Unbalanced trials and environment-specific variances

Muhammad Yaseen

2026-11-04

Table of contents

1 Abstract **1**

2 Missing plots and empty cells **1**
2.1 The idea 1
2.2 An example 2

3 Environment-specific error variances **5**
3.1 The model 5
3.2 The fit 6

4 Which variance model? **8**

5 References **10**

1 Abstract

Real trials lose plots, and some genotypes are not grown in every environment. The error variance also differs from one site to the next. This vignette shows how `bayes_ammi()` handles both: missing plots and empty cells by data augmentation, and a separate error variance for every environment with `.het = TRUE`. It ends with a comparison of the two variance models on `Maiz`.

2 Missing plots and empty cells

2.1 The idea

With B the largest number of plots in a cell, a cell (i, j) with $n_{ij} < B$ plots has $B - n_{ij}$ missing plots. A cell with $n_{ij} = 0$ is an empty cell. At every iteration

the sampler draws the missing plots from the model,

$$y_{ijl}^{\text{mis}} \mid \boldsymbol{\theta} \sim N(\eta_{ij}, \sigma_j^2), \qquad \eta_{ij} = \mu + \tau_i + \delta_j + \sum_{k=1}^{K} \lambda_k \alpha_{ik} \gamma_{jk},$$

and then draws the parameters from the conditional distributions of the balanced design (vignette 2) given the completed data. This is the data augmentation of Tanner and Wong (1987). The draws of the parameters come from their posterior given the observed plots only, so nothing is invented: a missing plot adds no information, and an empty cell is predicted from the row and column structure of the model. The missing plots are assumed to be missing at random (Little and Rubin 2019).

The sampler needs only the sums, sums of squares and counts of the observed plots in every cell, so the completed data are never stored.

2.2 An example

From `Maiz` we remove 60 plots at random and every plot of three genotype-environment combinations:

```
set.seed(seed = 21)
mz   <- Maiz[-sample(x = nrow(Maiz), size = 60), ]
gone <- (mz$entry == 2 & mz$site == 5) | (mz$entry == 7 & mz$site == 11)
 ↪  | (mz$entry == 9 & mz$site == 18)
mz   <- mz[!gone, ]
table(plots = table(factor(mz$entry):factor(mz$site)))
```

```
plots
  0   1   2   3   4
  3   1   5  47 124
```

```
fu <-
  bayes_ammi(
      .data   = mz
    , .y      = y
    , .gen    = entry
    , .env    = site
    , .rep    = rep
    , .nIter  = 3000
    , .k      = 2
    , .burnin = 500
    , .thin   = 1
    , .chains = 2
    , .priors = bam_priors(
          mu_mean     = 0
        , mu_var      = 1e+8
        , tau_mean    = 0
        , tau_var     = 1e+8
```

```
        , delta_mean  = 0
        , delta_var   = 1e+8
        , lambda_mean = 0
        , lambda_var  = 1e+8
        , nu          = 1
        , s2          = NULL
        )
    , .level  = 0.95
    , .seed   = 3
    , .align  = TRUE
    , .model  = "AMMI"
    , .het    = FALSE
    , .waic   = TRUE
    )
fu
```

```
Bayesian AMMI with 2 multiplicative term(s): 9 genotypes, 20 environments, 648 plots (3 empt
2 chain(s) x 3000 kept draws (iterations 3000 after a burn-in of 500, thinning 1); 2.4 s

Posterior means, standard deviations and 95% HPD intervals
    parameter       mean       sd      lower     upper  ess   rhat
 1:        mu   4865.378    34.85  4.797e+03   4931.90 4639 1.0023
 2:    sigma2 771133.151 45862.81  6.884e+05 866667.90 3493 1.0004
 3:    tau[1]   -235.197    95.45 -4.318e+02    -56.12 5393 1.0003
 4:    tau[2]   -287.087    99.35 -4.902e+02   -100.13 4427 1.0000
 5:    tau[3]    -16.772    97.76 -2.130e+02    165.23 5358 1.0002
 6:    tau[4]    368.080    96.72  1.815e+02    557.44 5813 1.0006
 7:    tau[5]    395.278    95.27  2.256e+02    600.17 5604 1.0004
 8:    tau[6]    439.592    99.69  2.415e+02    628.05 4629 1.0000
 9:    tau[7]   -197.581   105.05 -4.019e+02     10.22 3681 1.0001
10:    tau[8]   -528.930   100.93 -7.224e+02   -328.54 4393 1.0000
11:    tau[9]     62.618    99.87 -1.305e+02    260.72 4541 1.0007
12:  delta[1]  -1246.974   147.62 -1.550e+03   -973.11 5313 0.9998
13:  delta[2]   -603.360   155.02 -9.078e+02   -307.74 4487 1.0008
14:  delta[3]    287.881   144.60 -8.507e+00    548.59 5619 0.9999
15:  delta[4]    436.272   151.48  1.404e+02    727.13 4672 1.0000
16:  delta[5]    -21.712   154.67 -3.280e+02    270.00 4102 0.9999
17:  delta[6]   1476.541   147.84  1.181e+03   1755.92 5320 1.0010
18:  delta[7]  -1603.056   145.40 -1.891e+03  -1313.72 5601 1.0000
19:  delta[8]   -820.386   145.03 -1.102e+03   -538.55 5335 1.0000
20:  delta[9]     56.727   146.65 -2.301e+02    345.23 5396 1.0000
21: delta[10]  -1931.169   143.26 -2.212e+03  -1658.16 5660 1.0009
22: delta[11]    429.876   168.20  9.977e+01    752.36 3441 1.0018
23: delta[12]   2682.222   155.78  2.381e+03   2990.30 4191 1.0004
24: delta[13]   1547.696   157.58  1.240e+03   1852.64 3458 1.0004
25: delta[14]   1188.212   146.55  9.058e+02   1478.69 4743 0.9999
```

```
26: delta[15]    213.078   146.76 -7.575e+01    495.96 4750 0.9999
27: delta[16]    542.632   145.82  2.585e+02    829.31 5435 1.0004
28: delta[17]    -60.923   156.00 -3.637e+02    241.27 4233 1.0003
29: delta[18]   -408.313   163.28 -7.449e+02   -105.42 3649 1.0002
30: delta[19]  -2172.886   151.46 -2.468e+03  -1880.16 4538 1.0019
31: delta[20]      7.641   151.29 -2.878e+02    301.10 4954 1.0000
32: lambda[1]   5106.531   477.92  4.121e+03   6013.13 3157 1.0016
33: lambda[2]   1339.303   806.12  5.931e-02   2705.39  639 1.0065
    parameter       mean       sd      lower     upper  ess   rhat

Scores of the bilinear terms: x$summary (blocks alpha, gamma); biplot coordinates: x$scores

Sources: Crossa et al. (2011); Spiegelhalter et al. (2002); Watanabe (2010).
Full references: bam_cite(x).
```

The print shows the number of empty cells. `bam_predict()` gives the posterior of every cell mean η_{ij}, including the empty cells:

```
pu <-
  bam_predict(
      fit    = fu
    , .level = 0.95
    )
pu$cells[n == 0]
```

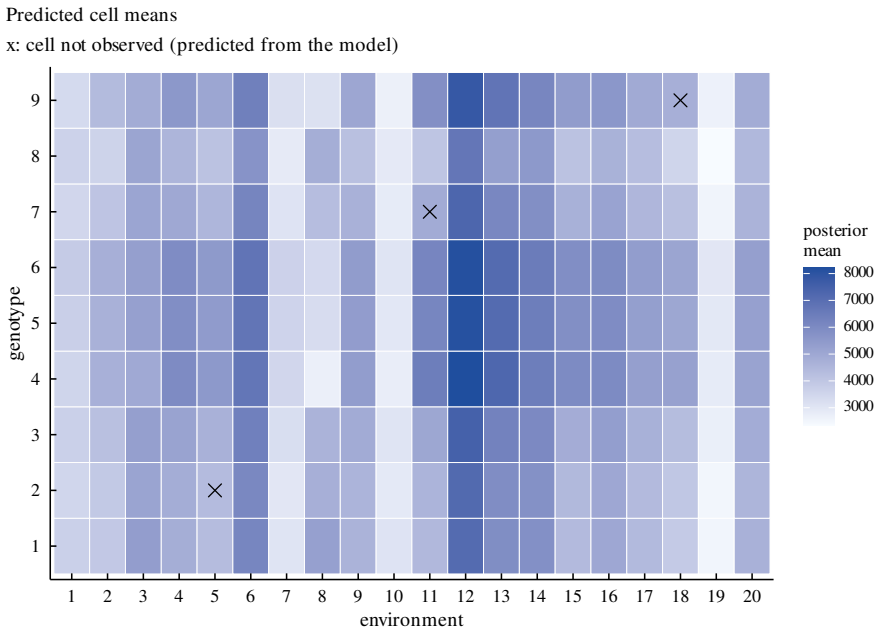
```
   Gen Env n observed     mean       sd    lower    upper
1:   2   5 0       NA 4369.859 279.9944 3794.720 4891.308
2:   7  11 0       NA 4903.347 322.4456 4269.447 5523.364
3:   9  18 0       NA 4761.894 311.5227 4165.312 5390.920
```

```
Maiz_cells <- data.table::as.data.table(x = Maiz)[, list(full = mean(y)),
 ↪  by = list(
    Gen = factor(entry)
  , Env = factor(site)
  )]
merge(x = pu$cells[n == 0], y = Maiz_cells, by = c("Gen", "Env"))[, list(
    Gen
  , Env
  , mean
  , lower
  , upper
  , full
  )]
```

```
   Gen Env     mean    lower    upper     full
1:   2   5 4369.859 3794.720 4891.308 5201.475
2:   7  11 4903.347 4269.447 5523.364 5051.575
3:   9  18 4761.894 4165.312 5390.920 5244.400
```

The column `full` holds the cell means of the complete data, which the fit never saw. 2 of the 3 HPD intervals contain them. The miss, genotype 2 in environment 5, is a cell whose interaction is not carried by the two bilinear terms: the model predicts an empty cell only from the pattern of the rest of the trial. The posterior standard deviations of the empty cells are on average 1.1 times those of the cells with all four plots.

```
plot(x = pu, which = "cells")
```



3 Environment-specific error variances

3.1 The model

With `.het = TRUE` the plots of environment j have their own variance, $e_{ijl} \sim N(0, \sigma_j^2)$, and every σ_j^2 has the scaled inverse chi-squared prior of the common variance. Write $w_j = 1/\sigma_j^2$. The conditional distributions change as follows.

- σ_j^2: scaled inverse chi-squared with $\nu + BG$ degrees of freedom and the residual sum of squares of environment j.
- τ_i: the data precision becomes $B \sum_j w_j$, the same for every genotype.
- μ and δ_j: no longer independent, so $(\mu, \mathbf{d})$ is drawn jointly from its multivariate normal conditional.
- λ_k: mean and variance weighted by w_j.
- $\boldsymbol{\alpha}_k$: still von Mises-Fisher, because the quadratic term $\sum_j w_j \gamma_{jk}^2$ does not

depend on $\boldsymbol{\alpha}_k$.

- $\boldsymbol{\gamma}_k$: the quadratic term $\sum_j w_j \gamma_{jk}^2$ now varies on the sphere. The conditional is a Bingham-von Mises-Fisher distribution (Hoff 2009), drawn one coordinate at a time in the eigenbasis of the quadratic form by slice sampling.

Perez-Elizalde et al. (2012) give the general linear-bilinear model with these conditionals.

3.2 The fit

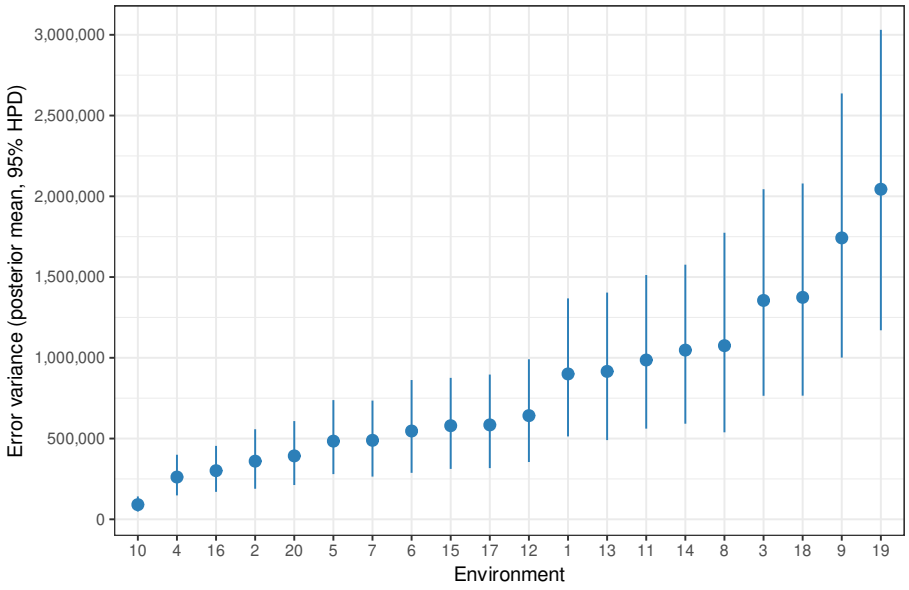
```
fh <-
  bayes_ammi(
      .data   = Maiz
    , .y      = y
    , .gen    = entry
    , .env    = site
    , .rep    = rep
    , .nIter  = 3000
    , .k      = 2
    , .burnin = 500
    , .thin   = 1
    , .chains = 2
    , .priors = bam_priors(
          mu_mean     = 0
        , mu_var      = 1e+8
        , tau_mean    = 0
        , tau_var     = 1e+8
        , delta_mean  = 0
        , delta_var   = 1e+8
        , lambda_mean = 0
        , lambda_var  = 1e+8
        , nu          = 1
        , s2          = NULL
        )
    , .level  = 0.95
    , .seed   = 4
    , .align  = TRUE
    , .model  = "AMMI"
    , .het    = TRUE
    , .waic   = TRUE
    )
fh$summary[block == "sigma2", list(parameter, mean, lower, upper, rhat)]
```

```
     parameter       mean      lower     upper      rhat
 1:  sigma2[1]  900202.55  513309.39 1367864.4 0.9999960
 2:  sigma2[2]  360015.63  189188.44  557743.7 0.9999301
 3:  sigma2[3] 1355210.60  764845.74 2043877.8 0.9998680
```

```
 4:  sigma2[4]  261692.05  148061.44  399759.5 1.0004044
 5:  sigma2[5]  484072.36  279891.01  738081.8 1.0004091
 6:  sigma2[6]  546696.45  287760.53  862390.4 1.0025834
 7:  sigma2[7]  489117.60  264488.24  734986.4 1.0003634
 8:  sigma2[8] 1075202.84  538702.93 1774320.1 1.0014234
 9:  sigma2[9] 1742349.83 1001193.02 2636884.7 1.0005600
10: sigma2[10]   90740.34   47911.21  141531.5 1.0032634
11: sigma2[11]  986252.96  561409.95 1512450.3 1.0005569
12: sigma2[12]  641984.58  354738.52  990922.7 0.9998358
13: sigma2[13]  915813.12  490491.10 1403805.4 1.0001763
14: sigma2[14] 1047612.38  591957.06 1576650.2 0.9998416
15: sigma2[15]  579620.73  312104.67  875774.2 1.0000445
16: sigma2[16]  301118.89  170076.30  454565.6 0.9999565
17: sigma2[17]  584613.06  317098.67  896266.5 1.0002354
18: sigma2[18] 1374336.60  765617.75 2079353.3 1.0016390
19: sigma2[19] 2043724.07 1170509.40 3030686.8 0.9999205
20: sigma2[20]  392646.55  212299.68  608151.6 1.0006141
```

The posterior means range from 90,740 to 2,043,724, a ratio of 23. One common variance cannot describe both extremes.

```
ggplot2::ggplot(
    data    = fh$summary[block == "sigma2"]
  , mapping = ggplot2::aes(x = reorder(label, mean), y = mean, ymin =
↪  lower, ymax = upper)
  ) +
  ggplot2::geom_pointrange(colour = "#2c7fb8") +
  ggplot2::scale_y_continuous(breaks = scales::pretty_breaks(n = 6),
↪  labels = scales::label_comma()) +
  ggplot2::labs(x = "Environment", y = "Error variance (posterior mean,
↪  95% HPD)") +
  ggplot2::theme_bw()
```



4 Which variance model?

```
fc <-
  bayes_ammi(
      .data   = Maiz
    , .y      = y
    , .gen    = entry
    , .env    = site
    , .rep    = rep
    , .nIter  = 3000
    , .k      = 2
    , .burnin = 500
    , .thin   = 1
    , .chains = 2
    , .priors = bam_priors(
          mu_mean     = 0
        , mu_var      = 1e+8
        , tau_mean    = 0
        , tau_var     = 1e+8
        , delta_mean  = 0
        , delta_var   = 1e+8
        , lambda_mean = 0
        , lambda_var  = 1e+8
        , nu          = 1
        , s2          = NULL
        )
```

```
    , .level  = 0.95
    , .seed   = 4
    , .align  = TRUE
    , .model  = "AMMI"
    , .het    = FALSE
    , .waic   = TRUE
    )
cp <-
  bam_compare(
      common          = fc
    , environment_var = fh
    )
cp
```

```
Comparison of Bayesian models (smaller is better)

             model terms       variances      DIC      pD     WAIC  p_waic
1: environment_var     2 per environment 11735.81 85.9708 11741.88 82.4746
2:          common     2          common 11862.02 67.9290 11865.68 65.9309
        se lambda_last_lower lambda_last_mean    dWAIC dWAIC_se
1: 41.9665          928.1938         2093.397   0.0000    0.000
2: 44.2329           63.7490         1686.287 123.8044   22.525

Sources: Spiegelhalter et al. (2002); Watanabe (2010); Vehtari et al. (2017).
Full references: bam_cite(x).
```

The environment-specific variances lower WAIC by 124, 5.5 times the standard error of the difference. They also sharpen the interaction: the lower HPD limit of λ_2 rises from 64 to 928, because the plots of the noisy environments no longer blur the second term. The posterior predictive check of the common-variance model tells the same story through the ratio of the largest to the smallest environment variance:

```
pp <-
  bam_ppc(
      fit    = fc
    , .draws = 500
    , .seed  = 1
    )
pp
```

```
Posterior predictive checks (500 replicated trials)

                                statistic  observed  mean_rep p_value
1:                  variance of cell means 1.842e+06 1.901e+06   0.682
2:                       largest cell mean 8.385e+03 8.608e+03   0.728
3:                      smallest cell mean 1.632e+03 1.743e+03   0.672
4:              variance of genotype means 1.247e+05 1.429e+05   0.662
```

```
5:                    first singular value 5.923e+03 5.902e+03   0.482
6:                    within-cell variance 7.559e+05 7.660e+05   0.526
7: largest / smallest environment variance 2.582e+01 2.924e+00   0.000
```

```
Sources: Gelman et al. (1996). Full references: bam_cite(x).
```

The observed ratio is 26, against 2.9 on average in the replicated trials of the common-variance model (posterior predictive p-value 0.000). The other statistics are reproduced well. Vignette 9 returns to these tools.

5 References

Crossa, J., Perez-Elizalde, S., Jarquin, D., Cotes, J. M., Viele, K., Liu, G. and Cornelius, P. L. (2011). Bayesian estimation of the additive main effects and multiplicative interaction model. *Crop Science*, 51(4), 1458–1469. doi:10.2135/cropsci2010.06.0343

Perez-Elizalde, S., Jarquin, D. and Crossa, J. (2012). A general Bayesian estimation method of linear-bilinear models applied to plant breeding trials with genotype x environment interaction. *Journal of Agricultural, Biological, and Environmental Statistics*, 17, 15–37. doi:10.1007/s13253-011-0063-9

Hoff, P. D. (2009). Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2), 438–456.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398), 528–540.

Little, R. J. A. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*, 3rd edn. Wiley, Hoboken.