Normal Score Transformation in Ultra High Dimensions

Hui Zou

University of Minnesota

February 9, 2017

Background

Box-Cox Model

Box and Cox (1964):

$$Y^{(\lambda)} = \sum_{j=1}^{p} X_j \beta_j + N(0, \sigma^2)$$

$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(Y) & \text{if } \lambda = 0 \end{cases}$$

Jointly estimate $\lambda, \boldsymbol{\beta}, \sigma^2$ by MLE (Bickel and Doksum 1981)

Maximum Rank Correlation Estimator (Han 1987, Sherman 1993, among others)

More recent examples

Semiparametric LDA (Lin and Jeon, 2003; Mai and Zou, 2015):

assume $(f_1(X_1), \ldots, f_p(X_p))|Y \sim N(\mu_Y, \mathbf{\Sigma})$ *where* $Y = 0, 1$,$f_1, \ldots, f_p$ *are unknown monotone increasing functions.*

Semiparametric Gaussian copula (Liu et al. 2009; Xue and Zou, 2012)

$(X_1, \cdots, X_p)$ *follows a semiparametric Gaussian copula distribution if*

$$(f_1(X_1), \cdots, f_p(X_p)) \sim N_p(0, \mathbf{\Sigma}),$$

where $f_1, \ldots, f_p$ *are unknown monotone increasing functions.*

Let $F_j(\cdot)$ be the CDF of X_j, then $f_j(x_j) = \Phi^{-1} \circ F_j(x_j) \sim N(0,1)$.

Would like to work on TTD: $(f_1(x_{i1}), \ldots, f_p(x_{ip}))^{\mathrm{T}}$

Let $F_j(\cdot)$ be the CDF of X_j, then $f_j(x_j) = \Phi^{-1} \circ F_j(x_j) \sim N(0,1)$.

Would like to work on TTD: $(f_1(x_{i1}), \ldots, f_p(x_{ip}))^{\mathrm{T}}$

Find an estimator of f_j, work on the ETD: $(\hat{f}_1(x_{i1}), \ldots, \hat{f}_p(x_{ip}))^{\mathrm{T}}$

or not

The truncated estimator

$$f_j(x_j) = \Phi^{-1} \circ F_j(x_j)$$

Let $\hat{F}_j$ be the usual CDF of X_j. Define $\hat{f}_j(\cdot) = \Phi^{-1} \circ \hat{F}_j^{\text{trunc.}}(\cdot)$

$$\hat{F}_j^{\text{trunc.}}(x) = \begin{cases} a_n, & \text{if } \hat{F}_j(x) < a_n, \\ \hat{F}_j(x), & \text{if } a_n < \hat{F}_j(x) < b_n, \\ b_n, & \text{if } \hat{F}_j(x) > b_n, \end{cases}$$

- $a_n = 1 - b_n = (4n^{1/4}\sqrt{\pi \log n})^{-1}$ in Liu et al. (2009) and $p = O(n^k)$
- $a_n = 1 - b_n = n^{-2}$ in Mai and Zou (2015) and $\log(p) = o(n^{1/3})$

Normal score transformation

Consider n observations $\{x_i\}_{i=1}^n$. The normal score transformation maps x_i to

$$\widehat{T}(x_i) = \Phi^{-1}\left(\frac{r_i - c}{n - 2c + 1}\right)$$

where r_i is the rank of x_i, Φ is the CDF of $N(0,1)$, and $c \geq 0$ is a constant. Popular choices of c are $0, 1/3, 3/8, 1/2$

Glass and Hopkins (1970), Wu et al. (2002), Anokhin et al. (2003), Dixon et al. (2007), Lambregts-Rommelse et al. (2008), Scuteri et al. (2007), Fan et al. (2013), Wang et al. (2015), Nansel et al. (2015); Berkowitz (2001), Peng et al. (2007), Cai et al. (2016)

This talk offers

Justifications of normal score transformations under ultra-high dimensions

Application 1: Semiparametric Gaussian Graphical Model

Conditional independence

$(X_1, \cdots, X_p)$ follows a semiparametric Gaussian copula distribution $(f_1(X_1), \cdots, f_p(X_p)) \sim N_p(0, \mathbf{\Sigma})$, $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$

$$\theta_{ij} = 0 \iff X_i \perp\!\!\!\perp X_j | \mathbf{X} \setminus \{X_i, X_j\}$$

Estimators

The ETD is $\widehat{\mathbf{T}}(\mathbf{X}_i) = (\widehat{T}_1(x_{i1}), \ldots, \widehat{T}_p(x_{ip}))^{\mathrm{T}}$

$$\widehat{\mathbf{\Sigma}}^{\mathrm{ETD}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathbf{T}}(\mathbf{X}_i)(\widehat{\mathbf{T}}(\mathbf{X}_i))^{\mathrm{T}}$$

$$\widehat{\mathbf{\Theta}} \Leftarrow \mathrm{SparseGraph}(\widehat{\mathbf{\Sigma}}^{\mathrm{ETD}})$$

Estimators

The ETD is $\widehat{\mathbf{T}}(\mathbf{X}_i) = (\widehat{T}_1(x_{i1}), \ldots, \widehat{T}_p(x_{ip}))^{\mathrm{T}}$

$$\widehat{\mathbf{\Sigma}}^{\mathrm{ETD}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathbf{T}}(\mathbf{X}_i)(\widehat{\mathbf{T}}(\mathbf{X}_i))^{\mathrm{T}}$$

$$\widehat{\mathbf{\Theta}} \Leftarrow \mathrm{SparseGraph}(\widehat{\mathbf{\Sigma}}^{\mathrm{ETD}})$$

As a reference, the TTD is $\mathbf{T}(\mathbf{X}_i) = (T_1(x_{i1}), \ldots, T_p(x_{ip}))^{\mathrm{T}}$

$$\widehat{\mathbf{\Sigma}}^{\mathrm{TTD}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{T}(\mathbf{X}_i)(\mathbf{T}(\mathbf{X}_i))^{\mathrm{T}}$$

$$\widehat{\mathbf{\Theta}}^{\mathrm{TTD}} \Leftarrow \mathrm{SparseGraph}(\widehat{\mathbf{\Sigma}}^{\mathrm{TTD}})$$

SparseGraph algorithms

- Neighborhood estimation (Meinshausen and Bühlmann 2006)
- Lasso penalized MLE (Li and Gui 2006; Yuan and Lin 2007; Banerjee et al. 2008; Friedman et al. 2008; Mazumder and Hastie 2012; Ma et al. 2012; Rothman et al. 2008; Ravikumar et al. 2011)
- Nonconvex penalized MLE (Fan et al. 2009)
- CLIME (Cai et al. 2011)
- Penalized D-Trace estimator (Zhang and Zou, 2014)

Penalized 'MLE':

$$\arg\min_{\mathbf{\Theta}\succ 0} \left\langle \mathbf{\Theta}, \hat{\mathbf{\Sigma}} \right\rangle - \log\det(\Theta) + \lambda \sum_{i\neq j} |\theta_{ij}|$$

The irrepresentable condition (Ravikumar et al. 2011)

$$\max_{e\in S^c} \|(\mathbf{\Sigma}^{\text{true}} \otimes \mathbf{\Sigma}^{\text{true}})_{e,S}((\mathbf{\Sigma}^{\text{true}} \otimes \mathbf{\Sigma}^{\text{true}})_{S,S})^{-1}\|_1 < 1.$$

Penalized D-Trace: (Zhang and Zou, 2014)

$$\arg\min_{\mathbf{\Theta}\succ 0} \frac{1}{2} \left\langle \Theta^2, \hat{\mathbf{\Sigma}} \right\rangle - \mathrm{tr}(\Theta) + \lambda \sum_{i\neq j} |\theta_{ij}|$$

$$\max_{e\in S^c} \|(\mathbf{\Sigma}^{\text{true}} \oplus \mathbf{\Sigma}^{\text{true}})_{e,S}((\mathbf{\Sigma}^{\text{true}} \oplus \mathbf{\Sigma}^{\text{true}})_{S,S})^{-1}\|_1 < 1.$$

CLIME (Cai et al. 2011)

$$\min_{\boldsymbol{\Theta}} \sum_{ij} |\theta_{ij}| \quad \text{s.t.} \quad \|\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}} - \mathbf{I}\|_{max} \leq \lambda_0.$$

Dantzig selector version of L_1 D-Trace est.

$$\min_{\boldsymbol{\Theta}} \sum_{ij} |\theta_{ij}| \quad \text{s.t.} \quad \|(\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}} + \widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta})/2 - \mathbf{I}\|_{max} \leq \lambda_0.$$

SparseGraph($\widehat{\mathbf{\Sigma}}$,$K$)

Input: $\widehat{\mathbf{\Sigma}}$

step 1: initialization by CLIME

$$\mathbf{\Theta}^0 = \arg\min_{\mathbf{\Theta}} \sum_{ij} |\theta_{ij}| \quad \text{s.t.} \quad \|\mathbf{\Theta}\widehat{\mathbf{\Sigma}} - \mathbf{I}\|_{max} \leq \lambda_0.$$

step 2: LLA updates: for $k = 1, 2, \ldots, K$

- $w_{ij} = P'_{\lambda^{\text{pen}}}(\theta_{ij}^{(k-1)})$ and solve

$$\widehat{\mathbf{\Theta}}^{(k)} = \arg\min_{\mathbf{\Theta} \succ 0} \left\langle \mathbf{\Theta}, \widehat{\mathbf{\Sigma}} \right\rangle - \log\det(\Theta) + \sum_{i \neq j} w_{ij}|\theta|_{ij}.$$

$$\Pr(\|\widehat{\mathbf{\Sigma}}^{\mathrm{TTD}} - \mathbf{\Sigma}\|_{\max} \geq \epsilon) \leq Cp^2 \exp(-Cn\epsilon^2)$$

Lemma

$$\Pr(\|\widehat{\mathbf{\Sigma}}^{\mathrm{ETD}} - \mathbf{\Sigma}\|_{\max} \geq \epsilon) \leq Cp^2 \exp(-\frac{Cn\epsilon^2}{\log^2 n})$$

Theorem

$\log(p) = o(n^{\alpha})$ *for any* $0 < \alpha < 1$. *Under the assumption* $\mathbf{\Theta}^{\text{true}}$ *is sparse and* $\|\mathbf{\Theta}^{\text{true}}\|_1$ *is bounded by M, with proper choices of* λ_0 *and* $\lambda^{\text{pen.}}$, *then w.h.p., SpareGraph(*$\widehat{\mathbf{\Sigma}}^{\text{ETD}}$*,*$K = 2$*) is equal to the oracle estimator of* $\mathbf{\Theta}^{\text{true}}$ *that assumes the support of* $\mathbf{\Theta}^{\text{true}}$ *was given.*

- Previously, $\alpha \leq 1/3$, unless we use the rank-correlation matrix estimator (Xue and Zou, 2012; Liu et al. 2012).
- The same result applies to other semiparametric copula models.

Application 2: improving model-free feature screening

Feature screening

$(y_i, X_i)_{i=1}^n$, $X = (x_1, \ldots, x_p)^{\mathrm{T}}$

1. For each variable X_j compute a ranking statistic R_j

2. Take the top $\hat{d}$ many variables according to the magnitude of ranking statistics

$$\hat{d} \ll p, \quad \hat{d} \approx n \text{ or } n/\log n$$

A data clean-up step: remove as many noise features as possible and keep all important ones–Sure Screening.

Malaria data

Data source: https://www.ncbi.nlm.nih.gov/sites/GDSbrowser.

22,283 probe sets as predictors and 71 samples of which 49 have been infected with malaria and 22 are healthy people.

Training/Test split: 50% vs 50% . Repeat 100 times.

Nearest Shrunken Centroid

NSC	Raw Data	Transformed Data
Error rate (%)	8.6(1.37)	5.7(0.70)

Model-free classification after screening

	Raw Data	Transformed Data
Random Forest	5.7% (0.38%)	2.9% (0.35%)

- SIS (correlation screening) (Fan and Lv, 2008)
- NIS (additive model screening) (Fan, Feng and Song, 2011)
- Distance correlation screening (Li, Zhong and Zhu, 2012)
- *t-test screening (Fan and Fan, 2008), Maximum marginal likelihood screening (Fan and Song, 2010), SIRS (Zhu, Li, Li and Zhu, 2011), Rank correlation screening (Li, Peng, Zhang and Zhu, 2013), empirical likelihood screening (Chang et al. 2013), quantile-adaptive screening (He et al. 2013), martingale difference correlation screening (Shao and Zhang, 2014), Covariance assisted screening (Ke, Jin and Fan, 2014), Kolmogorov filter (Mai and Zou, 2012), The fused Kolmogorov filter (Mai and Zou, 2014)*

  ...

Distance correlation

Székely, Rizzo and Bakirov (2007)

For two random vectors $\mathbf{u} \in \mathbb{R}^{d_\mathbf{u}}, \mathbf{v} \in \mathbb{R}^{d_\mathbf{v}}$,

$$\mathrm{dcov}^2(\mathbf{u}, \mathbf{v}) = \int_{\mathbb{R}^{d_\mathbf{u}+d_\mathbf{v}}} \|\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t}, \mathbf{s}) - \phi_\mathbf{u}(\mathbf{t})\phi_\mathbf{v}(\mathbf{s})\|^2 w(\mathbf{t}, \mathbf{s})\, \mathrm{d}\,\mathbf{t}\, \mathrm{d}\,\mathbf{s},$$

- $\phi_{\mathbf{u},\mathbf{v}}$ is the joint characteristic function of $\mathbf{u}$, $\mathbf{v}$;
- $\phi_\mathbf{u}$ and $\phi_\mathbf{v}$ are the characteristic functions of $\mathbf{u}$ and $\mathbf{v}$;
- $w(\mathbf{t}, \mathbf{s}) = \{c_{d_\mathbf{u}} c_{d_\mathbf{v}} \|\mathbf{t}\|_{d_\mathbf{u}}^{1+d_\mathbf{u}} \|\mathbf{s}\|_{d_\mathbf{v}}^{1+d_\mathbf{v}}\}^{-1}$ with $c_d = \dfrac{\pi^{(1+d)/2}}{\Gamma\{(1+d)/2\}}$.

$$\mathrm{dcov}(\mathbf{u}, \mathbf{v}) = 0 \iff \mathbf{u} \perp\!\!\!\perp \mathbf{v}$$

Another equivalent definition

$$\begin{aligned}\mathrm{dcov}^2(\mathbf{u},\mathbf{v}) &= \mathrm{E}[\|\mathbf{u}-\mathbf{u}'\|\|\mathbf{v}-\mathbf{v}'\|] + \mathrm{E}[\|\mathbf{u}-\mathbf{u}''\|]\,\mathrm{E}[\|\mathbf{v}-\mathbf{v}'\|] \\ &\quad -2\,\mathrm{E}[\|\mathbf{u}-\mathbf{u}'\|\|\mathbf{v}-\mathbf{v}''\|]\end{aligned}$$

where $(\mathbf{u},\mathbf{v})$, $(\mathbf{u}',\mathbf{v}')$ and $(\mathbf{u}'',\mathbf{v}'')$ are i.i.d. copies.

$$\mathrm{dcorr}(\mathbf{u},\mathbf{v}) = \frac{\mathrm{dcov}(\mathbf{u},\mathbf{v})}{\sqrt{\mathrm{dcov}(\mathbf{u})\,\mathrm{dcov}(\mathbf{v})}}.$$

Empirical distance covariance

Given data (U_i, V_i), $i = 1, 2 \ldots, n$.

$$\hat{S}_1 = \frac{1}{n^2} \sum_{j,k} \|U_j - U_k\| \|V_j - V_k\|$$

$$\hat{S}_2 = \frac{1}{n^2} \sum_{j,k} \|U_j - U_k\| \frac{1}{n^2} \sum_{j,k} \|V_j - V_k\|$$

$$\hat{S}_3 = \frac{1}{n^3} \sum_{i,j,k} \|U_i - U_k\| \|V_j - V_k\|$$

$$\widehat{\mathrm{dcov}}^2(\mathbf{u}, \mathbf{v}) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3$$

DCS

Li, Zhong and Zhu (2012)

Data (Y_i, X_i), $i = 1, \ldots, n$.

1. For each j in $1, \ldots, p$,

$$R_j^{\mathrm{dc}} = \widehat{\mathrm{dcorr}}^2(Y, X_j)$$

2. Pick the top d_n many variables according to the magnitude of R_j^{dc}.

The issue of malicious noise features

- $(y_i, X_i)_{i=1}^n$, $y_i = \pm 1$, $X = (X_1, \ldots, X_p)^{\mathrm{T}}$, p is large
- Efron (2008)–Ebay:

$$X|Y = y \sim N(\mu_y, \mathrm{Diag}(\sigma_1^2, \cdots, \sigma_p^2))$$

  see also Tibshirani et al. (2002)–NSC, Fan and Fan (2008)–FAIR
- A less perfect model:

$$g_i(X_j)|Y = y \sim \mu_{yj} + \epsilon_j, \quad E(\epsilon_j) = 0, g_j \text{ monotone}$$

A 'malicious' noise feature: $\mu_{+,j} = \mu_{-,j}$ and ϵ_j has very heavy tails.

NSC on the raw data

$n_1 = n_2 = n/2$, data standardized

1. Set a threshold level Δ (by CV)
2. Select $D = \{j : |\hat{\mu}_{+,j} - \hat{\mu}_{-,j}| = 2|\hat{\mu}_{+j}| \geq \Delta\}$
3. For $j \in D$, shrunken centroids mean by

$$\hat{\mu}'_{yj} = \mathrm{sgn}(\hat{\mu}_{yj})(|\hat{\mu}_{yj}| - \Delta/2)$$

4. The NSC classifier classifies x to class '+1' if

$$\sum_{j=1}^{p} x_j(\hat{\mu}'_{+,j} - \hat{\mu}'_{-,j}) > 0.$$

$$\eta_{yj} = \mathsf{E}(T(X_j)|Y = y) \quad \eta_{yj}^{\mathrm{ETD}} = \frac{1}{n_y} \sum_{y_i = y} \hat{T}(X_{ij})$$

$$\hat{\mathcal{D}} = \{j : |\hat{\eta}_{+j}^{\mathrm{ETD}} - \hat{\eta}_{-j}^{\mathrm{ETD}}| \geq \Delta\}$$

Proposition

Consider the "less perfect" model. The set of "interesting features" is $\mathcal{D} = \{j : \mu_{+j} \neq \mu_{-j}\}$. *With some proper choice of* Δ

$$\Pr(\hat{\mathcal{D}} = \mathcal{D}) \geq 1 - Cp \exp(-Cn\delta^2/\log n)$$

Back to the model-free setting

$$\text{dcov}(Y, X_k) = 0 \iff Y \perp\!\!\!\perp X_k \iff \text{dcov}(T_y(Y), T_k(X_k)) = 0$$

$$\omega_k = \text{dcov}(T_y(Y), T_k(X_k))$$

iRCDS:

1. For each variable X_j, compute

$$\hat{\omega}_j = \widehat{\text{dcov}}(\widehat{T}_y(Y), \widehat{T}_k(X_j))$$

2. iRCDS set is

$$\hat{\mathcal{D}}(d_n) = \{j : \hat{\omega}_j \text{ is among the } d_n\text{'th largest}\}.$$

Lemma

For $\epsilon \geq \frac{M\sqrt{\log n}}{\sqrt{n}}$,

$$\begin{aligned} & \Pr(|\widehat{\mathrm{dcov}}(\hat{T}_j(X_j), \hat{T}_y(Y)) - \widehat{\mathrm{dcov}}(T_j(X_j), T_y(Y))| \geq \epsilon) \\ \leq \quad & C\exp(-C\frac{n\epsilon^2}{\log^2 n}) \end{aligned}$$

$$\mathcal{D} = \{k : F(y \mid X_k) \text{ functionally depends on } X_k\}.$$

Theorem

There exists $\mathcal{S}$ such that $\mathcal{D} \subset \mathcal{S}$ and $\Delta_{\mathcal{S}} = \min_{j \in \mathcal{S}} \omega_j - \max_{j \in \mathcal{S}^c} \omega_j > cn^{-\kappa}$; then for any $d_n > |\mathcal{S}|$, we have

$$\Pr(\mathcal{D} \subset \hat{\mathcal{D}}(d_n)) \geq 1 - Cp[\exp(-Cn^{1-2(\kappa+\gamma)}) + n\exp(-Cn^{\gamma})].$$

γ is any constant in $(0, 1/2 - \kappa)$, e.g. $\gamma = (1 - 2\kappa)/3$.

Example 1

$$Y = (X_1 + X_2 + 1)^3 + N(0, 1),$$

$X_j \sim$ Cauchy independently, $n = 200, p = 5000$.

1. $\mathcal{M}$: the minimum model size to include all the true variables.

2. $\mathcal{P}_s$: the probability that an individual true variable is selected for a given model size $d_n = [n/\log(n)]$.

3. $\mathcal{P}_a$: the proportion that all true variables are selected for a given model size $d_n = [n/\log(n)]$.

Example 1

	$\mathcal{M}$					$\mathcal{P}_s$		$\mathcal{P}_a$
	5%	25%	50%	75%	95%	X1	X2	All
NIS	55.2	277.8	851.5	1788	3758	0.52	0.53	0.05
DCS	18.0	69.8	212.0	740.5	2285	0.58	0.56	0.15
iRDCS	2.0	2.0	2.0	2.0	2.0	1.00	1.00	1.00

Example 2

$$Y = X_1 + 0.8X_2 + 0.6X_3 + 0.4X_4 + 0.2X_5 + \exp(X_{20} + X_{21} + X_{22}) \cdot N(0, 1)$$

$(X_1, X_2, \ldots, X_p)^{\mathrm{T}}$ from multivariate normal distribution with zero mean and covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{p \times p}$. $n = 200, p = 2000$.

Example 2

	$\sigma_{ij} = 0.8^{\lvert i-j \rvert}$				
	5%	25%	50%	75%	95%
NIS	501.7	1293	1669	1863	1980
DCS	11.0	27.0	103.0	342.5	964.5
iRDCS	8.0	8.0	9.0	10.0	11.0

Example 3

$$Y = 5X_1 + 3(2X_2 - 1)^2 + 4g_3(X_3) + 6g_4(X_4) + \sqrt{1.74}\epsilon,$$

$$g_3(x) = \sin(2\pi x)/(2 - \sin(2\pi x))$$

$$\begin{aligned} g_4(x) &= 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) \\ &\quad +0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x) \end{aligned}$$

$X_j, j = 1, \cdots, p$ are independently Unif$(0, 1)$,
$n = 200, p = 2000$.

Example 3

	N(0,1) error					Cauchy error				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
NIS	4.0	4.0	4.0	4.0	7.0	41.8	557.5	1182	1717	1936
DCS	10.0	28.0	53.0	92.2	198.6	28.0	95.5	234.0	427.0	1072
iRDCS	5.0	13.0	24.0	47.0	106.2	10.0	43.0	77.5	142.2	367.7

Concluding remarks

- with normal score transformation, TTD $\approx$ ETD

- but you may not want to do the transformation

  1. A version of sparse PCA (Johnstone and Lu 2009)

  2. IF-PCA clustering (Jin and Wang 2016)

THANK YOU