EAM2025 XI Conference

23RD - 25TH JULY 2025

Spain Tenerife Canary Islands

European Association of Methodology

Estimating Context Effects in Small Samples while Controlling for Covariates: An Optimally Regularized Bayesian Estimator for Multilevel Latent Variable Models

Dr. Valerii Dashuk, MSH Medical School Hamburg

Universidad de La Laguna

Cabildo de Tenerife

Tagoro

cajasiete

Gobierno de Canarias
Consejería de Universidades, Ciencia e Innovación y Cultura
Agencia Canaria de Investigación, Innovación y Sociedad de la Información

ICI Instituto Canario de Igualdad

tea

hogrefe

Agenda:

- Motivation
- Theory & Estimator
- Simulation Insights
- The **MLOB** R Package
- Example Application
- Discussion & Conclusion

Motivation

- Estimate between-group effects in multilevel data.
- Small samples or low ICCs pose challenges.
- Maximum likelihood can be unstable or biased.
- We propose an MSE-optimal regularized Bayesian estimator.
- Supports covariates for realistic modelling.
- Available on CRAN via the MLOB R package.

Theory: Univariate Model Estimation

Start from the multilevel latent variable models as denoted by Zitzmann et al. (2020, 2021):

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_w X_{w,ij} + \varepsilon_{ij} \quad (1)$$

$$\text{Level 2: } \beta_{0j} = \alpha + \beta_b X_{b,j} + \delta_j \quad (2)$$

Maximum likelihood (ML): vs. Regularized Bayesian:

$$\hat{\beta}_{b,\mathrm{ML}} = \frac{\hat{\tau}_{YX}}{\hat{\tau}_X^2} \quad (3) \qquad \hat{\beta}_{b,\mathrm{B}} = \frac{\hat{\tau}_{YX}}{(1-\omega)\tau_0^2 + \omega\hat{\tau}_X^2} \quad (4)$$

Theory: Extended Model Estimation

Introduce covariates into the multivariate latent variable model

(Dashuk et al., 2025):

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_w X_{w,ij} + C_{w,ij}\gamma + \varepsilon_{ij} \quad (5)$$

$$\text{Level 2: } \beta_{0j} = \alpha + \beta_b X_{b,j} + C_{b,j}\gamma + \delta_j \quad (6)$$

Extended regularized Bayesian estimator (more details):

$$\tilde{\beta}_b = \frac{\hat{\tau}_{YX} - \gamma'\tau_{CX}}{(1-\omega)\tau_0^2 + \omega\hat{\tau}_X^2} \quad (7)$$

Theory: Advantages of New Estimator

- Data-driven prior selection
  - Prior parameters ω and τ_0^2 automatically chosen to minimize MSE.
- Incorporates covariates
  - Allows for more realistic and flexible models.
- Bias–variance trade-off
  - Accepts small bias to achieve lower variance and overall MSE.
- Robust in challenging settings
  - Optimized performance in small samples and low ICCs.

Simulation. Data-Generating Process

540 cases covered with 5000 replications each:

- ICC_X: Intra-Class Correlation, {0.05, 0.1, 0.3, 0.5}.
- J: Number of groups, {5, 10, 20, 30, 40}.
- n: Number of individuals, {5, 15, 30}.
- k: Number of covariates, {1}, with γ: true covariate parameter, {0.3}.
- β_b: True between-group parameter {0.2, 0.5, 0.6}.
- β_w: True within-group parameter {0.2, 0.5, 0.7}.

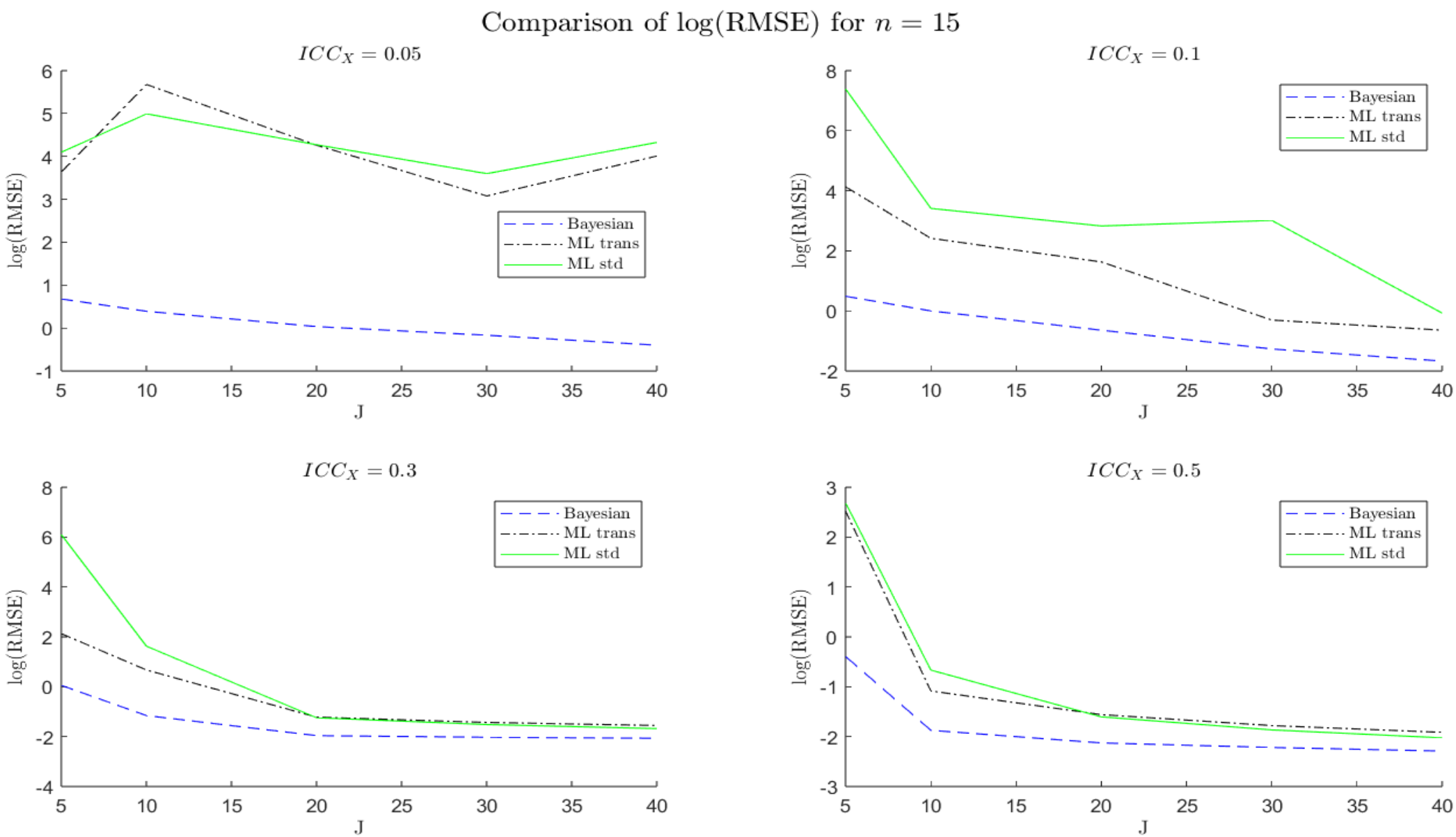
Simulation. Methods and Quality Evaluation

We estimated the generated data with 3 estimators:

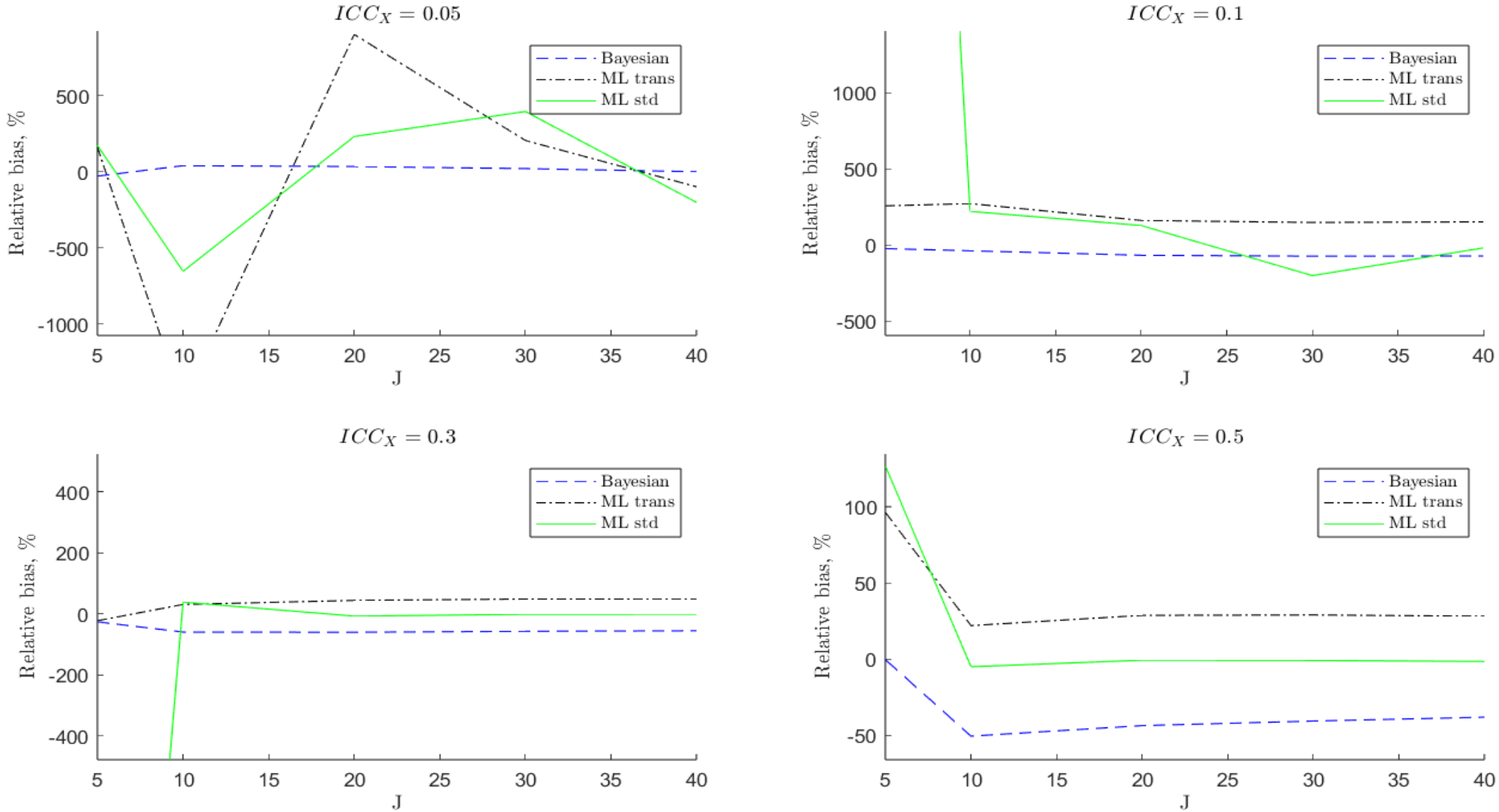
- Extended regularized Bayesian estimator.
- Standard ML estimator.
- Transformed ML estimator, applied to the transformed model.

Quality of estimators is measured with:

- Root mean squared error (RMSE).
- Coverage rate (CR).
- Relative bias (RB).
- Standard error ratio (SER).


Comparison of log(RMSE) for $n = 15$
$ICC_X = 0.05$
$ICC_X = 0.1$
$ICC_X = 0.3$
$ICC_X = 0.5$
log(RMSE)
J
Bayesian
ML trans
ML std

Comparison of relative bias in % for $n = 15$



MLOB Package: Key Features

To empower applied researchers use our estimator, we developed **MLOB** package in R:

- Estimates between-group effects using the extended regularized Bayesian approach.
- Handles unbalanced designs by effectively balancing it.
- Offers delete-d jackknife standard errors.
- Returns user-friendly output: estimates, SEs, CIs, Z-values, p-values.

MLOB Package: Usage in R

```
# Install

install.packages("MultiLevelOptimalBayes")

# Load

library(MultiLevelOptimalBayes)

# Fit model

result < - mlob(formula, data, group, balancing.limit, conf.level, jackknife,
punish.coeff)

summary(result)
```

Note: Dev. version of the MLOB package is available here:
https://github.com/MLOB-dev/MLOB

Real-World Application: Estimation

PASSNYC Dataset (1,272 schools, 32 Districts) – Estimating Math Proficiency

(https://www.kaggle.com/datasets/passnyc/data-science-for-good/data):

```
# Load and clean data

data <- read.table("2016 School Explorer.csv", sep=',', header=TRUE)

data_s <- data[data$Average.Math.Proficiency != 'N/A',]

data_s$math <- as.numeric(data_s$Average.Math.Proficiency)

data_s$ELA <- as.numeric(data_s$Average.ELA.Proficiency)

data_s$ENI <- as.numeric(data_s$Economic.Need.Index)

# Run extended regularized Bayesian model

result <- mlob(math ~ ENI + ELA + ENI:ELA, data = data_s, group = 'District’,
balancing.limit = 0.35)
```

Real-World Application: Results

```
➢ summary(result)

Call: mlob(math ~ ENI + ELA + ENI:ELA, data = data_subset, group = District, balancing.limit = 0.35)

Summary of Coefficients:

             Estimate Std. Error  Lower CI (95%) Upper CI (95%)  Z value   Pr(>|z|)    Significance
beta_b       -0.1169    0.01868    -0.1535        -0.0802        -6.2562   3.94e-10         ***
gamma_ELA     1.1324    0.21136     0.7182         1.5467         5.3580   8.41e-08         ***
gamma_ENI:ELA 0.3369    0.57599    -0.7920         1.4658         0.5849   5.58e-01

For comparison, summary of coefficients from unoptimized analysis (ML):

             Estimate Std. Error  Lower CI (95%) Upper CI (95%)  Z value   Pr(>|z|)    Significance
beta_b       -0.7721    0.3323     -1.4235        -0.1207        -2.3233   2.01e-02          *
gamma_ELA     1.1324    0.2113      0.7182         1.5467         5.3580   8.41e-08         ***
gamma_ENI:ELA 0.3369    0.5759     -0.7920         1.4658         0.5849   5.58e-01

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Limitations & Future Steps

Limitations:

- Advantage over ML diminishes with increasing sample size.
- Small risk of model misspecification under extreme conditions.
- Current model assumes two-level grouping only.

Future steps:

- Extend the model to more than two levels.
- Extend the model to handle directly unbalanced data.
- Extend the model to relax distributional assumptions.

Conclusion

- Introduced an **extended regularized Bayesian estimator** for between-group effects of multilevel latent variable model with covariates.
- Estimator is MSE-optimal: balances bias and variance.
- Handles small samples, low ICCs, and unbalanced data.
- Implemented in the **MLOB** R package with user-friendly interface.
- Validated with real-world data.
- **Try MLOB on your own data via CRAN!**

Sources & Funding

- Dashuk, V., Hecht, M., Lüdtke, O., Robitzsch, A. & Zitzmann, S. (2024). *An Optimally Regularized Estimator of Multilevel Latent Variable Models, with Improved MSE Performance*.
- Dashuk, V., Hecht, M., Lüdtke, O., Robitzsch, A. & Zitzmann, S. (2025). *Estimating Context Effects in Small Samples while Controlling for Covariates: An Optimally Regularized Bayesian Estimator for Multilevel Latent Variable Models*.
- Dashuk, V., Timilsina, B., Hecht, M. & Zitzmann, S. (2025). *MultiLevelOptimalBayes (MLOB): An R package for Regularized Bayesian Estimation of Multilevel Latent Variable Models with Covariates*.

- Funding: Funded by DFG (Project No. 471861494).

Thank you!

We invite and welcome your esteemed participation for any inquiries, reflections, and questions.

Contact: valerii.dashuk@research-development-innovation.de

Extended Regularized Bayesian Estimator 1

Model from Dashuk et al. (2025):

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_w X_{w,ij} + C_{w,ij}\gamma + \varepsilon_{ij} \quad (8)$$

$$\text{Level 2: } \beta_{0j} = \alpha + \beta_b X_{b,j} + C_{b,j}\gamma + \delta_j \quad (9)$$

Regress covariates C_ij on $X\_(b,j)$ and $X\_(w,ij)$:

$$\hat{C}_{ij,k} = \hat{\phi}_0 + \hat{\phi}_1 X_{b,j} + \hat{\phi}_2 X_{w,ij} \quad (10)$$

Compute residuals:

$$\tilde{C}_{ij,k} = C_{ij,k} - \hat{C}_{ij,k} \quad (11)$$

Extended Regularized Bayesian Estimator 2

Regress dependent variable Y_{ij} on all k residuals $\tilde{C}_{ij}$:

$$\hat{Y}_{ij} = \tilde{C}_{ij}\,\hat{\gamma} \tag{12}$$

Define $\tilde{Y}_{ij}$ as the difference:

$$\tilde{Y}_{ij} = Y_{ij} - \hat{Y}_{ij} \tag{13}$$

In the variable $\tilde{Y}_{ij}$ there is no variability associated with the covariates.

Substitute it for Y_{ij} in the two-level model:

$$\text{Level 1: } \tilde{Y}_{ij} = \beta_{0j} + \beta_w X_{w,ij} + \varepsilon_{ij} \tag{14}$$

$$\text{Level 2: } \beta_{0j} = \alpha + \beta_b X_{b,j} + \delta_j \tag{15}$$

Extended Regularized Bayesian Estimator 3

The model compress to the univariate case, therefore apply the regularized Bayesian estimator for the univariate model:

$$\tilde{\beta}_{b,\mathrm{B}} = \frac{\hat{\tau}_{\tilde{Y}X}}{(1-\omega)\tau_0^2 + \omega\hat{\tau}_X^2} = \frac{\hat{\tau}_{YX} - \gamma'\tau_{CX}}{(1-\omega)\tau_0^2 + \omega\hat{\tau}_X^2} \tag{16}$$

The estimator is derived, but how to choose priors ω and τ_0^2 to minimize MSE?

$$\mathrm{MSE}(\omega, \tau_0^2) = \mathrm{Var}\left(\tilde{\beta}_b(\omega, \tau_0^2)\right) + \left(\mathrm{Bias}\left(\tilde{\beta}_b(\omega, \tau_0^2)\right)\right)^2 \tag{17}$$

Extended Regularized Bayesian Estimator 4

Regularized Bayesian estimator expressed as F-distributed random variable:

$$\frac{\kappa_B(\omega,\tau_0^2)\theta_B(\omega,\tau_0^2)}{\kappa_2\theta_2}\,\tilde{\beta}_b \sim F(2\kappa_2, 2\kappa_B(\omega,\tau_0^2)) \qquad (18)$$

It depends on prior variance τ_0^2 and weight ω.

Priors are optimized by minimizing the MSE, expressed as a function of the prior parameters:

$$(\omega,\tau_0^2)_{optimal} = \arg\min_{(w,\tau_0^2)} \mathrm{MSE}(\tilde{\beta}_b\ (\omega,\tau_0^2)) \qquad (19)$$

■

←

Extended Regularized Bayesian Estimator 5

Sample covariances $\hat{\tau}_X^2$ and $\hat{\tau}_{\tilde{Y}X}$ can be calculated as:

$$\hat{\tau}_{\tilde{Y}X} = \frac{nJ-1}{(n-1)(J-1)J}\sum_{j=1}^{J}\bar{X}_j\,\bar{\tilde{Y}}_j - \frac{1}{n(n-1)J}\sum_{j=1}^{J}\sum_{i=1}^{n}X_{ij}\tilde{Y}_{ij} + \frac{J}{J-1}\,\bar{X}\,\bar{Y} \qquad (20)$$

$$\hat{\tau}_X^2 = \frac{nJ-1}{(n-1)(J-1)J}\sum_{j=1}^{J}\bar{X}_j{}^2 - \frac{1}{n(n-1)J}\sum_{j=1}^{J}\sum_{i=1}^{n}X_{ij}{}^2 + \frac{J}{J-1}\,\bar{X}^2 \qquad (21)$$

Extended Regularized Bayesian Estimator 6

The distributions of the sample covariances $\hat{\tau}_X^2$ and $\hat{\tau}_{\tilde{Y}X}$:

$$\hat{\tau}_X^2 \sim \text{Gamma}\left(\kappa_1 = \frac{\left(\sum_{i=1}^{nJ+J+1} \theta_{X,i}\right)^2}{2\sum_{i=1}^{nJ+J+1} \theta_{X,i}^2}, \theta_1 = \frac{\sum_{i=1}^{nJ+J+1} \theta_{X,i}^2}{\sum_{i=1}^{nJ+J+1} \theta_{X,i}}\right) \quad (22)$$

$$\hat{\tau}_{\tilde{Y}X} \sim \text{Gamma}\left(\kappa_2 = \frac{\left(\sum_{i=1}^{2(nJ+J+1)} \theta_{\tilde{Y}X,i}\right)^2}{2\sum_{i=1}^{2(nJ+J+1)} \theta_{\tilde{Y}X,i}^2}, \theta_2 = \frac{\sum_{i=1}^{2(nJ+J+1)} \theta_{\tilde{Y}X,i}^2}{\sum_{i=1}^{2(nJ+J+1)} \theta_{\tilde{Y}X,i}}\right) \quad (23)$$

With $\theta_{X,i}$ and $\theta_{\tilde{Y}X,i}$ - eigenvalues of transformed sums of $\hat{\tau}_X^2$ and $\hat{\tau}_{\tilde{Y}X}$

←

Extended Regularized Bayesian Estimator 7

The MSE of Bayesian estimation for between-group parameter β_b as a function of prior parameters τ_0^2 and ω:

$$MSE\ (\beta_b) = \frac{\kappa_2 * \theta_2^2 * (\kappa_B(\tau_0^2, \omega) + \kappa_2\ - 1)}{\theta_B^2(\tau_0^2, \omega)\ * (\kappa_B(\tau_0^2, \omega) - 1)^2 * (\kappa_B(\tau_0^2, \omega) - 2)} + \left(\frac{\kappa_2 * \theta_2}{(\kappa_B(\tau_0^2, \omega)\ - 1)\ * \theta_B(\tau_0^2, \omega)} - \beta_b\right)^2 \quad (24)$$

Equation (24) is further optimized for τ_0^2 and ω