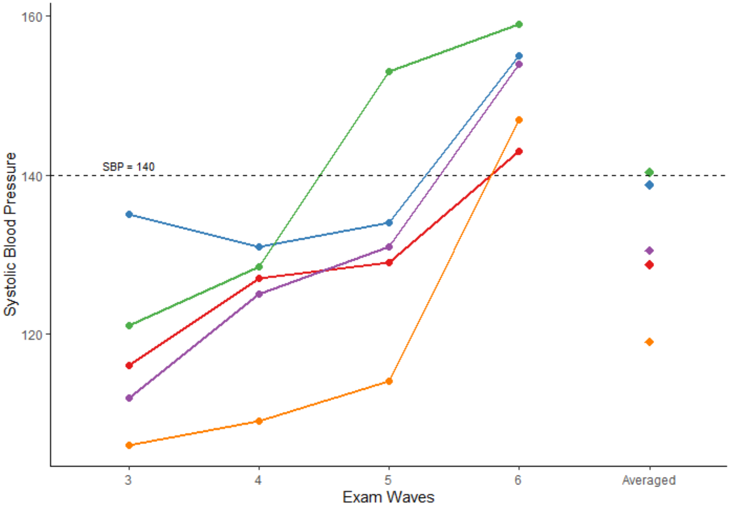
Variable Selection

Over **half million** SNPs, usually only **several** are important for certain disease, **simultaneous variable selection** is needed.

Why Quantile Model? A Toy Example.



Why Longitudinal Model?



Data

Longitudinal

1, 577 participants, over 4 waves, spanning 18 years

- 550,000+ SNPs.
  Ultra-high dimensional, $\log p = O\left(n^c\right)$ for some $0 < c < 1$
  (Fan & Lv, 2008).
- Time-varying phenotype factors: age (AGE), total cholesterol (TC), smoking status (SMK), body mass index (BMI), ventricular rate (VR), triglyceride (TRIG), and high density liptein cholesterol (HDL).
- Repeated measurements for systolic blood pressure (SBP) and diastolic blood pressure (DBP).

Summary Statistics

	Exam Waves			
Phenotypes	3	4	5	6
SBP ($\tau = 0.9$)	139	144	146	150
SBP ($\tau = 0.5$)	119	121	122	124
SBP ($\tau = 0.1$)	103	103	104	106
SBP	120.20	122.73	123.60	125.69
DBP ($\tau = 0.9$)	91	91	87	88
DBP ($\tau = 0.5$)	78	79	74.5	76
DBP ($\tau = 0.1$)	67	67	62	64.6
DBP	78.33	78.53	74.85	75.91
Age	44.84	48.30	51.97	56.03
TC	206.35	203.14	202.45	205.37
BMI	25.81	26.49	27.20	27.79
VR	64.66	62.89	64.61	63.19
TRIG	112.30	115.14	139.75	141.06
HDL	51.73	50.37	50.48	51.45

Model Highlights

- Quantile model.
- Longitudinal nature, correlation within subject.
- Simultaneous variable selection and modelling.

Longitudinal Quantile Model

For $\tau \in (0, 1)$ conditional quantile

$$\theta_\tau(\mathbf{Y}_i | \mathbf{X}_i) = \mathbf{X}_i \boldsymbol{\alpha}_\tau.$$

$\mathbf{X}_i$, phenotype and genotype risk factors.
$\mathbf{Y}_i$, blood pressure measurements over m time points.
$\boldsymbol{\alpha}$, coefficient vector.

Quantile Model

Quantile model to target specific level of **high** blood pressure.

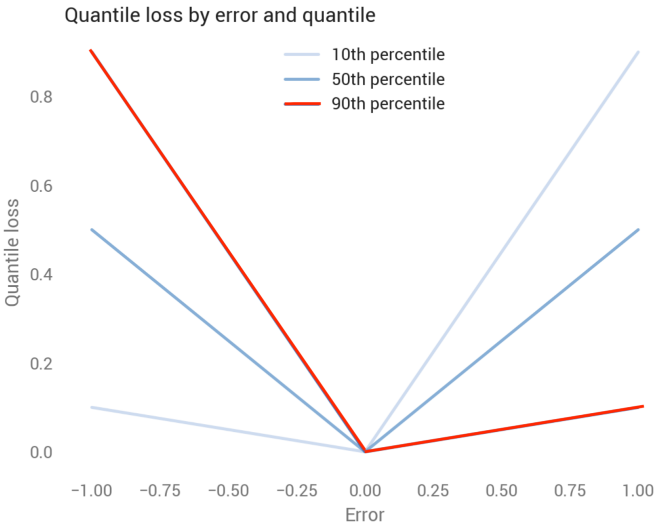
Quantile GEE

Quantile Generalized Estimating Equations

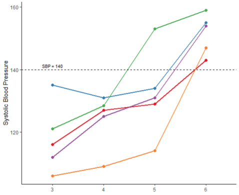
$$\mathbf{s}(\alpha) = \sum_i \mathbf{X}_i^{\mathrm{T}} \mathbf{\Gamma}_i \mathbf{R}^{-1} (\tau - I\{\mathbf{Y}_i \leq \mathbf{X}_i \alpha\})$$

$\tau - I\{\mathbf{Y}_i \leq \mathbf{X}_i \alpha\}$ is the subgradient of the check loss function,

Quantile Check Loss



GEE for Correlated Response



~

+

Employ Generalized Estimating Equations (GEE) for **correlated** response (blood pressure).

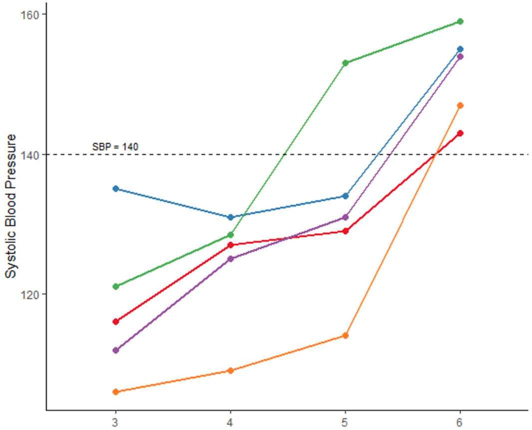
GEE for Correlated Response

Quantile Generalized Estimating Equations (GEE)

$$\mathbf{s}(\alpha) = \sum_i \mathbf{X}_i^{\mathrm{T}} \mathbf{\Gamma}_i \mathbf{R}^{-1} (\tau - I\{\mathbf{Y}_i \leq \mathbf{X}_i \alpha\})$$

$\mathbf{\Gamma}_i$ is the $m \times m$ diagonal matrix of the conditional density of error given $\mathbf{X}_{ij}$ at 0, which are considered weights for the estimating equations,
R is the $m \times m$ working correlation matrix.

Working Correlation Matrix R



Independence

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

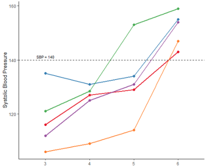
Exchangeable

$$\begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

AR-1

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Penalized Method



$\sim$

$+$

$$
\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{im} \end{pmatrix} \sim \begin{pmatrix} X_{i11}, X_{i12}, \cdots\cdots\cdots\cdots, X_{i1p_n} \\ X_{i21}, X_{i22}, \cdots\cdots\cdots\cdots, X_{i2p_n} \\ \vdots \\ X_{im1}, X_{im2}, \cdots\cdots\cdots\cdots, X_{imp_n} \end{pmatrix}
$$

Incorporate penalized method to achieve **simultaneous variable selection and modeling.**

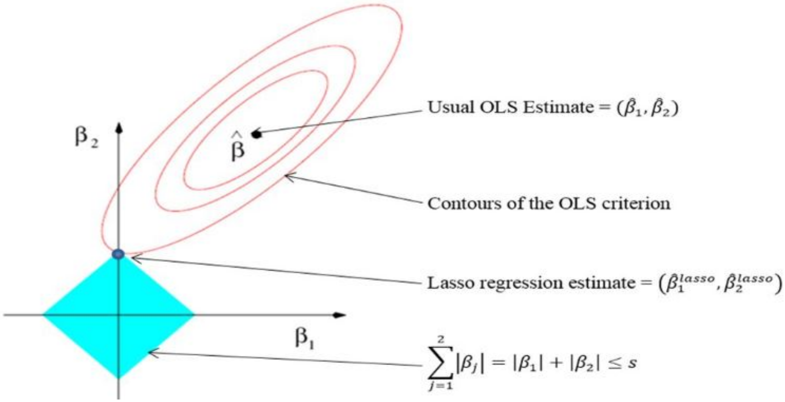
Quantile Penalized GEE

Quantile Penalized Generalized Estimating Equations

$$\mathbf{S}^P(\alpha) = \sum_i \mathbf{X}_i^{\mathrm{T}} \mathbf{\Gamma}_i \mathbf{R}^{-1}(\tau - I\{\mathbf{Y}_i \leq \mathbf{X}_i\alpha\}) - n\mathbf{q}_\lambda(|\alpha|)\mathbf{sgn}(\alpha)$$

$\mathbf{q}_\lambda(|\alpha|)$ is the SCAD penalty term (Fan and Li 2001).

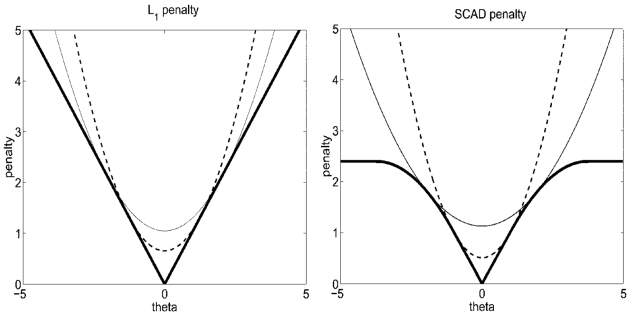
Variable Selection



SCAD Penalty

Smoothly Clipped Absolute Deviation Penalty (SCAD) (Fan and Li, 2001)

- Oracle property but non-convexity



Challenges

$$\mathbf{S}^P(\alpha) = \sum_i \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\Gamma}_i \mathbf{R}^{-1} \Big(\tau - \underbrace{\boxed{I\{\mathbf{Y}_i \leq \mathbf{X}_i\alpha\}}}_{\text{discontinuous}}\Big) - n \underbrace{\boxed{\mathbf{q}_\lambda(|\alpha|)\mathbf{sgn}(\alpha)}}_{\text{discontinuous}}$$

$\tau - I\{\mathbf{Y}_i \leq \mathbf{X}_i\alpha\}$ is the subgradient of the check loss function,
$\mathbf{q}_\lambda(|\alpha|)$ is the first derivative SCAD penalty term.

Challenges

Induced smoothing for discontinuity from the quantile check function

Local linear approximation to handle the penalty terms:

$$\sum_i \mathbf{X}_i^{\mathrm{T}} \mathbf{\Gamma}_i \mathbf{R}^{-1} \left(\underbrace{\boxed{\Phi\left(\frac{\mathbf{Y}_i - \mathbf{X}_i \widehat{\alpha}}{h}\right)}}_{\text{smoothed}} - (1-\tau) \right) - n \underbrace{\boxed{\mathbf{q}_\lambda(|\widehat{\alpha}|)\mathbf{sgn}(\widehat{\alpha})\frac{|\widehat{\alpha}|}{\epsilon + |\widehat{\alpha}|}}}_{\text{approximated}}$$

High-dimensional BIC

$$HBIC(\lambda) = \log\left(\sum_{i=1}^{n} \rho_{\tau}\left(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\alpha}_{\lambda}\right)\right) + p_{1,\lambda}\frac{\log(n)}{2n}C_n$$

$p_{1,\lambda}$ is the number of non-zero parameters.

C_n set to $\log\left(\log(p_n)\right)$.

Phenotypes Results

	$\tau = 0.9$ quantile	$\tau = 0.5$ median	mean PGEE
AGE	0.197	0.152	0.197
	(0.033)	(0.014)	(0.012)
TC	0.12	0.195	0.222
	(0.061)	(0.042)	(0.041)
BMI	0.165	0.203	0.243
	(0.039)	(0.028)	(0.019)
VR	0.117	0.089	0.136
	(0.033)	(0.02)	(0.02)
TRIG	0.496	0.523	0.292
	(0.717)	(0.61)	(0.289)
HDL	0.068	0.043	0.033
	(0.025)	(0.015)	(0.022)

- Six phenotypes are selected throughout and literature also confirms the importance of these phenotypes[3].
- All selected phenotype variables, age, TC, BMI, VR, TRIG, and HDL, are positively associated.

[3] Dua et al., “Body mass index relates to blood pressure among adults”; Kannel, “Risk factors in hypertension.”

SNP Selection Results

	$\tau = 0.9$ quantile	$\tau = 0.5$ median	mean PGEE
SNP Selected:	rs13149993	rs13149993	rs13149993
	rs1957563	rs1957563	rs1957563
	rs634094*	rs634094	rs634094
	rs12692453*	rs12692453	rs12692453
	rs11725914*	rs11725914	rs11725914
	rs4472152*	rs4472152	rs4472152
	rs4523751**	rs4523751	rs4523751
	rs10853618**	rs10853618	rs10853618
	rs10853619**	rs10853619	rs10853619
	rs2825442*	rs2825442	rs2825442
	rs1275988	rs1275988	
	rs7467853	rs7467853	
	rs2820000*	rs2820000	
	rs6056810*	rs6056810	
		rs12739022	rs12739022
	rs16823124	rs8120907	rs10937395
	rs9874923	rs11222084	rs11736662
	rs16893776	rs10832255	rs1491840
	rs2384550	rs7105670	rs1563920
	rs956006		rs7997664
	rs4987082		rs17754634
	rs4712545*		rs17045014
	rs17747515**		rs204873
			rs54211

- The symbol * indicates newly detected SNP.
- the symbol ** indicates newly detected SNP with a plausible association.

SNP Selection Results

	$\tau = 0.9$ quantile	$\tau = 0.5$ median	mean PGEE
SNP Selected:	rs13149993	rs13149993	rs13149993
	rs1957563	rs1957563	rs1957563
	rs634094*	rs634094	rs634094
	rs12692453*	rs12692453	rs12692453
	rs11725914*	rs11725914	rs11725914
	rs4472152*	rs4472152	rs4472152
	*rs*4523751**	*rs*4523751	*rs*4523751
	*rs*10853618**	*rs*10853618	*rs*10853618
	rs10853619**	rs10853619	rs10853619
	rs2825442*	rs2825442	rs2825442
	rs1275988	rs1275988	
	rs7467853	rs7467853	
	rs2820000*	rs2820000	
	rs6056810*	rs6056810	
		rs12739022	rs12739022
	rs16823124	rs8120907	rs10937395
	rs9874923	rs11222084	rs11736662
	rs16893776	rs10832255	rs1491840
	rs2384550	rs7105670	rs1563920
	rs956006		rs7997664
	rs4987082		rs17754634
	rs4712545*		rs17045014
	*rs*17747515**		rs204873
			rs54211

Our novel approach to blood pressure analysis identifies new plausible SNP pathways for high blood pressure, such as:

- rs4523751 is located in chromosome 12 LINC00615 gene while LINC00615 has been found as one of top SNPs that associated with blood pressure
- rs17747515 and rs10853618 are located in chromosome 18 colorectal carcinoma (DCC) genes that have been found significantly associated with hypertension via netrin signaling pathway.

SNP Selection Results

	$\tau = 0.9$ quantile	$\tau = 0.5$ median	mean PGEE
SNP Selected:	*rs*13149993	*rs*13149993	*rs*13149993
	rs1957563	rs1957563	rs1957563
	rs634094*	rs634094	rs634094
	rs12692453*	rs12692453	rs12692453
	rs11725914*	rs11725914	rs11725914
	rs4472152*	rs4472152	rs4472152
	rs4523751**	rs4523751	rs4523751
	rs10853618**	rs10853618	rs10853618
	rs10853619**	rs10853619	rs10853619
	rs2825442*	rs2825442	rs2825442
	*rs*1275988	*rs*1275988	
	rs7467853	rs7467853	
	rs2820000*	rs2820000	
	rs6056810*	rs6056810	
		rs12739022	rs12739022
	rs16823124	rs8120907	rs10937395
	rs9874923	rs11222084	rs11736662
	rs16893776	rs10832255	rs1491840
	*rs*2384550	rs7105670	rs1563920
	rs956006		rs7997664
	rs4987082		rs17754634
	rs4712545*		rs17045014
	rs17747515**		rs204873
			rs54211

Our novel approach also confirms SNPs previously identified in literature, such as:

- rs13149993 is previous detected by Kelly et al. (2013).
- rs1275988 is confirmed by Manichaikul et al. (2016).
- rs2384550 is confirmed by Levy et al. (2009).

SNP Selection Results

	$\tau = 0.9$ quantile	$\tau = 0.5$ median	mean PGEE
SNP Selected:	rs13149993	rs13149993	rs13149993
	rs1957563	rs1957563	rs1957563
	rs634094*	*rs634094*	*rs634094*
	rs12692453*	*rs12692453*	*rs12692453*
	rs11725914*	*rs11725914*	*rs11725914*
	rs4472152*	*rs4472152*	*rs4472152*
	rs4523751**	rs4523751	rs4523751
	rs10853618**	rs10853618	rs10853618
	rs10853619**	rs10853619	rs10853619
	rs2825442*	*rs2825442*	*rs2825442*
	rs1275988	rs1275988	
	rs7467853	rs7467853	
	rs2820000*	*rs2820000*	
	rs6056810*	*rs6056810*	
		rs12739022	rs12739022
	rs16823124	rs8120907	rs10937395
	rs9874923	rs11222084	rs11736662
	rs16893776	rs10832255	rs1491840
	rs2384550	rs7105670	rs1563920
	rs956006		rs7997664
	rs4987082		rs17754634
	rs4712545*		rs17045014
	rs17747515**		rs204873
			rs54211

- Newly detected SNPs.
- May deserve further investigation.

Tianhai Zu
zuti@mail.uc.edu

Simulation

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\alpha} + \left(1 + \kappa \left|\mathbf{X}_1\right|\right) \mathbf{e_i},$$

$\boldsymbol{\alpha}_0 = \left(1, 1, 1, 1, 1, 0, \ldots, 0\right)$.
$\kappa = 0$, homoscedastic
$\kappa = 1$, heteroscedastic
Errors $\mathbf{e_i}$, MVN, exchangeable correlation structure, $\rho = 0.8$.

Synthetic Gene Simulations

	Proposed Method		
	TP	FP	F_1
# of SNPs = 2000	4.98	0.47	0.9629
# of SNPs = 100,000	4.96	0.51	0.9563
	Univariate GWAS		
	TP	FP	F_1
# of SNPs = 2000	4.93	4.68	0.6890
# of SNPs = 100,000	4.95	213.87	0.0450

Simulations

		Independence			Exchangeable			AR-1		
p_n	τ	TP	FP	MSE	TP	FP	MSE	TP	FP	MSE
10	0.1	5	0.00	0.0129	5	0.00	0.0044	5	0.00	0.0082
		(0)	(0.00)	(0.0081)	(0)	(0.00)	(0.0051)	(0)	(0.00)	(0.0062)
	0.5	5	0.00	0.0071	5	0.00	0.0040	5	0.00	0.0045
		(0)	(0.00)	(0.0040)	(0)	(0.00)	(0.0025)	(0)	(0.00)	(0.0030)
	0.9	5	0.00	0.0125	5	0.00	0.0039	5	0.00	0.0081
		(0)	(0.00)	(0.0076)	(0)	(0.00)	(0.0040)	(0)	(0.00)	(0.0060)
200	0.1	5	0.11	0.0150	5	0.03	0.0090	5	0.07	0.0120
		(0)	(0.35)	(0.0122)	(0)	(0.17)	(0.0089)	(0)	(0.26)	(0.0112)
	0.5	5	0.00	0.0075	5	0.00	0.0042	5	0.00	0.0048
		(0)	(0.00)	(0.0042)	(0)	(0.00)	(0.0027)	(0)	(0.00)	(0.0029)
	0.9	5	0.11	0.0160	5	0.04	0.0081	5	0.04	0.0110
		(0)	(0.37)	(0.0122)	(0)	(0.20)	(0.0074)	(0)	(0.20)	(0.0083)
2000	0.1	5	0.15	0.0213	5	0.04	0.0126	5	0.10	0.0186
		(0)	(0.39)	(0.0142)	(0)	(0.20)	(0.0122)	(0)	(0.30)	(0.0163)
	0.5	5	0.01	0.0100	5	0.00	0.0069	5	0.00	0.0075
		(0)	(0.10)	(0.0063)	(0)	(0.00)	(0.0040)	(0)	(0.00)	(0.0046)
	0.9	5	0.11	0.0202	5	0.06	0.0127	5	0.06	0.0164
		(0)	(0.35)	(0.0151)	(0)	(0.28)	(0.0138)	(0)	(0.28)	(0.0137)

Homoscedastic Model

$n = 200$ and $m = 5$

True correlation structure is exchangeable.

Simulations

p_n	τ	Independence TP	Independence FP	Independence MSE	Exchangeable TP	Exchangeable FP	Exchangeable MSE	AR-1 TP	AR-1 FP	AR-1 MSE
10	0.1	5	0.04	0.0716	5	0.01	0.0569	5	0.01	0.0605
		(0)	(0.18)	(0.0481)	(0)	(0.10)	(0.0447)	(0)	(0.10)	(0.0466)
	0.5	5	0.00	0.0351	5	0.00	0.0249	5	0.00	0.0270
		(0)	(0.07)	(0.0242)	(0)	(0.00)	(0.0178)	(0)	(0.00)	(0.0200)
	0.9	5	0.04	0.2283	5	0.01	0.0519	5	0.01	0.0539
		(0)	(0.20)	(0.1342)	(0)	(0.12)	(0.0362)	(0)	(0.12)	(0.0381)
200	0.1	5	1.89	0.3196	5	0.52	0.1182	5	0.84	0.1534
		(0)	(0.83)	(0.1844)	(0)	(0.87)	(0.1356)	(0)	(1.49)	(0.1681)
	0.5	5	0.50	0.0838	5	0.06	0.0391	5	0.14	0.0446
		(0)	(0.74)	(0.0665)	(0)	(0.26)	(0.0355)	(0)	(0.38)	(0.0386)
	0.9	5	1.83	0.3051	5	0.49	0.1257	5	0.89	0.1472
		(0)	(0.71)	(0.1775)	(0)	(0.82)	(0.1584)	(0)	(1.38)	(0.1613)
2000	0.1	5	1.83	0.3313	5	0.52	0.1380	5	0.74	0.1554
		(0)	(0.70)	(0.2023)	(0)	(0.72)	(0.1787)	(0)	(1.08)	(0.1527)
	0.5	5	0.51	0.0763	5	0.08	0.0353	5	0.18	0.0420
		(0)	(0.72)	(0.0539)	(0)	(0.27)	(0.0265)	(0)	(0.46)	(0.0313)
	0.9	5	1.19	0.3687	5	0.65	0.1653	5	0.88	0.2083
		(0)	(0.80)	(0.2221)	(0)	(0.83)	(0.2145)	(0)	(1.65)	(0.2405)

Heteroscedastic Model

$n = 200$ and $m = 5$

True correlation structure is exchangeable.

Theoretical Properties I

Theorem 1 . Let $r_n = C\sqrt{p_1/n}$ and $c_n = C\sqrt{\log n}$. Assume (C1)-(C3) hold, and that $\sqrt{p_1/n} + \sqrt{\log(p \vee n)/n} << \lambda << \min_{j \le p_1} |\alpha_{0j}|$. When $p_1^{3/2} + \sqrt{np_1^2 r_n \log n} + p_1^{3/2}\log^{3/2} n = o(\sqrt{n})$, for any $\epsilon_n \to 0$, there exists $\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{\alpha}}_{(1)}, \widehat{\boldsymbol{\alpha}}_{(2)})$ that satisfies

(i) $\widehat{\boldsymbol{\alpha}}$ is an ϵ_n-approximate local minimizer of $\left\|\mathbf{S}^P_{(1)}(\widehat{\boldsymbol{\alpha}})\right\|$ and also an $o(\sqrt{n})$-approximate root.

(ii) $\widehat{\boldsymbol{\alpha}}_{(2)} = \mathbf{0}$.

Theoretical Properties II

Theorem 1 . Continued.

(iii) $\widehat{\boldsymbol{\alpha}}_{(1)}$ has the asymptotic normality property

$$\sqrt{n}\mathbf{a}^{\mathrm{T}}\boldsymbol{\Psi}_2^{-1/2}\boldsymbol{\Psi}_1\left(\widehat{\boldsymbol{\alpha}}_{(1)} - \boldsymbol{\alpha}_{0(1)}\right) \xrightarrow{d} N(0,1),$$

where $\mathbf{a}$ is any p_1-dimensional unit vector, $\boldsymbol{\Psi}_1 = E\left[\mathbf{X}_{i(1)}^{\mathrm{T}}\boldsymbol{\Gamma}_i\mathbf{R}^{-1}\boldsymbol{\Gamma}_i\mathbf{X}_{i(1)}\right]$ and $\boldsymbol{\Psi}_2 = E\left[\mathbf{X}_{i(1)}^{\mathrm{T}}\boldsymbol{\Gamma}_i\mathbf{R}^{-1}\mathbf{V}_0\mathbf{R}^{-1}\boldsymbol{\Gamma}_i\mathbf{X}_{i(1)}\right]$, and $\mathbf{V}_0$ is the true covariance matrix of $I\left\{\mathbf{Y}_i \leq \mathbf{X}_i\boldsymbol{\alpha}_0\right\}$.

(iv) $\left\|\mathbf{S}_{(2)}^{P}\left(\widehat{\boldsymbol{\alpha}}\right)\right\| = \mathbf{0}$ and $\widehat{\boldsymbol{\alpha}}_{(2)}$ has the zero crossing property: $\lim_{a\downarrow 0}\mathbf{S}_j^P\left(\widehat{\boldsymbol{\alpha}} + a\mathbf{1}_j\right)\mathbf{S}_j^P\left(\widehat{\boldsymbol{\alpha}} - a\mathbf{1}_j\right) < 0$ for $j > p_1$, where $\mathbf{1}_j$ denotes the binary p-vector with a single one at position j.

Theoretical Properties

Theorem 2 . Under the same assumptions of Theorem 1 and that $\sqrt{p_1 h} + \sqrt{n} h^2 = o(1)$, there exists $\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{\alpha}}_{(1)}, \widehat{\boldsymbol{\alpha}}_{(2)})$ that satisfies

(i) $\left\| \mathbf{S}_{\Phi}^{P}(\boldsymbol{\alpha}) \right\| = 0$.

(ii) $\widehat{\boldsymbol{\alpha}}_{(1)}$ has the asymptotic normality property as in Theorem 1.

(iii) $\widehat{\boldsymbol{\alpha}}_{(2)}$ has the zero crossing property: $\lim_{a \downarrow 0} \mathbf{S}_j^P(\widehat{\boldsymbol{\alpha}} + a\mathbf{1}_j)\mathbf{S}_j^P(\widehat{\boldsymbol{\alpha}} - a\mathbf{1}_j) < 0$ for $j > p_1$.

High Dimensional BIC

Theorem 3 . In addition to the assumptions used in Theorems 1 and 2, assume that we search over λ with the resulting support $S_\lambda \subseteq \{1, \ldots, p\}$ having size bounded by a constant p_{max} which is a known constant upper bound of p_1. Furthermore, we assume $p = O\left(n^\kappa\right)$ for some constant $\kappa > 0$ and $\min_{j \leq p_1} \left|\alpha_{0j}\right| >> \sqrt{C_n \log n / n}$. Then for $\widehat{\lambda}$ that minimizes HBIC, we have $P\left(S_{\widehat{\lambda}} = S_0\right) \rightarrow 1$.

Contributions in Application

- First to identify important risk factors and simultaneous modeling

  1) 500,000+ ultra-high dimensional genotype variables (SNPs)
  2) time-varying phenotype variables
  3) high quantiles of blood pressure
  4) longitudinal data

- First even for cross-sectional data

Contributions in Application

- Uncover SNPs, some have plausible pathways for hypertension
- Confirm SNPs previously identified in scientific literature
- Provide a comprehensive picture for heterogeneous data
- Different risk factors for different high quantiles, potentially for targeted treatments
- Joint modeling vs. traditional univariate GWAS analysis

Contributions in Methodology

- Methodology: First quantile longitudinal penalized GEE in ultra-high dimensions
- Computation: Efficient algorithms along with detailed implementations
- Theory:

  1) Consistency
  2) Asymptotic normality
  3) Oracle property in ultra-high dimensions
  4) Consistency for HBIC (polynomial order)

Thank You!

More details are also available on my website:
zzz1990771.github.io

This paper is in press in the Journal of the American Statistical Association.