Normal Score Transformation in Ultra High Dimensions

Hui Zou
University of Minnesota

February 9, 2017
Background
Box-Cox Model

Box and Cox (1964):

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

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

Jointly estimate \( \lambda, \beta, \sigma^2 \) 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):

\[ \text{assume } (f_1(X_1), \ldots, f_p(X_p))|Y \sim N(\mu_Y, \Sigma) \text{ where } \]

\[ Y = 0, 1, f_1, \ldots, f_p \text{ are unknown monotone increasing functions.} \]

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

\[(X_1, \cdots, X_p) \text{ follows a semiparametric Gaussian copula distribution if} \]

\[ (f_1(X_1), \cdots, f_p(X_p)) \sim N_p(0, \Sigma), \]

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: $(f_1(x_{i1}), \ldots, f_p(x_{ip}))^T$
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: $(f_1(x_{i1}), \ldots, f_p(x_{ip}))^T$

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

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 $\hat{f}_j(\cdot) = \Phi^{-1} \circ \hat{F}_j^{\text{trunc.}}(\cdot)$

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

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

$\triangleright a_n = 1 - b_n = n^{-2}$ 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

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

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

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
\( (f_1(X_1), \cdots, f_p(X_p)) \sim N_p(0, \Sigma), \Theta = \Sigma^{-1} \)

\[ \theta_{ij} = 0 \iff X_i \perp X_j \mid X \setminus \{X_i, X_j\} \]
Estimators

The ETD is \( \hat{T}(X_i) = (\hat{T}_1(x_{i1}), \ldots, \hat{T}_p(x_{ip}))^T \)

\[ \hat{\Sigma}_{\text{ETD}} = \frac{1}{n} \sum_{i=1}^{n} \hat{T}(X_i)(\hat{T}(X_i))^T \]

\( \hat{\Theta} \leftarrow \text{SparseGraph}(\hat{\Sigma}_{\text{ETD}}) \)
Estimators

The ETD is $\hat{T}(X_i) = (\hat{T}_1(x_{i1}), \ldots, \hat{T}_p(x_{ip}))^T$

$$\hat{\Sigma}^{ETD} = \frac{1}{n} \sum_{i=1}^{n} \hat{T}(X_i)(\hat{T}(X_i))^T$$

$$\hat{\Theta} \leftarrow \text{SparseGraph}(\hat{\Sigma}^{ETD})$$

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

$$\hat{\Sigma}^{TTD} = \frac{1}{n} \sum_{i=1}^{n} T(X_i)(T(X_i))^T$$

$$\hat{\Theta}^{TTD} \leftarrow \text{SparseGraph}(\hat{\Sigma}^{TTD})$$
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’:

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

The irrepresentable condition (Ravikumar et al. 2011)

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

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

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

$$\max_{e \in S^c} \| (\Sigma^{\text{true}} \oplus \Sigma^{\text{true}})_e, S ( (\Sigma^{\text{true}} \oplus \Sigma^{\text{true}})_{S, S} )^{-1} \|_1 < 1.$$
CLIME (Cai et al. 2011)

\[
\min_{\Theta} \sum_{ij} |\theta_{ij}| \quad \text{s.t.} \quad \|\Theta \hat{\Sigma} - I\|_{max} \leq \lambda_0.
\]

Dantzig selector version of $L_1$ D-Trace est.

\[
\min_{\Theta} \sum_{ij} |\theta_{ij}| \quad \text{s.t.} \quad \|(\Theta \hat{\Sigma} + \hat{\Sigma} \Theta)/2 - I\|_{max} \leq \lambda_0.
\]
SparseGraph\((\hat{\Sigma}, K)\)

Input: \(\hat{\Sigma}\)

step 1: initialization by CLIME

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

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

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

\[
\hat{\Theta}^{(k)} = \arg \min_{\Theta > 0} \left\langle \Theta, \hat{\Sigma} \right\rangle - \log \det(\Theta) + \sum_{i \neq j} w_{ij} |\theta_{ij}|.
\]
Lemma

\[ \Pr(\|\hat{\Sigma}^{\text{TTD}} - \Sigma\|_{\text{max}} \geq \epsilon) \leq C\rho^2 \exp(-Cn\epsilon^2) \]

\[ \Pr(\|\hat{\Sigma}^{\text{ETD}} - \Sigma\|_{\text{max}} \geq \epsilon) \leq C\rho^2 \exp(-\frac{Cn\epsilon^2}{\log^2 n}) \]
Theorem

\[ \log(p) = o(n^\alpha) \text{ for any } 0 < \alpha < 1. \]

Under the assumption \( \Theta^{\text{true}} \) is sparse and \( \|\Theta^{\text{true}}\|_1 \) is bounded by \( M \), with proper choices of \( \lambda_0 \) and \( \lambda^{\text{pen.}} \), then w.h.p., \( \text{SpareGraph}(\hat{\Sigma}^{\text{ETD}}, K = 2) \) is equal to the oracle estimator of \( \Theta^{\text{true}} \) that assumes the support of \( \Theta^{\text{true}} \) 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

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

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


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

<table>
<thead>
<tr>
<th>NSC</th>
<th>Raw Data</th>
<th>Transformed Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error rate (%)</td>
<td>8.6(1.37)</td>
<td>5.7(0.70)</td>
</tr>
</tbody>
</table>

Model-free classification after screening

<table>
<thead>
<tr>
<th></th>
<th>Raw Data</th>
<th>Transformed Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>5.7% (0.38%)</td>
<td>2.9% (0.35%)</td>
</tr>
</tbody>
</table>
- 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

Szekely, Rizzo and Bakirov (2007)

For two random vectors $u \in \mathbb{R}^{d_u}$, $v \in \mathbb{R}^{d_v}$,

\[ \text{d} \text{cov}^2(u, v) = \int_{\mathbb{R}^{d_u+d_v}} \| \phi_{u,v}(t, s) - \phi_u(t)\phi_v(s) \|^2 w(t, s) \, dt \, ds, \]

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

\[ \text{d} \text{cov}(u, v) = 0 \iff u \perp \perp v \]
Another equivalent definition

\[
dcov^2(u, v) = E[\|u - u'\| \|v - v'\|] + E[\|u - u''\|]E[\|v - v'\|]
- 2 E[\|u - u'\| \|v - v''\|]
\]

where \((u, v), (u', v')\) and \((u'', v'')\) are i.i.d. copies.

\[
dcorr(u, v) = \frac{dcov(u, v)}{\sqrt{dcov(u) dcov(v)}}.
\]
Empirical distance covariance

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

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

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

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

\[
\hat{d_{cov}}^2(u, v) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3
\]
Li, Zhong and Zhu (2012)

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

1. For each \(j\) in \(1, \ldots, p\),

\[ R_j^{dc} = \text{dcorr}^2 (Y, X_j) \]

2. Pick the top \(d_n\) many variables according to the magnitude of \(R_j^{dc}\).
The issue of malicious noise features

- $(y_i, X_i)_{i=1}^n, y_i = \pm 1, X = (X_1, \ldots, X_p)^T, p$ is large
- Efron (2008)–Ebay:
  \[ X|Y = y \sim N(\mu_y, \text{Diag}(\sigma_1^2, \cdots, \sigma_p^2)) \]
  
  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 $\epsilon_j$ has very heavy tails.
NSC on the raw data

\( n_1 = n_2 = n/2 \), data standardized

1. Set a threshold level \( \Delta \) (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

\[
\hat{\mu}'_{yj} = \text{sgn}(\hat{\mu}_{yj})(|\hat{\mu}_{yj}| - \Delta/2)
\]

4. The NSC classifier classifies \( x \) to class ‘+1’ if

\[
\sum_{j=1}^{p} x_j(\hat{\mu}'_{+,j} - \hat{\mu}'_{-,j}) > 0.
\]
\[ \eta_{yj} = E(T(X_j) | Y = y) \quad \eta_{yj}^{\text{ETD}} = \frac{1}{n_y} \sum_{y_i = y} \hat{T}(X_{ij}) \]

\[ \hat{D} = \{ j : |\hat{\eta}_{+j}^{\text{ETD}} - \hat{\eta}_{-j}^{\text{ETD}}| \geq \Delta \} \]

**Proposition**

Consider the “less perfect” model. The set of “interesting features” is \( \hat{D} = \{ j : \mu_{+j} \neq \mu_{-j} \} \). With some proper choice of \( \Delta \)

\[ \Pr(\hat{D} = D) \geq 1 - Cp \exp(-Cn\delta^2 / \log n) \]
Back to the model-free setting

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

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

iRCDS:

1. For each variable \( X_j \), compute

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

2. iRCDS set is

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

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

$$\Pr(\left|\hat{\text{dco}v}(\hat{T}_j(X_j), \hat{T}_y(Y)) - \text{dco}v(T_j(X_j), T_y(Y))\right| \geq \epsilon) \leq C \exp\left(-C \frac{n\epsilon^2}{\log^2 n}\right)$$
\[ \mathcal{D} = \{ k : F(y \mid X_k) \text{ functionally depends on } X_k \}. \]

**Theorem**

There exists $S$ such that $\mathcal{D} \subset S$ and
\[ \Delta_S = \min_{j \in S} \omega_j - \max_{j \in S^c} \omega_j > cn^{-\kappa}; \text{ then for any } d_n > |S|, \text{ we have} \]
\[ \Pr(\mathcal{D} \subset \hat{\mathcal{D}}(d_n)) \geq 1 - Cp[\exp(-Cn^{1-2(\kappa+\gamma)}) + n \exp(-Cn^\gamma)]. \]

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

\[ Y = (X_1 + X_2 + 1)^3 + N(0, 1), \]

\[ X_j \sim \text{Cauchy independently, } n = 200, \rho = 5000. \]

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

<table>
<thead>
<tr>
<th></th>
<th>$\mathcal{M}$</th>
<th></th>
<th></th>
<th></th>
<th>$\mathcal{P}_s$</th>
<th>$\mathcal{P}_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5%</td>
<td>25%</td>
<td>50%</td>
<td>75%</td>
<td>95%</td>
<td>X1</td>
</tr>
<tr>
<td>NIS</td>
<td>55.2</td>
<td>277.8</td>
<td>851.5</td>
<td>1788</td>
<td>3758</td>
<td>0.52</td>
</tr>
<tr>
<td>DCS</td>
<td>18.0</td>
<td>69.8</td>
<td>212.0</td>
<td>740.5</td>
<td>2285</td>
<td>0.58</td>
</tr>
<tr>
<td>iRDCS</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Example 2

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

\((X_1, X_2, \ldots, X_p)^\top\) from multivariate normal distribution with zero mean and covariance matrix \(\Sigma = (\sigma_{ij})_{p \times p}\). \(n = 200, p = 2000\).
Example 2

\[ \sigma_{ij} = 0.8^{|i-j|} \]

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIS</td>
<td>501.7</td>
<td>1293</td>
<td>1669</td>
<td>1863</td>
<td>1980</td>
</tr>
<tr>
<td>DCS</td>
<td>11.0</td>
<td>27.0</td>
<td>103.0</td>
<td>342.5</td>
<td>964.5</td>
</tr>
<tr>
<td>iRDCS</td>
<td>8.0</td>
<td>8.0</td>
<td>9.0</td>
<td>10.0</td>
<td>11.0</td>
</tr>
</tbody>
</table>
Example 3

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

\[ g_3(x) = \frac{\sin(2\pi x)}{(2 - \sin(2\pi x))} \]

\[ g_4(x) = 0.1 \sin(2\pi x) + 0.2 \cos(2\pi x) + 0.3 \sin^2(2\pi x) + 0.4 \cos^3(2\pi x) + 0.5 \sin^3(2\pi x) \]

\[ X_j, j = 1, \cdots, p \text{ are independently Unif}(0, 1), \]
\[ n = 200, p = 2000. \]
## Example 3

<table>
<thead>
<tr>
<th></th>
<th>N(0,1) error</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Cauchy error</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5%</td>
<td>25%</td>
<td>50%</td>
<td>75%</td>
<td>95%</td>
<td>5%</td>
<td>25%</td>
<td>50%</td>
<td>75%</td>
<td>95%</td>
<td>5%</td>
<td>25%</td>
<td>50%</td>
<td>75%</td>
<td>95%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIS</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>7.0</td>
<td>41.8</td>
<td>557.5</td>
<td>1182</td>
<td>1717</td>
<td>1936</td>
<td>10.0</td>
<td>28.0</td>
<td>53.0</td>
<td>92.2</td>
<td>198.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCS</td>
<td>10.0</td>
<td>28.0</td>
<td>53.0</td>
<td>92.2</td>
<td>198.6</td>
<td>28.0</td>
<td>95.5</td>
<td>234.0</td>
<td>427.0</td>
<td>1072</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iRDCS</td>
<td>5.0</td>
<td>13.0</td>
<td>24.0</td>
<td>47.0</td>
<td>106.2</td>
<td>10.0</td>
<td>43.0</td>
<td>77.5</td>
<td>142.2</td>
<td>367.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Concluding remarks

- with normal score transformation, $\text{TTD} \approx \text{ETD}$

- 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