

# AMBARTI models for agricultural experiments

Prado, E. B., Santos, A. A. L

*Hamilton Institute & Dept. of Mathematics and Statistics*



February 18th, 2021 - Group meeting

# Agenda

- ▶ Additive Main Effect interaction (AMMI) models
- ▶  $\text{AMMI} + \text{BART} = \text{AMBARTI}$
- ▶ Simulation (AMMI and AMBARTI)
- ▶ Next steps
- ▶ Appendix

# Additive Main effects and Multiplicative Interactions (AMMI)

Linear–bilinear models are frequently used to analyse two-way data such as genotype-by-environment data.

An example of this class of models is the AMMI model:

$$y_{ij} = \mu + g_i + e_j + \sum_{q=1}^Q \lambda_q \gamma_{iq} \delta_{jq} + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \tau^{-1}),$$

where  $g_i$  is the effect of genotype and  $e_j$  the effect of environment.

## Additive Main effects and Multiplicative Interactions (AMMI)

The following priors are assumed in its Bayesian version (Josse et al, JABES, 2014):

$$\mu \sim N\left(m, s_{\mu}^2\right),$$

$$g_i \sim N\left(0, s_g^2\right),$$

$$e_j \sim N\left(0, s_e^2\right),$$

$$(\lambda_q)_{q=1,\dots,Q} \sim \text{ordered sample of } Q \text{ independent } N^+\left(0, s_{\lambda}^2\right),$$

$$\gamma_{1q} \sim N^+(0, 1) \quad \text{for } q = 1, \dots, Q,$$

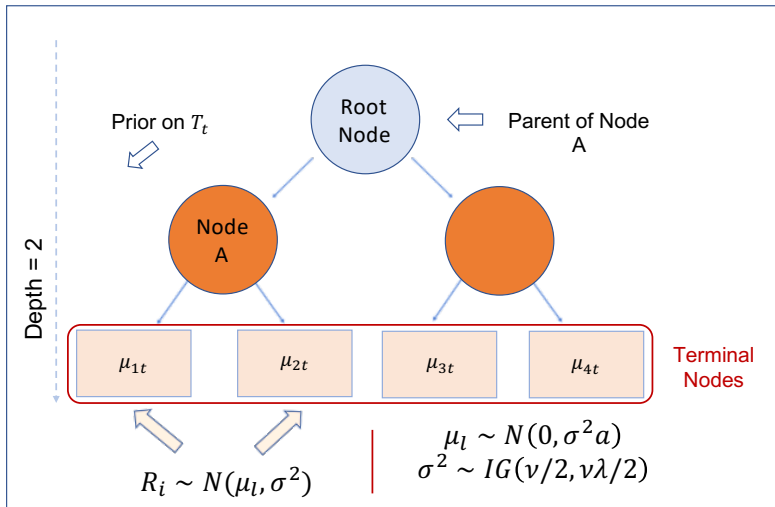
$$\gamma_{iq} \sim N(0, 1) \quad \text{for } i > 1 \text{ and } q = 1, \dots, Q,$$

$$\delta_{jq} \sim N(0, 1) \quad \text{for } j \geq 1 \text{ and } q = 1, \dots, Q,$$

$$\sigma_E \sim U\left(0, S_{ME}\right).$$

# Additive Main Effect Bayesian Additive Regression Tree Interaction models (AMBARTI)

# BART



## AMMI + BART

BART is a flexible tree-based method that can be used for predicting when there are interactions and non-linear relationships.

$$y_{ij}|\mathbf{x}_{ij}, \mathcal{T}, \mathcal{M}, \Theta, \sigma^2 \sim \text{N} \left( g_i + e_j + \sum_{t=1}^T h(\mathbf{x}_{ij}, \mathcal{M}_t, \mathcal{T}_t), \sigma^2 \right),$$

where  $y_{ij}$  is the yield for genotype  $i$  and environment  $j$ , and  $g_i$  and  $e_j$  are the genotype and environment effects, respectively.

$$\mu_{t\ell}|\mathcal{T}_t \sim \text{N}(\mu_\mu = 0, \sigma_\mu^2),$$

$$g_i|\mathcal{T}_t \sim \text{N}(\mu_g, \sigma_g^2),$$

$$e_j|\mathcal{T}_t \sim \text{N}(\mu_e, \sigma_e^2),$$

$$\sigma_g^2 \sim \text{IG}(a_g, b_g),$$

$$\sigma_e^2 \sim \text{IG}(a_e, b_e),$$

$$\sigma^2 \sim \text{IG}(a, b).$$

Interesting fact...



## Bayesian AMMI - Postprocessing (Josse et al, 2014)

Recall that

$$\mu_{ij} = \hat{y}_{ij} = \hat{\mu} + \hat{g}_i + \hat{e}_j + \sum_{q=1}^Q \hat{\lambda}_q \hat{\gamma}_{iq} \hat{\delta}_{jq}.$$

*"This means that, concretely,  $S$  matrices of size  $I \times J$  are available as draws from the posterior distributions of the  $\mu_{ij}$ . Thus, it is possible to **apply a postprocessing on each matrix** ( $s = 1, \dots, S$ ) performing the classical procedure (in accordance with the chosen constraints): each matrix is centered by row and by column, and an SVD is applied on the resulting matrix. Consequently, for each  $s$ , **new parameters ( $\mu, g, e, \gamma, \delta, \lambda_q$ ) meeting the constraints are available**. Consequently, draws in the posterior distribution of the parameters (taking the  $S$  new values) are available. Such a postprocessing makes it easier to interpret the results."*

## Simulation: AMMI scenarios (20 combinations)

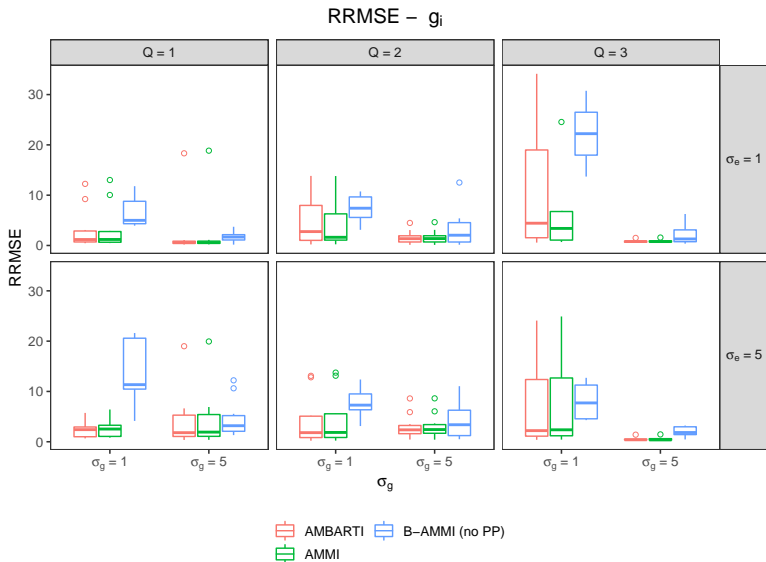
We consider the following setting to generate a set of simulated data sets:

$$y_{ij} = \mu + g_i + e_j + \sum_{q=1}^Q \lambda_q \gamma_{iq} \delta_{jq} + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \tau^{-1}),$$

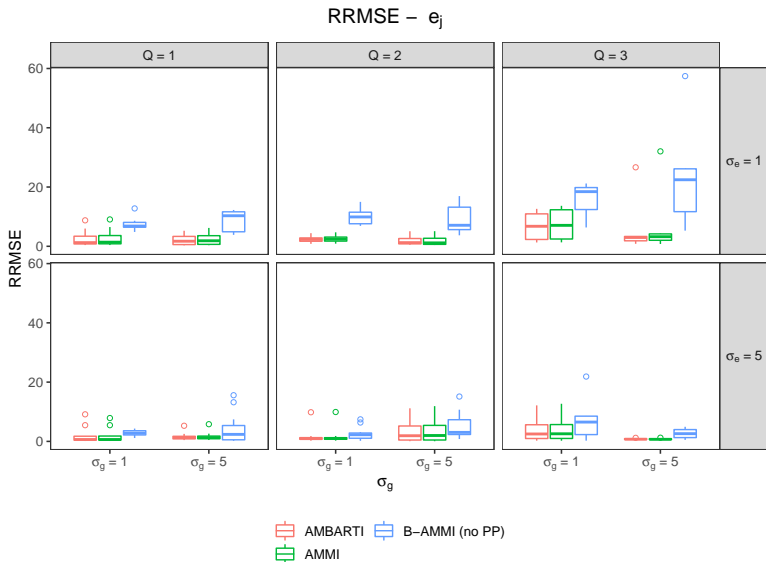
where

- ▶  $\mu = 100, \tau = 1$ .
- ▶  $I = J = 10$  (without repetitions).
- ▶  $e_j \sim \mathcal{N}(0, s_e)$ , with  $s_e = c(1, 5)$ .
- ▶  $g_i \sim \mathcal{N}(0, s_g)$ , with  $s_g = c(1, 5)$ .
- ▶  $Q = c(1, 2, 3)$ , with  $\lambda = c(8, 12, [8, 12], [10, 12], [8, 10, 12])$ .
- ▶ We used RMSE (Root Mean Squared Error) and RRMSE (Relative Root Mean Squared Error).

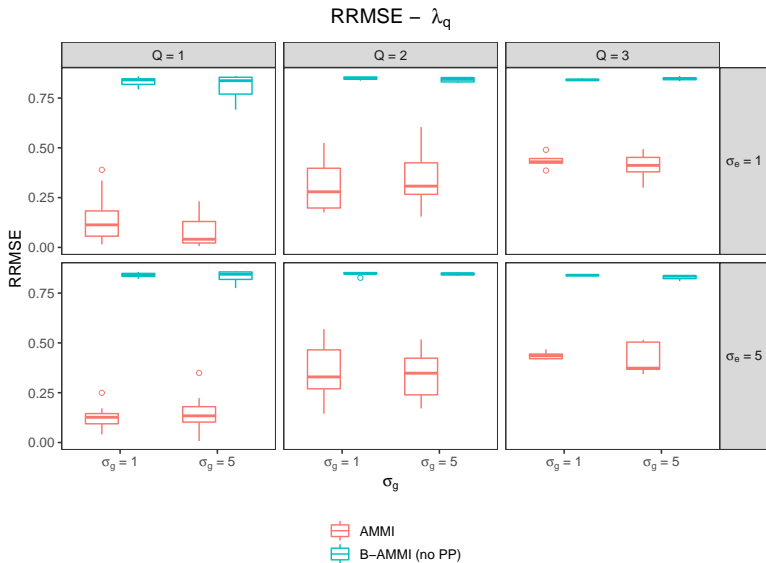
# Simulation AMMI ( $y_{ij} = \mu + \textcolor{red}{g}_i + e_j + \sum_{q=1}^Q \lambda_q \gamma_{iq} \delta_{jq}$ )



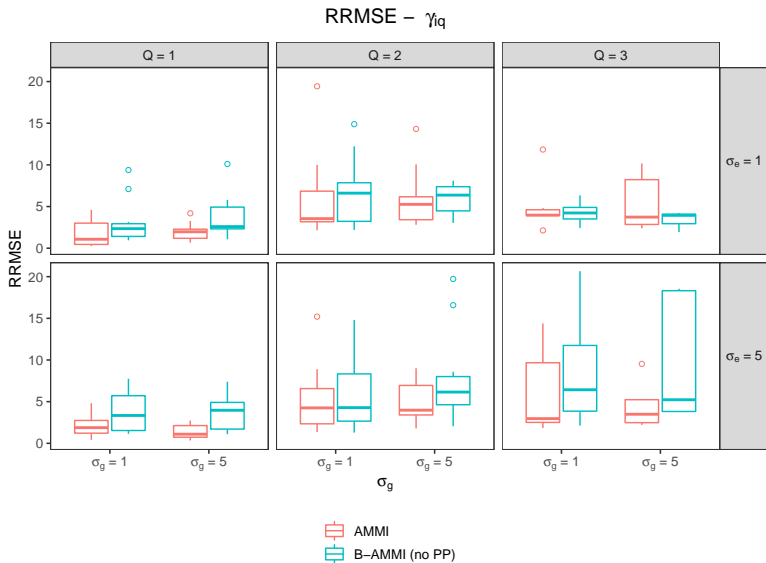
# Simulation AMMI ( $y_{ij} = \mu + g_i + e_j + \sum_{q=1}^Q \lambda_q \gamma_{iq} \delta_{jq}$ )



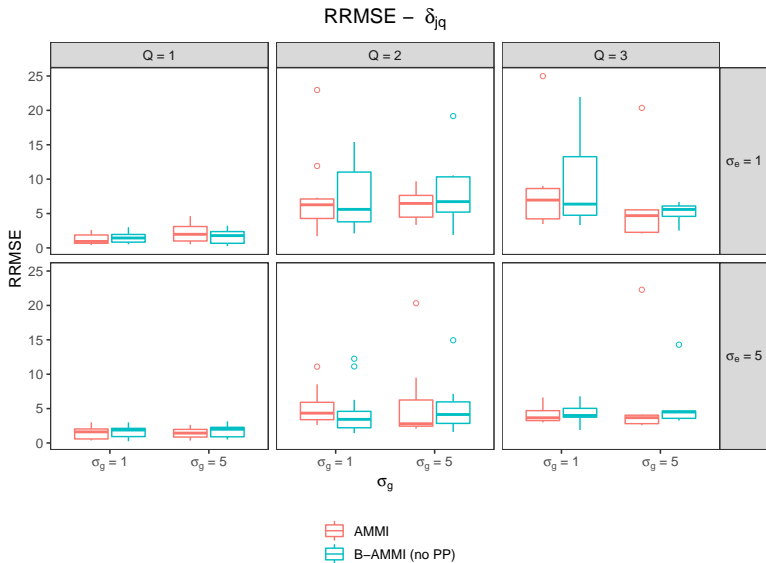
# Simulation AMMI ( $y_{ij} = \mu + g_i + e_j + \sum_{q=1}^Q \lambda_q \gamma_{iq} \delta_{jq}$ )



# Simulation AMMI ( $y_{ij} = \mu + g_i + e_j + \sum_{q=1}^Q \lambda_q \gamma_{iq} \delta_{jq}$ )

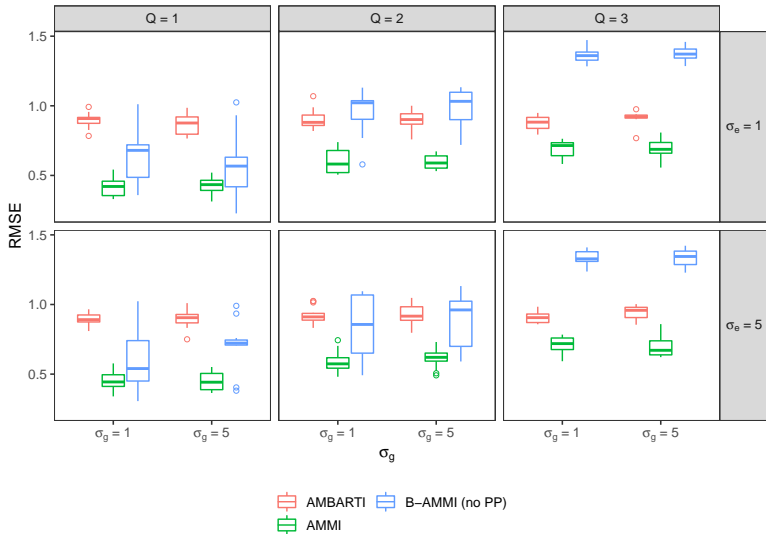


# Simulation AMMI ( $y_{ij} = \mu + g_i + e_j + \sum_{q=1}^Q \lambda_q \gamma_{iq} \delta_{jq}$ )



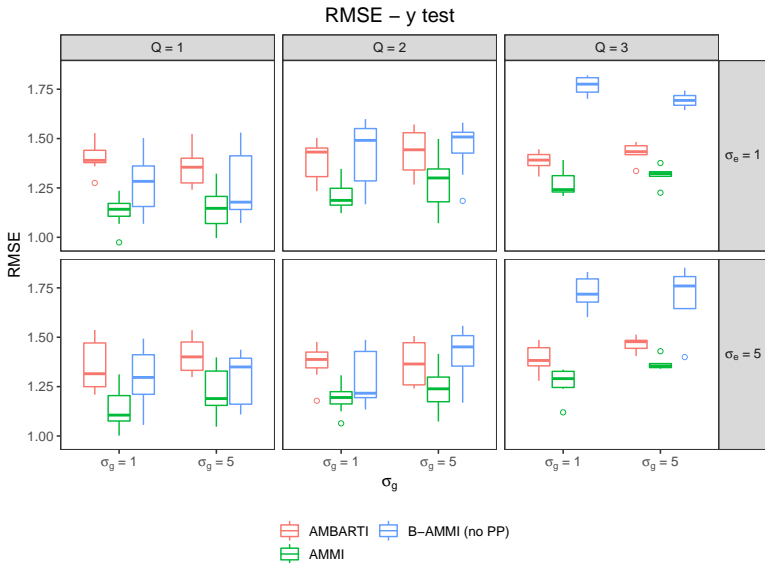
# Simulation AMMI ( $y_{ij} = \mu + g_i + e_j + \sum_{q=1}^Q \lambda_q \gamma_{iq} \delta_{jq}$ )

RMSE – Bilinear part





# Simulation AMMI ( $y_{ij} = \mu + g_i + e_j + \sum_{q=1}^Q \lambda_q \gamma_{iq} \delta_{jq}$ )



## Simulation AMBARTI (4 combinations)

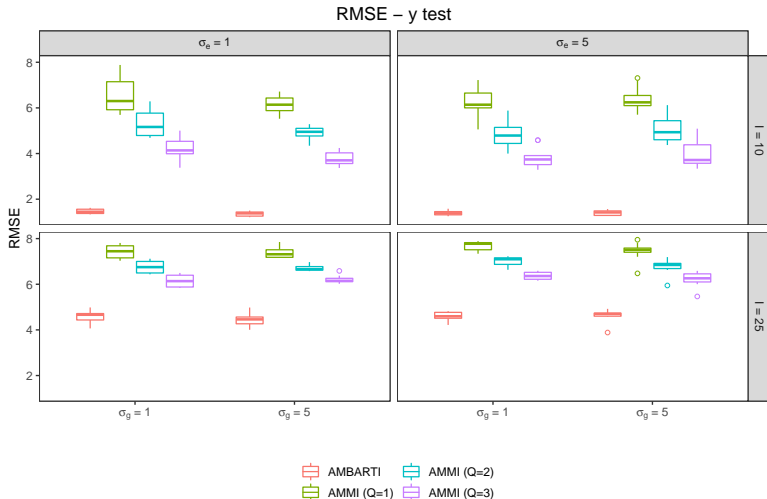
We consider the following setting to generate a set of simulated data sets:

$$y_{ij}|\mathbf{x}_{ij}, \mathcal{T}, \mathcal{M}, \Theta, \sigma^2 \sim \text{N} \left( g_i + e_j + \sum_{t=1}^T h(\mathbf{x}_{ij}, \mathcal{M}_t, \mathcal{T}_t), \sigma^2 \right),$$

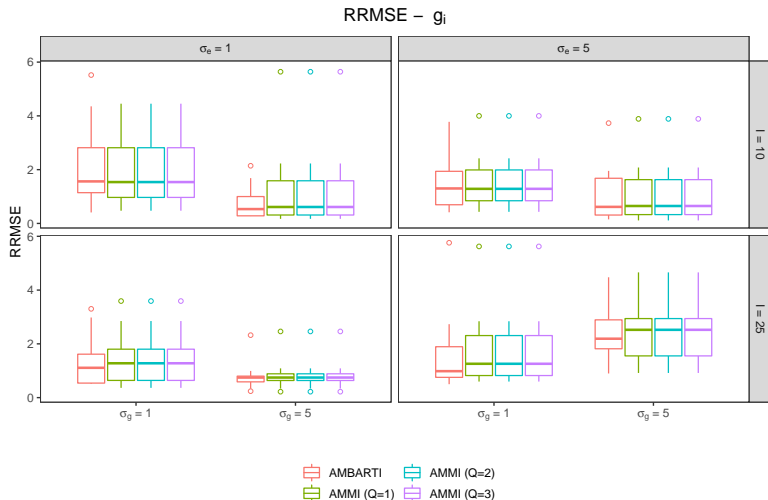
where  $y_{ij}$  is the yield for genotype  $i$  and environment  $j$ , and  $g_i$  and  $e_j$  are the genotype and environment effects, respectively.

- ▶  $\sigma^2 = 1$ ,  $T = 200$ .
- ▶  $I = J = 10$  (without repetitions).
- ▶  $\mu_{t\ell}|\mathcal{T}_t \sim \text{N}(\mu_\mu = 0, \sigma_\mu^2 = 3)$ ,
- ▶  $e_j \sim \text{N}(0, s_e)$ , with  $s_e = c(1, 5)$ .
- ▶  $g_i \sim \text{N}(0, s_g)$ , with  $s_g = c(1, 5)$ .

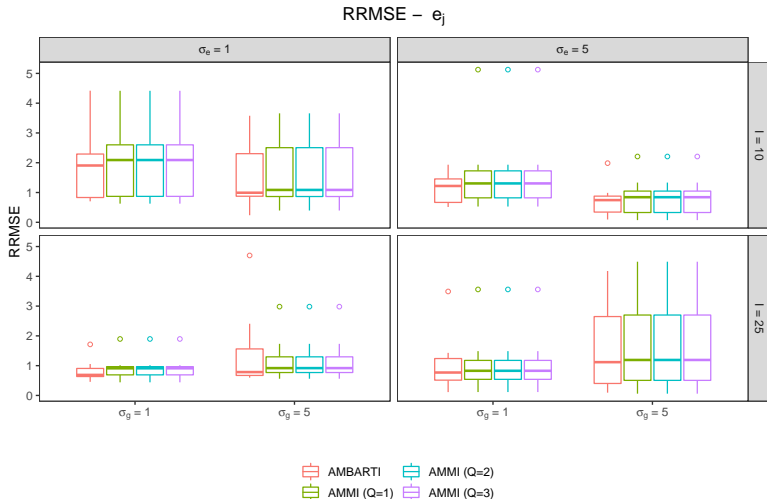
# Simulation AMBARTI ( $y_{ij} = g_i + e_j + \sum_{t=1}^T h(\mathbf{x}_{ij}, \mathcal{M}_t, \mathcal{T}_t)$ )



# Simulation AMBARTI ( $y_{ij} = \textcolor{red}{g}_i + e_j + \sum_{t=1}^T h(\mathbf{x}_{ij}, \mathcal{M}_t, \mathcal{T}_t)$ )



# Simulation AMBARTI ( $y_{ij} = g_i + e_j + \sum_{t=1}^T h(\mathbf{x}_{ij}, \mathcal{M}_t, \mathcal{T}_t)$ )



Real data

Innovar project: wheat data

## Next steps

1. Adapt our implementation to simulate large ( $> 15$ )  $I$  and  $J$ .
2. Perform more simulations considering large  $I$  and  $J$ .
3. Analyse real data sets.



That's all, folks! Thank you!

This work was supported by a Science Foundation Ireland Career Development Award grant number: 17/CDA/4695



## Appendix

## Simulation: full model (4 combinations)

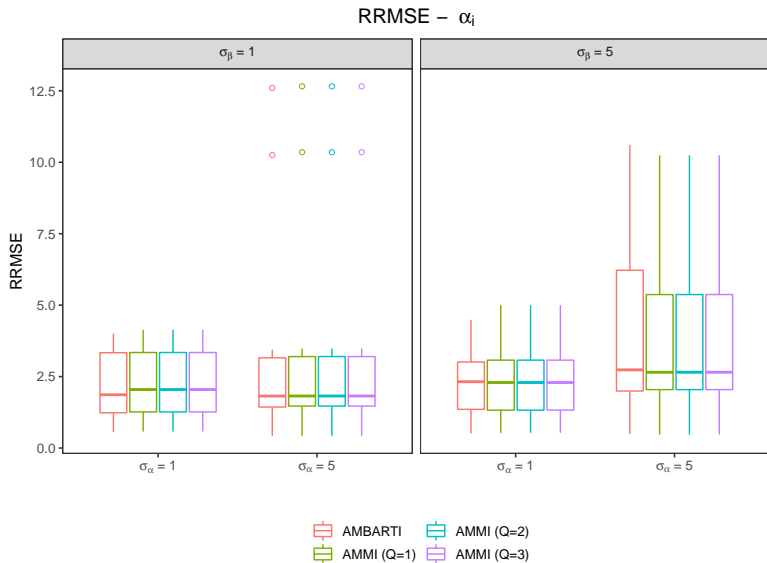
We consider the following setting to generate a set of simulated data sets:

$$y_{ij} = \mu + g_i + e_j + g_i \times e_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \tau^{-1}),$$

where

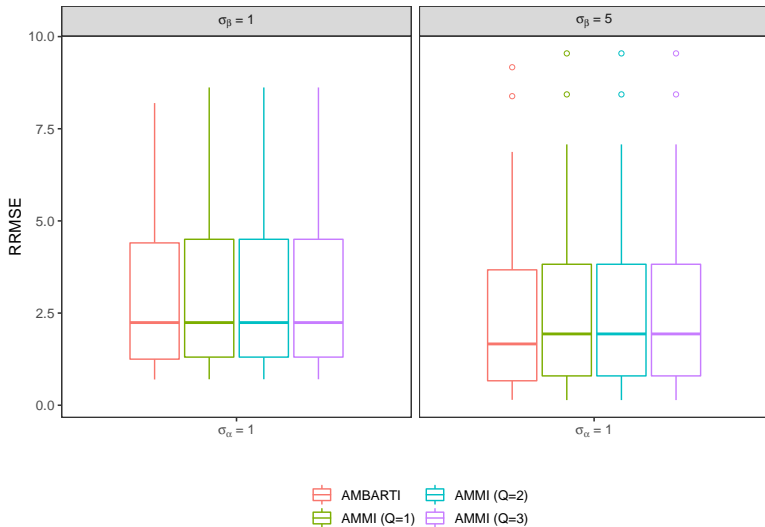
- ▶  $\mu = 100, \tau = 1$ .
- ▶  $I = J = 10$  (without repetitions).
- ▶  $e_j \sim \mathcal{N}(0, s_e)$ , with  $s_e = c(1, 5)$ .
- ▶  $g_i \sim \mathcal{N}(0, s_g)$ , with  $s_g = c(1, 5)$ .

# Simulation full model ( $y_{ij} = g_i + e_j + g_i \times e_j$ )



# Simulation full model ( $y_{ij} = g_i + e_j + g_i \times e_j$ )

RRMSE –  $\beta_j$



# Simulation full model ( $y_{ij} = g_i + e_j + g_i \times e_j$ )

