AMBARTI models for agricultural experiments

Prado, E. B., Santos, A. A. L
Hamilton Institute & Dept. of Mathematics and Statistics

February 23rd, 2021 - Group meeting

Agenda

- Additive Main Effect interaction (AMMI) models
- AMMI + BART = AMBARTI
- Simulation (AMMI and AMBARTI)
- Real data sets
- 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:

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

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

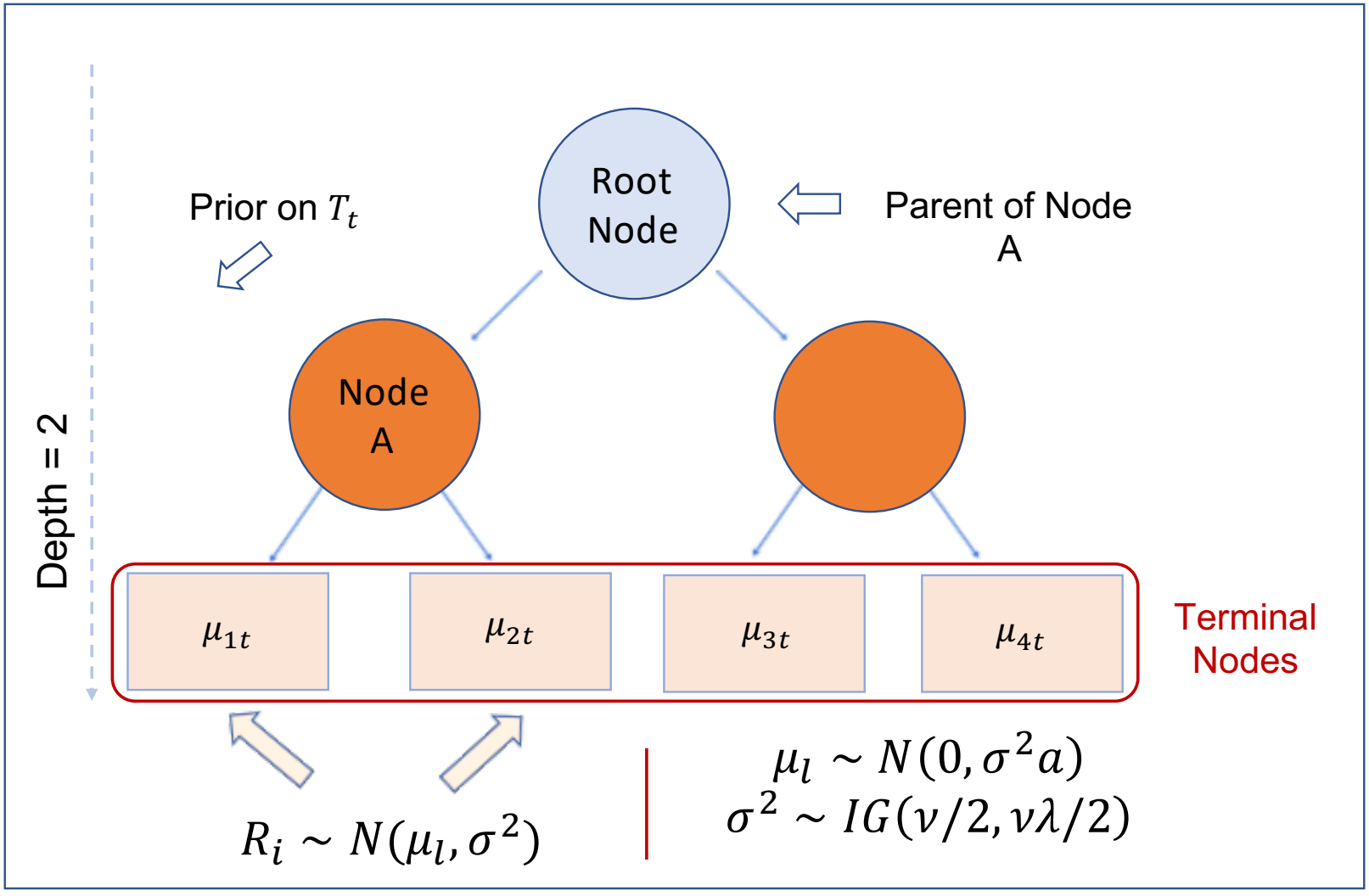
Additive Main effects and Multiplicative Interactions (AMMI)

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

$$
\begin{aligned}
\mu &\sim \mathrm{N}\left(m, s_\mu^2\right), \\
g_i &\sim \mathrm{N}\left(0, s_g^2\right), \\
e_j &\sim \mathrm{N}\left(0, s_e^2\right), \\
(\lambda_q)_{q=1,\ldots,Q} &\sim \text{ ordered sample of } Q \text{ independent } \mathrm{N}^+\left(0, s_\lambda^2\right), \\
\gamma_{1q} &\sim \mathrm{N}^+(0,1) \quad \text{for } q = 1, \ldots, Q, \\
\gamma_{iq} &\sim \mathrm{N}(0,1) \quad \text{for } i > 1 \text{ and } q = 1, \ldots, Q, \\
\delta_{jq} &\sim \mathrm{N}(0,1) \quad \text{for } j \geq 1 \text{ and } q = 1, \ldots, Q, \\
\sigma_E &\sim \mathrm{U}\left(0, S_{\mathrm{ME}}\right).
\end{aligned}
$$

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

BART


Prior on T_t
Root Node
Parent of Node A
Node A
Depth = 2
μ_{1t}
μ_{2t}
μ_{3t}
μ_{4t}
Terminal Nodes
$R_i \sim N(\mu_l, \sigma^2)$
$\mu_l \sim N(0, \sigma^2 a)$
$\sigma^2 \sim IG(\nu/2, \nu\lambda/2)$

AMMI + BART

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

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

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.

$$\begin{aligned}
\mu_{t\ell}|\mathcal{T}_t &\sim \mathsf{N}(\mu_\mu = 0, \sigma^2_\mu), \\
g_i|\mathcal{T}_t &\sim \mathsf{N}(\mu_g, \sigma^2_g), \\
e_j|\mathcal{T}_t &\sim \mathsf{N}(\mu_e, \sigma^2_e), \\
\sigma^2_g &\sim \mathsf{IG}(a_g, b_g), \\
\sigma^2_e &\sim \mathsf{IG}(a_e, b_e), \\
\sigma^2 &\sim \mathsf{IG}(a, b).
\end{aligned}$$

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 μ_{ij}. Thus, it is possible to apply a postprocessing on each matrix ($s = 1, ..., S$) 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 (μ, g, e, γ, δ, λ_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)

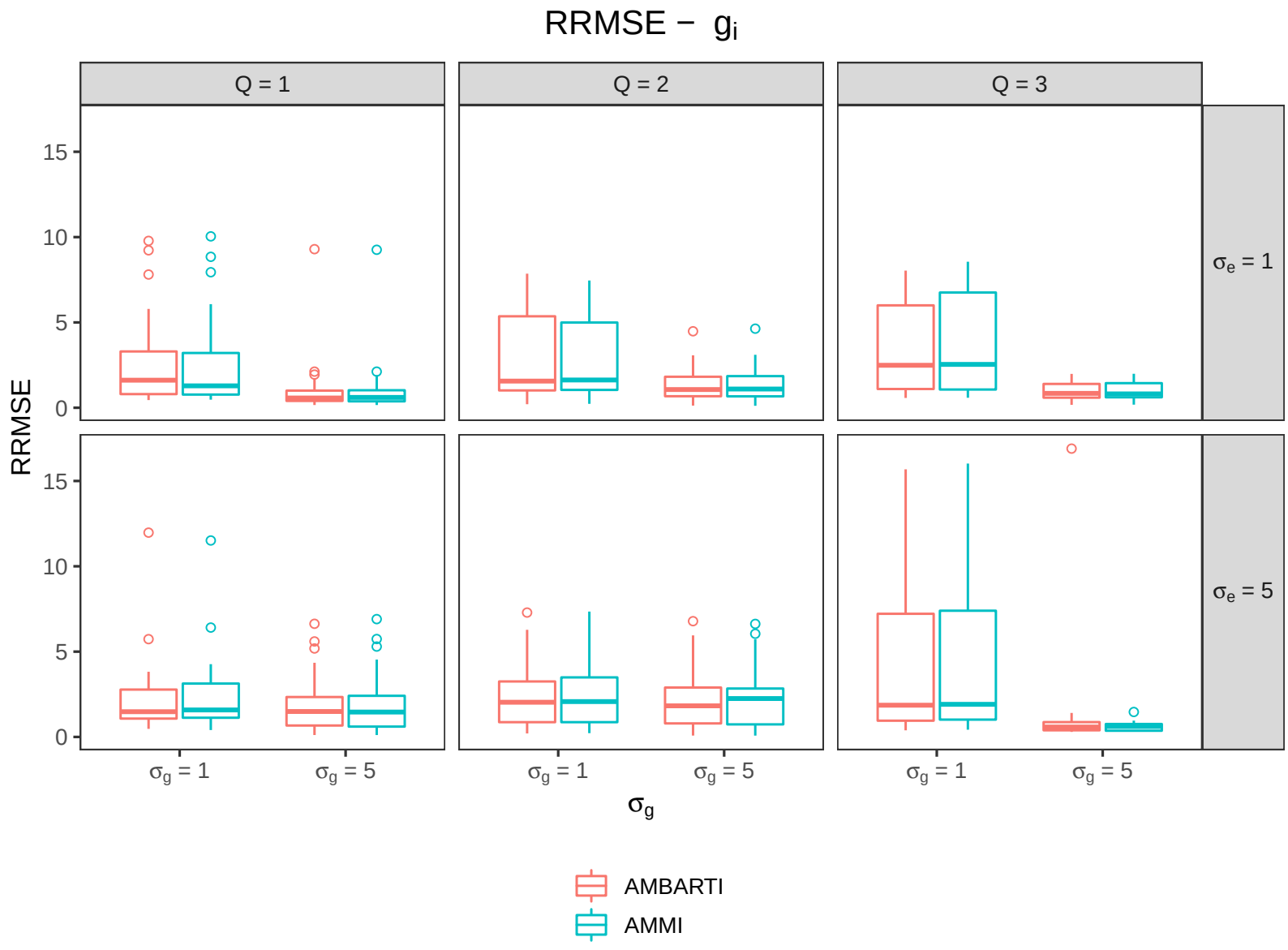
We consider the following setting to generate a set of simulated data sets:

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

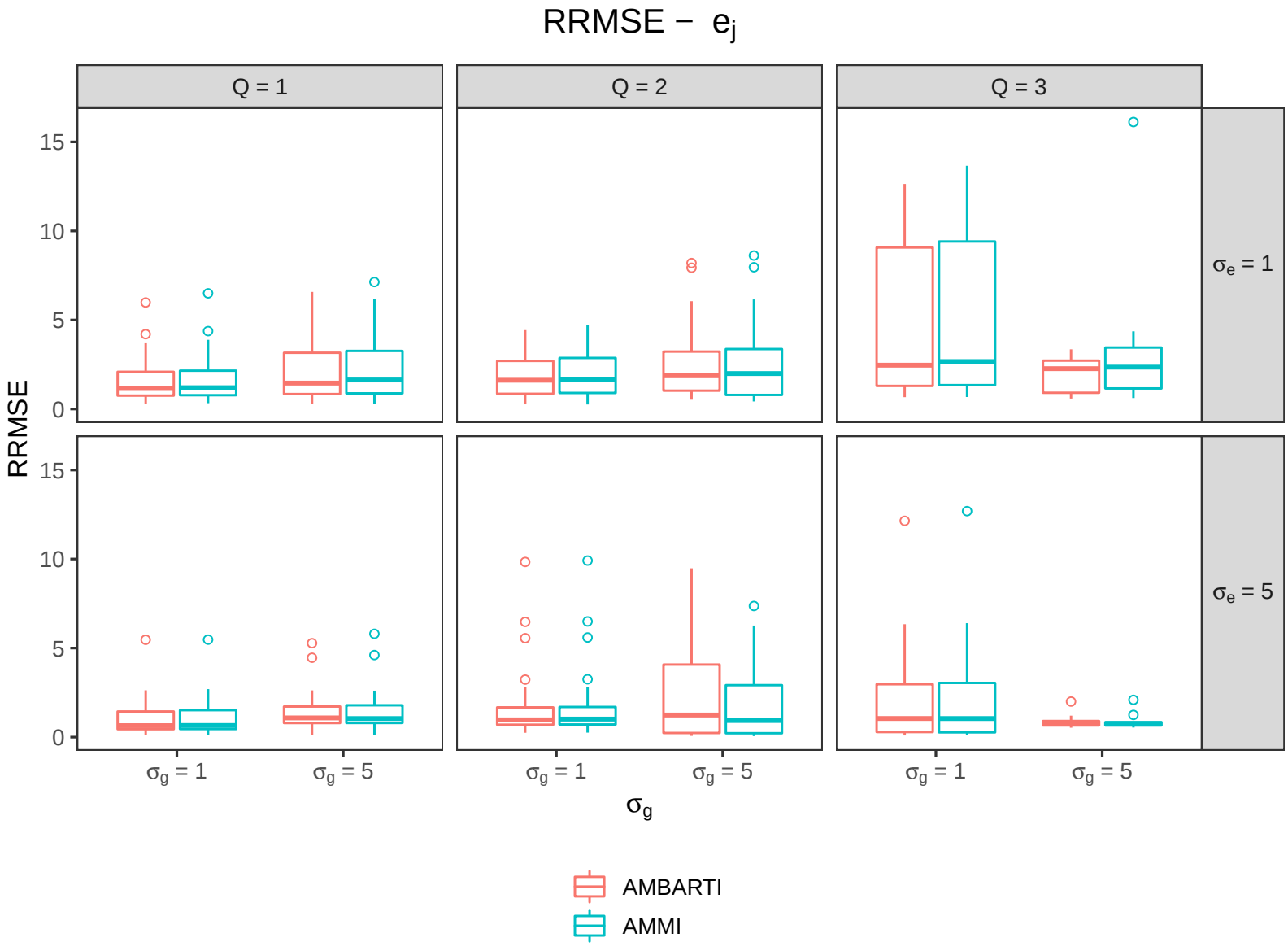
where

- $\mu = 100$, $\tau = 1$.
- $I = J = c(10, 25)$ (without repetitions).
- $e_j \sim \mathrm{N}(0, s_e)$, with $s_e = c(1, 5)$.
- $g_i \sim \mathrm{N}(0, s_g)$, 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).

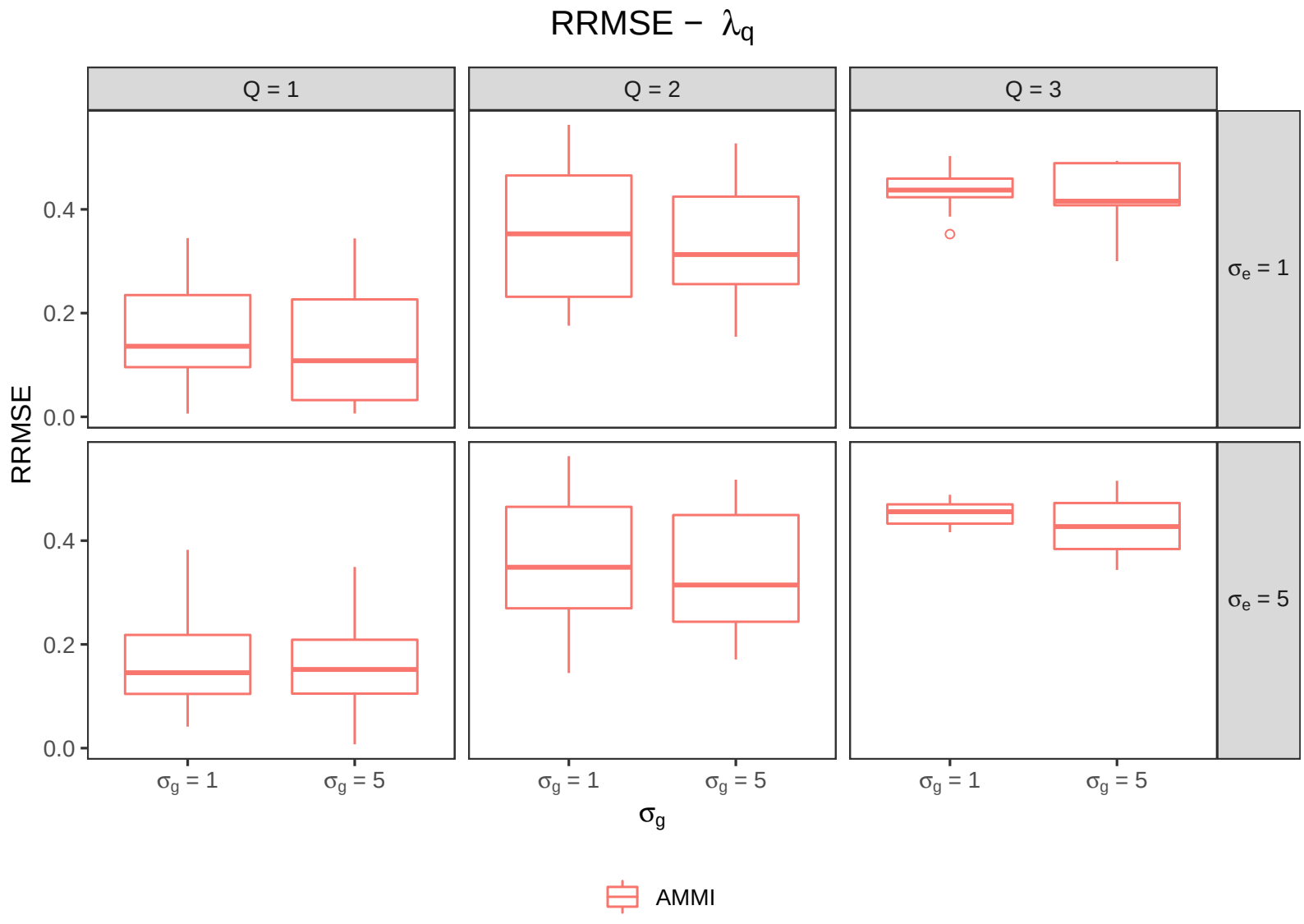
Simulation AMMI ($y_{ij} = \mu + g_i + e_j + \sum_{q=1}^{Q} \lambda_q \gamma_{iq} \delta_{jq}$)



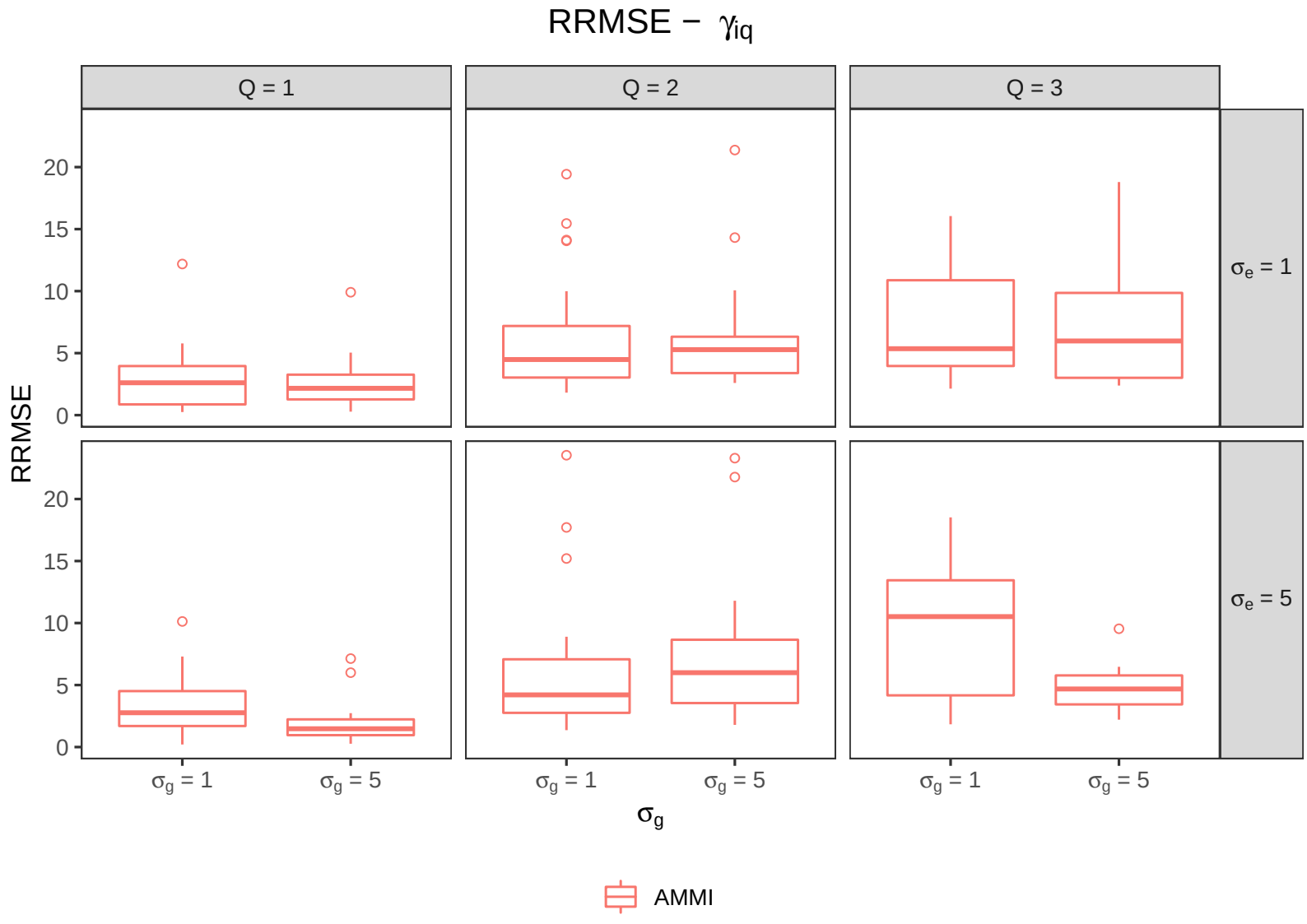
Simulation AMMI ($y_{ij} = \mu + g_i + e_j + \sum_{q=1}^{Q} \lambda_q \gamma_{iq} \delta_{jq}$)



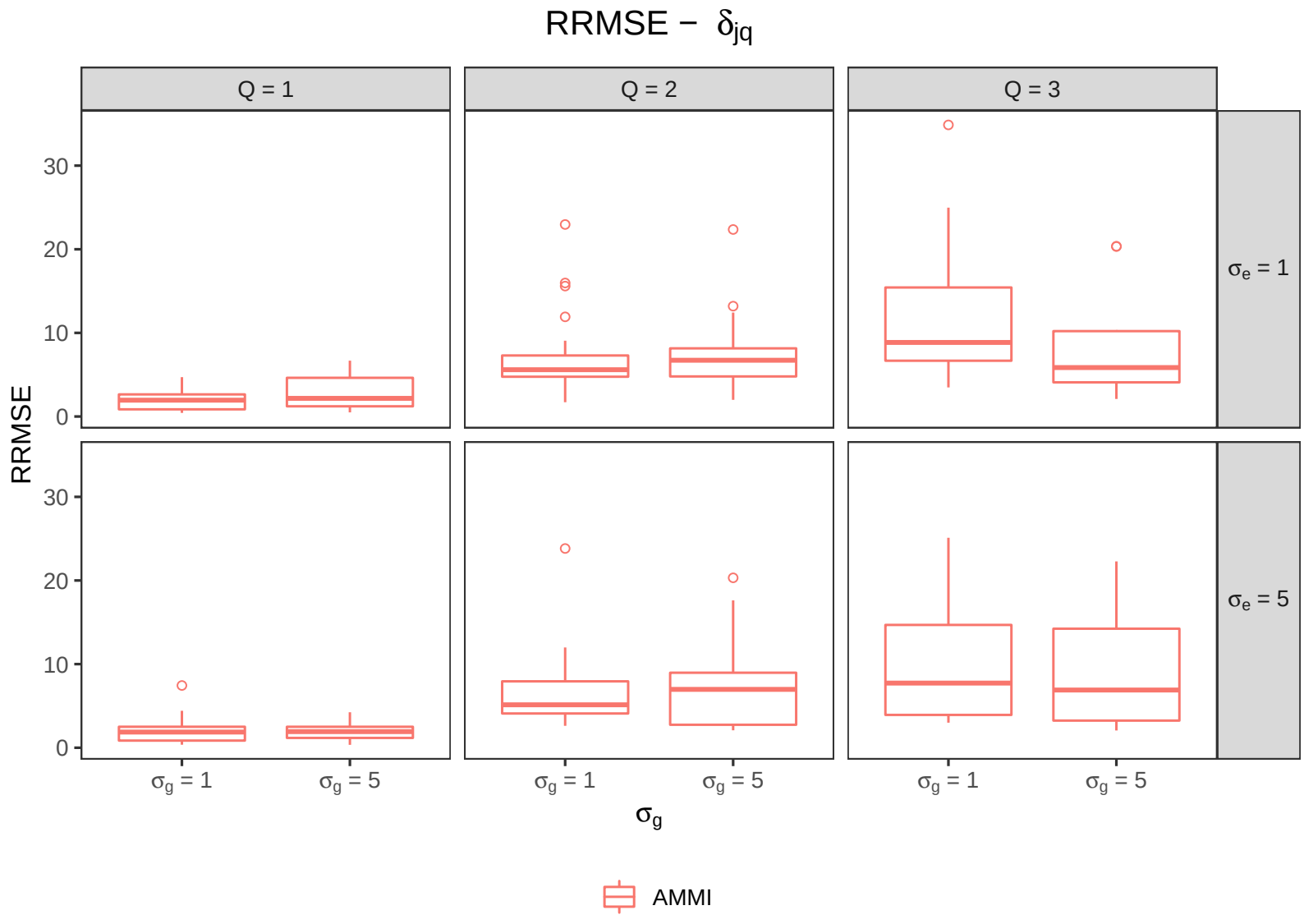
Simulation AMMI ($y_{ij} = \mu + g_i + e_j + \sum_{q=1}^{Q} \lambda_q \gamma_{iq} \delta_{jq}$)



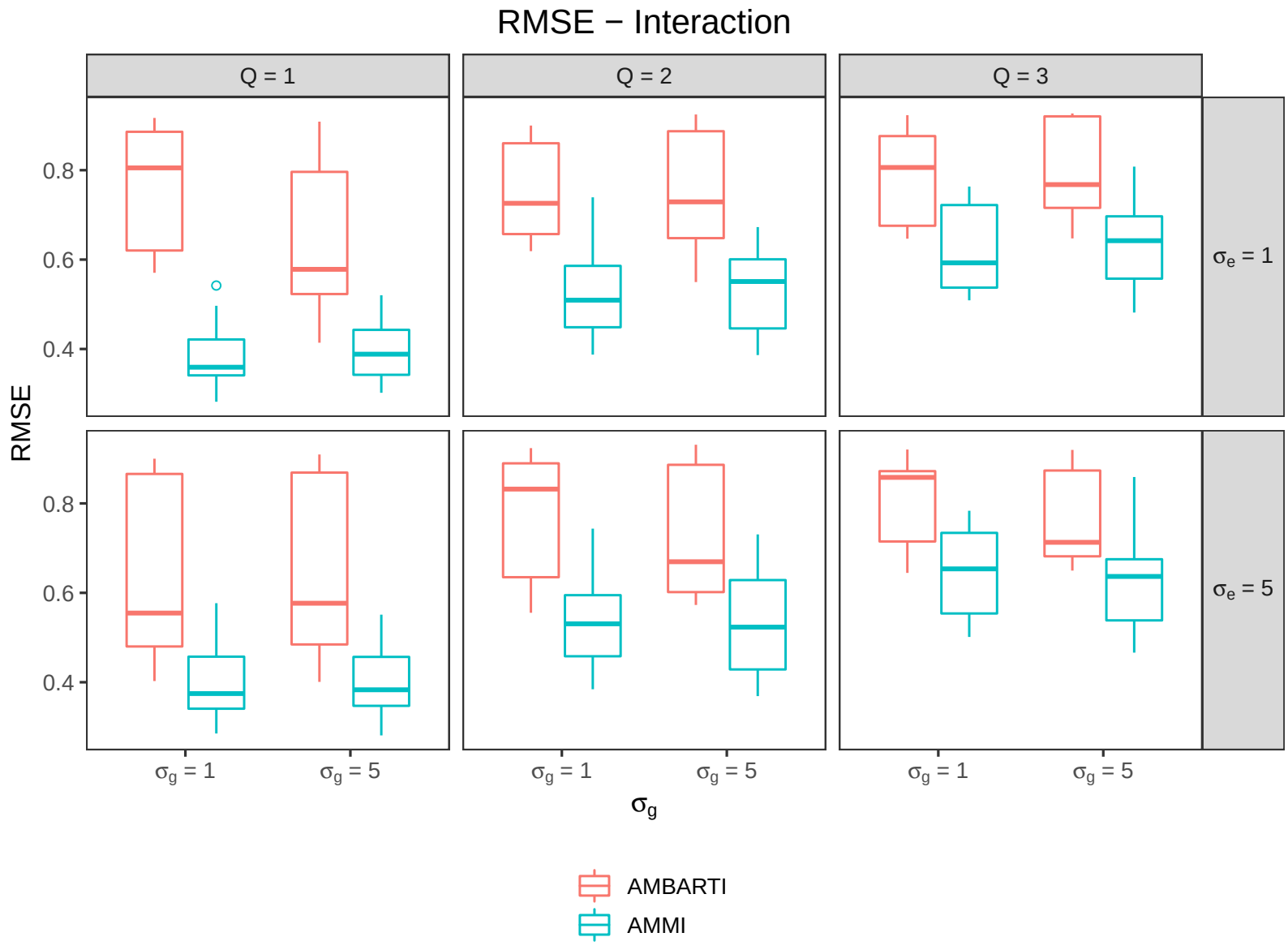
Simulation AMMI ($y_{ij} = \mu + g_i + e_j + \sum_{q=1}^{Q} \lambda_q \gamma_{iq} \delta_{jq}$)



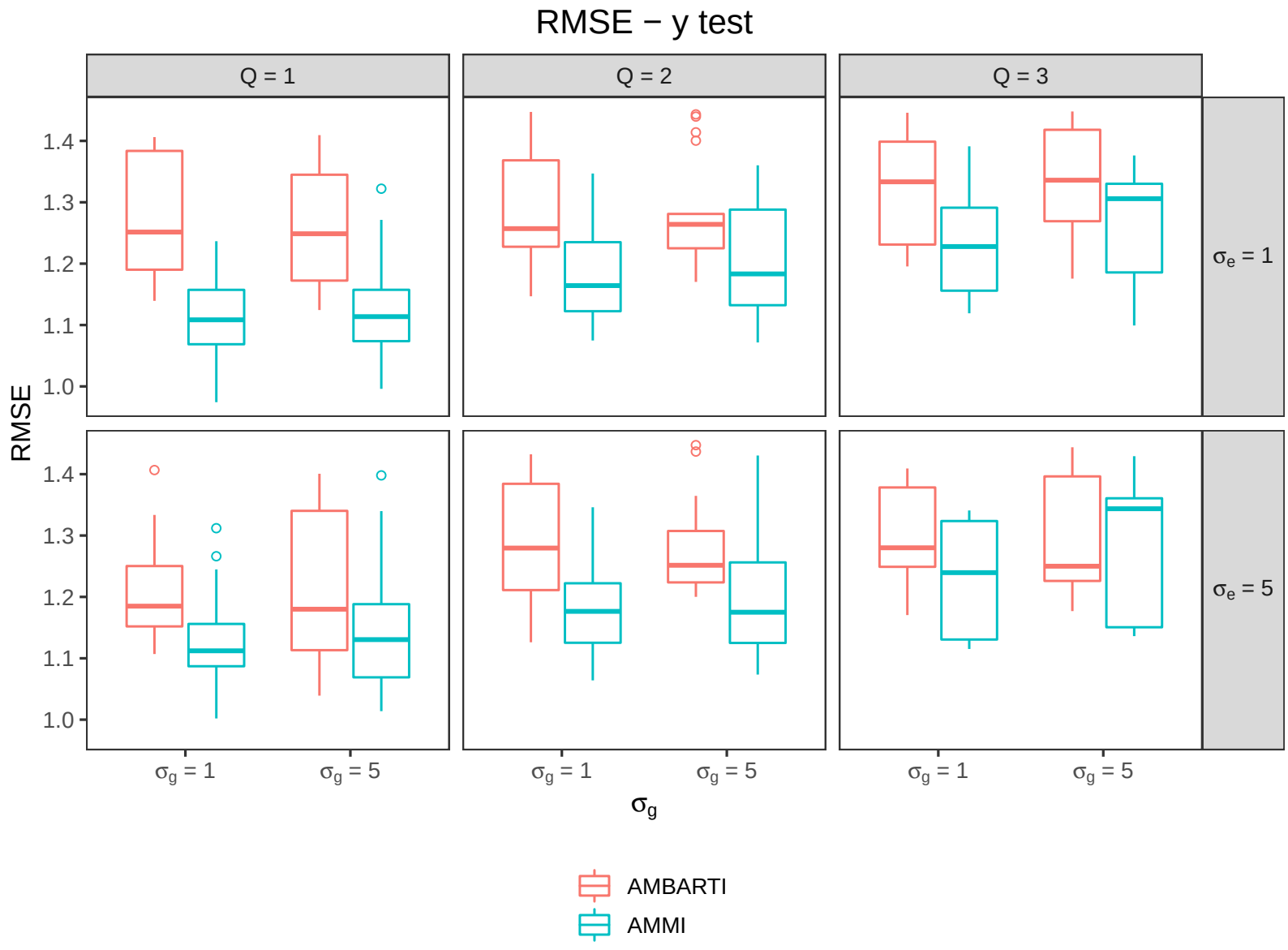
Simulation AMMI ($y_{ij} = \mu + g_i + e_j + \sum_{q=1}^{Q} \lambda_q \gamma_{iq} \delta_{jq}$)



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



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



Simulation AMBARTI (8 combinations)

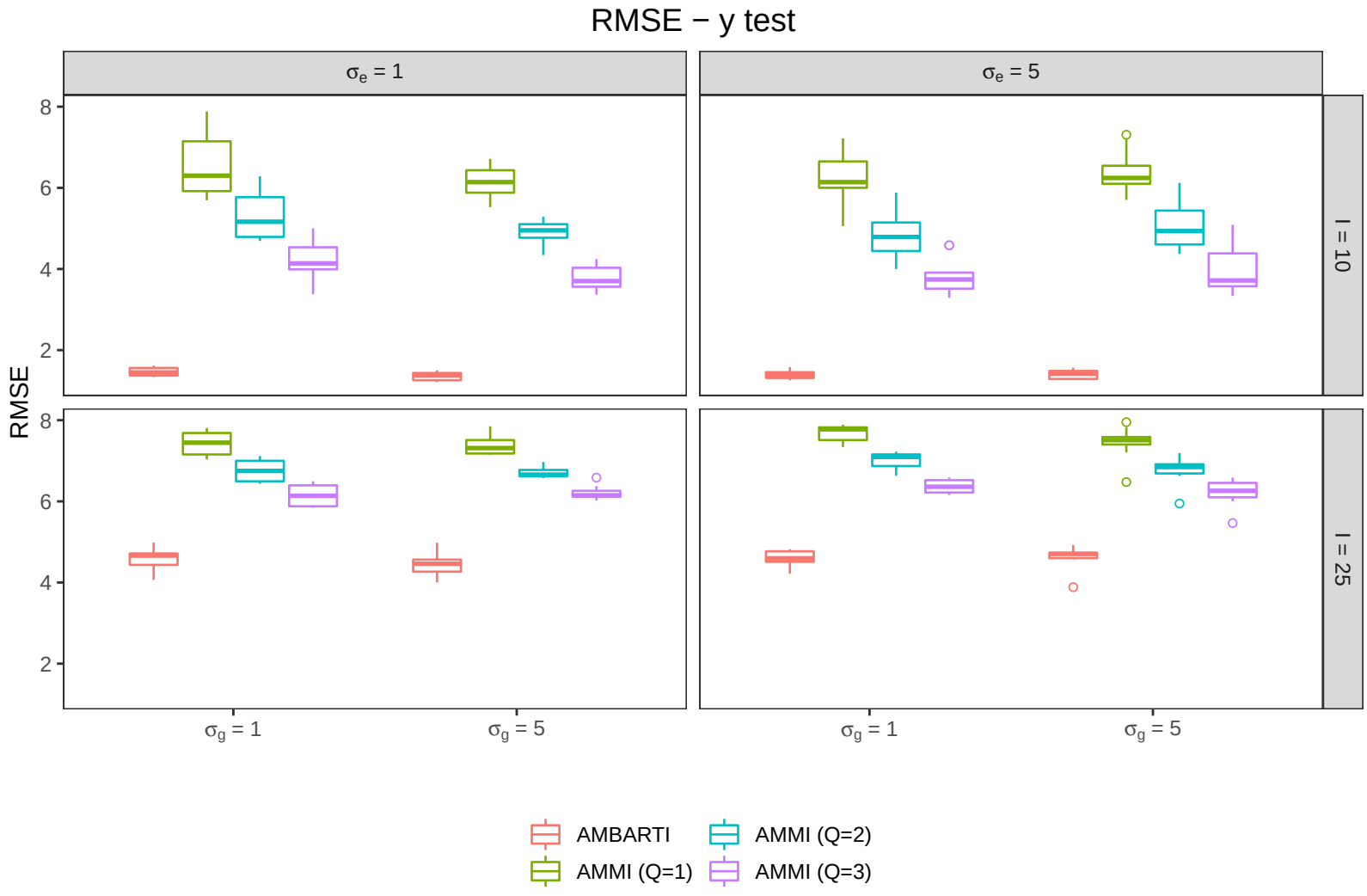
We consider the following settings to generate a set of simulated data sets:

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

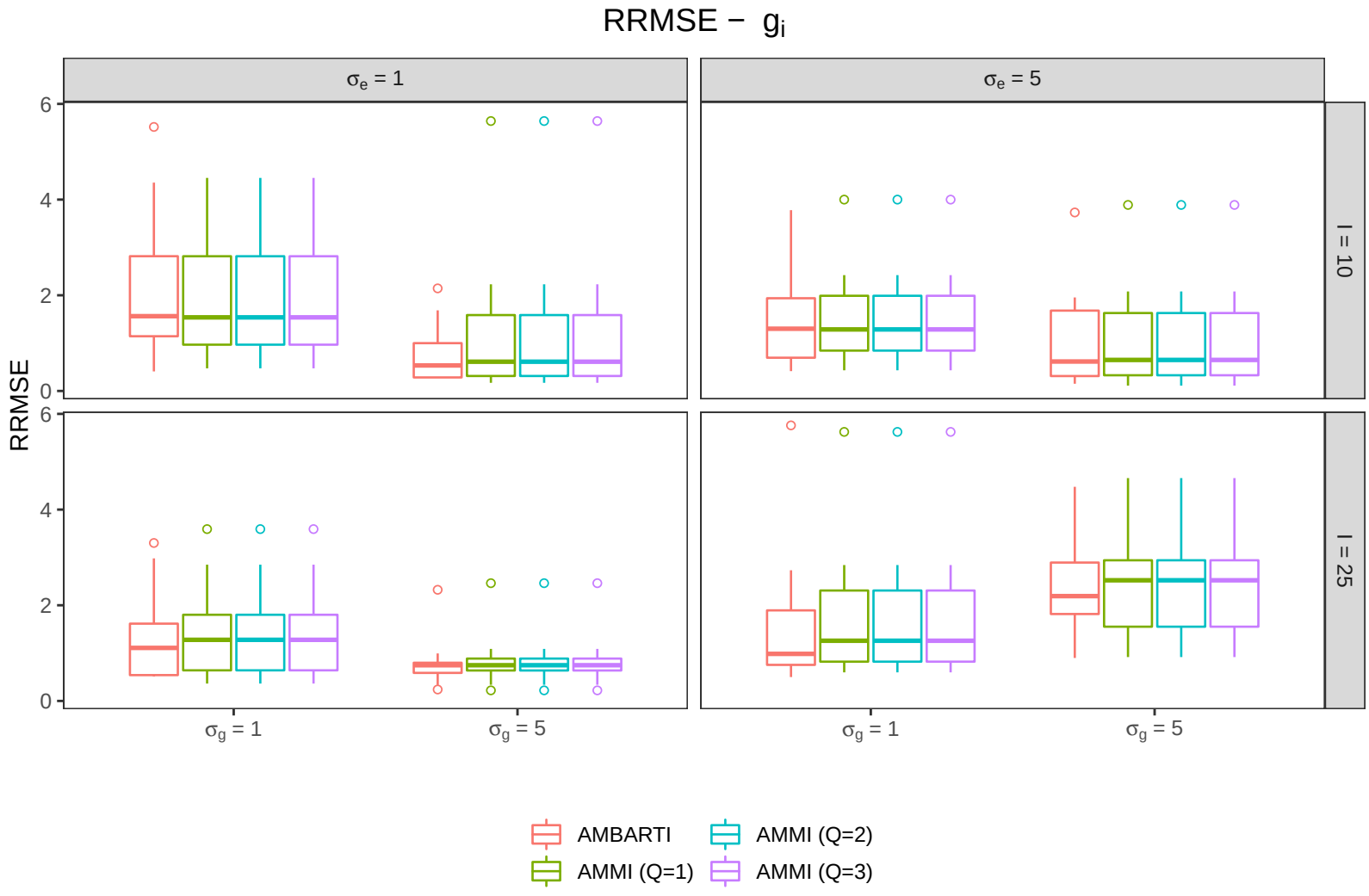
where y_{ij} is the yield for genotype i and environment j, and g_i and e_j are the genotype and environment effects, respectively.

- $\sigma^2 = 1,\ T = 200.$
- $I = J = c(10, 25)$ (without repetitions).
- $\mu_{t\ell}|\mathcal{T}_t \sim \mathsf{N}(\mu_\mu = 0, \sigma^2_\mu = 3)$,
- $e_j \sim \mathsf{N}(0, s_e)$, with $s_e = c(1, 5)$.
- $g_i \sim \mathsf{N}(0, s_g)$, with $s_g = c(1, 5)$.

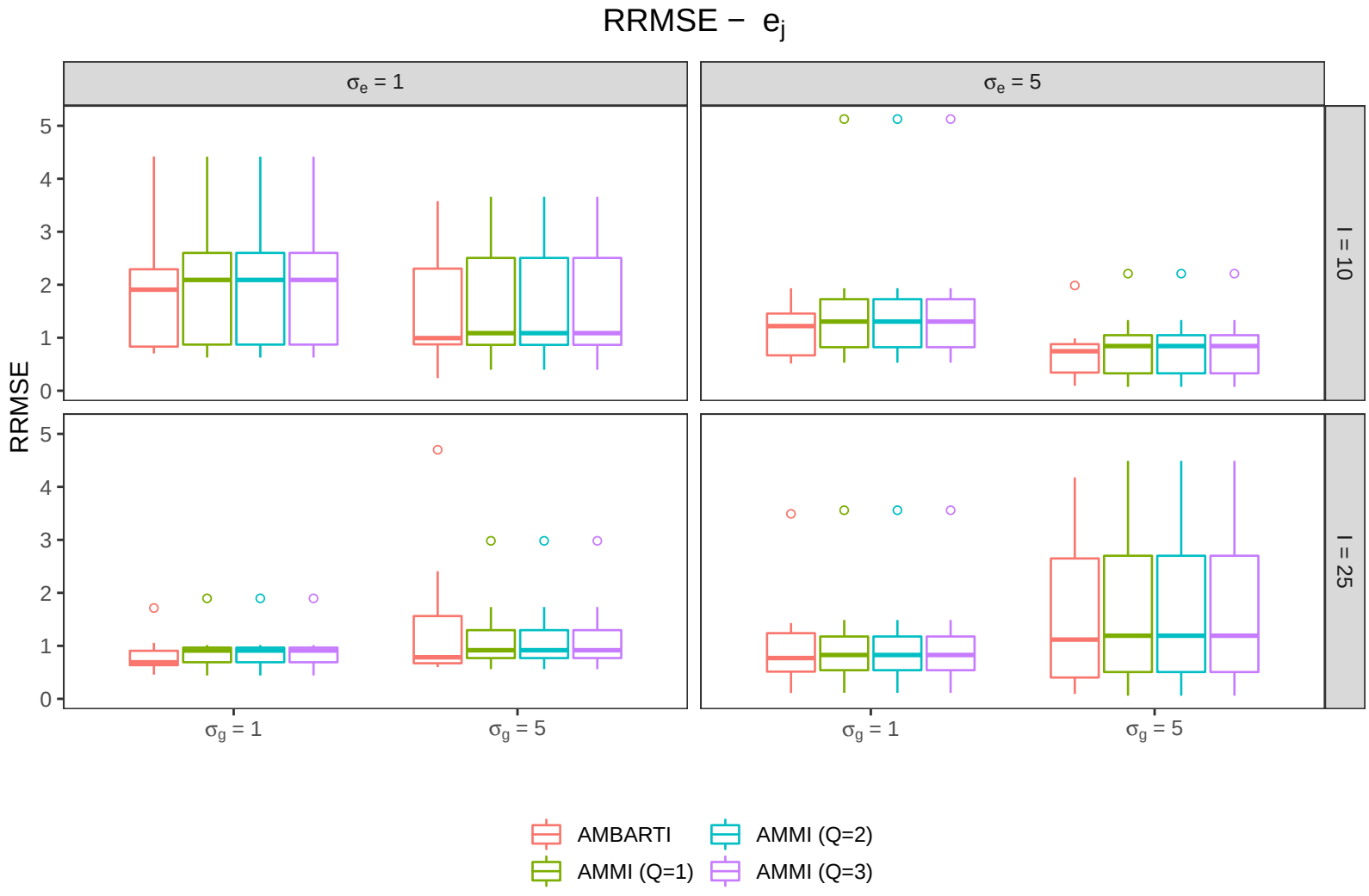
Simulation AMBARTI ($y_{ij} = g_i + e_j + \sum_{t=1}^{T} h(\mathbf{x}_{ij}, \mathcal{M}_t, \mathcal{T}_t)$)



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



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

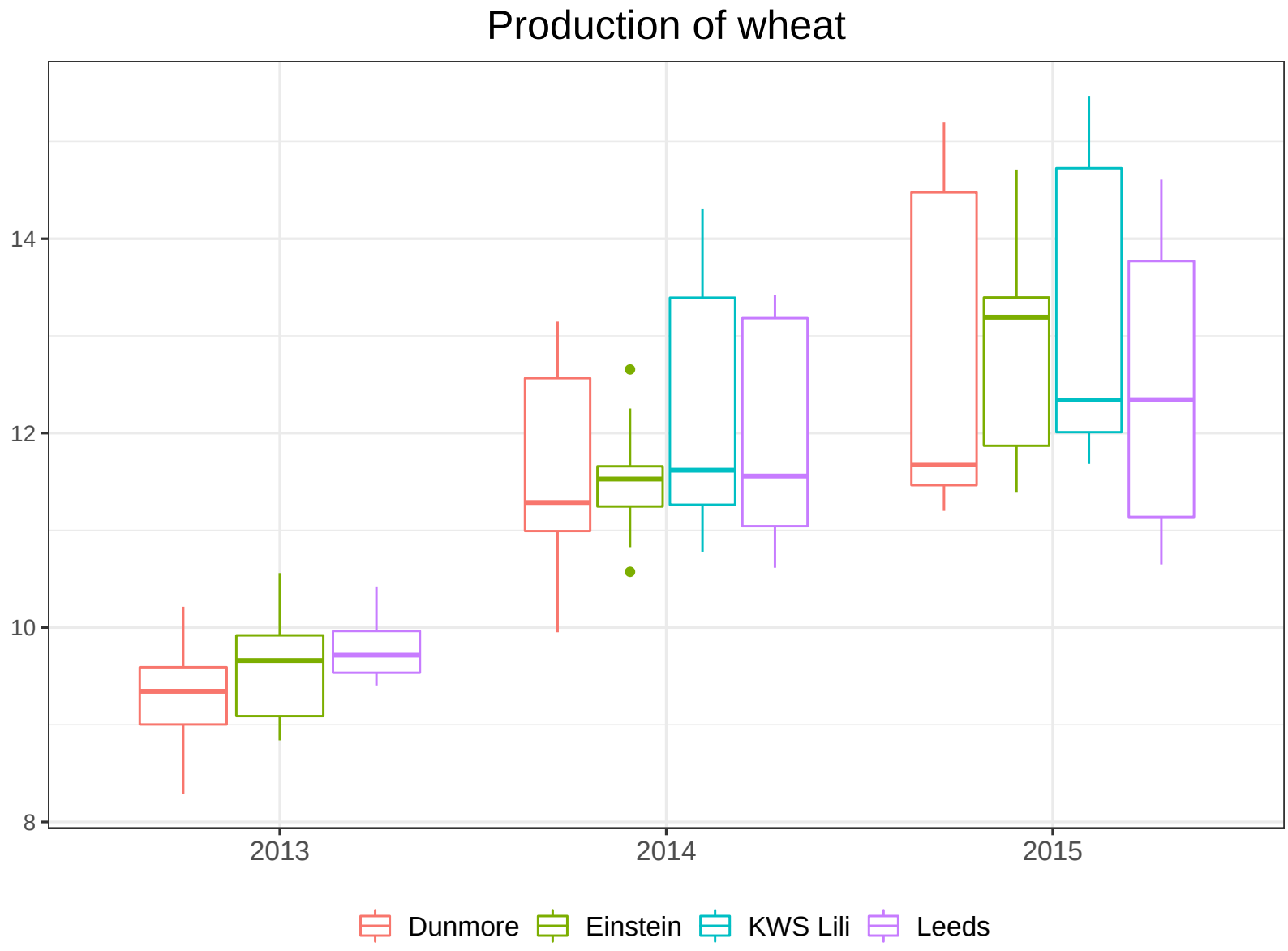


Real data

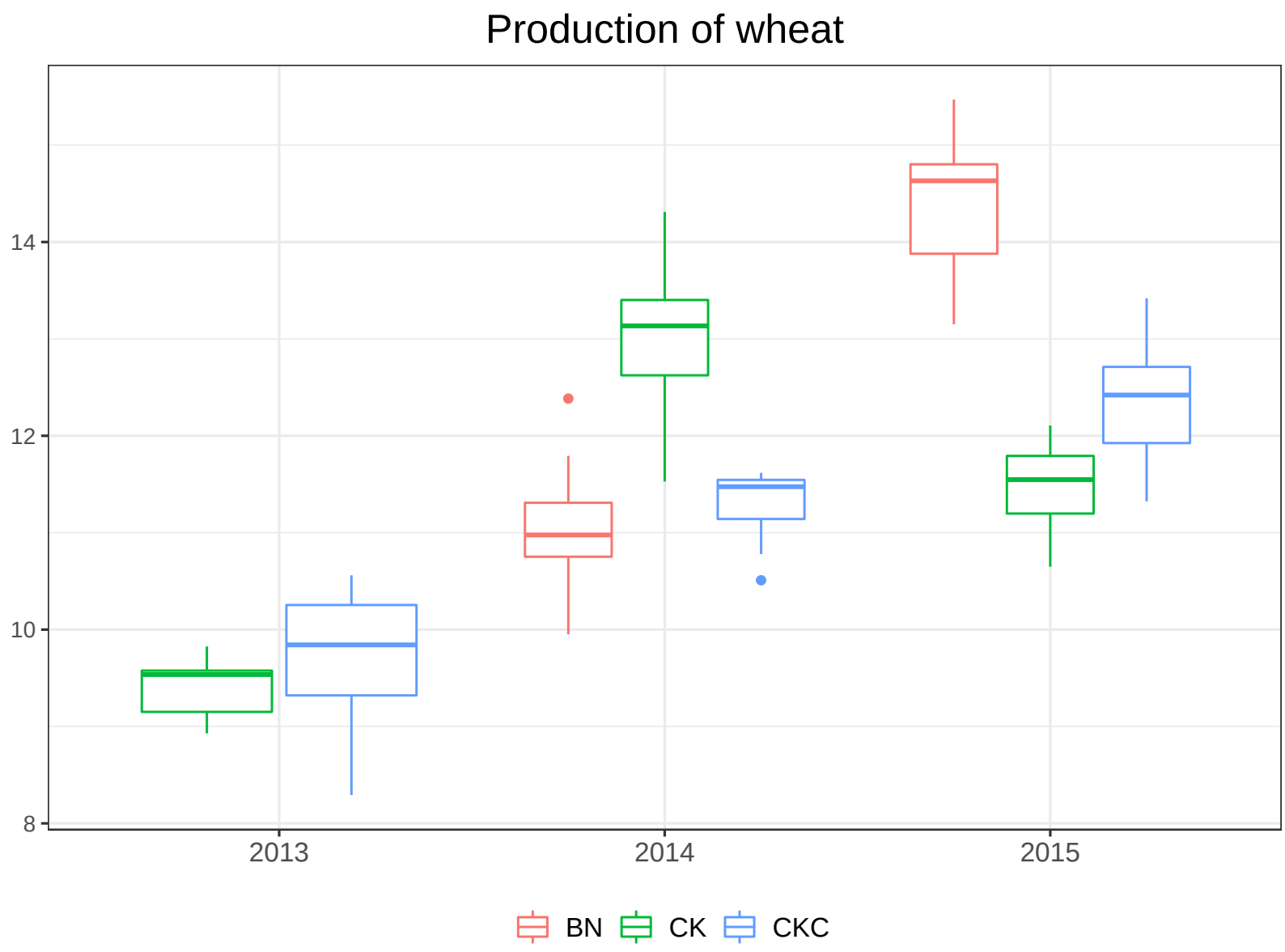
INNOVAR project: wheat data

- European Union Horizon 2020 project INNOVAR: improve the efficacy and accuracy of European crop variety testing and on-farm decision-making.
- INNOVAR is a collaboration between 22 different partners across Europe.
- Each year has ~ 20 genotypes and ~ 8 environments (4 replicates).
- The response variable is the production of wheat in tonnes per hectare.

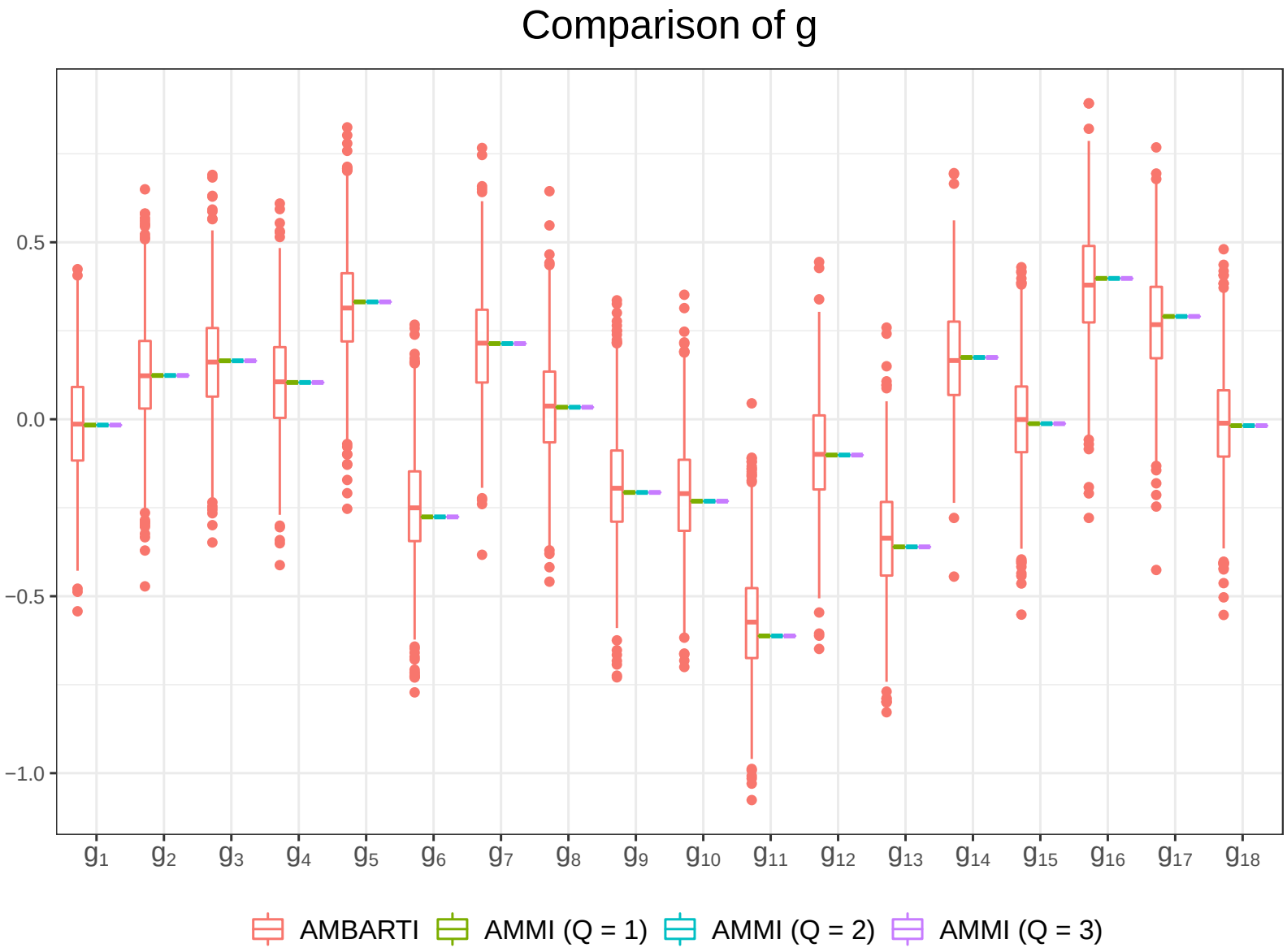
Results (2015)



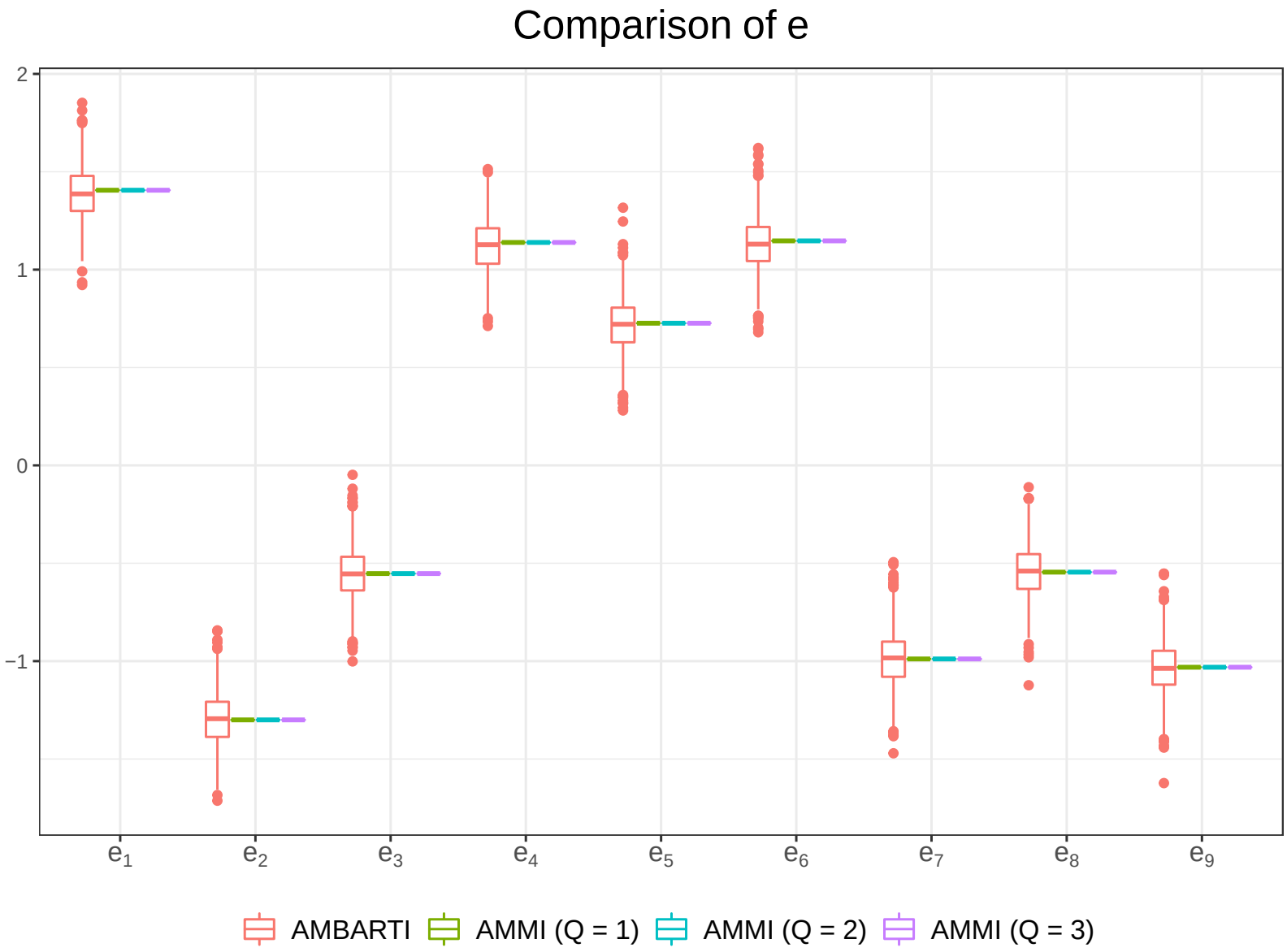
Results (2015)



Results (2015)



Results (2015)



Results

Year	AMMI	AMBARTI
2010	0.5030	0.4814
2011	0.4539	0.4263
2012	0.3755	0.3029
2013	0.4416	0.4247
2014	0.4644	0.4386
2015	0.3811	0.3389
2016	0.4205	0.3830
2017	0.4175	0.3928
2018	0.5659	0.5532
2019	0.4509	0.4010

Table 1: RMSE for $\hat{y}$ on test data.

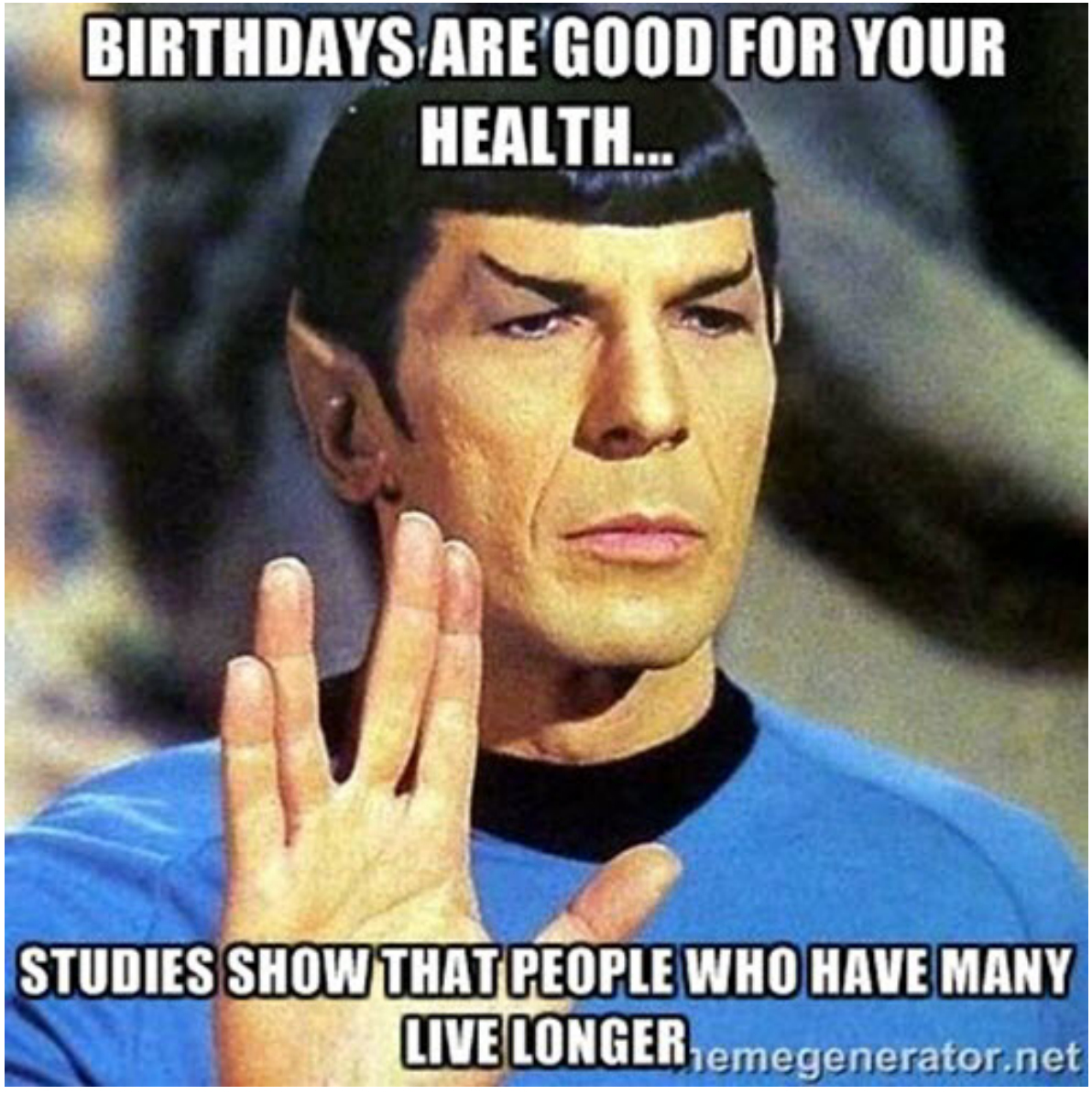
Next steps

1. Starting writing the paper.

Guess what!


ROCKET LEAGUE
ROCKET LEAGUE
ROCKET LEAGUE
ROCKET LEAGUE
HAPP
BIRTHDAY

Guess what!


BIRTHDAYS ARE GOOD FOR YOUR HEALTH...
STUDIES SHOW THAT PEOPLE WHO HAVE MANY LIVE LONGER
emegenerator.net

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)

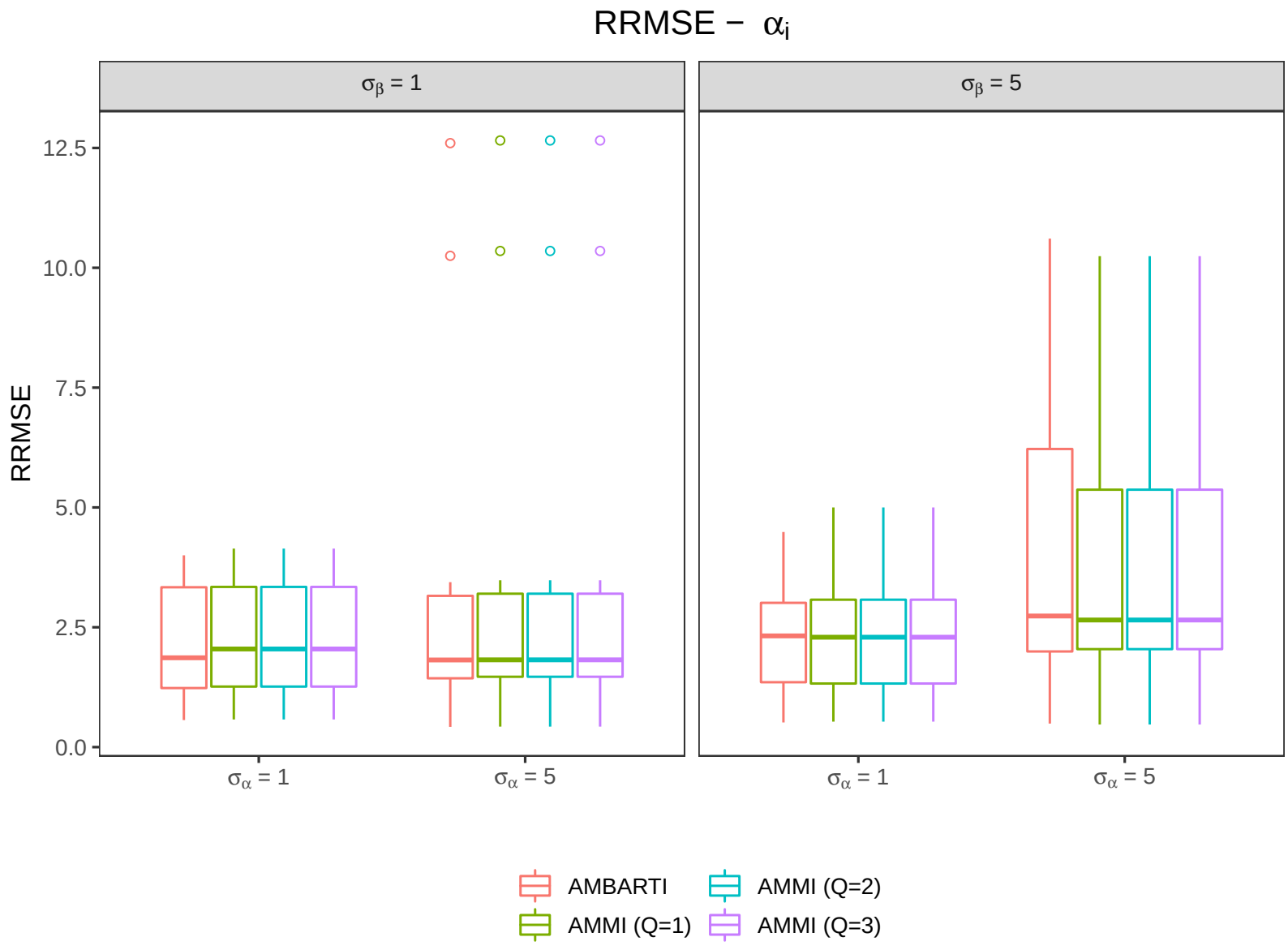
We consider the following settings to generate a set of simulated data sets:

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

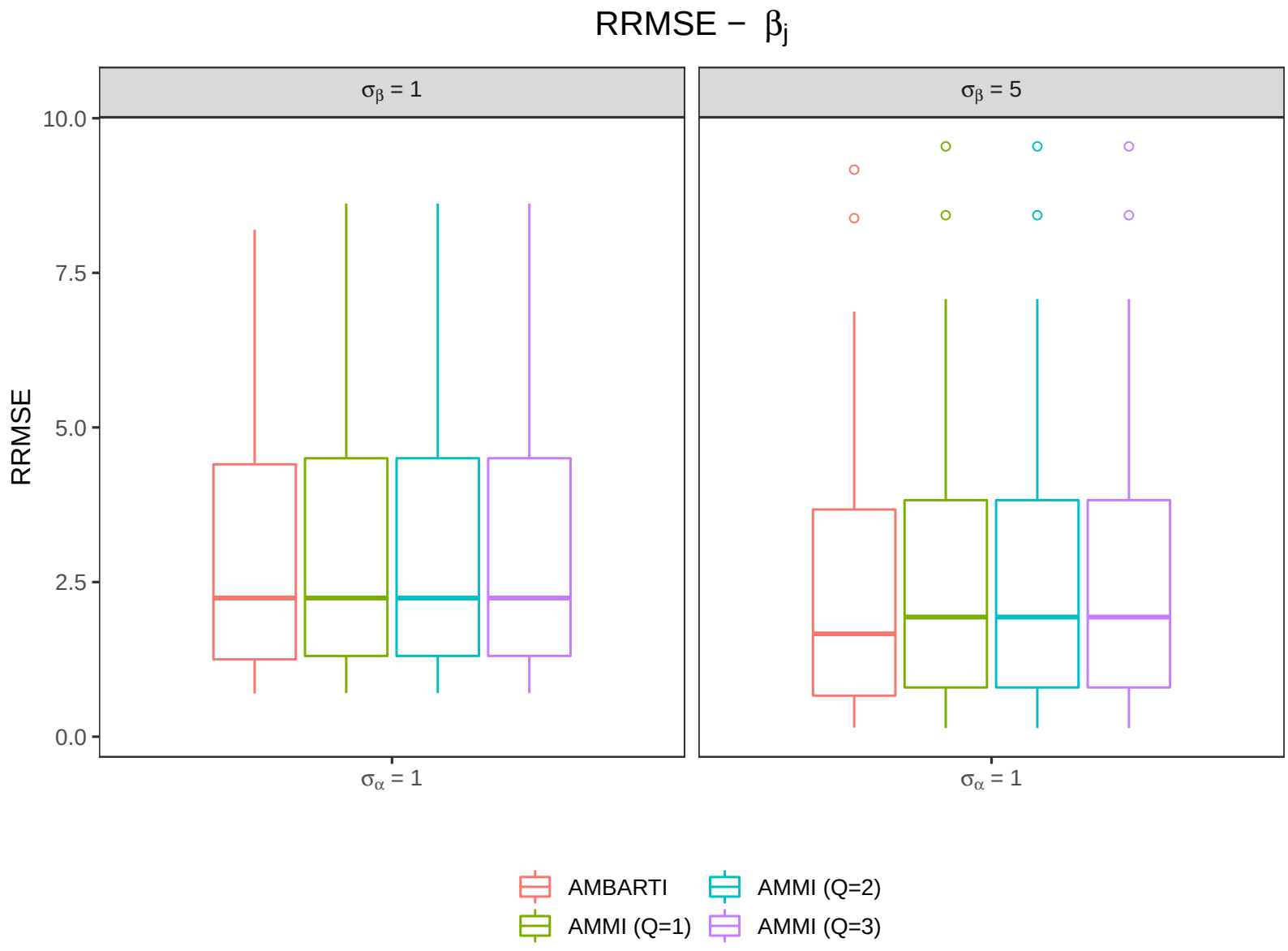
where

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

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



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



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