An Asynchronous Distributed Expectation Maximization Algorithm For Massive Data: The DEM Algorithm

Sanvesh Srivastava [*], Glen DePalma [†] & Chuanhai Liu [‡]

June 14, 2018

1 Proof of Theorems in Sections 4.2 and 4.3

Our theoretical setup has the following assumptions:

1. Θ is a subset in the P-dimensional Euclidean space $\mathbb{R}^P$.

2. The set $\Pi_{\theta_0} \otimes \Theta_{\theta_0} = \{(\widetilde{p}, \theta) \in \Pi \otimes \Theta : F(\widetilde{p}, \theta) \geq F(\tilde{p}_0, \theta_0)\}$ is compact for any starting point of the $(\tilde{p}_t,\ \theta_t)$ sequence, denoted as $(\tilde{p}_0, \theta_0)$, that satisfies $\mathcal{L}(\theta_0) > -\infty$ and $\tilde{p}_0 = \prod_{k=1}^{K} h(Y_k \mid Z_k, \theta_0)$.

3. $F(\tilde{p}, \theta)$ is continuous in $\Pi \otimes \Theta$ and differentiable in the interior of $\Pi \otimes \Theta$.

4. $\Pi_{\theta_0} \otimes \Theta_{\theta_0}$ is in the interior of $\Pi \otimes \Theta$ for any $\theta_0 \in \Theta$.

[*]Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, Iowa 52242, sanvesh-srivastava@uiowa.edu

[†]Department of Statistics, Purdue University, West Lafayette, Indiana 47907, glen.depalma@gmail.com

[‡]Department of Statistics, Purdue University, West Lafayette, Indiana 47907, chuanhai@purdue.edu

5. The first order differential $\partial Q(\theta \mid \theta_t, \theta_{t_{(N+1)}}, \ldots, \theta_{t_K})/\partial\theta$ is continuous in $(\theta, \theta_t, \theta_{t_{(N+1)}}, \ldots, \theta_{t_K})$.

6. Worker k returns Q_k to the manager infinitely often for $t = 0, \ldots, \infty$ and $k = 1, \ldots, K$.

1.1 Proof of Theorem 4.1

The proof uses arguments similar to Theorems 1 and 2 of Neal & Hinton (1998). First, the E step of DEM at the $(t+1)$-th iteration updates $\widetilde{p}_{t,k} = h(Y_k \mid Z_k, \theta_{t_k})$ to $\widetilde{p}_{(t+1),k} = h(Y_k \mid Z_k, \theta_t)$ for worker k if $k \in U_{(t+1)}$; otherwise, $\widetilde{p}_{(t+1),k} = h(Y_k|Z_k, \theta_{t_k})$. Define $\widetilde{p}_{(t+1)} = \prod_{k_1 \in U_{t+1}} \widetilde{p}_{(t+1),k_1} \prod_{k_2 \in U_{t+1}^c} \widetilde{p}_{(t+1),k_2}$. Theorem 1 in Neal & Hinton (1998) implies that $F(\widetilde{p}_t, \theta_t) \leq F(\widetilde{p}_{(t+1)}, \theta_t)$ for a given θ_t. Second, the M step of DEM at the $(t+1)$-th iteration updates θ_t to $\theta_{(t+1)}$ and increases F from $F(\widetilde{p}_{(t+1)}, \theta_t)$ to $F(\widetilde{p}_{(t+1)}, \theta_{(t+1)})$ for fixed $\widetilde{p}_{(t+1)}$. At the end of $(t+1)$-th iteration of DEM, $F(\widetilde{p}_t, \theta_t) \leq F(\widetilde{p}_{(t+1)}, \theta_t) \leq F(\widetilde{p}_{(t+1)}, \theta_{(t+1)})$, where the first and last equality follow from Theorem 1 in Neal & Hinton (1998). Because t is a generic iteration, DEM maintains the monotone ascent of $F(\widetilde{p}, \theta)$ at every iteration and $\{F(\widetilde{p}_t, \theta_t), t \geq 0\}$ sequence converges because $F(\widetilde{p}, \theta)$ is upper bounded by our assumption. Theorem 2 in Neal & Hinton (1998) implies that if $(\hat{\widetilde{p}}, \hat{\theta})$ is a fixed point of $F(\widetilde{p}_t, \theta_t)$ sequence, then $\hat{\mathcal{L}} = \mathcal{L}(\hat{\theta})$ is a fixed point of $\mathcal{L}(\theta_t)$ sequence. This implies that there exists a monotone subsequence of $\mathcal{L}(\theta_t)$ converging to $\hat{\mathcal{L}}$.

1.2 Proof of Theorem 4.2

To prove this theorem, we require the definition of a closed map. A point-to-set mapping A is closed on a set X if $x_k \to x$, $x_k \in X$, and $y_k \to y$, $y_k \in A(x_k)$, then $y \in A(x)$ for every $x \in X$; see Luenberger & Ye (2008, pp 203) for details. If A is continuous, then it is closed.

The proof is based on Theorems 1, 2, 4, and 5 in Wu (1983). Our assumptions imply that the point-to-set map $(\tilde{p}_t, \theta_t) \mapsto (\tilde{p}_{t+1}, \theta_{t+1})$ is continuous, thus closed, on $\Pi \otimes \Theta \cap (\mathcal{S} \cup \mathcal{M})^c$; see Theorem 2 in Wu (1983). Theorem 2 in Neal & Hinton (1998) implies that $F(\widetilde{p}_t, \theta_t) \leq$

$F(\widetilde{p}_{t+1}, \theta_{t+1})$ for every $(\widetilde{p}_t, \theta_t) \in \Pi \otimes \Theta \cap (\mathcal{S} \cup \mathcal{M})^c$, so F is our ascent function. The global convergence theorem in Wu (1983) implies that all limit points of $(\widetilde{p}_t, \theta_t)$ sequence lie in $\mathcal{S} \cup \mathcal{M}$ and $F(\widetilde{p}_t, \theta_t)$ converges monotonically to $\hat{F} = F(\hat{\widetilde{p}}, \hat{\theta})$ for some $(\hat{\widetilde{p}}, \hat{\theta}) \in \mathcal{S} \cup \mathcal{M}$.

If $\mathcal{S}(\hat{F})$ (respectively $\mathcal{M}(\hat{F})) = \{(\hat{\widetilde{p}}, \hat{\theta})\}$, then there cannot be two different stationary points (respectively local maxima) with the same $\hat{F}$. This implies that $(\widetilde{p}_t, \theta_t) \to (\hat{\widetilde{p}}, \hat{\theta})$ and $\theta_t \to \hat{\theta}$ using coordinate-wise convergence. The first part of the theorem is proved.

Assumption 2 implies that $(\widetilde{p}_t, \theta_t)$ is a bounded sequence, so Theorem 5 in Wu (1983) implies that the set of limit points of the sequence $(\widetilde{p}_t, \theta_t)$ with $\|(\tilde{p}_{t+1}, \theta_{t+1}) - (\tilde{p}_t, \theta_t)\|_{\Pi \otimes \Theta} \to 0$ as $t \to \infty$ is connected and compact. Since $\mathcal{S}(\hat{F})$ and $\mathcal{M}(\hat{F})$ are discrete, the only connected and compact components of the stationary points (respectively local maxima) are singletons. All the limit points of $(\widetilde{p}_t, \theta_t)$ are in $\mathcal{S}(\hat{F}) \cup \mathcal{M}(\hat{F})$, so $(\widetilde{p}_t, \theta_t) \to (\hat{\widetilde{p}}, \hat{\theta})$ and the second part of the theorem is also proved.

1.3 Proof of Theorem 4.3

Recall that

$$i_{\text{com},\bar{\psi}} = \sum_{i=N+1}^{K} (i_{\text{com},\psi})_{kk}, \quad i_{\text{obs},\bar{\psi}} = \sum_{k=N+1}^{K} (i_{\text{obs},\psi})_{kk}, \quad i_{\text{com}} = i_{\text{com},\theta} + i_{\text{com},\bar{\psi}}, \quad i_{\text{obs}} = i_{\text{obs},\theta} + i_{\text{obs},\bar{\psi}}.$$

Define $C_{\theta,\bar{\psi}} = i_{\text{com},\theta}^{-1} i_{\text{com},\bar{\psi}}$ and $O_{\bar{\psi}} = i_{\text{com}}^{-1} i_{\text{obs},\bar{\psi}}$ and substitute them in

$$S_{\text{EM}} = i_{\text{com}}^{-1} i_{\text{obs}}, \; i_{\text{obs}} = -\frac{\partial^2 \log g(Z_{1:K}|\theta)}{\partial\theta \cdot \partial\theta^T}\Big|_{\theta=\hat{\theta}^E}, \; i_{\text{com}} = -\,\mathbb{E}_Y \left\{ \frac{\partial^2 \log f(Y_{1:K}, Z_{1:K}|\theta)}{\partial\theta \cdot \partial\theta^T} \mid Z_{1:K}, \theta \right\} \Big|_{\theta=\hat{\theta}^E},$$

where i_{obs} and i_{com} are the observed-data and complete-data information matrices, to obtain that

$$S_{\text{EM}} = i_{\text{com}}^{-1} i_{\text{obs}} = (I + i_{\text{com},\theta}^{-1} i_{\text{com},\bar{\psi}})^{-1} i_{\text{com},\theta}^{-1} i_{\text{obs},\theta} + O_{\bar{\psi}} = (I + C_{\theta,\bar{\psi}})^{-1} S_{\text{DEM}} + O_{\bar{\psi}}.$$

Simplifying the equality in the above display yields

$$\lambda_{\min}(S_{\text{EM}}) \overset{(i)}{\leq} \lambda_{\max}\{(I + C_{\theta,\bar{\psi}})^{-1} S_{\text{DEM}}\} + \lambda_{\min}(O_{\bar{\psi}}) \overset{(ii)}{\leq} \lambda_{\max}\{(I + C_{\theta,\bar{\psi}})^{-1}\}\lambda_{\min}\{S_{\text{DEM}}\} + \lambda_{\min}(O_{\bar{\psi}})$$

$$= \{1 + \lambda_{\min}(C_{\theta,\bar{\psi}})\}^{-1}\lambda_{\min}\{S_{\text{DEM}}\} + \lambda_{\min}(O_{\bar{\psi}}) \tag{1}$$

$$\lambda_{\min}(S_{\text{EM}}) \overset{(iii)}{\geq} \lambda_{\min}\{(I + C_{\theta,\bar{\psi}})^{-1} S_{\text{DEM}}\} + \lambda_{\min}(O_{\bar{\psi}}) \overset{(iv)}{\geq} \lambda_{\min}\{(I + C_{\theta,\bar{\psi}})^{-1}\}\lambda_{\min}(S_{\text{DEM}}) + \lambda_{\min}(O_{\bar{\psi}})$$

$$= \{1 + \lambda_{\max}(C_{\theta,\bar{\psi}})\}^{-1}\lambda_{\min}(S_{\text{DEM}}) + \lambda_{\min}(O_{\bar{\psi}}), \tag{2}$$

where inequalities (i), (ii), (iii), and (iv) follow from Problem III.6.5 in Bhatia (1997); therefore,

$$\frac{\lambda_{\min}(S_{\text{DEM}})}{1 + \lambda_{\max}(C_{\theta,\bar{\psi}})} + \lambda_{\min}(O_{\bar{\psi}}) \leq \lambda_{\min}(S_{\text{EM}}) \leq \frac{\lambda_{\min}(S_{\text{DEM}})}{1 + \lambda_{\min}(C_{\theta,\bar{\psi}})} + \lambda_{\min}(O_{\bar{\psi}}).$$

2 Additional experimental results from Section 5

Recall the linear mixed effects model used for experiments. Let p, q, m, n, and n_i be the number of fixed effects, number of random effects, sample size, total number of observations, and total number of observations for sample i ($i = 1, \ldots, m$) so that $n = \sum_{i=1}^{m} n_i$. If $\mathbf{y}_i \in \mathbb{R}^{n_i}$ is the observation for sample i for $i = 1, \ldots, m$, then

$$\mathbf{y}_i = X_i\,\boldsymbol{\beta} + Z_i\,\mathbf{b}_i + \mathbf{e}_i, \quad \mathbf{b}_i \sim N_q(\mathbf{0}, \Sigma),\ \Sigma = \tau^2 D, \quad \mathbf{e}_i \sim N_{n_i}(\mathbf{0}, \tau^2 I_{n_i}), \tag{3}$$

where $X_i \in \mathbb{R}^{n_i \times p}$ and $Z_i \in \mathbb{R}^{n_i \times q}$ are known matrices of fixed and random effects covariates, respectively, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the fixed effects parameter vector, τ^2 is the error variance parameter, D is a symmetric positive definite matrix, $\mathbf{b}_i \in \mathbb{R}^q$ is the random effects vector for sample i that follows a q-dimensional Gaussian distribution with mean $\mathbf{0}$ and covariance parameter $\Sigma = \tau^2 D$, and I_{n_i} is n_i-by-n_i identity matrix. The parameter vector is $\theta = \{\boldsymbol{\beta}, \Sigma, \tau^2\}$.

The linear mixed-effects model in (3) satisfies Assumptions 1–5 in Theorem 4.2. Let LL^T be the Cholesky decomposition of Σ, where L is lower triangular, and vech(L) be the lower triangular part of L arranged in a $q(q+1)/2$-dimensional vector. Our parameter vector can be also defined as $\theta = \{\boldsymbol{\beta}, \text{vech}(L), \tau^2\}$ and we assume that the parameter space Θ is a compact subset of the $(p + q(q+1)/2 + 1)$-dimensional Euclidean space. This verifies Assumption 1. In our simulation and real data analysis, we fix $\boldsymbol{\beta}_0 = \mathbf{0}$, $L_0 = I_q$, and $\tau_0^2 = 10$ as the starting point of DEM iterations. The conditional distribution of missing data $\mathbf{b}_i$ in (3) is also Gaussian with mean $\hat{\mathbf{b}}_i$ and covariance matrix $\hat{C}_i$ $(i = 1, \ldots, m)$; see Equation 3.6 in van Dyk (2000) for the analytic forms of $\hat{\mathbf{b}}_i$ and $\hat{C}_i$. Define Π in Assumption 2 to be a compact set of continuous distributions with density $\tilde{p}$, finite $\text{KL}\{\tilde{p}, N(\hat{\mathbf{b}}_i, \hat{C}_i)\}$ for every i, and finite $\int \log\{\tilde{p}(y)\}\tilde{p}(y)dy$. For any such θ_0, $\Pi_{\theta_0} \otimes \Theta_{\theta_0}$ is a compact subset of $\Pi \otimes \Theta$, which verifies Assumption 4. Assumption 2 is true because the likelihood function is finite at $\theta_0 = \{\boldsymbol{\beta}_0, \text{vech}(L_0), \tau_0^2\}$. The likelihood for θ in (3) is based on a Gaussian density and is differentiable in the interior of Θ, which verifies Assumption 3. Equation 3.4 in van Dyk (2000) shows that $Q_k(\theta \mid \theta_{t_k})$ is differentiable for every k. The Q-function in DEM is the sum of $Q_1(\theta \mid \theta_{t_1}), \ldots, Q_K(\theta \mid \theta_{t_K})$, so it is also differentiable, which verifies Assumption 5. Our implementation ensures that $Q_k(\theta \mid \theta_{t_k})$ is returned to the manager for every k before convergence is declared, satisfying Assumption 6.

The accuracy of every algorithm in parameter estimation was judged using errors defined as

$$\text{err}_{\boldsymbol{\beta}}^2 = p^{-1} \sum_{i=1}^{p} \left(\hat{\beta}_i - \hat{\beta}_i^{\text{EM}}\right)^2, \quad \text{err}_{\tau^2}^2 = (\hat{\tau}^2 - \hat{\tau}^{2\text{EM}})^2, \quad \text{err}_{\text{var}}^2 = q^{-1} \sum_{i=1}^{q} \left(\hat{\Sigma}_{ii} - \hat{\Sigma}_{ii}^{\text{EM}}\right)^2,$$
$$\text{err}_{\text{cov}}^2 = 2q^{-1}(q-1)^{-1} \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \left(\hat{\Sigma}_{ij} - \hat{\Sigma}_{ij}^{\text{EM}}\right)^2, \tag{4}$$

where $\{\hat{\boldsymbol{\beta}}^{\text{EM}}, \hat{\Sigma}^{\text{EM}}, \hat{\tau}^{2\text{EM}}\}$ and $\{\hat{\boldsymbol{\beta}}, \hat{\Sigma}, \hat{\tau}^2\}$ respectively were the parameter estimates of ECME_0 and its competitor, including lme4, Meta-lme4, IEM, or DEM. If err_i represented the error in replication i of the experiment, the root mean square error (RMSE) over R replications was defined as

$$\mathrm{RMSE}_{\boldsymbol{\beta}}^2 = R^{-1}\sum_{i=1}^{R}\mathrm{err}_{\boldsymbol{\beta}\,i}^2, \quad \mathrm{RMSE}_{\tau^2}^2 = R^{-1}\sum_{i=1}^{R}\mathrm{err}_{\tau^2\,i}^2,$$
$$\mathrm{RMSE}_{\mathrm{var}}^2 = R^{-1}\sum_{i=1}^{R}\mathrm{err}_{\mathrm{var}\,i}^2, \quad \mathrm{RMSE}_{\mathrm{cov}}^2 = R^{-1}\sum_{i=1}^{R}\mathrm{err}_{\mathrm{cov}\,i}^2. \tag{5}$$

The smaller the RMSE, the closer are the results to the benchmark ECME_0 algorithm.

Table 1: Root mean square error (5) in estimation of fixed effects ($\boldsymbol{\beta}$) averaged across simulation replications. The maximum Monte Carlo error is of the order 10^{-4}

	$K = 10$							
	$m = 10^4, n = 10^6$				$m = 10^5, n = 10^7$			
	$q = 3$		$q = 6$		$q = 3$		$q = 6$	
	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$
lme4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Meta-lme4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
IEM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.3$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.5$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.7$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$K = 20$							
	$m = 10^4, n = 10^6$				$m = 10^5, n = 10^7$			
	$q = 3$		$q = 6$		$q = 3$		$q = 6$	
	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$
lme4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Meta-lme4	0.0001	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
IEM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.3$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.5$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.7$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

References

Bhatia, R. (1997), *Matrix Analysis*, Vol. 169, Springer-Verlag, New York.

Luenberger, D. G. & Ye, Y. (2008), *Linear and nonlinear programming*, Vol. 116, Springer.

Neal, R. M. & Hinton, G. E. (1998), A view of the EM algorithm that justifies incremental, sparse, and other variants, *in* 'Learning in graphical models', Springer, pp. 355–368.

van Dyk, D. A. (2000), 'Fitting mixed-effects models using efficient EM-type algorithms', *Journal of Computational and Graphical Statistics* **9**(1), 78–98.

Wu, C. (1983), 'On the convergence properties of the EM algorithm', *The Annals of Statistics* **11**(1), 95–103.

Table 2: Root mean square error (5) in the estimation of τ^2 and Σ averaged across simulation replications. The maximum Monte Carlo errors are of the order 10^{-4}, 10^{-2}, and 10^{-3} for the error variances, variances of the random effects, and covariances of random effects

	Error variance (τ^2)							
	$K = 10$							
	$m = 10^4, n = 10^6$				$m = 10^5, n = 10^7$			
	$q = 3$		$q = 6$		$q = 3$		$q = 6$	
	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$
lme4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Meta-lme4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
IEM	0.0001	0.0071	0.0004	0.0002	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.3$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.5$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.7$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$K = 20$							
	$m = 10^4, n = 10^6$				$m = 10^5, n = 10^7$			
	$q = 3$		$q = 6$		$q = 3$		$q = 6$	
	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$
lme4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Meta-lme4	0.0001	0.0001	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000
IEM	0.0001	0.0002	0.0008	0.0018	0.0001	0.0002	0.0001	0.0001
DEM ($\gamma = 0.3$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.5$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.7$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

	Variances of random effects (diagonal elements of Σ)							
	$K = 10$							
	$m = 10^4, n = 10^6$				$m = 10^5, n = 10^7$			
	$q = 3$		$q = 6$		$q = 3$		$q = 6$	
	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$
lme4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Meta-lme4	0.0008	0.0008	0.0018	0.0015	0.0001	0.0001	0.0001	0.0002
IEM	0.0001	0.0017	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.3$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.5$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.7$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$K = 20$							
	$m = 10^4, n = 10^6$				$m = 10^5, n = 10^7$			
	$q = 3$		$q = 6$		$q = 3$		$q = 6$	
	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$
lme4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Meta-lme4	0.0009	0.0009	0.0022	0.0024	0.0001	0.0001	0.0003	0.0003
IEM	0.0001	0.0001	0.0002	0.0171	0.0000	0.0001	0.0000	0.0000
DEM ($\gamma = 0.3$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.5$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.7$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

	Covariances of random effects (off-diagonal elements of Σ)							
	$K = 10$							
	$m = 10^4, n = 10^6$				$m = 10^5, n = 10^7$			
	$q = 3$		$q = 6$		$q = 3$		$q = 6$	
	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$
lme4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Meta-lme4	0.0006	0.0006	0.00010	0.0009	0.0001	0.0001	0.0001	0.0001
IEM	0.0000	0.0010	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.3$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.5$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.7$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$K = 20$							
	$m = 10^4, n = 10^6$				$m = 10^5, n = 10^7$			
	$q = 3$		$q = 6$		$q = 3$		$q = 6$	
	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$
lme4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Meta-lme4	0.0006	0.0007	0.0015	0.0014	0.0001	0.0001	0.0002	0.0002
IEM	0.0000	0.0000	0.0001	0.0040	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.3$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.5$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.7$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 3: The ratio of IEM or DEM and ECME_0 log likelihoods averaged over simulation replications. Monte Carlo errors are in parenthesis

	IEM or DEM log likelihood / ECME_0 log likelihood							
	$K = 10$							
	$m = 10^4, n = 10^6$				$m = 10^5, n = 10^7$			
	$q = 3$		$q = 6$		$q = 3$		$q = 6$	
	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$
IEM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
DEM ($\gamma = 0.3$)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
DEM ($\gamma = 0.5$)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
DEM ($\gamma = 0.7$)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
	$K = 20$							
	$m = 10^4, n = 10^6$				$m = 10^5, n = 10^7$			
	$q = 3$		$q = 6$		$q = 3$		$q = 6$	
	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$	$p = 10$	$p = 20$
IEM	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
DEM ($\gamma = 0.3$)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
DEM ($\gamma = 0.5$)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
DEM ($\gamma = 0.7$)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)

Table 4: Root mean square error (5) in estimation of fixed effects ($\boldsymbol{\beta}$), variances of random effects (diagonal elements of Σ), error variance (τ^2), and covariances of random effects (off-diagonal elements of Σ) averaged over all replications. The maximum Monte Carlo errors are of the order 10^{-3}, 10^{-4}, 10^{-3}, and 10^{-4}, respectively. The subscripts $1, \ldots, 6$ represent *Action*, *Children − Action*, *Comedy − Action*, *Drama − Action*, *popularity*, and *previous* predictors

	β_{Action}	$\beta_{\text{Children − Action}}$	$\beta_{\text{Comedy − Action}}$	$\beta_{\text{Drama − Action}}$	$\beta_{\text{popularity}}$	β_{previous}
lme4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Meta-lme4	0.0000	0.0002	0.0001	0.0001	0.0000	0.0000
IEM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.3$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.5$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.7$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

	σ^2_{Action}	$\sigma^2_{\text{Children − Action}}$	$\sigma^2_{\text{Comedy − Action}}$	$\sigma^2_{\text{Drama − Action}}$	$\sigma^2_{\text{popularity}}$	$\sigma^2_{\text{previous}}$	τ^2
lme4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0006
Meta-lme4	0.0001	0.0002	0.0001	0.0000	0.0000	0.0001	0.0007
IEM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0010
DEM ($\gamma = 0.3$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0008
DEM ($\gamma = 0.5$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0007
DEM ($\gamma = 0.7$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0006

	σ_{12}	σ_{13}	σ_{14}	σ_{15}	σ_{16}	σ_{23}	σ_{24}	σ_{25}
lme4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Meta-lme4	0.0001	0.0001	0.0000	0.0000	0.0001	0.0001	0.0001	0.0000
IEM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.3$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.5$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DEM ($\gamma = 0.7$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	σ_{26}	σ_{34}	σ_{35}	σ_{36}	σ_{45}	σ_{46}	σ_{56}	
lme4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
Meta-lme4	0.0001	0.0000	0.0000	0.0000	0.0001	0.0001	0.0000	
IEM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
DEM ($\gamma = 0.3$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
DEM ($\gamma = 0.5$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
DEM ($\gamma = 0.7$)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	

Table 5: The ratio of IEM or DEM log likehood over ECME_0 log likehood averaged over all replications. Monte Carlo errors are in parenthesis

IEM	DEM ($\gamma = 0.30$)	DEM ($\gamma = 0.50$)	DEM ($\gamma = 0.70$)
1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)