Estimating Mean and Covariance Structure with Reweighted Least Squares

Bang Quan Zheng

August 31, 2019

UCLA

Research Question

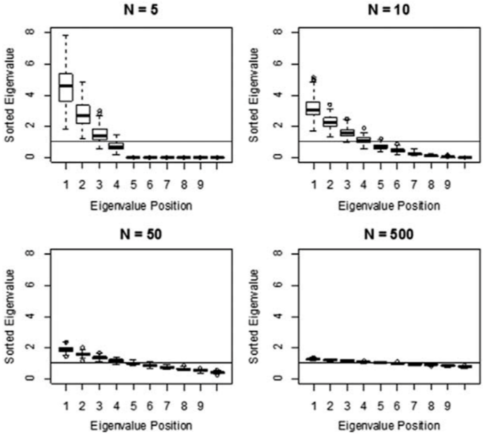
Does Reweighted Least Squares method (RLS) perform better in small samples than maximum likelihood (ML) method for mean and covariance structure?

UCLA

Introduction

- Normal-theory methods (e.g. ML) in SEM rely on the assumption of asymptotic behavior of statistics $N \to \infty$
- When N is large enough, the model follows a χ^2 distribution
- Real-world applications of SEM often have small or modest sample sizes
- When $P > N$, the sample covariance matrix is not invertible
- When the ratio of $\frac{p}{N}$ is slightly less than 1, even though the sample covariance structure is invertible, but it's numerically ill conditioned.

10-dimensional Multivariate-normal Simulation Illustration



Existing Methods in the Literature

Reweighted Least Squares (Browne, 1985)

$$RLS^* = tr[(\mathbf{S} - \Sigma(\theta))\hat{\Sigma}_{ML}^{-1}]^2$$

Regularized GLS (Arruda and Bentler, 2017)

$$RGLS^* = tr[(\mathbf{S} - \Sigma(\theta))\hat{\Sigma}_{REG}^{-1}]^2$$

A puzzle here: Does Reweighted Least Squares also work in mean and covariance structure?

UCLA

A Covariance & Mean Structure Model

$$\boldsymbol{T} = \underbrace{F}_{\text{Covariance}} + \underbrace{(\bar{X} - \mu)' \hat{\Sigma}^{-1} (\bar{X} - \mu)}_{\text{Mean Structure}}$$

$\boldsymbol{T}$ function reflects how closely the sample covariance matrix $\mathbf{S}$ is reproduced by the estimated model covariance matrix $\hat{\Sigma}$, as well as how closely the sample mean vector $\bar{X}$ is reproduced by the estimated model mean vector μ. Therefoe, a model may fit badly if the means are modeled poorly, or if the covariances are modeled poorly, or both.

UCLA

How to derive structured means?

$$X = \mu_x + \Lambda\xi + \epsilon$$

Taking expectation $E(\xi) = \mu_\xi$, $E(\epsilon) = 0$, and assume $\mu_x = 0$, then

$$\bar{X} = \Lambda\mu_\xi$$

We can re-construct the population means as

$$\hat{\mu} = \Lambda\mu_\xi$$

UCLA

We can simplify the notation in this way

$$\boldsymbol{T} = \underbrace{F}_{\text{Covariance}} + \underbrace{(\bar{X} - \mu)'\hat{\Sigma}^{-1}(\bar{X} - \mu)}_{\text{Mean Structure}}$$

$$\downarrow$$

$$\boldsymbol{T} = F + (\bar{X} - \Lambda\mu_{\xi})'\hat{\Sigma}^{-1}(\bar{X} - \Lambda\mu_{\xi})$$

UCLA

Reweighted Least Squares

$$F_{GLS} = \frac{1}{2}tr[\{(S - \Sigma(\theta))V^{-1}\}]^2$$

- V^{-1} is a biased weight matrix
- Replace V^{-1} with $\hat{\Sigma}_{ML}^{-1}$ that derives from ML

UCLA

Why is T_{RLS} unbiased?

$$T_{ML} = T_{RLS} + n \sum_{k=3}^{\infty} \frac{1}{k} tr \left\{ I_p - S\hat{\Sigma}^{-1} \right\}^k$$

- When n > > p, the second term will become 0
- When $p > n$, the second term will be positive
- $\rightarrow T_{ML}$ will be too large, and T_{RLS} will be about right

UCLA

How do we derive $\hat{\Sigma}_{ML}$?

$$F_{ML} = log|\Sigma(\theta)| - log|S_N| + tr(S_N \Sigma(\theta)^{-1}) - p$$

$$\hat{\theta}_{ML} = argmin\, F_{ML}(\theta)$$

Therefore,

$$\Sigma(\hat{\theta}_{ML}) = \hat{\Lambda}\hat{\Phi}\hat{\Lambda}' + \hat{\Psi}$$

$$\hat{\Sigma}_{ML} = \Sigma(\hat{\theta}_{ML})$$

UCLA

How do we derive RLS?

Once we derive $\hat{\Sigma}_{ML}$, we can fit it into the T_{RLS} function, and obtain the test-statistics.

$$F_{GLS} = \frac{1}{2} tr[\{(S - \Sigma(\theta))V^{-1}\}]^2$$

$$\downarrow$$

$$T_{RLS} = \frac{1}{2} tr[(S - \hat{\Sigma}(\theta))\hat{\Sigma}_{ML}^{-1}]^2$$

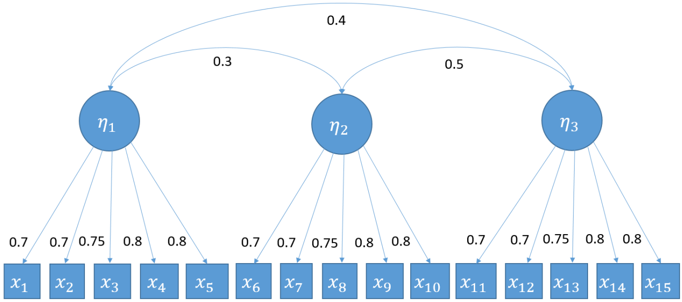
UCLA

RLS for Mean and Covariance Structure

$$\boldsymbol{T} = T_{RLS} + (\bar{X} - \Lambda\mu_\xi)' \hat{\Sigma}_{ML}^{-1} (\bar{X} - \Lambda\mu_\xi)$$

UCLA

Population Model Path-Diagram



UCLA

For Covariance Structure

- 5 indicators per factor
- 3 latent factors
- 120 data points
- 33 free parameters
- 87 df

For Covariance and Mean Structure

- 5 indicators per factor
- 3 latent factors
- 135 data points
- 36 free parameters
- 99 df

Test Statistics

Table 1. Test Statistics and Standard Deviation by Sample Size

	Test Statistics					Standard Deviation				
N	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS
50	102.32	76.71	87.65	115.92	97.13	15.53	10.51	12.04	16.58	12.43
80	96.39	80.70	87.22	108.47	97.34	14.71	11.78	13.36	15.31	12.38
100	93.66	82.21	87.10	106.12	97.16	14.77	12.32	12.55	15.18	13.67
200	89.86	83.86	87.42	103.07	98.97	13.99	12.97	13.71	15.04	13.81
300	89.55	85.50	87.40	101.56	98.71	14.06	12.65	12.81	14.66	14.68
400	88.59	85.89	87.31	100.41	98.69	13.10	12.62	13.16	13.79	13.93
500	88.54	86.27	87.23	100.42	99.19	13.76	13.04	12.79	14.08	14.27
800	87.45	86.20	87.14	100.53	98.43	13.57	12.68	13.52	14.18	13.72
1,000	88.06	86.21	86.99	99.83	99.68	13.51	12.87	13.06	13.94	14.03
2,000	87.40	87.07	87.13	98.99	99.34	12.86	12.96	12.87	13.72	13.82
5,000	87.38	87.12	87.07	98.94	99.04	13.44	12.94	12.96	14.28	14.46
10,000	86.58	87.01	86.54	98.75	98.56	12.95	12.71	13.00	14.61	13.93

Note: GLS contains 37 non-convergence when N=50. 1 non-convergence when N=80

Test Statistics

Table 1. Test Statistics and Standard Deviation by Sample Size

	Test Statistics					Standard Deviation				
N	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS
50	102.32	76.71	87.65	115.92	97.13	15.53	10.51	12.04	16.58	12.43
80	96.39	80.70	87.22	108.47	97.34	14.71	11.78	13.36	15.31	12.38
100	93.66	82.21	87.10	106.12	97.16	14.77	12.32	12.55	15.18	13.67
200	89.86	83.86	87.42	103.07	98.97	13.99	12.97	13.71	15.04	13.81
300	89.55	85.50	87.40	101.56	98.71	14.06	12.65	12.81	14.66	14.68
400	88.59	85.89	87.31	100.41	98.69	13.10	12.62	13.16	13.79	13.93
500	88.54	86.27	87.23	100.42	99.19	13.76	13.04	12.79	14.08	14.27
800	87.45	86.20	87.14	100.53	98.43	13.57	12.68	13.52	14.18	13.72
1,000	88.06	86.21	86.99	99.83	99.68	13.51	12.87	13.06	13.94	14.03
2,000	87.40	87.07	87.13	98.99	99.34	12.86	12.96	12.87	13.72	13.82
5,000	87.38	87.12	87.07	98.94	99.04	13.44	12.94	12.96	14.28	14.46
10,000	86.58	87.01	86.54	98.75	98.56	12.95	12.71	13.00	14.61	13.93

Note: GLS contains 37 non-convergence when N=50. 1 non-convergence when N=80

UCLA

Test Statistics

Table 1. Test Statistics and Standard Deviation by Sample Size

	Test Statistics					Standard Deviation				
N	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS
50	102.32	76.71	87.65	115.92	97.13	15.53	10.51	12.04	16.58	12.43
80	96.39	80.70	87.22	108.47	97.34	14.71	11.78	13.36	15.31	12.38
100	93.66	82.21	87.10	106.12	97.16	14.77	12.32	12.55	15.18	13.67
200	89.86	83.86	87.42	103.07	98.97	13.99	12.97	13.71	15.04	13.81
300	89.55	85.50	87.40	101.56	98.71	14.06	12.65	12.81	14.66	14.68
400	88.59	85.89	87.31	100.41	98.69	13.10	12.62	13.16	13.79	13.93
500	88.54	86.27	87.23	100.42	99.19	13.76	13.04	12.79	14.08	14.27
800	87.45	86.20	87.14	100.53	98.43	13.57	12.68	13.52	14.18	13.72
1,000	88.06	86.21	86.99	99.83	99.68	13.51	12.87	13.06	13.94	14.03
2,000	87.40	87.07	87.13	98.99	99.34	12.86	12.96	12.87	13.72	13.82
5,000	87.38	87.12	87.07	98.94	99.04	13.44	12.94	12.96	14.28	14.46
10,000	86.58	87.01	86.54	98.75	98.56	12.95	12.71	13.00	14.61	13.93

Note: GLS contains 37 non-convergence when N=50. 1 non-convergence when N=80

UCLA

Test Statistics

Table 1. Test Statistics and Standard Deviation by Sample Size

	Test Statistics					Standard Deviation				
N	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS
50	102.32	76.71	87.65	115.92	97.13	15.53	10.51	12.04	16.58	12.43
80	96.39	80.70	87.22	108.47	97.34	14.71	11.78	13.36	15.31	12.38
100	93.66	82.21	87.10	106.12	97.16	14.77	12.32	12.55	15.18	13.67
200	89.86	83.86	87.42	103.07	98.97	13.99	12.97	13.71	15.04	13.81
300	89.55	85.50	87.40	101.56	98.71	14.06	12.65	12.81	14.66	14.68
400	88.59	85.89	87.31	100.41	98.69	13.10	12.62	13.16	13.79	13.93
500	88.54	86.27	87.23	100.42	99.19	13.76	13.04	12.79	14.08	14.27
800	87.45	86.20	87.14	100.53	98.43	13.57	12.68	13.52	14.18	13.72
1,000	88.06	86.21	86.99	99.83	99.68	13.51	12.87	13.06	13.94	14.03
2,000	87.40	87.07	87.13	98.99	99.34	12.86	12.96	12.87	13.72	13.82
5,000	87.38	87.12	87.07	98.94	99.04	13.44	12.94	12.96	14.28	14.46
10,000	86.58	87.01	86.54	98.75	98.56	12.95	12.71	13.00	14.61	13.93

Note: GLS contains 37 non-convergence when N=50. 1 non-convergence when N=80

Test Statistics

Table 1. Test Statistics and Standard Deviation by Sample Size

	Test Statistics					Standard Deviation				
N	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS
50	102.32	76.71	87.65	115.92	97.13	15.53	10.51	12.04	16.58	12.43
80	96.39	80.70	87.22	108.47	97.34	14.71	11.78	13.36	15.31	12.38
100	93.66	82.21	87.10	106.12	97.16	14.77	12.32	12.55	15.18	13.67
200	89.86	83.86	87.42	103.07	98.97	13.99	12.97	13.71	15.04	13.81
300	89.55	85.50	87.40	101.56	98.71	14.06	12.65	12.81	14.66	14.68
400	88.59	85.89	87.31	100.41	98.69	13.10	12.62	13.16	13.79	13.93
500	88.54	86.27	87.23	100.42	99.19	13.76	13.04	12.79	14.08	14.27
800	87.45	86.20	87.14	100.53	98.43	13.57	12.68	13.52	14.18	13.72
1,000	88.06	86.21	86.99	99.83	99.68	13.51	12.87	13.06	13.94	14.03
2,000	87.40	87.07	87.13	98.99	99.34	12.86	12.96	12.87	13.72	13.82
5,000	87.38	87.12	87.07	98.94	99.04	13.44	12.94	12.96	14.28	14.46
10,000	86.58	87.01	86.54	98.75	98.56	12.95	12.71	13.00	14.61	13.93

Note: GLS contains 37 non-convergence when N=50. 1 non-convergence when N=80

UCLA

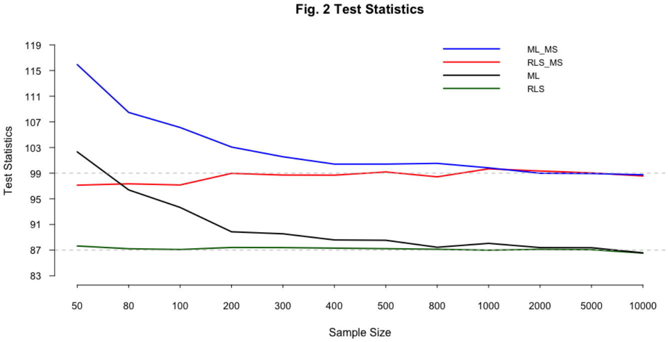
Test Statistics

Table 1. Test Statistics and Standard Deviation by Sample Size

	Test Statistics					Standard Deviation				
N	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS
50	102.32	76.71	87.65	115.92	97.13	15.53	10.51	12.04	16.58	12.43
80	96.39	80.70	87.22	108.47	97.34	14.71	11.78	13.36	15.31	12.38
100	93.66	82.21	87.10	106.12	97.16	14.77	12.32	12.55	15.18	13.67
200	89.86	83.86	87.42	103.07	98.97	13.99	12.97	13.71	15.04	13.81
300	89.55	85.50	87.40	101.56	98.71	14.06	12.65	12.81	14.66	14.68
400	88.59	85.89	87.31	100.41	98.69	13.10	12.62	13.16	13.79	13.93
500	88.54	86.27	87.23	100.42	99.19	13.76	13.04	12.79	14.08	14.27
800	87.45	86.20	87.14	100.53	98.43	13.57	12.68	13.52	14.18	13.72
1,000	88.06	86.21	86.99	99.83	99.68	13.51	12.87	13.06	13.94	14.03
2,000	87.40	87.07	87.13	98.99	99.34	12.86	12.96	12.87	13.72	13.82
5,000	87.38	87.12	87.07	98.94	99.04	13.44	12.94	12.96	14.28	14.46
10,000	86.58	87.01	86.54	98.75	98.56	12.95	12.71	13.00	14.61	13.93

Note: GLS contains 37 non-convergence when N=50. 1 non-convergence when N=80

Test Statistics



UCLA

Empirical Rejection Frequency

Table 2. Simulation Concerning Emprical Rejection Rates

	Average P-values					Number of P< 0.05					Empirical Rejection Frequency				
N	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS
50	0.226	0.726	0.485	0.218	0.537	310	0	50	309	27	0.310	0.000	0.050	0.309	0.027
80	0.316	0.638	0.499	0.327	0.531	188	9	53	165	25	0.188	0.009	0.053	0.165	0.025
100	0.370	0.607	0.497	0.367	0.536	140	22	43	129	37	0.140	0.022	0.043	0.129	0.037
200	0.441	0.570	0.495	0.420	0.501	82	34	60	90	50	0.083	0.034	0.060	0.090	0.050
300	0.448	0.531	0.493	0.450	0.504	83	36	54	77	45	0.071	0.036	0.054	0.077	0.045
400	0.465	0.526	0.495	0.471	0.506	62	40	56	53	48	0.062	0.040	0.056	0.053	0.048
500	0.470	0.515	0.492	0.471	0.498	75	44	44	61	64	0.075	0.044	0.044	0.061	0.064
800	0.494	0.515	0.496	0.471	0.512	63	39	50	66	44	0.063	0.039	0.050	0.066	0.044
1,000	0.479	0.517	0.499	0.481	0.487	61	42	48	50	54	0.061	0.042	0.048	0.050	0.054
2,000	0.494	0.495	0.494	0.497	0.494	56	46	44	47	48	0.056	0.046	0.044	0.047	0.048
5,000	0.495	0.497	0.499	0.503	0.500	63	49	51	51	56	0.063	0.049	0.051	0.051	0.056
10,000	0.512	0.499	0.507	0.508	0.509	53	44	47	59	49	0.053	0.044	0.047	0.059	0.049

Note: GLS had 37 non-convergences when N=50. 1 non-convergence when N=80

UCLA

Empirical Rejection Frequency

Table 2. Simulation Concerning Emprical Rejection Rates

	Average P-values					Number of P< 0.05					Empirical Rejection Frequency				
N	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS
50	0.226	0.726	0.485	0.218	0.537	310	0	50	309	27	0.310	0.000	0.050	0.309	0.027
80	0.316	0.638	0.499	0.327	0.531	188	9	53	165	25	0.188	0.009	0.053	0.165	0.025
100	0.370	0.607	0.497	0.367	0.536	140	22	43	129	37	0.140	0.022	0.043	0.129	0.037
200	0.441	0.570	0.495	0.420	0.501	82	34	60	90	50	0.083	0.034	0.060	0.090	0.050
300	0.448	0.531	0.493	0.450	0.504	83	36	54	77	45	0.071	0.036	0.054	0.077	0.045
400	0.465	0.526	0.495	0.471	0.506	62	40	56	53	48	0.062	0.040	0.056	0.053	0.048
500	0.470	0.515	0.492	0.471	0.498	75	44	44	61	64	0.075	0.044	0.044	0.061	0.064
800	0.494	0.515	0.496	0.471	0.512	63	39	50	66	44	0.063	0.039	0.050	0.066	0.044
1,000	0.479	0.517	0.499	0.481	0.487	61	42	48	50	54	0.061	0.042	0.048	0.050	0.054
2,000	0.494	0.495	0.494	0.497	0.494	56	46	44	47	48	0.056	0.046	0.044	0.047	0.048
5,000	0.495	0.497	0.499	0.503	0.500	63	49	51	51	56	0.063	0.049	0.051	0.051	0.056
10,000	0.512	0.499	0.507	0.508	0.509	53	44	47	59	49	0.053	0.044	0.047	0.059	0.049

Note: GLS had 37 non-convergences when N=50. 1 non-convergence when N=80

Empirical Rejection Frequency

Table 2. Simulation Concerning Emprical Rejection Rates

	Average P-values					Number of P< 0.05					Empirical Rejection Frequency				
N	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS
50	0.226	0.726	0.485	0.218	0.537	310	0	50	309	27	0.310	0.000	0.050	0.309	0.027
80	0.316	0.638	0.499	0.327	0.531	188	9	53	165	25	0.188	0.009	0.053	0.165	0.025
100	0.370	0.607	0.497	0.367	0.536	140	22	43	129	37	0.140	0.022	0.043	0.129	0.037
200	0.441	0.570	0.495	0.420	0.501	82	34	60	90	50	0.083	0.034	0.060	0.090	0.050
300	0.448	0.531	0.493	0.450	0.504	83	36	54	77	45	0.071	0.036	0.054	0.077	0.045
400	0.465	0.526	0.495	0.471	0.506	62	40	56	53	48	0.062	0.040	0.056	0.053	0.048
500	0.470	0.515	0.492	0.471	0.498	75	44	44	61	64	0.075	0.044	0.044	0.061	0.064
800	0.494	0.515	0.496	0.471	0.512	63	39	50	66	44	0.063	0.039	0.050	0.066	0.044
1,000	0.479	0.517	0.499	0.481	0.487	61	42	48	50	54	0.061	0.042	0.048	0.050	0.054
2,000	0.494	0.495	0.494	0.497	0.494	56	46	44	47	48	0.056	0.046	0.044	0.047	0.048
5,000	0.495	0.497	0.499	0.503	0.500	63	49	51	51	56	0.063	0.049	0.051	0.051	0.056
10,000	0.512	0.499	0.507	0.508	0.509	53	44	47	59	49	0.053	0.044	0.047	0.059	0.049

Note: GLS had 37 non-convergences when N=50. 1 non-convergence when N=80

Empirical Rejection Frequency

Table 2. Simulation Concerning Emprical Rejection Rates

	Average P-values					Number of P< 0.05					Empirical Rejection Frequency				
N	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS
50	0.226	0.726	0.485	0.218	0.537	310	0	50	309	27	0.310	0.000	0.050	0.309	0.027
80	0.316	0.638	0.499	0.327	0.531	188	9	53	165	25	0.188	0.009	0.053	0.165	0.025
100	0.370	0.607	0.497	0.367	0.536	140	22	43	129	37	0.140	0.022	0.043	0.129	0.037
200	0.441	0.570	0.495	0.420	0.501	82	34	60	90	50	0.083	0.034	0.060	0.090	0.050
300	0.448	0.531	0.493	0.450	0.504	83	36	54	77	45	0.071	0.036	0.054	0.077	0.045
400	0.465	0.526	0.495	0.471	0.506	62	40	56	53	48	0.062	0.040	0.056	0.053	0.048
500	0.470	0.515	0.492	0.471	0.498	75	44	44	61	64	0.075	0.044	0.044	0.061	0.064
800	0.494	0.515	0.496	0.471	0.512	63	39	50	66	44	0.063	0.039	0.050	0.066	0.044
1,000	0.479	0.517	0.499	0.481	0.487	61	42	48	50	54	0.061	0.042	0.048	0.050	0.054
2,000	0.494	0.495	0.494	0.497	0.494	56	46	44	47	48	0.056	0.046	0.044	0.047	0.048
5,000	0.495	0.497	0.499	0.503	0.500	63	49	51	51	56	0.063	0.049	0.051	0.051	0.056
10,000	0.512	0.499	0.507	0.508	0.509	53	44	47	59	49	0.053	0.044	0.047	0.059	0.049

Note: GLS had 37 non-convergences when N=50. 1 non-convergence when N=80

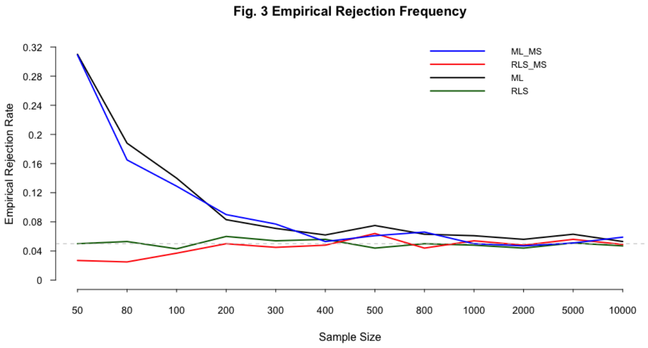
Empirical Rejection Frequency

Table 2. Simulation Concerning Emprical Rejection Rates

	Average P-values					Number of P< 0.05					Empirical Rejection Frequency				
N	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS	ML	GLS	RLS	ML MS	RLS MS
50	0.226	0.726	0.485	0.218	0.537	310	0	50	309	27	0.310	0.000	0.050	0.309	0.027
80	0.316	0.638	0.499	0.327	0.531	188	9	53	165	25	0.188	0.009	0.053	0.165	0.025
100	0.370	0.607	0.497	0.367	0.536	140	22	43	129	37	0.140	0.022	0.043	0.129	0.037
200	0.441	0.570	0.495	0.420	0.501	82	34	60	90	50	0.083	0.034	0.060	0.090	0.050
300	0.448	0.531	0.493	0.450	0.504	83	36	54	77	45	0.071	0.036	0.054	0.077	0.045
400	0.465	0.526	0.495	0.471	0.506	62	40	56	53	48	0.062	0.040	0.056	0.053	0.048
500	0.470	0.515	0.492	0.471	0.498	75	44	44	61	64	0.075	0.044	0.044	0.061	0.064
800	0.494	0.515	0.496	0.471	0.512	63	39	50	66	44	0.063	0.039	0.050	0.066	0.044
1,000	0.479	0.517	0.499	0.481	0.487	61	42	48	50	54	0.061	0.042	0.048	0.050	0.054
2,000	0.494	0.495	0.494	0.497	0.494	56	46	44	47	48	0.056	0.046	0.044	0.047	0.048
5,000	0.495	0.497	0.499	0.503	0.500	63	49	51	51	56	0.063	0.049	0.051	0.051	0.056
10,000	0.512	0.499	0.507	0.508	0.509	53	44	47	59	49	0.053	0.044	0.047	0.059	0.049

Note: GLS had 37 non-convergences when N=50. 1 non-convergence when N=80

Empirical Rejection Frequency



Conclusion

- Based test statistics and empirical rejection frequency, this study finds that both T_{ML} and T_{RLS} work fine when samples are large.
- When sample sizes are less than 500, the estimates of T_{ML} become increasingly inaccurate.
- Reweighted least squares produces consistent parameter estimates across all sample sizes.

UCLA

Thank You