Supplemental Material for "Methods to Account for Uncertainty in Latent Class Assignments when using Latent Classes as Predictors in Regression Models, with Application to Acculturation Strategy Measures"

EM Algorithm Details. Here we review the EM algorithm for the joint model, for the linear and binary outcomes.

The EM algorithm consists of a expectation step, where the expectation of complete-data score equation is obtained conditional on the most recent estimates of parameters, and a maximization step, where the score equation is solved, replacing the complete data with the conditional expectation.

First, note that the complete data log-likelihood can be factored into three components: one involving the overall probability of class membership, one involving the probability of the manifest variables given class membership, and the standard regression model that treats the latent variables as observed. Hence we can carry out the maximizations separately. For the overall probability of class assignment, the score equation is given by

$$\begin{pmatrix} \sum_i I(C_i = 1)(\theta_1)^{-1} \\ \vdots \\ \sum_i I(C_i = J)(\theta_J)^{-1} \end{pmatrix} = 0$$

Taking expectation yields

$$\begin{pmatrix} \sum_i p_{ij}(\theta_1)^{-1} \\ \vdots \\ \sum_i p_{ij}(\theta_J)^{-1} \end{pmatrix} = 0$$

where $p_{ij} = P(C_i = j \mid Z_{i1}, ..., Z_{iL}, Y_i, X_{i1}, ..., X_{iM})$ is defined below. Using standard binomial results, under the constraint that $\sum_j \theta_j = 1$, a closed form maximization step is obtained as $\theta_j^{(t)} = n^{-1} \sum_i p_{ij}^{(t-1)}$, where $p_{ij}^{(s)}$ is obtained by replacing the parameters in p_{ij} with their estimates from the sth iteration of the algorithm. A similar derivation yields $\pi^{(t)}_{klj} = \frac{\sum_i p_{ij}^{(t-1)} I(z_{il}=k)}{\sum_i p_{ij}^{(t-1)}}$ for the class-specific probabilities associated with the manifest variables.

For the regression parameters, note that the score equation for a generalized linear model with a canonical link function g is given by

$$\sum\nolimits_{i=1}^{n} \left(y_i - g^{-1}(\mathbf{X}_i^T \beta)\right) \mathbf{X}_i = 0$$

where for linear regression, $g(u) = u$ is the identify function, and for logistic regression, $g(u) = \log\left(\frac{u}{1-u}\right)$. For the complete data linear model (taking the first latent class as the reference class, so that $\alpha_1 = 0$), this yields

$$\sum\nolimits_{i=1}^{n} \left(y_i - (\beta_0 + \sum_{j=2}^{J} I(C_i = j)\alpha_j + \sum_m \beta_m X_{im}) \right) \begin{pmatrix} 1 \\ I(C_i = 2) \\ \vdots \\ I(C_i = J) \\ X_{i1} \\ \vdots \\ X_{iJ} \end{pmatrix} = 0$$

Since $I^2(u) = I(u)$ for the indicator function, taking conditional expectation yields

$$\sum_{i=1}^{n} \begin{pmatrix} y_i - (\beta_0 + \sum_m \beta_m X_{im} + \sum_{j=2}^{J} p_{ij}\alpha_j) \\ (y_i - (\beta_0 + \sum_m \beta_m X_{im} + \alpha_2))p_{i2} \\ \vdots \\ (y_i - (\beta_0 + \sum_m \beta_m X_{im} + \alpha_J))p_{iJ} \\ (y_i - (\beta_0 + \sum_m \beta_m X_{im} + \sum_{j=2}^{J} p_{ij}\alpha_j))X_{i1} \\ \vdots \\ (y_i - (\beta_0 + \sum_m \beta_m X_{im} + \sum_{j=2}^{J} p_{ij}\alpha_j))X_{iM} \end{pmatrix} = 0$$

This can be solved for $\alpha^{(\mathbf{t})}$ and $\beta^{(\mathbf{t})}$ by replacing p_{ij} with $p_{ij}^{(t-1)}$ and taking the derivative of this expected score and using a Newtonian method, or by use of a gradient-finding algorithm; here we used the optim function in R. For the variance parameter, the complete data score equation is given by

$$\sum_i \sum_j I(C_i = j)\frac{1}{2\sigma^2} r_{ij}^2 = 0$$

where $r_{ij} = y_i - (\beta_0 + \sum_m \beta_m X_{im} + \alpha_j)$ is the residual when $C_i = j$. The maximizer is in closed form: $\sigma^{(t-1)^2} = n^{-1} \sum_i \sum_j p_{ij} r_{ij}^{(t-1)^2}$. The conditional probability for a given observation belonging to a latent class in the linear model is then given by

$$p_{ij} = P(C_i = j \mid Z_{i1}, ..., Z_{iL}, Y_i, X_{i1}, ..., X_{iM}) = \frac{\theta_j \prod_l \prod_k \pi_{klj}^{Z_{ikl}} \exp(-\frac{1}{2\sigma^2}(y_i - \sum_m \beta_m X_{im} - \alpha_j)^2)}{\sum_{j=1}^{J} \theta_j \prod_l \prod_k \pi_{klj}^{Z_{ikl}} \exp(-\frac{1}{2\sigma^2}(y_i - \sum_m \beta_m X_{im} - \alpha_j)^2)}$$

For the logistic model, the score equation is given by

$$\sum_{i=1}^{n} \left(y_i - \frac{\exp\left(\beta_0 + \sum_{j=2}^{J} I(C_i = j)\alpha_j + \sum_m \beta_m X_{im}\right)}{1 + \exp\left(\beta_0 \sum_{j=2}^{J} I(C_i = j)\alpha_j + \sum_m \beta_m X_{im}\right)} \right) \begin{pmatrix} 1 \\ I(C_i = 2) \\ \vdots \\ I(C_i = J) \\ X_{i1} \\ \vdots \\ X_{iJ} \end{pmatrix} = 0$$

Taking conditional expectation yields

$$\sum_i \begin{pmatrix} y_i - \sum_j p_{ij} \frac{\exp\left(\beta_0+\alpha_j+\sum_m \beta_m X_{im}\right)}{1+\exp\left(\beta_0+\alpha_j+\sum_m \beta_m X_{im}\right)} \\ (y_i - \frac{\exp\left(\beta_0+\alpha_2+\sum_m \beta_m X_{im}\right)}{1+\exp\left(\beta_0+\alpha_2+\sum_m \beta_m X_{im}\right)})p_{i2} \\ \vdots \\ (y_i - \frac{\exp\left(\beta_0+\alpha_J+\sum_m \beta_m X_{im}\right)}{1+\exp\left(\beta_0+\alpha_J+\sum_m \beta_m X_{im}\right)})p_{iJ} \\ (y_i - \sum_j p_{ij} \frac{\exp\left(\beta_0+\alpha_j+\sum_m \beta_m X_{im}\right)}{1+\exp\left(\beta_0+\alpha_j+\sum_m \beta_m X_{im}\right)})X_{i1} \\ \vdots \\ (y_i - \sum_j p_{ij} \frac{\exp\left(\beta_0+\alpha_j+\sum_m \beta_m X_{im}\right)}{1+\exp\left(\beta_0+\alpha_j+\sum_m \beta_m X_{im}\right)})X_{im} \end{pmatrix} = 0$$

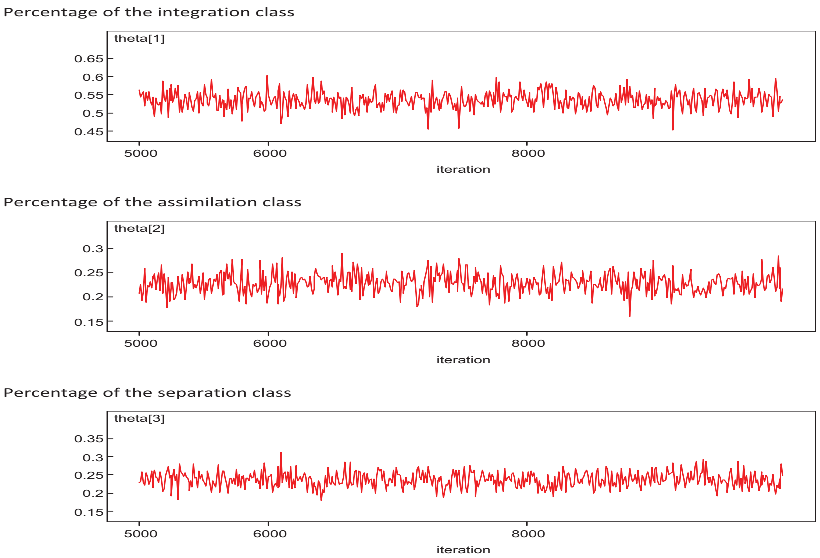
where again $\alpha^{(\mathbf{t})}$ and $\beta^{(\mathbf{t})}$ are obtained by replacing p_{ij} with $p_{ij}^{(t-1)}$ and using either a Newton-Raphson type algorithm or a a gradient-finding algorithm. The conditional probability for a given observation belonging to a latent class in the logistic model is then given by

$$p_{ij} = P(C_i = j \mid Z_{i1}, ..., Z_{iL}, Y_i, X_{i1}, ..., X_{iM}) =$$

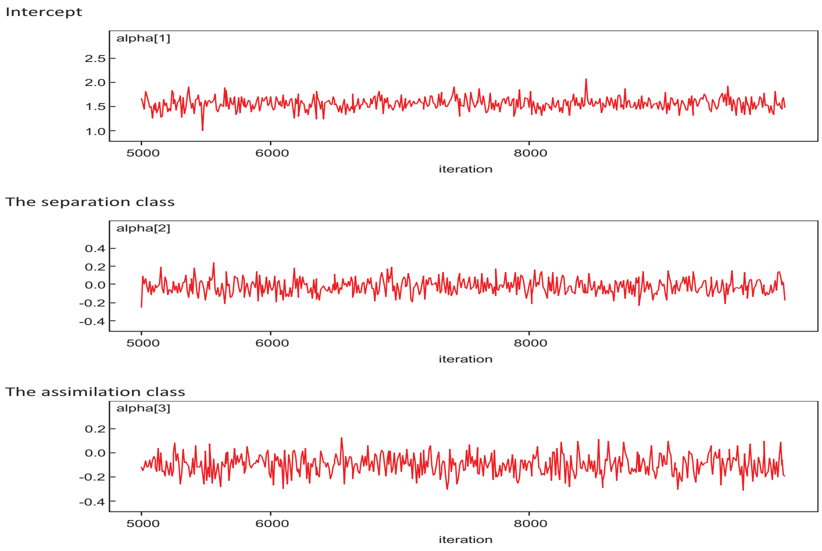
$$\frac{\theta_j \prod_l \prod_k \pi_{klj}^{Z_{ikl}} \frac{\exp\left(\beta_0+\alpha_j+\sum_m \beta_m X_{im}\right)^{Y_i}}{1+\exp\left(\beta_0+\alpha_j+\sum_m \beta_m X_{im}\right)}}{\sum_{j=1}^{J} \theta_j \prod_l \prod_k \pi_{klj}^{Z_{ikl}} \frac{\exp\left(\beta_0+\alpha_j+\sum_m \beta_m X_{im}\right)^{Y_i}}{1+\exp\left(\beta_0+\alpha_j+\sum_m \beta_m X_{im}\right)}}$$

MCMC Convergence for MASALA Analysis.

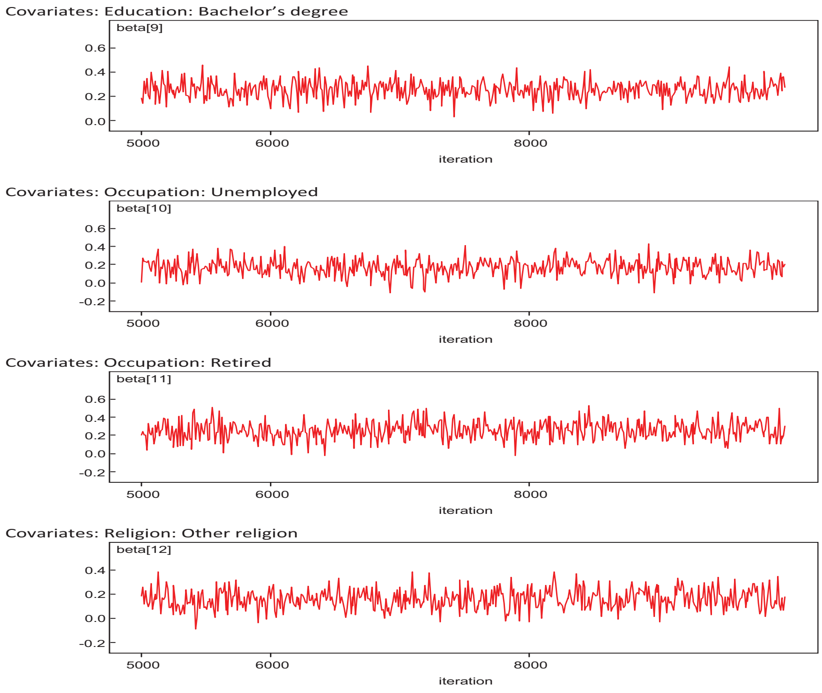
Here we show the trace plots for the marginal probability of the class membership θ (Figures 1 and 7 and the regression parameters on the depression outcome α (Figures 2 and 8 for the assimilation latent classes) and β (Figures 3-6 and 9-11 for the control variables), for both the linear and logistic model of the MASALA analysis.


Percentage of the integration class
Percentage of the assimilation class
Percentage of the separation class

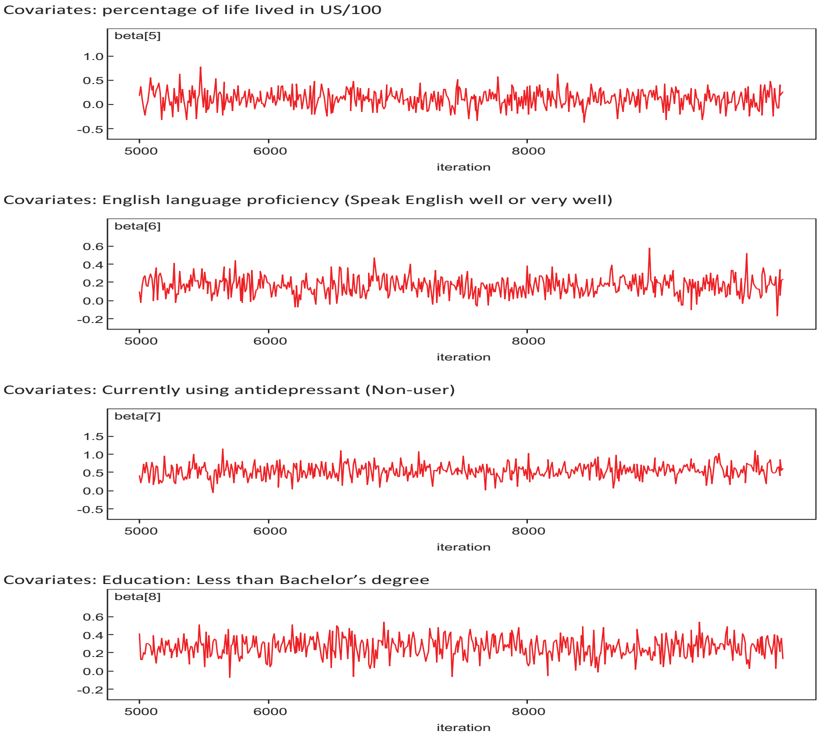
eFigure 1: Trace plots for linear regression proportion of latent class membership: MASALA Analysis.



eFigure 2: Trace plots for linear regression parameters associated with latent class membership: MASALA Analysis.



eFigure 3: Trace plots for linear regression parameters associated with control covariates: MASALA Analysis.



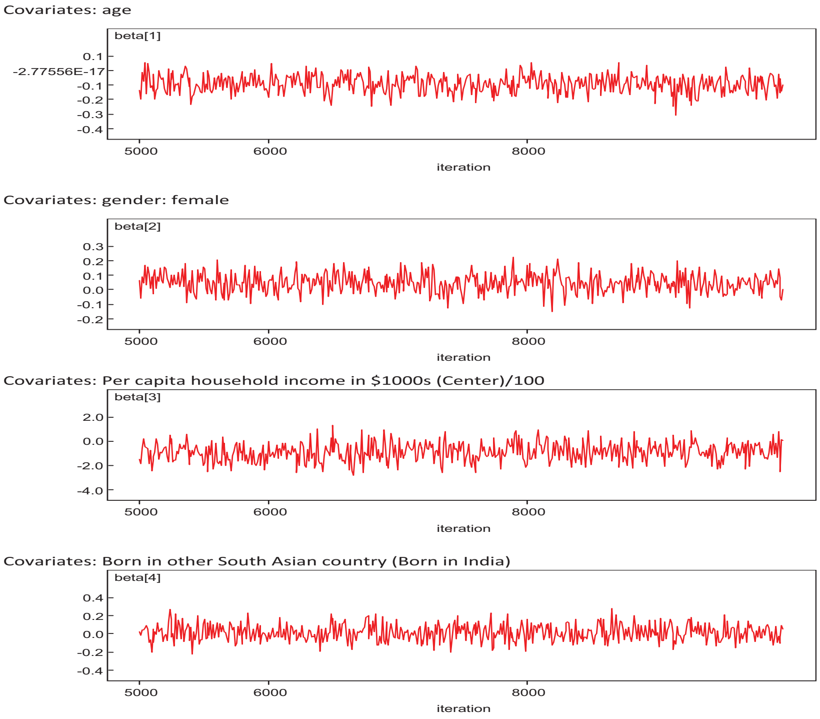
Covariates: percentage of life lived in US/100

Covariates: English language proficiency (Speak English well or very well)

Covariates: Currently using antidepressant (Non-user)

Covariates: Education: Less than Bachelor's degree

eFigure 4: Trace plots for linear regression parameters associated with control covariates: MASALA Analysis.

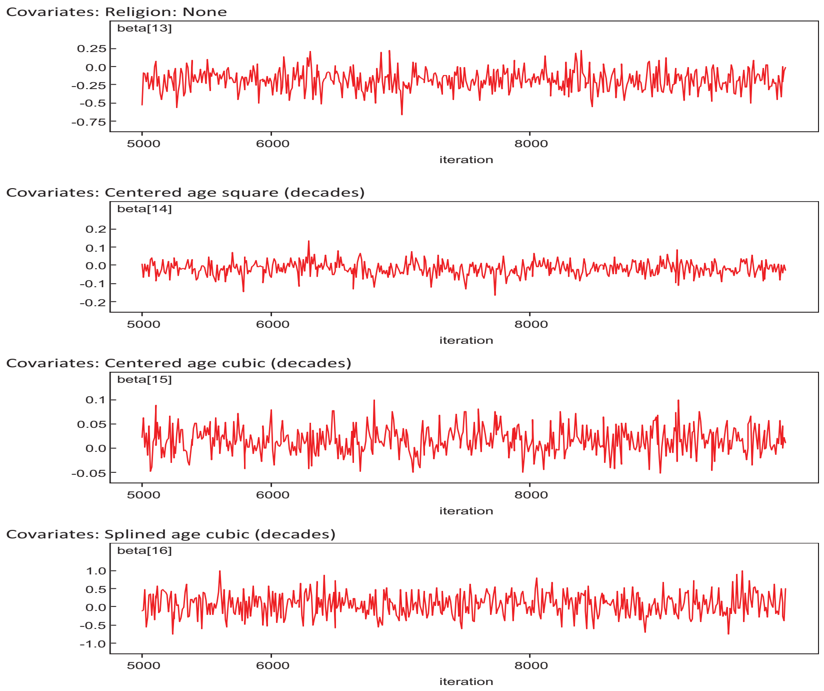

Covariates: age

Covariates: gender: female

Covariates: Per capita household income in $1000s (Center)/100

Covariates: Born in other South Asian country (Born in India)

eFigure 5: Trace plots for linear regression parameters associated with control covariates: MASALA Analysis.



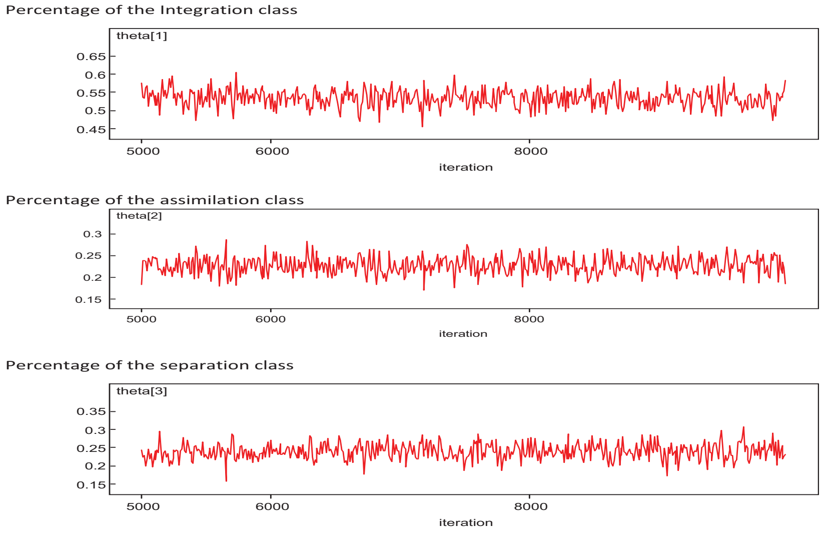
Covariates: Religion: None

Covariates: Centered age square (decades)

Covariates: Centered age cubic (decades)

Covariates: Splined age cubic (decades)

eFigure 6: Trace plots for linear regression parameters associated with control covariates: MASALA Analysis.

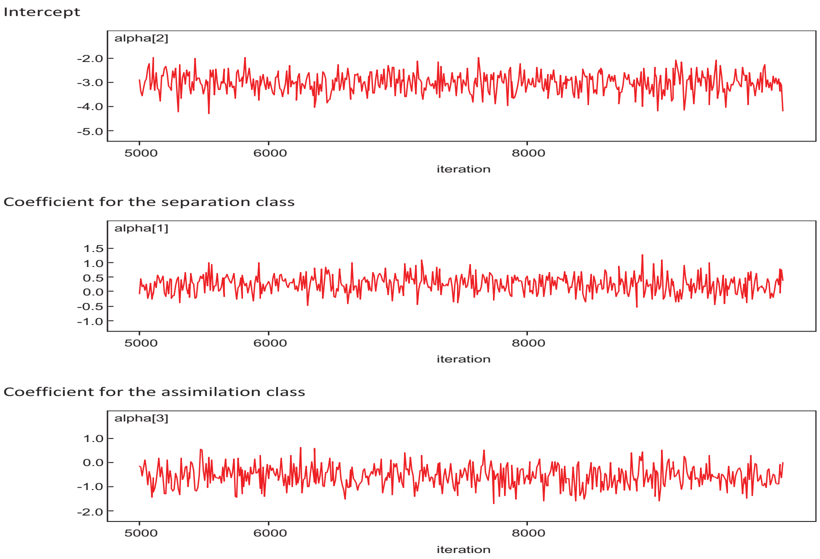


Percentage of the Integration class

Percentage of the assimilation class

Percentage of the separation class

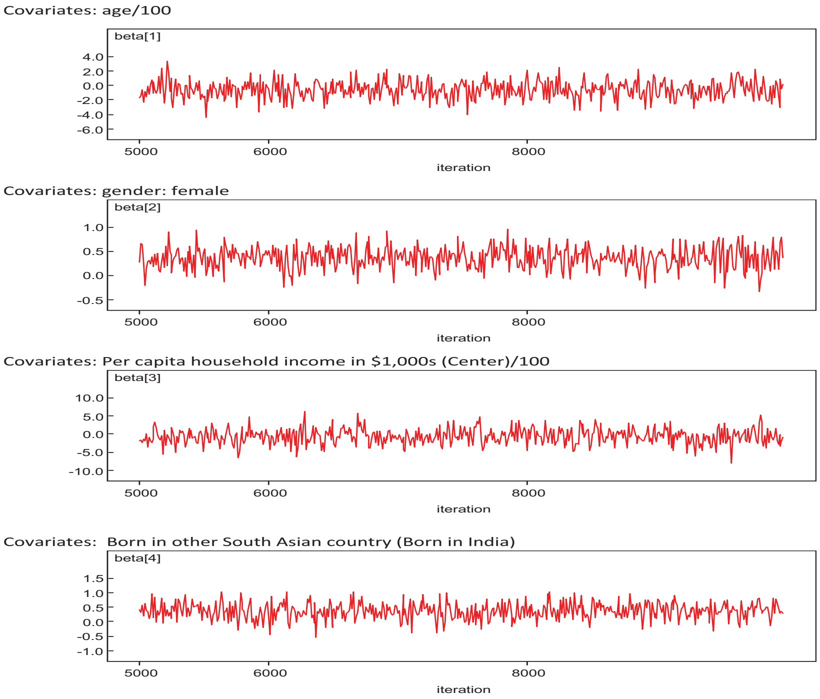
eFigure 7: Trace plots for logistic regression proportion of latent class membership: MASALA Analysis.


Intercept

Coefficient for the separation class

Coefficient for the assimilation class

eFigure 8: Trace plots for logistic regression parameters associated with latent class membership: MASALA Analysis.

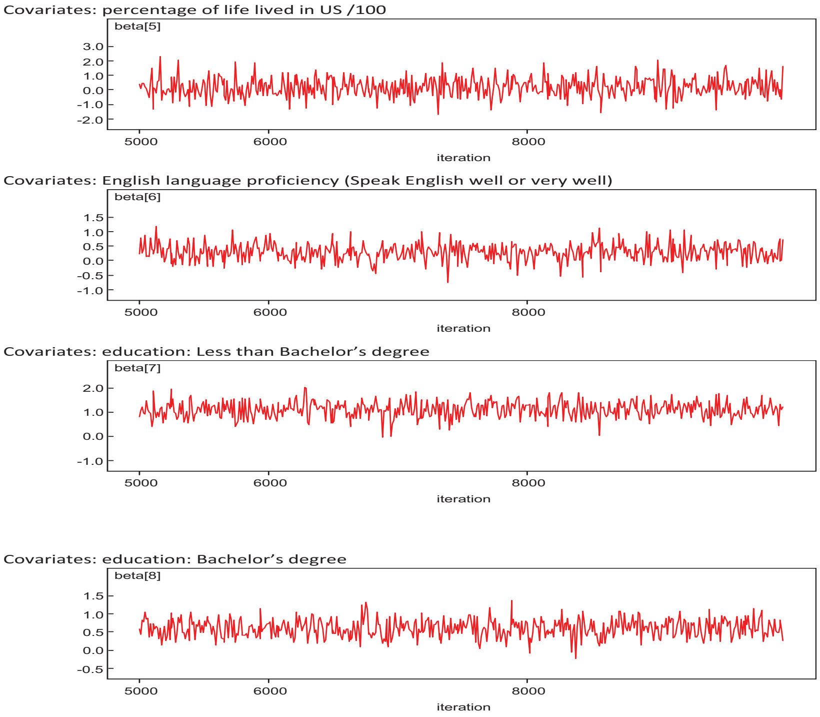

Covariates: age/100

Covariates: gender: female

Covariates: Per capita household income in $1,000s (Center)/100

Covariates: Born in other South Asian country (Born in India)

eFigure 9: Trace plots for logistic regression parameters associated with control covariates: MASALA Analysis.



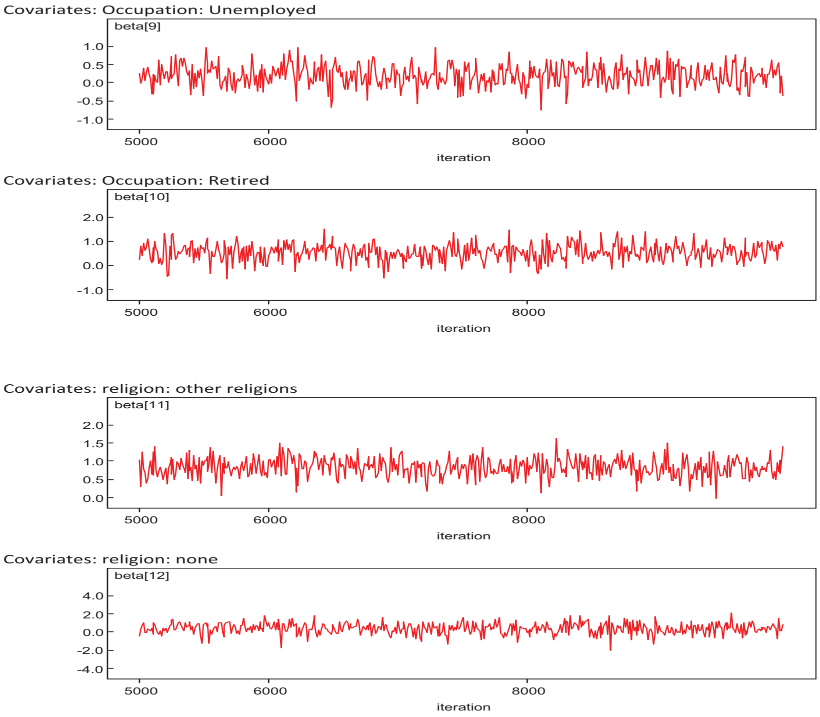
Covariates: percentage of life lived in US /100

Covariates: English language proficiency (Speak English well or very well)

Covariates: education: Less than Bachelor’s degree

Covariates: education: Bachelor’s degree

eFigure 10: Trace plots for logistic regression parameters associated with control covariates: MASALA Analysis.



Covariates: Occupation: Unemployed

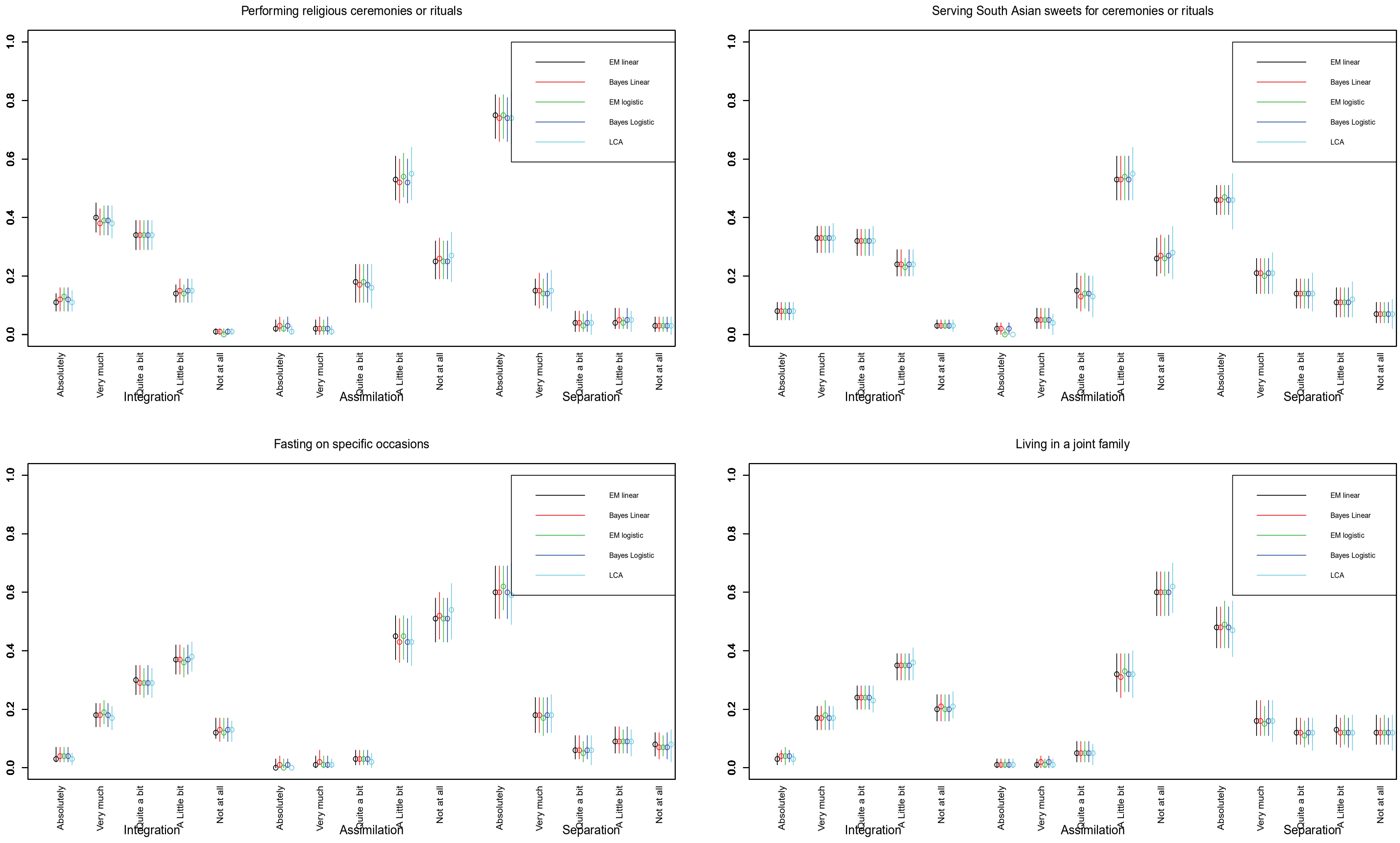
Covariates: Occupation: Retired

Covariates: religion: other religions

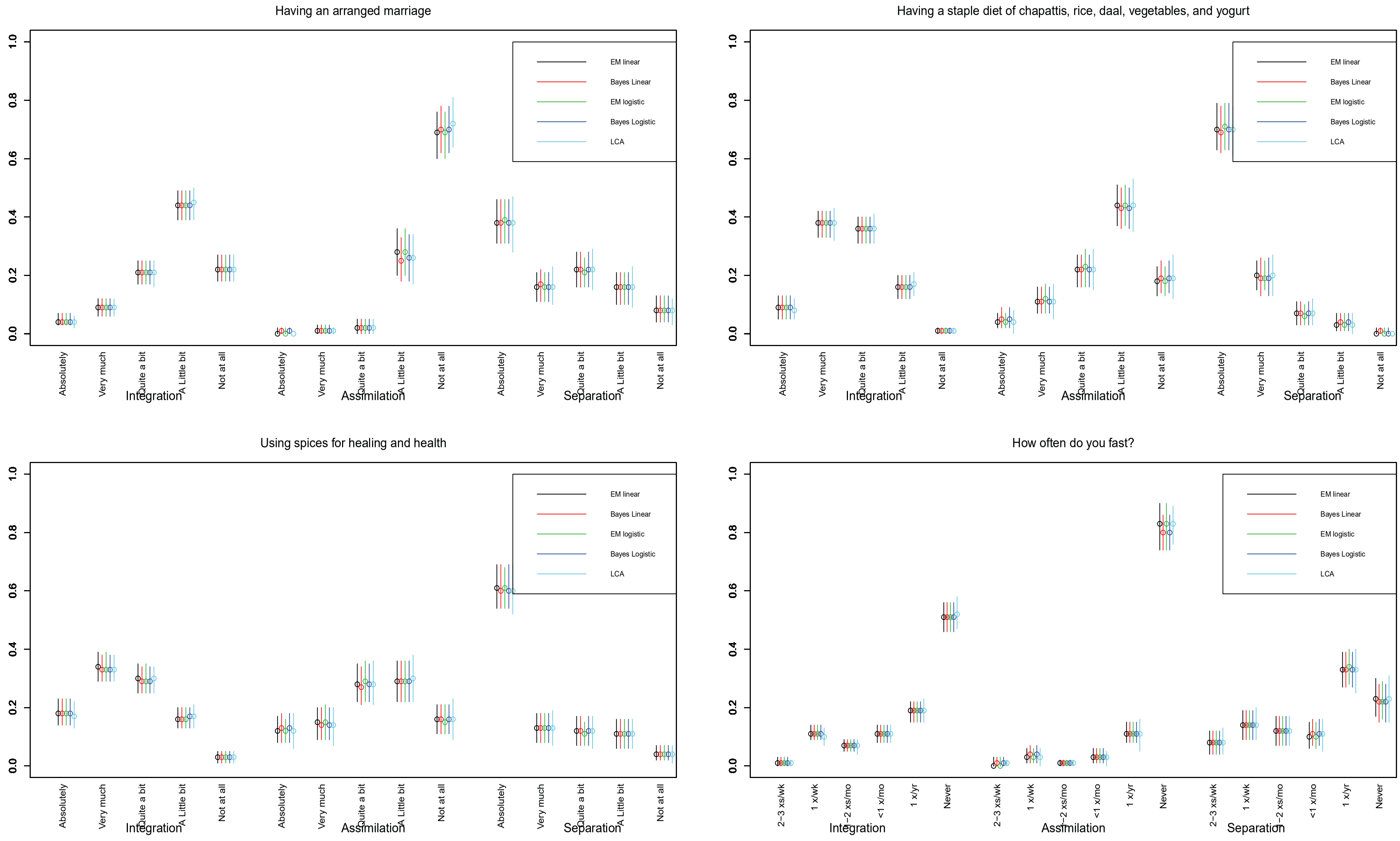
Covariates: religion: none

eFigure 11: Trace plots for logistic regression parameters associated with control covariates: MASALA Analysis.

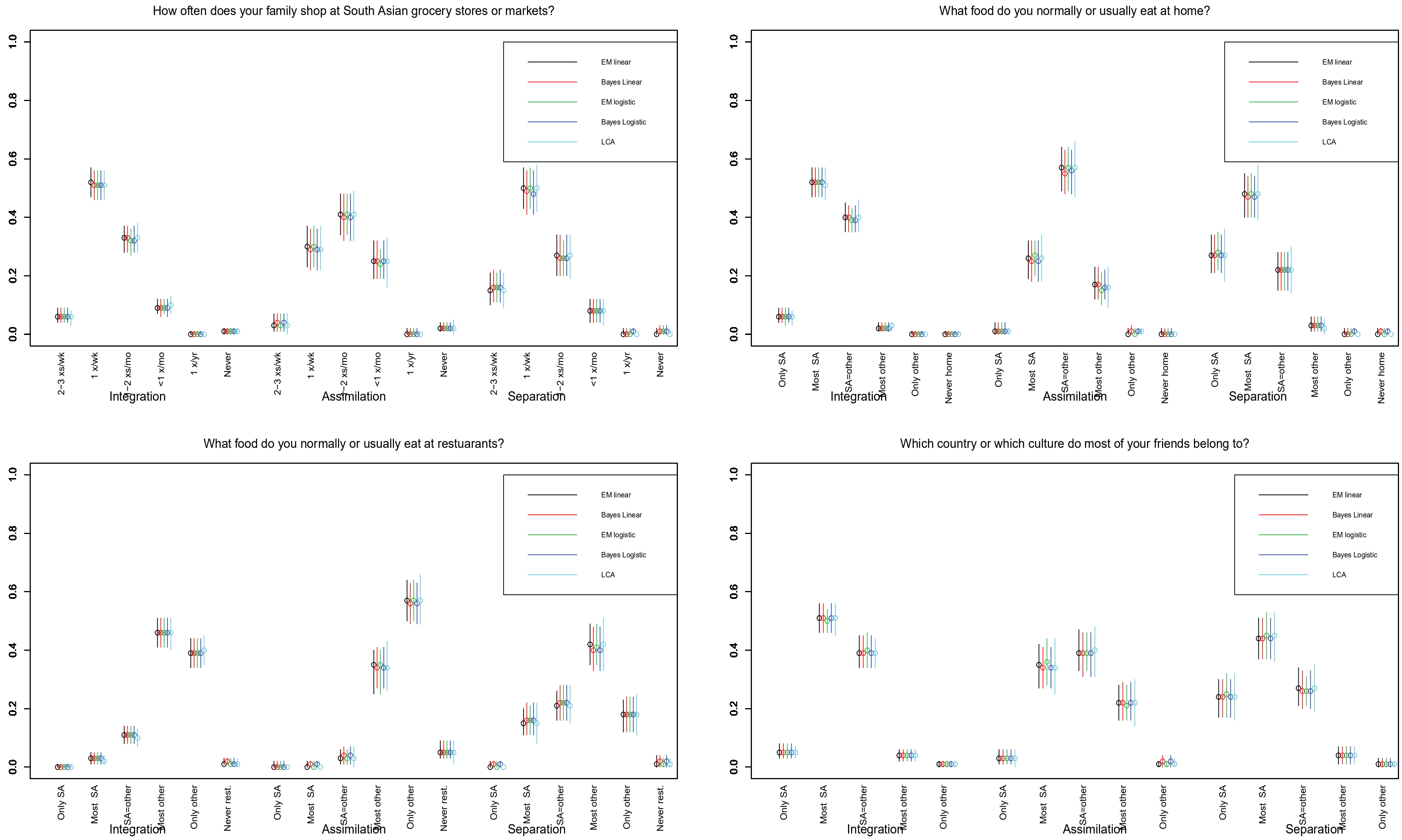
Distribution of manifest variable defining acculturation clusters for MASALA Analysis.



eFigure 12: Distribution of manifest variable defining acculturation clusters (with 95% confidence interval); LCA=latent class model alone (no inclusion in distal outcome model).



eFigure 13: Distribution of manifest variable defining acculturation clusters (with 95% confidence interval); LCA=latent class model alone (no inclusion in distal outcome model).



eFigure 14: Distribution of manifest variable defining acculturation clusters (with 95% confidence interval); LCA=latent class model alone (no inclusion in distal outcome model).