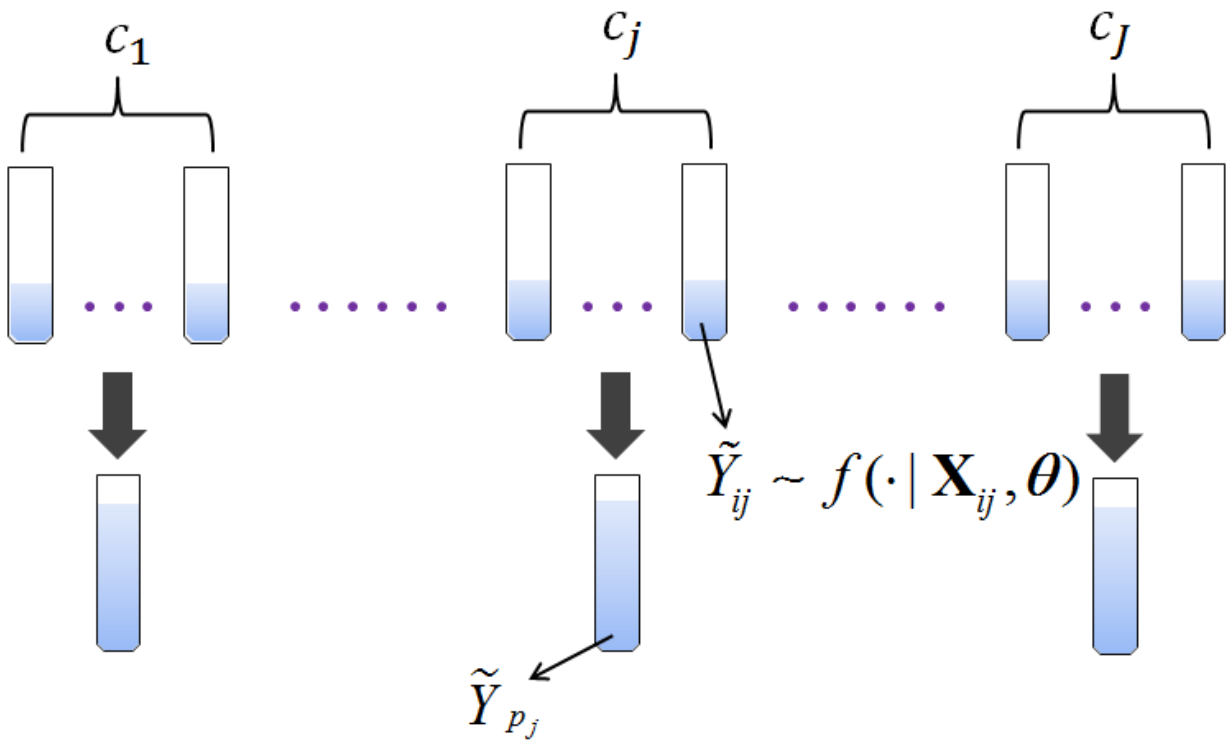
A general framework for the regression analysis of pooled biomarker assessments

Colin M. Gallagher

joint work with Yan Liu and Christopher S. McMahan
Clemson University

Settings and Notation



- $\tilde{Y}_{p_j}$: true biomarker level for the jth pooled specimen which is assumed to be $\tilde{Y}_{p_j} = \sum_{i=1}^{c_j} \tilde{Y}_{ij}/c_j$
- Y_{p_j} : observed biomarker level for the jth pooled specimen which follows a conditional distribution with pdf $f_\varepsilon(\cdot \mid \tilde{Y}_{p_j})$
- $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$

What is observed? What is latent?

- Assume $\widetilde{Y}_{p_j} = \sum_{i=1}^{c_j} \widetilde{Y}_{ij}/c_j$ (equal pooling amounts)
- We assume a parametric model for the *latent* individual biomarker distribution.
- This uniquely determines the distribution of the average.
- If we assumed a distribution for the average, it would not determine the distribution of the summands.
- Y_{p_j} : observed biomarker level for the jth pooled specimen. Allowing for measurement error we assume $Y_{p_j} \sim f_\varepsilon(\cdot|\widetilde{Y}_{p_j})$

Maximum Likelihood Estimate

Let $g_j(Y_{p_j}|\mathbf{x}_j, \boldsymbol{\theta})$ denote the pdf of Y_{p_j} which is given by

$$g_j(Y_{p_j}|\mathbf{x}_j, \boldsymbol{\theta}) = \int \cdots \int f_\varepsilon(Y_{p_j}|\tilde{y}_{p_j}) h_j(\tilde{\mathbf{y}}_j|\mathbf{x}_j, \boldsymbol{\theta}) d\tilde{\mathbf{y}}_j,$$

where $h_j(\tilde{\mathbf{y}}_j|\mathbf{x}_j, \boldsymbol{\theta}) = \prod_{i=1}^{c_j} f(\tilde{y}_{ij}|\mathbf{x}_{ij}, \boldsymbol{\theta})$ is the joint pdf of $\tilde{\mathbf{Y}}_j$

The maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ can be obtained by

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{j=1}^{J} \log\left\{g_j(Y_{p_j}|\mathbf{x}_j, \boldsymbol{\theta})\right\}$$

Approximation of log likelihood function

Consider the distribution of Y_{p_j}, which is given by

$$\begin{aligned} g_j(Y_{p_j}|\mathbf{x}_j,\boldsymbol{\theta}) &= E\left[f_\varepsilon(Y_{p_j}|\widetilde{y}_{p_j})\right] \\ &= \int\cdots\int f_\varepsilon(Y_{p_j}|\widetilde{y}_{p_j})h_j(\widetilde{\boldsymbol{y}}_j|\mathbf{x}_j,\boldsymbol{\theta})d\widetilde{\boldsymbol{y}}_j \end{aligned}$$

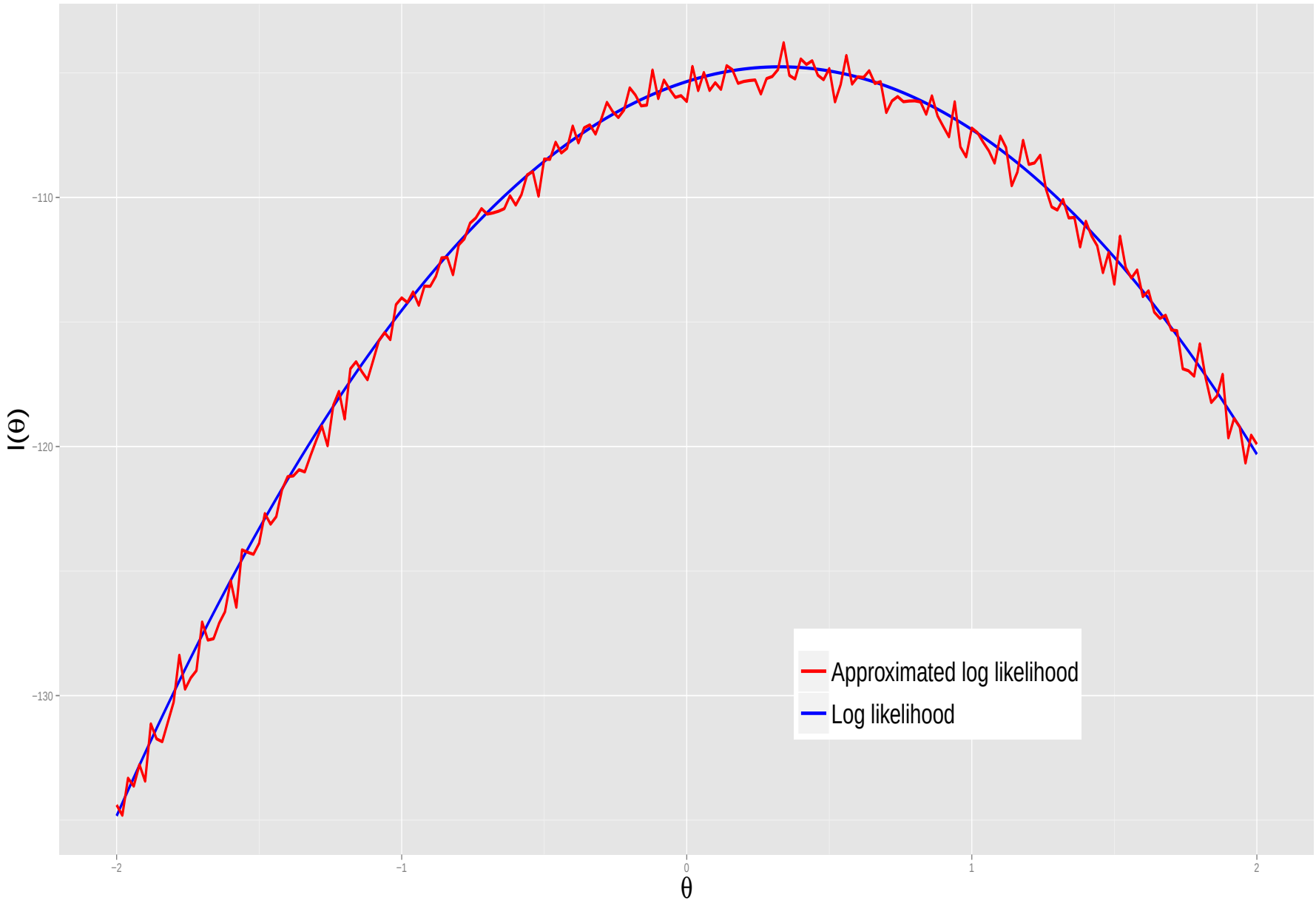
Let $(\widetilde{\boldsymbol{Y}}_j^1, \widetilde{\boldsymbol{Y}}_j^2, \ldots, \widetilde{\boldsymbol{Y}}_j^M)$ be a random sample from the distribution function $h_j(\cdot|\mathbf{x}_j,\boldsymbol{\theta})$, then

$$g_j(Y_{p_j}|\mathbf{x}_j,\boldsymbol{\theta}) \approx g_j^M(Y_{p_j}|\mathbf{x}_j,\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M} f_\varepsilon(Y_{p_j}|\widetilde{Y}_{p_j}^m),$$

where $\widetilde{Y}_{p_j}^m = \sum_{i=1}^{c_j}\widetilde{Y}_{ij}^m/c_j$. Then the MCMLE of $\boldsymbol{\theta}$ can be obtained by

$$\widehat{\boldsymbol{\theta}}_M = \arg\max_{\boldsymbol{\theta}} \sum_{j=1}^{J} \log\left\{g_j^M(Y_{p_j}|\mathbf{x}_j,\boldsymbol{\theta})\right\}$$

Oh no!


−110
−120
−130
I(Θ)
−2
−1
0
1
2
θ
Approximated log likelihood
Log likelihood

Smooth Version

- Problem: the above approximate likelihood results from a different random sample for each theta and is thus non-smooth.
- We need a smooth function in order to numerically optimize
- The solution is from Robert and Casella (2004): use importance sampling
- We select an importance distribution h^* with similar support and shape as h
- Our error in approximating the unobservable likelihood is a function of M and h^*; In the paper we discuss selection of M based on a given h^* and provide guidance in selecting h^*.

Monte Carlo approximation using an importance distribution

Follow the method of Robert and Casella (2004), consider

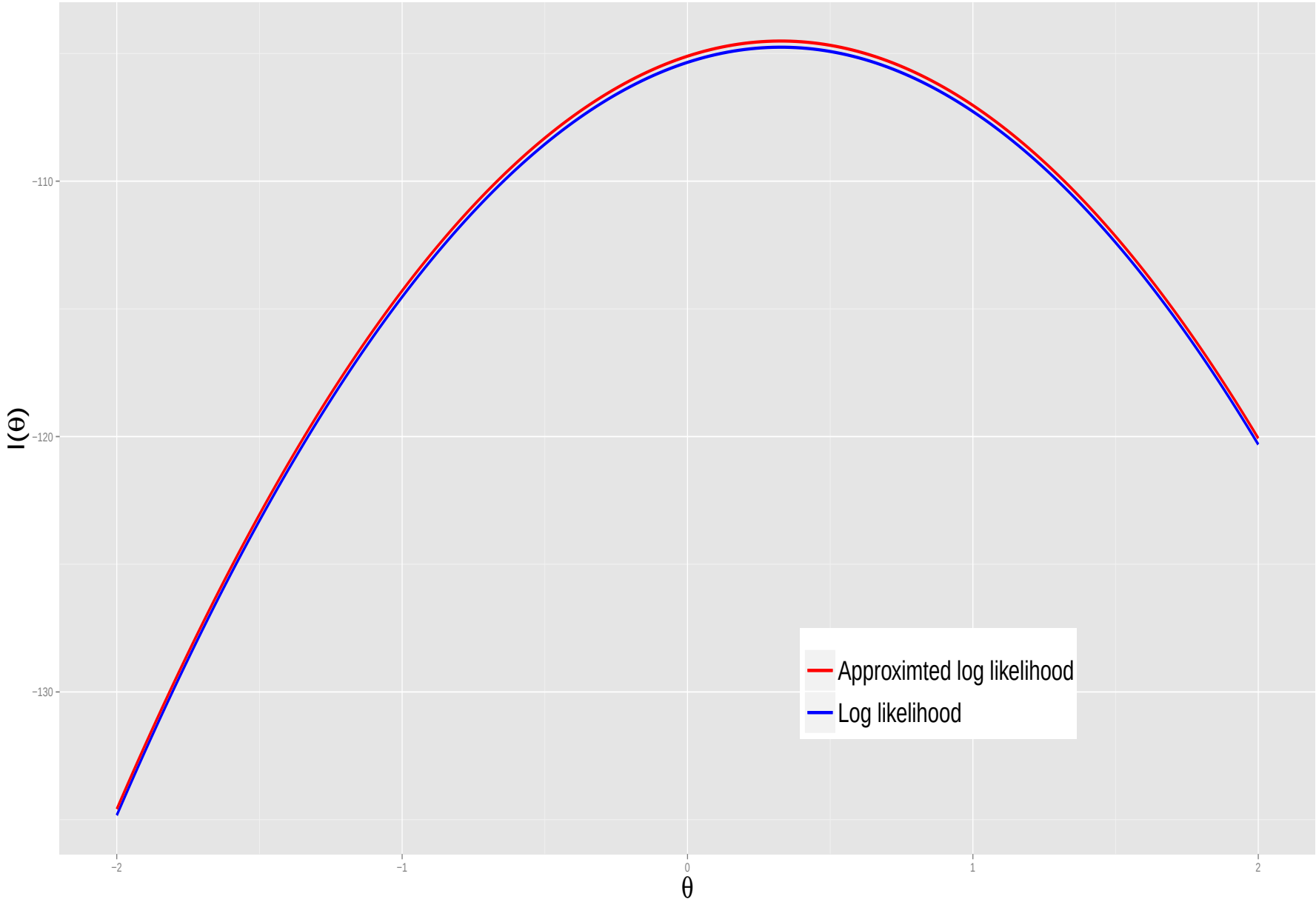
$$g_j(Y_{p_j}|\mathbf{x}_j,\boldsymbol{\theta}) = \int \cdots \int f_\varepsilon(Y_{p_j}|\tilde{y}_{p_j}) \frac{h_j(\tilde{\mathbf{y}}_j|\mathbf{x}_j,\boldsymbol{\theta})}{h_j^*(\tilde{\mathbf{y}}_j|\mathbf{x}_j,\boldsymbol{\theta}^*)} h_j^*(\tilde{\mathbf{y}}_j|\mathbf{x}_j,\boldsymbol{\theta}^*) d\tilde{\mathbf{y}}_j \tag{1}$$

Let $\tilde{\boldsymbol{Y}}_j^1, \ldots, \tilde{\boldsymbol{Y}}_j^M$ be a random sample from $h_j^*(\cdot|\mathbf{x}_j,\boldsymbol{\theta}^*)$

$$g_j^M(Y_{p_j}|\mathbf{x}_j,\boldsymbol{\theta},\boldsymbol{\theta}^*) = \frac{1}{M}\sum_{m=1}^{M} f_\varepsilon(Y_{p_j}|\tilde{Y}_{p_j}^m) \frac{h_j(\tilde{\boldsymbol{Y}}_j^m|\mathbf{x}_j,\boldsymbol{\theta})}{h_j^*(\tilde{\boldsymbol{Y}}_j^m|\mathbf{x}_j,\boldsymbol{\theta}^*)}$$

Then the MLE of $\boldsymbol{\theta}$ can be obtained by

$$\widehat{\boldsymbol{\theta}}_M = \arg\max_{\boldsymbol{\theta}} \sum_{j=1}^{J} \log\left\{ g_j^M(Y_{p_j}|\mathbf{x}_j,\boldsymbol{\theta},\boldsymbol{\theta}^*) \right\}$$


−110
−120
−130
l(θ)
−2
−1
0
1
2
θ
Approximted log likelihood
Log likelihood

Monte Carlo Maximum Likelihood Estimates (MCMLEs)

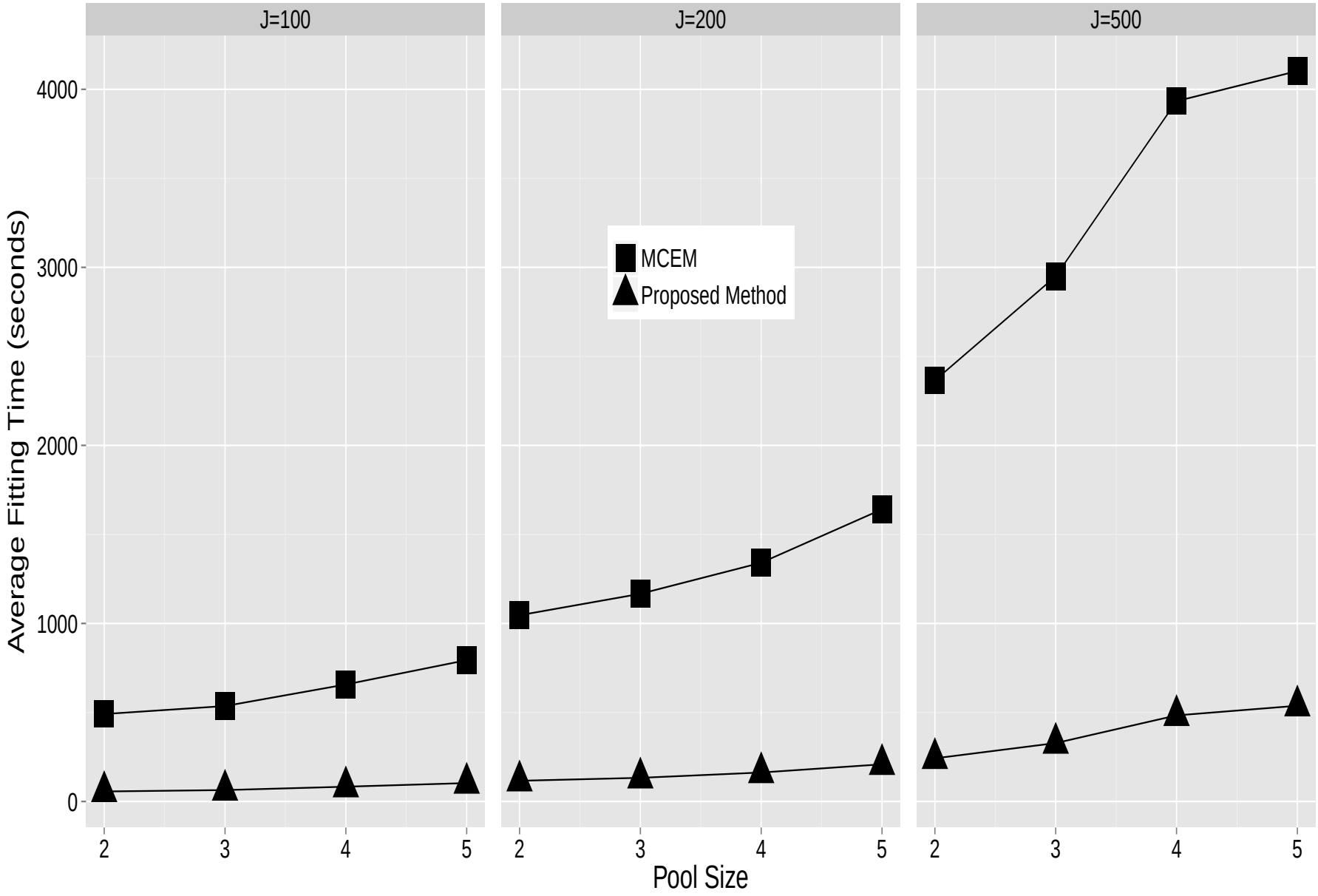
Consistency and asymptotic normality of $\widehat{\theta}_M$ have been established yeilding:

- An approach that allows us to specify the Monte Carlo sample size M
- A method for estimating standard error of mcmle

simulation results for lognormal

Competing method: Monte Carlo Expectation Maximization (MCEM; Mitchell et.al, 2014)

Model		Measure	$c=1$	$c=2$	$c=3$	$c=4$	$c=5$
M2(MCMLE)	β_0	Bias(SD)	0.00(0.07)	0.00(0.05)	0.00(0.04)	0.00(0.04)	0.00(0.04)
		Cov(SE)	0.94(0.07)	0.95(0.05)	0.95(0.04)	0.94(0.04)	0.94(0.03)
	β_1	Bias(SD)	0.00(0.05)	0.00(0.04)	0.00(0.03)	0.00(0.03)	0.00(0.02)
		Cov(SE)	0.95(0.05)	0.94(0.04)	0.95(0.03)	0.94(0.03)	0.94(0.02)
	β_2	Bias(SD)	0.01(0.10)	0.00(0.08)	0.00(0.06)	0.00(0.05)	0.00(0.05)
		Cov(SE)	0.95(0.10)	0.93(0.07)	0.96(0.06)	0.97(0.05)	0.96(0.05)
M2(MCEM)	β_0	Bias(SD)	0.00(0.07)	0.00(0.05)	0.00(0.04)	0.00(0.04)	0.00(0.04)
		Cov(SE)	0.94(0.07)	0.95(0.05)	0.95(0.04)	0.93(0.04)	0.94(0.03)
	β_1	Bias(SD)	0.00(0.05)	0.00(0.04)	0.00(0.03)	0.00(0.03)	0.00(0.02)
		Cov(SE)	0.95(0.05)	0.94(0.04)	0.94(0.03)	0.94(0.03)	0.94(0.02)
	β_2	Bias(SD)	0.01(0.10)	0.00(0.08)	0.00(0.06)	0.00(0.05)	0.00(0.05)
		Cov(SE)	0.95(0.10)	0.92(0.07)	0.96(0.06)	0.97(0.05)	0.96(0.05)


J=100
J=200
J=500
Average Fitting Time (seconds)
0
1000
2000
3000
4000
MCEM
Proposed Method
2
3
4
5
Pool Size

Features of data

- Data was collected by the Collaborative Perinatal Project (CPP)
  - Monocyte chemotactic protein 1 (MCP1) was measured from both individual and pooled specimens
  - All samples were assayed in duplicate
- Individual dataset
  - 752 individual measurements
  - Estimated standard deviation of measurement error is 0.044
- Pooled dataset
  - 81 individual and 350 pooled (pool size 2) measurements
  - Estimated standard deviation of measurement error is 0.05

Data Analysis

- Covariates:
  - age (standardized; denoted as x_1)
  - race (1=Africa American/0=others; denoted as x_2)
  - spontaneous abortion (SA) status (1=yes/0=no; denoted as x_3)
- Competing method: Monte Carlo Expectation Maximization (MCEM; Mitchell et.al, 2014)
- Full model

$$\begin{aligned} \log(\tilde{Y}_{ij}) &= \beta_0 + \sum_{i=1}^{3} \beta_i x_i + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1^2 + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim N(0, \sigma^2) \end{aligned}$$

Variable Selection

- Four configurations were considered
  - Measurements are from individual or pooled specimens
  - Measurement error exists or not
- Best subset selection were conducted based on AIC
  - 128 models were fitted under each configuration
  - Selected model: Age, SA, Race$*$SA
- Time to finish variable selection:
  - Our method : **7 hours**
  - MCEM : **74 hours**

Distribution	Dataset		Age	SA	Race∗SA
Log-normal	Individual	Est(P-value)	0.111(0.005)	0.271(0.002)	-0.508(0.000)
	Individual*	Est(P-value)	0.117(0.002)	0.220(0.007)	-0.409(0.001)
	Pool	Est(P-value)	0.105(0.112)	0.314(0.007)	-0.822(0.000)
	Pool*	Est(P-value)	0.155(0.018)	0.262(0.023)	-0.681(0.006)

Distribution	Dataset		Age	SA	Race∗SA
Log-normal	Individual	Est(P-value)	0.111(0.005)	0.271(0.002)	-0.508(0.000)
	Individual*	Est(P-value)	0.117(0.002)	0.220(0.007)	-0.409(0.001)
	Pool	Est(P-value)	0.105(0.112)	0.314(0.007)	-0.822(0.000)
	Pool*	Est(P-value)	0.155(0.018)	0.262(0.023)	-0.681(0.006)

Distribution	Dataset		Age	SA	Race∗SA
Log-normal	Individual	Est(P-value)	0.111(0.005)	0.271(0.002)	-0.508(0.000)
	Individual*	Est(P-value)	0.117(0.002)	0.220(0.007)	-0.409(0.001)
	Pool	Est(P-value)	0.105(0.112)	0.314(0.007)	-0.822(0.000)
	Pool*	Est(P-value)	0.155(0.018)	0.262(0.023)	-0.681(0.006)

Distribution	Dataset		Age	SA	Race∗SA
Log-normal	Individual	Est(P-value)	0.111(0.005)	0.271(0.002)	-0.508(0.000)
	Individual*	Est(P-value)	0.117(0.002)	0.220(0.007)	-0.409(0.001)
	Pool	Est(P-value)	0.105(0.112)	0.314(0.007)	-0.822(0.000)
	Pool*	Est(P-value)	0.155(0.018)	0.262(0.023)	-0.681(0.006)
	HPD*	Est(P-value)	0.110(0.007)	0.220(0.014)	-0.360(0.005)

Distribution	Dataset		Age	SA	Race∗SA
Log-normal	Individual	Est(P-value)	0.111(0.005)	0.271(0.002)	-0.508(0.000)
	Individual*	Est(P-value)	0.117(0.002)	0.220(0.007)	-0.409(0.001)
	Pool	Est(P-value)	0.105(0.112)	0.314(0.007)	-0.822(0.000)
	Pool*	Est(P-value)	0.155(0.018)	0.262(0.023)	-0.681(0.006)
	HPD*	Est(P-value)	0.110(0.007)	0.220(0.014)	-0.360(0.005)

Conclusion

The characteristics of the proposed method:

- Accurate and computationally efficient
- Accounts for measurement error
- Broadly applicable to analysis pooled data under many common regression models