Big Data Analysis via Multivariate Confidence Distribution

Ching-Wei Cheng

Department of Statistics
Purdue University

October 6, 2015
(Group Meeting)

A joint work with Prof. Guang Cheng

- "Big Data" are characterized by (ultra)high dimensionality and (extraordinarily) large sample size, which introduce unique computational and statistical challenges (Fan *et al.*, 2014)
  - Massive sample size arises scalability issue
  - Aggregating data from different sources creates heterogeneity
  - Heterogeneity may lead to high dimensionality
  - Increasing dimensionality causes noise accumulation
- We propose a Divide-and-Conquer (D&C) approach using the Confidence Distribution (CD) to make efficient statistical inference of Big Data
- Taking individual participant data (IPD)[1] method as oracle
- When IPD data is not available, aggregate data (AD) method (the average estimator) is the most commonly used combining approach

[1] The "gold standard" in meta-analysis

Outline

Review of the General CD Framework

Inference for single parameter

- $\theta \in \mathbb{R}$, the parameter of interest
- $\boldsymbol{X}_n = \{X_1, \ldots, X_n\}$, observed data
- $\widehat{\theta} \equiv \widehat{\theta}(\boldsymbol{X}_n)$, an estimate of θ

Definition 1

A function $H(\cdot) \equiv H(\widehat{\theta}, \cdot) = H(\boldsymbol{X}_n, \cdot)$ on $\mathcal{X} \times \Theta \to [0, 1]$ is called a *confidence distribution (CD)* for a parameter θ, if

(R1) For each given $\boldsymbol{X}_n \in \mathcal{X}$, $H(\cdot)$ is a sample-dependent continuous CDF on Θ;

(R2) When $\theta = \theta_0$ the true parameter value, $H(\theta_0) \equiv H(\boldsymbol{X}_n, \theta_0)$ follows *Uniform*$(0, 1)$.

The function $H(\cdot)$ is an *asymptotic confidence distribution (aCD)*, if the *Uniform*$(0, 1)$ requirement is true only asymptotically. The function $h(\theta) = H'(\theta)$ is called the *(a)CD density* if it exists.

Review of the General CD Framework

(Xie and Singh, 2013)

- A distribution estimator of θ rooted from Fisher's fiducial reasoning
- The CD concept is related to normalized likelihood functions, bootstrap distributions, p-value/significance functions, fiducial inference, and Bayesian mapproaches
- (R1) furnishes a CD with CDF properties over the parameter space
- (R2) makes a CD endowed with inferential ability
  - H^{-1} can be used to construct confidence intervals
  - $H(C) = \int_C \mathrm{d}H(\theta)$ can be used as a p-value for testing $K_0 : \theta \in C$
  - $M = H^{-1}(1/2)$ gives a median estimator of θ
- Xie *et al.* (2011) propose a unified CD approach for meta-analysis (combining infoemation from independent sources)
- However, extension for multi-parameter is non-trivial

Review of the General CD Framework

Example 1

Suppose we observe $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ and $(\bar{X}, \hat{\sigma}^2)$ is an bivariate estimator of (μ, σ^2). It is known that $(\mu - \bar{X})/\sigma \sim \mathcal{N}(0, 1)$ and $(n-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$ independently. Define

$$H_{t,\chi^2}(\mu, \sigma^2) = \Phi\left(\frac{\mu - \bar{X}}{\sigma}\right)\left[1 - F_{\chi^2_{n-1}}\left(\frac{(n-1)\hat{\sigma}^2}{\sigma^2}\right)\right],$$

where the second multiplicative term on the right hand side is to ensure $H_{t,\chi^2}(\mu, \sigma^2)$ is increasing in each coordinates. Clearly, $H_{\mathcal{N},\chi^2}(\mu, \sigma^2)$ is a proper bivariate CDF and thus satisfies (R1). However, at the true parameter values, $H_{\mathcal{N},\chi^2}(\mu_0, \sigma_0^2) \overset{d}{=} U_1 U_2 \not\sim \textit{Uniform}(0, 1)$, where "$\overset{d}{=}$" represents the equality in distribution and $U_1, U_2 \overset{iid}{\sim} \textit{Uniform}(0, 1)$.

- Pivot functions and probability integral transform (PIT) are important for building CD functions!
- (R2) fails due to multivariate PIT

Review of the General CD Framework

Review of CD-based approaches for multivariate meta-analysis

To the best of our knowledge, only two papers are dedicated to CD-based approaches for multivariate meta-analysis with S heterogeneous studies:

- Yang *et al.* (2014) employ the bootstrap distribution idea under random-effect model settings
  - Let $\boldsymbol{\xi}_s$ denote multivariate normal CD (MNCD) random vectors having multivariate Gaussian CDF $H_s(\boldsymbol{\theta}_s; \widehat{\boldsymbol{\theta}}_s)$
  - Define $\boldsymbol{\xi}$ using a weighted linear combination of $\boldsymbol{\xi}_s$, and thus $\boldsymbol{\xi}$ is also a MNCD random vector
- Liu *et al.* (2015) make use of the normalized likelihood function idea under fixed-effect model settings
  - Study-level CD densities $h_s(\boldsymbol{\theta}_s; \widehat{\boldsymbol{\theta}}_s)$
  - Combined CD density $h^{\mathrm{Liu}}(\boldsymbol{\theta}) = \prod_{s=1}^{S} h_s(\boldsymbol{\theta}_s; \widehat{\boldsymbol{\theta}}_s)$
  - $\widehat{\boldsymbol{\theta}}_{\mathrm{Liu}} = \arg\max_{\boldsymbol{\theta}} h^{\mathrm{Liu}}(\boldsymbol{\theta})$ is asymptotically as efficient as IPD estimator
- Both methods are motivated by the CD framework, yet none of them provide multivariate CD (MCD) functions satisfying (R2)
- The number of studies S is fixed

A Genuine MCD Approach

Generic notations

- $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\mathsf{T} \in \Theta \subset \mathbb{R}^p$ the parameter of interest with $p = \dim(\Theta)$
- $\boldsymbol{X}_N = \{X_1, \dots, X_N\}$, observed data of size N, iid from a population $F_{\boldsymbol{\theta}}$
- IPD M-estimator

$$\widehat{\boldsymbol{\theta}}_{\mathrm{IPD}} = \underset{\boldsymbol{\theta} \in \Theta}{\arg\min} \sum_{i=1}^{N} \varrho(X_i; \boldsymbol{\theta}), \tag{1}$$

  where ϱ is a known loss function
- $\boldsymbol{\Sigma}_{\mathrm{IPD}}$ denotes the covariance matrix of $\widehat{\boldsymbol{\theta}}_{\mathrm{IPD}}$

A Genuine MCD Approach

Generic notations

For D&C approaches when N is massive, partition the full data into S chunks...

- $\boldsymbol{\theta}_s = \boldsymbol{Q}_s(\boldsymbol{\theta}) \in \Theta_s$ denote the parameter in the s^{th} subpopulation with $p_s = \dim(\Theta_s)$ for some measurable mappings $\boldsymbol{Q}_s(\cdot)$
- $\boldsymbol{X}_s = \{X_{s,1}, \ldots, X_{s,n_s}\} \in \mathcal{X}_s$, the s^{th} subsample of size $n_s \equiv n = N/S$
- Define the subsample M-estimators as

$$\widehat{\boldsymbol{\theta}}_s = \underset{\boldsymbol{\theta}_s \in \Theta_s}{\arg\min} \sum_{i=1}^{n} \varrho(X_{s,i}; \boldsymbol{\theta}_s), \quad s = 1, \ldots, S. \tag{2}$$

  with covariance matrices $\boldsymbol{\Sigma}_s$[2]
- Assume $\widehat{\boldsymbol{\theta}}_s$ are elliptically distributed
- IPD M-estimator (1) can be rewritten as

$$\widehat{\boldsymbol{\theta}}_{\text{IPD}} = \underset{\boldsymbol{\theta} \in \Theta}{\arg\min} \sum_{s=1}^{S} \sum_{i=1}^{n} \varrho(X_{s,i}; \boldsymbol{Q}_s(\boldsymbol{\theta})) \tag{3}$$

[2] Note that the scale matrix, which is proportional to the covariance matrix, is used to describe and standardize an elliptical distribution, while statistical modeling usually comes up with covariance matrices.

MCD Procedure for Homogeneous Case

Recall that normalization is important for CD function construction...

Theorem 1

Assume regularity conditions (B1)–(B8) in He and Shao (1996) applicable to all partitions of a homogeneous massive data. Let $S = N^\gamma, \gamma \in (0,1)$. Then

(a) *is asymptotically normal, that is,*

$$\frac{1}{\sqrt{S}} \sum_{s=1}^{S} \boldsymbol{\Sigma}_s^{-1/2} (\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p); \tag{4}$$

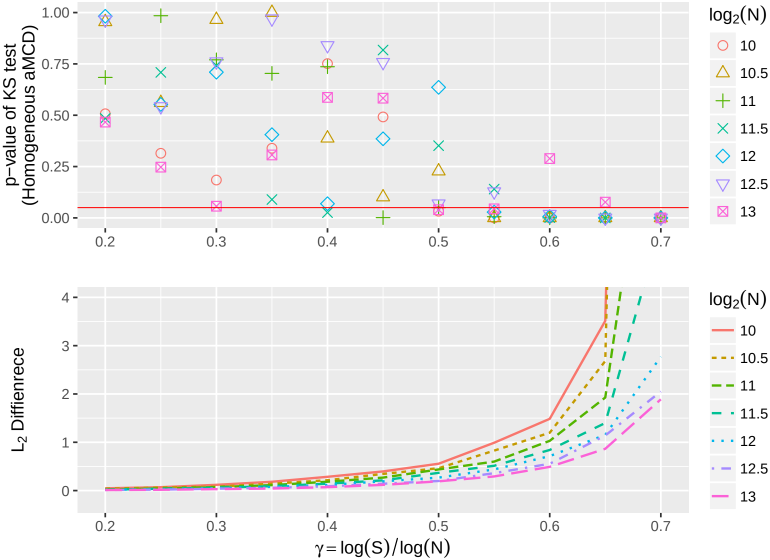
(b) *is asymptotically equivalent to the normalized IPD estimator, that is,*

$$\boldsymbol{\Sigma}_{\mathrm{IPD}}^{-1/2} (\widehat{\boldsymbol{\theta}}_{\mathrm{IPD}} - \boldsymbol{\theta}_0) - \frac{1}{\sqrt{S}} \sum_{s=1}^{S} \boldsymbol{\Sigma}_s^{-1/2} (\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_0) = o_{a.s.}(1) \tag{5}$$

if $\gamma < r/(2+r)$, where $r > 0$ is a constant which measures the smoothness of the score function.

MCD Procedure for Homogeneous Case

Empirical veryfication of the Theorem 1



MCD Procedure for Homogeneous Case

Construct aMCD function

- Theorem 1 (a) immediately leads to

$$\frac{1}{S}\left[\sum_{s=1}^{S}\boldsymbol{\Sigma}_s^{-1/2}(\widehat{\boldsymbol{\theta}}_s-\boldsymbol{\theta}_0)\right]^{\mathsf{T}}\left[\sum_{s=1}^{S}\boldsymbol{\Sigma}_s^{-1/2}(\widehat{\boldsymbol{\theta}}_s-\boldsymbol{\theta}_0)\right]\xrightarrow{D}\chi_p^2.$$

- Define an asymptotic MCD (aMCD) function as

$$H^a(\boldsymbol{\theta})=F_{\chi_p^2}\left(\frac{1}{S}\left[\sum_{s=1}^{S}\boldsymbol{\Sigma}_s^{-1/2}(\widehat{\boldsymbol{\theta}}_s-\boldsymbol{\theta})\right]^{\mathsf{T}}\left[\sum_{s=1}^{S}\boldsymbol{\Sigma}_s^{-1/2}(\widehat{\boldsymbol{\theta}}_s-\boldsymbol{\theta})\right]\right), \quad (6)$$

  where $F_{\chi_p^2}$ denotes the χ_p^2 CDF
- Define the corresponding asymptotic multivariate confidence density (aMCd) by switching the CDF to density, i.e.,

$$h^a(\boldsymbol{\theta})=f_{\chi_p^2}\left(\frac{1}{S}\left[\sum_{s=1}^{S}\boldsymbol{\Sigma}_s^{-1/2}(\widehat{\boldsymbol{\theta}}_s-\boldsymbol{\theta})\right]^{\mathsf{T}}\left[\sum_{s=1}^{S}\boldsymbol{\Sigma}_s^{-1/2}(\widehat{\boldsymbol{\theta}}_s-\boldsymbol{\theta})\right]\right), \quad (7)$$

  where $f_{\chi_p^2}$ is the χ_p^2 density function

MCD Procedure for Homogeneous Case

Construct eMCD function

Exact MCD (eMCD) functions can be made in similar fashion...

- $\boldsymbol{g}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}_s) = \kappa_s \boldsymbol{\Sigma}_s^{-1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_s)$ denote the pivot functions, where $\kappa_s > 0$ s.t. $\kappa_s \boldsymbol{\Sigma}_s^{-1/2}$ are scale matrices
- For $j = 1, \ldots, p$, denote the j^{th} component of $\boldsymbol{g}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}_s)$ by $g_j(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}_s)$ and the corresponding marginal distributions by $G_{s,j}$, which are free of $\boldsymbol{\theta}$
- For notational simplicity,
  - $\boldsymbol{G}_s(\boldsymbol{g}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_s)) = (G_{s,1}(g_1(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_s)), \ldots, G_{s,p}(g_p(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_s)))$
  - $\boldsymbol{\Phi}^{-1}(u_1, \ldots, u_p) = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p))$
- For elliptical distributions, Pearson uncorrelatedness is invariant under monotonic transformations (Lindskog *et al.*, 2003)
  - Pearson correlation $\rho(X, Y) = 0$ iff Kendall's tau $\tau(X, Y) = 0$
  - Rank-based correlations are invariant under monotonic transformations

MCD Procedure for Homogeneous Case

Construct eMCD function

- Finally we have $\mathbf{\Phi}^{-1}(\boldsymbol{G}_s(\boldsymbol{g}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}_s))) \overset{iid}{\sim} \mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p), s = 1, \ldots, S$
- Define an eMCD by

$$H^e(\boldsymbol{\theta}) = F_{\chi^2_p} \left(\frac{1}{S} \left[\sum_{s=1}^{S} \mathbf{\Phi}^{-1}(\boldsymbol{G}_s(\boldsymbol{g}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}_s))) \right]^{\mathsf{T}} \left[\sum_{s=1}^{S} \mathbf{\Phi}^{-1}(\boldsymbol{G}_s(\boldsymbol{g}(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}_s))) \right] \right) \tag{8}$$

and let $h^e(\boldsymbol{\theta})$ denote the associated eMCd.

MCD Procedure for Heterogeneous Case

- $\boldsymbol{\theta}_s = (\boldsymbol{\alpha}_s, \boldsymbol{\beta}) \in \mathcal{A}_s \times \mathcal{B}, s = 1, \ldots, S$
  - $\boldsymbol{\beta} \in \mathcal{B}$ is the commonality across the all subpopulations
  - $\boldsymbol{\alpha}_s \in \mathcal{A}_s$ are the subpopulation-specific heterogeneity
- Therefore $\boldsymbol{\theta} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_S, \boldsymbol{\beta})$
- Without loss of generality, consider $\dim(\mathcal{A}_s) \equiv p_h$ and let $\dim(\mathcal{B}) = p_c$.
- Recall that $\boldsymbol{\theta}_s = \boldsymbol{Q}_s(\boldsymbol{\theta})$

MCD Procedure for Heterogeneous Case

Corollary 1

Assume regularity conditions (B1)–(B8) in He and Shao (1996) applicable to the group M-estimates $\widehat{\boldsymbol{\theta}}_s$ of $\boldsymbol{\theta}_s = \boldsymbol{Q}_s(\boldsymbol{\theta}), s = 1, \ldots, S$. Let $S = N^\gamma$ and $\gamma \in (0, 1)$. Then we have

$$\frac{1}{\sqrt{S}} \sum_{s=1}^{S} \boldsymbol{\Sigma}_s^{-1/2} (\widehat{\boldsymbol{\theta}}_s - \boldsymbol{Q}_s(\boldsymbol{\theta}_0)) \xrightarrow{D} \mathcal{N}(\mathbf{0}_{p_c+p_h}, \boldsymbol{I}_{p_c+p_h}) \tag{9}$$

if $\gamma < r/(2+r)$, where $r > 0$ is a constant which measures the smoothness of the score function.

- IPD information cannot be fully recovered
- Only asymptotic normality is needed for defining MCD functions

MCD Procedure for Heterogeneous Case

A one-stage MCD functions

- A one-stage aMCD is defined by

$$
\begin{aligned}
&H^a(\boldsymbol{\theta}) \\
&= F_{\chi^2_{p_c+p_h}}\left(\frac{1}{S}\left[\sum_{s=1}^{S}\boldsymbol{\Sigma}_s^{-1/2}(\widehat{\boldsymbol{\theta}}_s-\boldsymbol{Q}_s(\boldsymbol{\theta}))\right]^{\mathsf{T}}\left[\sum_{s=1}^{S}\boldsymbol{\Sigma}_s^{-1/2}(\widehat{\boldsymbol{\theta}}_s-\boldsymbol{Q}_s(\boldsymbol{\theta}))\right]\right)
\end{aligned} \tag{10}
$$

- A one-stage eMCD is defined by

$$
\begin{aligned}
&H^e(\boldsymbol{\theta}) \\
&= F_{\chi^2_{p_c+p_h}}\left(\frac{1}{S}\left[\sum_{s=1}^{S}\boldsymbol{\Phi}^{-1}(\boldsymbol{G}_s(\boldsymbol{g}(\boldsymbol{Q}_s(\boldsymbol{\theta});\widehat{\boldsymbol{\theta}}_s)))\right]^{\mathsf{T}}\left[\sum_{s=1}^{S}\boldsymbol{\Phi}^{-1}(\boldsymbol{G}_s(\boldsymbol{g}(\boldsymbol{Q}_s(\boldsymbol{\theta});\widehat{\boldsymbol{\theta}}_s)))\right]\right).
\end{aligned} \tag{11}
$$

MCD Procedure for Heterogeneous Case

A two-stage MCD procedure

Drawbacks of one-stage MCD functions...

- $H^a(\boldsymbol{\theta})$ and $H^a(\boldsymbol{\theta})$ can only make inference for the whole $\boldsymbol{\theta}$ but not for $\boldsymbol{\beta}$
- In practice, people make inference for commonality and heterogeneity separately
- One-stage MCD estimators of $\boldsymbol{\alpha}_s$ are not efficient (shown later by simulation)
- Efficiency boosting technique for heterogeneity estimates
  - Obtain a MCD estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{MCD}}$ for commonality $\boldsymbol{\beta}$ first
  - Plug $\widehat{\boldsymbol{\beta}}_{\mathrm{MCD}}$ back to the model and estimate $\boldsymbol{\alpha}_s$ separately

MCD Procedure for Heterogeneous Case

A two-stage MCD procedure

Stage 1: Estimate commonality $\boldsymbol{\beta}$ via MCD

- Decompose subsample estimates by

$$\widehat{\boldsymbol{\theta}}_s = \begin{pmatrix} \widehat{\boldsymbol{\alpha}}_s \\ \widehat{\boldsymbol{\beta}}_s \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_s = \begin{bmatrix} \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\alpha}}_s \widehat{\boldsymbol{\alpha}}_s} & \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\alpha}}_s \widehat{\boldsymbol{\beta}}_s} \\ \boldsymbol{\Sigma}^{\mathsf{T}}_{\widehat{\boldsymbol{\alpha}}_s \widehat{\boldsymbol{\beta}}_s} & \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}_s \widehat{\boldsymbol{\beta}}_s} \end{bmatrix},$$

- Since $\mathcal{B} \subset \Theta_s$ for all $s = 1, \ldots, S$, we are back to homogeneous case
- aMCD for $\boldsymbol{\beta}$

$$H^a(\boldsymbol{\beta}) = F_{\chi^2_{p_c}} \left(\frac{1}{S} \left[\sum_{s=1}^{S} \boldsymbol{\Sigma}^{-1/2}_{\widehat{\boldsymbol{\beta}}_s \widehat{\boldsymbol{\beta}}_s} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s) \right]^{\mathsf{T}} \left[\sum_{s=1}^{S} \boldsymbol{\Sigma}^{-1/2}_{\widehat{\boldsymbol{\beta}}_s \widehat{\boldsymbol{\beta}}_s} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s) \right] \right) \tag{12}$$

- eMCD for $\boldsymbol{\beta}$

$$H^e(\boldsymbol{\beta}) = F_{\chi^2_{p_c}} \left(\frac{1}{S} \left[\sum_{s=1}^{S} \boldsymbol{\Phi}^{-1} (\boldsymbol{G}_s(\boldsymbol{g}(\boldsymbol{\beta}; \widehat{\boldsymbol{\beta}}_s))) \right]^{\mathsf{T}} \left[\sum_{s=1}^{S} \boldsymbol{\Phi}^{-1} (\boldsymbol{G}_s(\boldsymbol{g}(\boldsymbol{\beta}; \widehat{\boldsymbol{\beta}}_s))) \right] \right), \tag{13}$$

- Denote the MCD estimate of $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}_{\mathrm{MCD}}$

MCD Procedure for Heterogeneous Case

A two-stage MCD procedure

Stage 2: Efficiency boosting

- Estimate $\boldsymbol{\alpha}_s$ by

$$\breve{\boldsymbol{\alpha}}_s = \arg\min_{\boldsymbol{\alpha}_s \in \mathcal{A}_s} \sum_{i=1}^{n} \varrho(X_{s,i}; \boldsymbol{\alpha}_s, \widehat{\boldsymbol{\beta}}_{\mathrm{MCD}}), \quad s = 1, \ldots, S. \tag{14}$$

Theorem 2 (Under construction...)

Assume suitable regularity conditions. We hope to prove

$$\breve{\boldsymbol{\alpha}}_s - \boldsymbol{\alpha}_0 \xrightarrow{D} \mathcal{N}(\mathbf{0}_{p_h}, \boldsymbol{\Sigma}_{\breve{\boldsymbol{\alpha}}_s}), \tag{15}$$

where $\boldsymbol{\Sigma}_{\breve{\boldsymbol{\alpha}}_s}$ *is not larger than* $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\alpha}}_s} = \mathrm{Var}(\widehat{\boldsymbol{\alpha}}_s)$.

- Conceptually, $\widehat{\boldsymbol{\beta}}_{\mathrm{MCD}}$ is $\sqrt{N}$-consistent, which is way closer to $\boldsymbol{\beta}_0$ than $\widehat{\boldsymbol{\beta}}_s$ which are $\sqrt{n}$-consistent
- Plugging $\widehat{\boldsymbol{\beta}}_{\mathrm{MCD}}$ back into the model is like we know $\boldsymbol{\beta}_0$
- Estimating $\boldsymbol{\alpha}_s$ when knowing $\boldsymbol{\beta}_0$ should be more accurate

Statistical Inference via MCD Functions

- Methodological comparisons among the existing MCD-related approaches:

Features	Liu *et al.* (2015)	Yang *et al.* (2014)	Proposed
Idea	Normalized likelihood	Bootstrap distribution	p-value function
Subsample CD	Yes	Yes	No
Satisfy (R2)	No	No	Yes
Need point est. for CR?	Yes	Yes	No

- Based on the proposed D&C MCD approach...
  - Test hypotheses and construct confidence regions using the resulting MCD functions
    - Without having a combined point estmiate
    - Significant savings on computational cost will be shown in simulation
  - Point estimation can be done using MCd maximizer
    - Covariance estimates can be obtained via inverse of Hessian matrix
    - Two-stage MCD procedures can boost the efficiency of heterogeneity estimators

Simulation Study

- A fixed-effect simulation model:

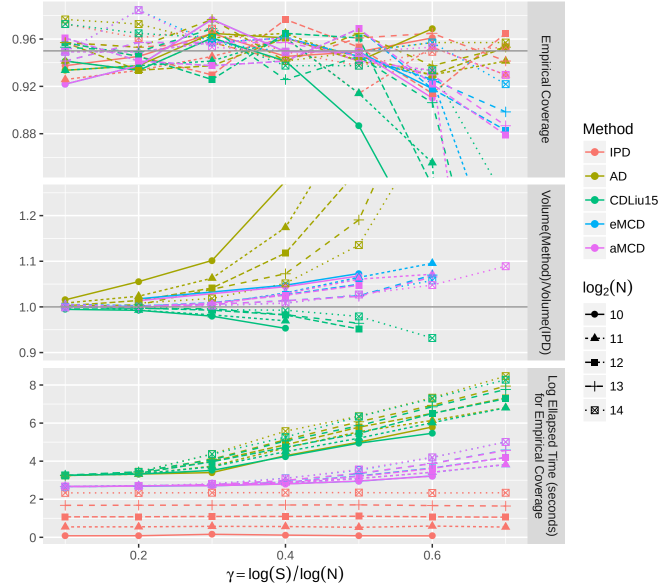
$$Y_{s,j} = \alpha_s + X_{s,i}^{\mathsf{T}} \boldsymbol{\beta}_0 + \epsilon_{s,i}, \quad s = 1, \ldots, S, i = 1, \ldots, n, \tag{16}$$

  - $\boldsymbol{\alpha}_s \equiv 1$ if homogeneous; $\boldsymbol{\alpha}_s \overset{iid}{\sim}$ *Uniform*$(-10, 2)$ if heterogeneous
  - $\boldsymbol{\beta}_0 = (0.5, -0.7, -1.1, 0.2, -0.5)^{\mathsf{T}}$
  - $\epsilon_{s,i} \overset{iid}{\sim} \mathcal{N}(0, 0.5)$
  - $X_{s,i}$ were generated from $\mathcal{N}((\mu_1, \ldots, \mu_5)^{\mathsf{T}}, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_5^2))$, where $\mu_k \overset{iid}{\sim} \mathcal{N}(-0.5, 1.5)$ and $\sigma_k \overset{iid}{\sim}$ *Gamma*$(1.2, 0.7)$
  - 256 simulation replicates
- Computation environment:
  - R 3.1.0 on the Radon computer cluster operated by ITaP Research Computing[3].
  - Up to 32 cores (4 full nodes, each with 8 cores) were used for parallel computing.

[3]See https://www.rcac.purdue.edu/compute/radon/ for more information.

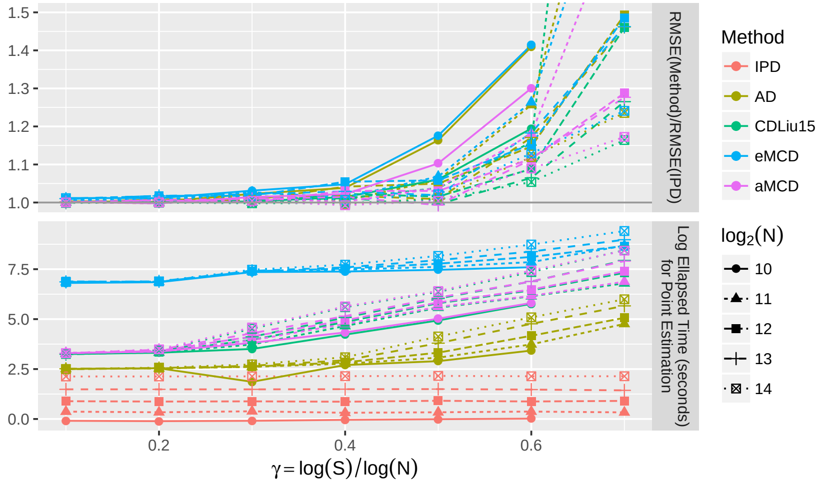
Simulation Study

Confidence region construction for homogeneous case



Simulation Study

Point estimation for homogeneous case



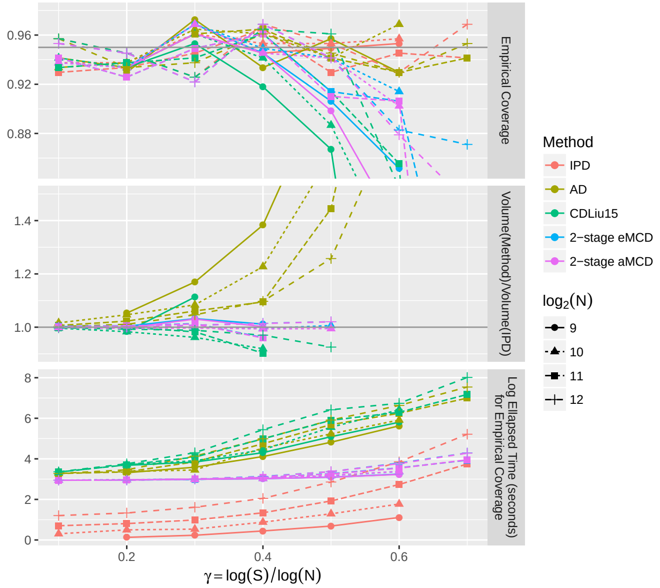
Simulation Study

Summary for homogeneous case

- 95% confidence region construction
  - Empirical Coverage
    - CDLiu15 starts getting deteriorated at smaller γ
    - MCD approaches' coverages are still acceptable at $\gamma = 0.6$
    - AD keeps decent coverage even at $\gamma = 0.7$ (traded with volume)
  - Volume (provided coverage in $[0.925, 0.975]$)
    - AD produces particular larger volumes as S increases
    - Other methods are similar
  - Computation time
    - MCD approaches are similar
    - CDLiu15 and AD are similar
    - Obviously, MCD approaches outspeed CDLiu15 and AD
- Point estimation:
  - RMSE
    - All methods have similar performance
  - Computation time
    - AD is the fastest D&C approach
    - aMCD and CDLiu15 are almost the same

Simulation Study

95% confidence region for heterogeneous case



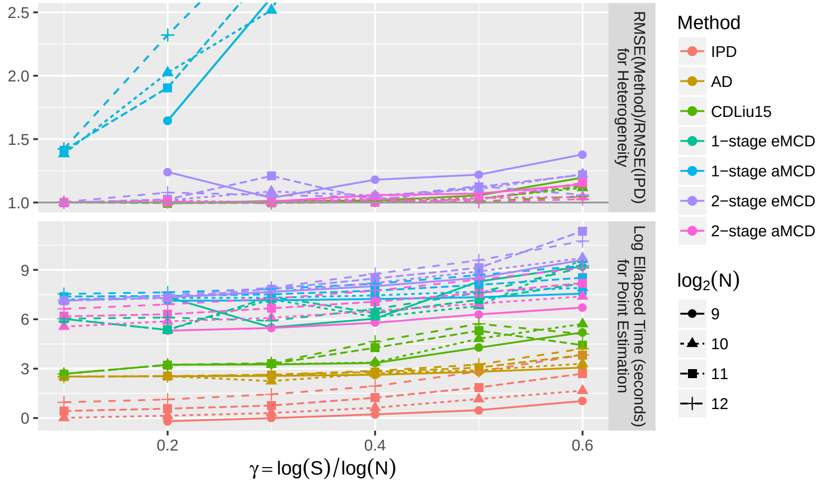
Simulation Study

Confidence region construction for heterogeneous case

- Empirical Coverage
  - CDLiu15 still gets deteriorated at smaller γ
  - MCD methods start becoming worse at $\gamma = 0.5$
  - AD still keeps great coverage all the time (still traded with volume)
- Volume (provided coverage in $[0.925, 0.975]$)
  - AD produces particular larger volumes
  - Other methods are similar
- Computation time
  - Obviously, MCD approaches still outspeed CDLiu15 and AD

Simulation Study

Point estimation for heterogeneous case



Simulation Study

Point estimation for heterogeneous case

- RMSE for commonality estimation
  - All methods are very similar (thus not shown)
- RMSE for heterogeneity estimation
  - One-stage MCD methods and AD are similar (upper, overlapping lines)
  - Two-stage MCD methods, CDLiu15 and IPD are similar (lower lines)
  - Empirical evidence of efficiency boosting for two-stage MCD approaches
- Computation time
  - MCD approaches still are the slowest ones...

Simulation Study

Summary for the proposed MCD approaches

- Outperform AD and CDLiu15 in confidence region construction
  - Significantly faster
  - Allow γ slightly larger than 0.5, while CDLiu15 starts getting worse around $\gamma = 0.4$
  - Produce comparable volumes of confidence regions
- Do not have obvious advantage on point estimation than AD and CDLiu15
  - Efficiency boosting is shown effective in two-stage MCD procedures
  - Still too time-consuming

Remark: Logistic and Poisson regression results are similar

A Real Data Illustration

The hourly Bike Sharing dataset[a]:

- Consisting of a two-year (2011 and 2012) record with $17{,}379$ instances and 17 features is used for real-data illustration
- See Fanaee-T and Gama (2014) for detailed descriptions
- Response: `casual` (count of casual users)
- Continuous explanatory variables:
  - `temp` (temperature)
  - `atemp` (feeling temperature)
  - `hum` (humidity)
  - `windspeed` (wind speed)

[a]Available online at https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

A Real Data Illustration

Evaluation criteria

- Train and test
  - First year ($8{,}645$ instances) is used for training
  - Second year ($8{,}734$ instances) is used for testing
- 95% confidence region construction for commonality (four explanatory variables) [train]
  - Volume
- Point estimation [train]
  - Parameter estimates with standard errors
- Prediction performance
  - Root mean square predicted error (RMSPE)
  - Relative absolute error (RAE)
  - Relative root square error (RRSE)
  - Correlation between predicted and actual counts

  (RAE, RRSE and correlation are used in Fanaee-T and Gama (2014))

A Real Data Illustration

Heterogeneous analysis (`weekday` as grouping variable)

	IPD	AD	CDLiu15	eMCD	aMCD
10^5 Volume of 95% CR	7.33	11.28	7.52	N/A	2.58
Estimtes	*Mean±SE*				
`weekday0`	3.05±0.0124	3.08±0.0233	3.05±0.0125	2.95±0.0640	3.01±0.0640
`weekday1`	2.35±0.0070	2.12±0.0372	2.37±0.0131	2.25±0.0750	2.31±0.0750
`weekday2`	2.06±0.0077	1.92±0.0407	2.06±0.0135	1.96±0.0804	2.02±0.0804
`weekday3`	1.95±0.0081	1.41±0.0465	1.95±0.0139	1.85±0.0834	1.91±0.0834
`weekday4`	1.97±0.0079	1.77±0.0352	1.98±0.0135	1.88±0.0817	1.93±0.0817
`weekday5`	2.28±0.0071	2.61±0.0337	2.29±0.0127	2.18±0.0763	2.24±0.0763
`weekday6`	3.01±0.0058	3.04±0.0244	3.01±0.0121	2.91±0.0640	2.97±0.0640
`temp`	0.25±0.0745	0.30±0.0830	0.22±0.0757	0.30±N/A	0.20±0.0294
`atemp`	3.42±0.0841	3.43±0.0934	3.44±0.0855	3.43±N/A	3.49±0.0332
`hum`	-1.83±0.0116	-1.77±0.0127	-1.83±0.0117	-1.77±N/A	-1.80±0.0045
`windspeed`	0.24±0.0168	0.33±0.0189	0.25±0.0168	0.33±N/A	0.27±0.0065
RMSPE	63.46	45.01	44.83	44.97	44.91
RAE	0.96	0.69	0.67	0.67	0.67
RRSE	1.11	0.79	0.79	0.79	0.79
Correlation	0.35	0.63	0.65	0.65	0.65

A Real Data Illustration

- aMCD outperforms others in terms of estimation variation
- Point estimates from CD approaches are close to IPD
- All D&C approaches are similar in terms of prediction
  - Though, why IPD is bad?
- N/A in eMCD due to computation issue...
  - Non-existence of volume coincides with non-existence of covariance estimates, though it does not need point estimates for confidence regions
  - N/A covariance estimates result from the Hessian matrix is not positive definite

Conclusion and Future Work

- The proposed D&C MCD method is a desent alternative when IPD is not available
  - It is a genuine CD approach rooted from IPD p-value function recovery
  - The number of data partitions S can diverge with total sample size N
  - Achieves oracle properties if the growth rate of S is controlled
  - Produces confidence regions/performs hypothesis testing without having a combined point estimate/test statistic, thus fast
  - Allow larger S than CDLiu15
- The real data example shows our aMCD
  - is as good as CDLiu15 in terms of point estimation and prediction
  - has tremendous power on variation estimation
- The only obvious drawback is that it is time-consuming on point estimation

Conclusion and Future Work

Future Work:

- Extend to Big Time Series Data
  - Consider (auto)correlation among subsample estimates
  - Subsampling

References I

Fan, J., Han, F. and Liu, H. (2014) Challenges of big data analysis. *National Science Review*, **1**, 293–314. URL http://nsr.oxfordjournals.org/content/1/2/293.abstract.

Fanaee-T, H. and Gama, J. (2014) Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, **2**, 113–127. URL http://dx.doi.org/10.1007/s13748-013-0040-3.

He, X. and Shao, Q.-M. (1996) A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, **24**, 2608–2630. URL http://dx.doi.org/10.1214/aos/1032181172.

Lindskog, F., McNeil, A. and Schmock, U. (2003) Kendall's tau for elliptical distributions. In *Credit Risk* (eds. G. Bol, G. Nakhaeizadeh, S. T. Rachev, T. Ridder and K.-H. Vollmer), Contributions to Economics, 149–156. Physica-Verlag HD. URL http://dx.doi.org/10.1007/978-3-642-59365-9_8.

Liu, D., Liu, R. Y. and Xie, M. (2015) Multivariate meta-analysis of heterogeneous studies using only summary statistics: Efficiency and robustness. *Journal of the American Statistical Association*, **110**, 326–340. URL http://dx.doi.org/10.1080/01621459.2014.899235.

References II

Xie, M. and Singh, K. (2013) Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, **81**, 3–39. URL http://dx.doi.org/10.1111/insr.12000.

Xie, M., Singh, K. and Strawderman, W. E. (2011) Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association*, **106**, 320–333. URL http://dx.doi.org/10.1198/jasa.2011.tm09803.

Yang, G., Liu, D., Liu, R. Y., Xie, M. and Hoaglin, D. C. (2014) Efficient network meta-analysis: A donfidence distribution approach. *Statistical Methodology*, **20**, 105–125. URL http://www.sciencedirect.com/science/article/pii/S1572312714000136.