Nonlinear Shrinkage Estimation in Quadratic Inference Function Analysis for Correlated Data

Clifford Lam

One-day workshop on Modern Statistical Methods in Health and Environment
Brunel University, June 9, 2017

Outline of the Talk

- GEE and QIF - In a Nutshell
- Motivation for Covariance Matrix Regularization
- Linear and Nonlinear Shrinkage
- Nonlinear Shrinkage in QIF
- Simulations and Epileptic Seizure Data Analysis
- Summary and Future Research

The Data

We deal with data that can be grouped into independent 'clusters'.

- **Longitudinal studies**: Independent subjects are observed over time. Within subject correlation.
  - E.g. - Alzheimer patients with different treatments are observed over time.
- **Clustered Randomized Controlled Trials** : Subjects are clustered into independent subgroups.
  - E.g. - Assessing the benefit of different teaching methods in different schools. Students observed within the same school are correlated.

GEE - Generalized Estimating Equations

- Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}}$, $i = 1, \ldots, n$, be independent response vectors.
- Let $\boldsymbol{\mu}_i = E(\mathbf{y}_i)$, and $g(\mu_{ij}) = \mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0$, $j = 1, \ldots, n_i$, where $g(\cdot)$ is a known link function, and $\boldsymbol{\beta}_0$ is d-dimensional.
- Define $\mathbf{A}_i(\boldsymbol{\beta}) = \partial \boldsymbol{\mu}_i(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}^{\mathrm{T}}$, $\mathbf{V}_i = \mathbf{D}_i^{1/2} \mathbf{R}_i \mathbf{D}_i^{1/2}$, where $\mathbf{R}_i$ is an $n_i \times n_i$ **working correlation matrix**, $\mathbf{D}_i$ is a diagonal matrix of **working marginal variances**.
- GEE find $\boldsymbol{\beta}$ to solve

$$\sum_{i=1}^{n} \mathbf{A}_i^{\mathrm{T}}(\boldsymbol{\beta}) \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}.$$

★ Can misspecify $\mathbf{R}_i$, losing efficiency, although still consistent.

QIF - Quadratic Inference Functions

- Proposed in Qu et al (2000). Decomposing

$$\mathbf{R}_i^{-1} \approx \sum_{r=1}^{m} \gamma_{ri} \mathbf{M}_{ri},$$

GEE is then

$$\sum_{r=1}^{m} \sum_{i=1}^{n} \gamma_{ri} \mathbf{A}_i^{\mathrm{T}}(\boldsymbol{\beta}) \mathbf{D}_i^{-1/2} \mathbf{M}_{ri} \mathbf{D}_i^{-1/2} (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}))$$
$$=: \sum_{r=1}^{m} \sum_{i=1}^{n} \gamma_{ri} \mathbf{g}_{ri}(\boldsymbol{\beta}).$$

QIF - Quadratic Inference Functions

- QIF utilizes **extended score equations**,

$$\bar{\mathbf{g}}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\beta}) = \begin{pmatrix} n^{-1}\sum_{i=1}^{n}\mathbf{g}_{1i}(\boldsymbol{\beta}) \\ \vdots \\ n^{-1}\sum_{i=1}^{n}\mathbf{g}_{mi}(\boldsymbol{\beta}) \end{pmatrix}.$$

- QIF then minimizes, with respect to $\boldsymbol{\beta}$,

$$E\left(\frac{\partial\bar{\mathbf{g}}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\right)^{\mathrm{T}}\mathbf{R}_n^{-1}(\boldsymbol{\beta})\bar{\mathbf{g}}(\boldsymbol{\beta}),\ \ \mathbf{R}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\beta})\mathbf{g}_i^{\mathrm{T}}(\boldsymbol{\beta}).$$

★ $\mathrm{Var}(\bar{\mathbf{g}}(\boldsymbol{\beta}_0)) = E(\mathbf{R}_n(\boldsymbol{\beta}_0))$. For fixed d, m, as $n \to \infty$, $\mathbf{R}_n(\boldsymbol{\beta}_0) - E(\mathbf{R}_n(\boldsymbol{\beta}_0)) \xrightarrow{\mathcal{P}} \mathbf{0}$.

QIF - Quadratic Inference Functions

- QIF optimally weights the m group of GEE equations when
  - $\boldsymbol{\beta}$ is close to $\boldsymbol{\beta}_0$;
  - ★ $\mathbf{R}_n(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{P}} E(\mathbf{R}_n(\boldsymbol{\beta}_0))$.
- QIF can be more efficient than GEE if working correlations are misspecified given the above.
- Initial value of $\boldsymbol{\beta}$ can be obtained from GEE, say $\tilde{\boldsymbol{\beta}}$, which is consistent as $n \to \infty$.
- ★ In many studies, n is not large. The dimension dm of the sample covariance $\mathbf{R}_n(\boldsymbol{\beta})$ cannot be assumed fixed relative to n. $\mathbf{R}_n(\boldsymbol{\beta}_0) - E(\mathbf{R}_n(\boldsymbol{\beta}_0)) \not\to \mathbf{0}$!

Curse of Dimensionality in Covariance Matrix Estimation

- Given stationary $\{\mathbf{y}_t\}_{1\le t\le n}$. Let $E(\mathbf{y}_t)=\mathbf{0}$, and
$$\mathbf{Y}=(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n)^{\mathrm{T}}.$$
- Sample covariance matrix $\mathbf{S}=n^{-1}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$.
- If $p/n\to c>0$ and $\{\mathbf{y}_t\}$ is i.i.d.$(\mathbf{0},\sigma^2\mathbf{I}_p)$ with independent components satisfying certain moment conditions, can show almost surely
$$\lambda_1\to\sigma^2(1+\sqrt{c})^2,\quad \lambda_N\to\min(\sigma^2(1-\sqrt{c})^2,0).$$
- If $p=25$, $n=500$, then the largest and smallest eigenvalues are 50% larger and 40% smaller than the corresponding true ones respectively.

Rotation-equivariant Estimator

- In Ledoit and Wolf's (2012) Annals paper, they consider the class of **rotation-equivariant estimators** $\mathbf{\Sigma(D) = PDP^T}$ in solving

$$\min_{\mathbf{D}} \|\mathbf{PDP^T} - \mathbf{\Sigma}\|_F^2,$$

with $\mathbf{S = PD_{sam}P^T}$, where $\mathbf{D_{sam}}$ contains the eigenvalues of $\mathbf{S}$.

- Solution is $d_i = \mathbf{p}_i^T \mathbf{\Sigma} \mathbf{p}_i$. How to estimate them?

★ They are NOT converging to the true eigenvalues!

Data Splitting and Regularization

- Split $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$, $\mathbf{Y}_1$ indep. of $\mathbf{Y}_2$, with size $p \times m$ and $p \times (n-m)$ resp.
- $\widetilde{\mathbf{\Sigma}}_1 = m^{-1}\mathbf{Y}_1\mathbf{Y}_1^{\mathrm{T}} = \mathbf{P}_1\mathbf{D}_1\mathbf{P}_1^{\mathrm{T}}$, $\widetilde{\mathbf{\Sigma}}_2 = (n-m)^{-1}\mathbf{Y}_2\mathbf{Y}_2^{\mathrm{T}}$.
- Write $\mathbf{P}_1 = (\mathbf{p}_{11}, \ldots, \mathbf{p}_{1p})$. Lam (2016) shows

Almost sure convergence of $\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}$

Under appropriate moment conditions, if $p/n \to c > 0$ and $\sum_{n\geq 1}(n-m)^{-3} < \infty$,

$$\frac{1}{p}\sum_{i=1}^{p}\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}\mathbf{1}_{\{\lambda_{1i}\leq x\}} - \frac{1}{p}\sum_{i=1}^{p}\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{p}_{1i}\mathbf{1}_{\{\lambda_{1i}\leq x\}} \overset{a.s.}{\to} 0,$$

where λ_{1i} is the ith largest eigenvalue of $\widetilde{\mathbf{\Sigma}}_1$.

- Hence, estimate $\mathbf{\Sigma}$ by $\widehat{\mathbf{\Sigma}}_m = \mathbf{P}_1\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}$.

Data Splitting and Regularization

Efficiency of $\widehat{\mathbf{\Sigma}}_m$

Under appropriate conditions, if $p/n, p/m \to c > 0$ and $\sum_{n \geq 1} p(n-m)^{-5} < \infty$,

$$\frac{\|\mathbf{P}\text{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{P})\mathbf{P}^{\mathrm{T}} - \mathbf{\Sigma}_p\|_F}{\|\mathbf{P}_1\text{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}} - \mathbf{\Sigma}_p\|_F} \xrightarrow{a.s.} 1.$$

★ For QIF setting, when n is not large, then better assume dm grows with n. Can we do the same for non-identically distributed data of different lengths?

Problem Formulation

- We want to find a rotation-equivariant estimator to be as close to $E(\mathbf{R}_n(\boldsymbol{\beta}_0))$ as possible. Consider

$$\min_{\mathbf{D}} \|\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}} - E(\mathbf{R}_n(\boldsymbol{\beta}_0))\|_F,$$

  where $\mathbf{R}_n(\boldsymbol{\beta}_0) = \mathbf{P}\mathbf{D}_{\mathsf{sam}}\mathbf{P}^{\mathrm{T}}$.

- Solution is

$$\mathbf{D} = \mathsf{diag}(\mathbf{P}^{\mathrm{T}} E(\mathbf{R}_n(\boldsymbol{\beta}_0))\mathbf{P}),$$
$$\Longrightarrow \boldsymbol{\Sigma}_{\mathsf{Ideal}} = \mathbf{P}\mathsf{diag}(\mathbf{P}^{\mathrm{T}} E(\mathbf{R}_n(\boldsymbol{\beta}_0))\mathbf{P})\mathbf{P}^{\mathrm{T}}.$$

Data Splitting in QIF Regularization

- Partition $\{1,\ldots,n\} = \{S_1,\ldots,S_L\}$. Each S_j is "small". Define

$$\mathbf{R}_{(-j)}(\boldsymbol{\beta}) = n^{-1}\sum_{i\in S_j^c}\mathbf{g}_i(\boldsymbol{\beta})\mathbf{g}_i^{\mathrm{T}}(\boldsymbol{\beta}) =: \mathbf{P}_1^{(j)}(\boldsymbol{\beta})\mathbf{D}^{(j)}\mathbf{P}_1^{(j)\mathrm{T}}(\boldsymbol{\beta}),$$

$$\mathbf{R}_{(j)}(\boldsymbol{\beta}) = n^{-1}\sum_{i\in S_j}\mathbf{g}_i(\boldsymbol{\beta})\mathbf{g}_i^{\mathrm{T}}(\boldsymbol{\beta}).$$

- Can write

$$\boldsymbol{\Sigma}_{\mathsf{Ideal}} = \sum_{j=1}^{L}\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}E(\mathbf{R}_{(j)}(\boldsymbol{\beta}_0))\mathbf{P})\mathbf{P}^{\mathrm{T}}.$$

- With a consistent $\widetilde{\boldsymbol{\beta}}$ and writing $\mathbf{P}_1^{(j)} = \mathbf{P}_1^{(j)}(\widetilde{\boldsymbol{\beta}})$,

$$\widehat{\boldsymbol{\Sigma}}_m = \sum_{j=1}^{L}\mathbf{P}_1^{(j)}\mathrm{diag}(\mathbf{P}_1^{(j)\mathrm{T}}\mathbf{R}_{(j)}(\widetilde{\boldsymbol{\beta}})\mathbf{P}_1^{(j)})\mathbf{P}_1^{(j)\mathrm{T}}.$$

Data Splitting in QIF Regularization

Efficiency of $\widehat{\boldsymbol{\Sigma}}_m$

If $\mathbf{R}_n(\widetilde{\boldsymbol{\beta}})$ has distinct eigenvalues, $dm \cdot \max_i n_i = o(n^{1/2})$, and $|S_j|$ has order of $n^{1/2}$ for all $j = 1, \ldots, L$, then

$$\max_{j=1,\ldots,L} \|\mathbf{P}_1^{(j)} - \mathbf{P}\| \xrightarrow{\mathcal{P}} 0, \quad \|\boldsymbol{\Sigma}_{\text{Ideal}} \widehat{\boldsymbol{\Sigma}}_m^{-1} - \mathbf{I}_{dm}\| \xrightarrow{\mathcal{P}} 0.$$

- Can assume $\max_i n_i < \infty$ with $dm/n \to c > 0$, but need to assume certain eigenvalue conditions on $\partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}^{\mathrm{T}}$ and $\mathbf{V}_i$ for all $i = 1, \ldots, n$.
- Can permute data M times to obtain $\widehat{\boldsymbol{\Sigma}}_m^{(1)}, \ldots, \widehat{\boldsymbol{\Sigma}}_m^{(M)}$. Define

$$\widehat{\boldsymbol{\Sigma}}_{m,M} = M^{-1} \sum_{s=1}^{M} \widehat{\boldsymbol{\Sigma}}_m^{(s)}.$$

Iterative Algorithm

- Define $\mathbf{G}(\boldsymbol{\beta}) = E(\partial \bar{\mathbf{g}}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^{\mathrm{T}})$. Propose to solve $\mathbf{G}(\widetilde{\boldsymbol{\beta}})\widehat{\boldsymbol{\Sigma}}_{m,M}^{-1}\bar{\mathbf{g}}(\boldsymbol{\beta}) = \mathbf{0}$ for $\boldsymbol{\beta}$.
- With $\widetilde{\boldsymbol{\beta}}$ as initial estimate $\boldsymbol{\beta}^{(0)}$, let $\mathbf{V}(\widetilde{\boldsymbol{\beta}}) = \mathbf{G}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}})\widehat{\boldsymbol{\Sigma}}_{m,M}^{-1}\mathbf{G}(\widetilde{\boldsymbol{\beta}})$. Run

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} - \mathbf{V}^{-1}(\widetilde{\boldsymbol{\beta}})\mathbf{G}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}})\widehat{\boldsymbol{\Sigma}}_{m,M}^{-1}\bar{\mathbf{g}}(\boldsymbol{\beta}^{(r)}).$$

Asymptotic normality

Let $\widehat{\boldsymbol{\beta}}$ be the root of $\mathbf{G}(\widetilde{\boldsymbol{\beta}})\widehat{\boldsymbol{\Sigma}}_{m,M}^{-1}\bar{\mathbf{g}}(\boldsymbol{\beta}) = \mathbf{0}$. For $\mathbf{B} \in \mathbb{R}^{a\times d}$,

$$n^{1/2}\mathbf{S}^{-1/2}\mathbf{B}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}_a),$$

$$\mathbf{S} = \mathbf{B}\mathbf{V}^{-1}(\widetilde{\boldsymbol{\beta}})\mathbf{G}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}})\widehat{\boldsymbol{\Sigma}}_{m,M}^{-1}E(\mathbf{R}_n(\boldsymbol{\beta}_0))\widehat{\boldsymbol{\Sigma}}_{m,M}^{-1}\mathbf{G}(\widetilde{\boldsymbol{\beta}})\mathbf{V}^{-1}(\widetilde{\boldsymbol{\beta}})\mathbf{B}^{\mathrm{T}}.$$

Asymptotic Covariance Matrix

- We estimate $\mathbf{S}$ by $\widehat{\mathbf{S}} = \mathbf{B}\mathbf{V}^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{B}^{\mathrm{T}}$.
- Practical covariance matrix for $\widehat{\boldsymbol{\beta}}$ is $\mathbf{V}^{-1}(\widehat{\boldsymbol{\beta}})/n$.

Simulation Study

- Matlab code is used for the QIF with our covariance regularization adapted.

- Compared to GEE, and **linear shrinkage** of Westgate and Braun (2013).

- 1. **Gaussian data**; $\boldsymbol{\beta}_0 = (1, 0.5, -0.5, 2, 3)^{\mathrm{T}}$. Length of $\mathbf{y}_i$ random between 3 and 6.
  - Compound Symmetry $\mathbf{R}_{\mathsf{CS}}$: $(\mathbf{R})_{ij} = 0.8$ for $i \neq j$.
  - Toeplitz $\mathbf{R}_{\mathsf{TP}}$: From first to fifth off-diagonal : $(0.8, 0.6, 0.4, 0.2, 0.1)$.

- $\mathbf{X} = \mathsf{Intercept} + (1.5\mathbf{R}_{\mathsf{CS}})^{1/2}(\mathbf{z}_2, \ldots, \mathbf{z}_5)$, $\mathbf{z}_i \sim N(\mathbf{0}, \mathbf{I})$.

- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + (1.5\mathbf{R})^{1/2}\mathbf{z}$, $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$.

Simulation Study

- 2. **Poisson data**; $\boldsymbol{\beta}_0 = (0.2, 0.5, -0.5, 0.2, -0.3)^{\mathrm{T}}$. Length of $\mathbf{y}_i$ random between 3 and 6.
  - Toeplitz $\mathbf{R}_{\mathsf{TP}}$: From first to fifth off-diagonal : $(0.8, 0.6, 0.4, 0.2, 0.1)$. Is only **approximate** because of computational time to simulate data with EXACT correlation. R package `PoisNor`.
- $\mathbf{X} = \mathsf{Intercept} + (1.5\mathbf{R}_{\mathsf{CS}})^{1/2}(\mathbf{z}_2, \ldots, \mathbf{z}_5)$, $\mathbf{z}_i \sim N(\mathbf{0}, \mathbf{I})$; Mean $\boldsymbol{\mu} = \exp(\mathbf{X}\boldsymbol{\beta}_0)$.
- Generate $\mathbf{y}$ using the **approximate** Toeplitz correlation structure, with corresponding mean.

Epileptic Seizure Data Analysis

- $\mathbf{y}$: Seizure Counts of 59 patients, observed in 4 successive 2-week intervals.
- x_{1i} : 1 - Received drug, 0 - Received placebo.
- x_{2i} : $\log(C/4)$, where C =seizure counts in a 8-week prior period.
- x_{3i} : $\log$(Subject's age).
- x_{4i} : 1-First 2-week interval, 2-second, and so on.

Summary and Future Research

- When dimension is large, sample covariance matrix is not a good estimator. Eigenvalues are heavily biased.
- In QIF, novel nonlinear shrinkage performs better than original QIF and linear shrinkage QIF.
- Most important is still the choice of working correlation structures. Can combine with penalized estimation, penalizing 'useless' equations? Empirical likelihood approach.

References

- Lam, C., 2016. Nonparametric Eigenvalue-Regularized Precision or Covariance Matrix Estimator. *Annals of Statistics*, **44(3)**, 928-953.
- Ledoit, O. and Wolf, M., 2012. NONLINEAR SHRINKAGE ESTIMATION OF LARGE-DIMENSIONAL COVARIANCE MATRICES. *Annals of Statistics*, **40(2)**, 1024-1060.
- Qu, A., Lindsay, B.G. and Li, B., 2000. Improving generalised estimating equations using quadratic inference functions. *Biometrika*, **87**, 823âĂŞ836.
- Westgate, P. M. and Braun, T. M., 2013. An improved quadratic inference function for parameter estimation in the analysis of correlated data. *Statist. Med.*, **32** 3260-3273.