Identification de variétés robustes aux stress environnementaux sur la base de la distribution de leur valeur génétique

Salma Guennouni & Adrien Sardi

31 Mai 2024

Encadrants: Tristan Mary-Huard, INRAE-AgroParisTech & Laurence Moreau, INRAE, UMR GQE-Moulon

Problématique

Développement de variétés de plantes **performantes** et **robustes** face aux aléas climatiques.

Plan

Plan

Données phénotypiques

Individu	Environnement	Rendement
B73	Bol12R	3.159023
PH207	Cam12R	6.234565
Oh43	Cam12R	7.349568
W64A	Deb11W	4.857412
B73	Ber14W	3.367456

Table: Exemple de données phénotypiques

Données phénotypiques

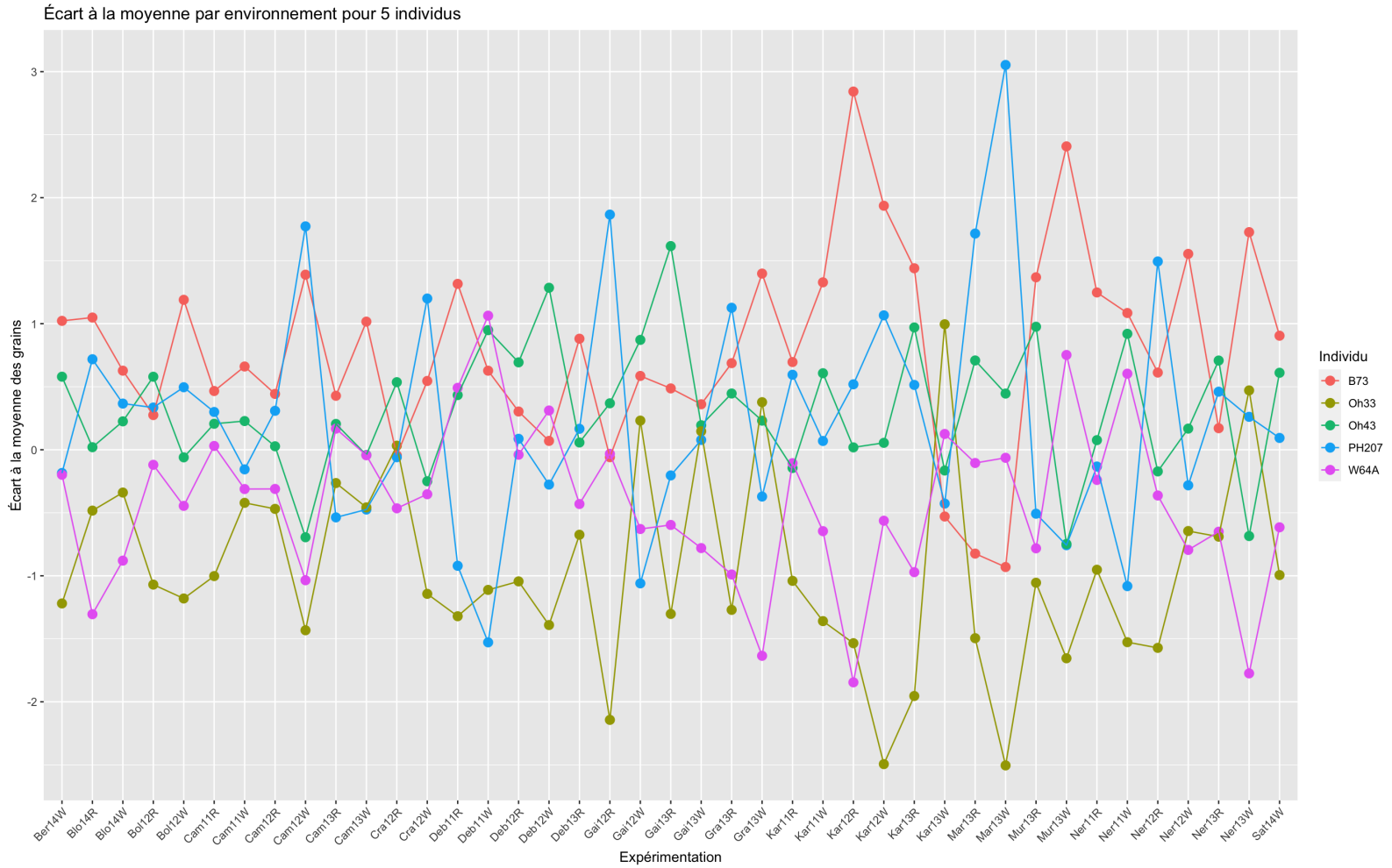


Figure: Écart à la moyenne de l'environnement pour différents individus

Données génotypiques

Individu	SNP1	SNP2	SNP3	SNP4
B73	2	2	2	0
PH207	0	2	2	0
Oh43	0	0	0	2
W64A	0	0	0	0
B73	2	2	2	0

Table: Exemple de données génotypiques

Données génotypiques

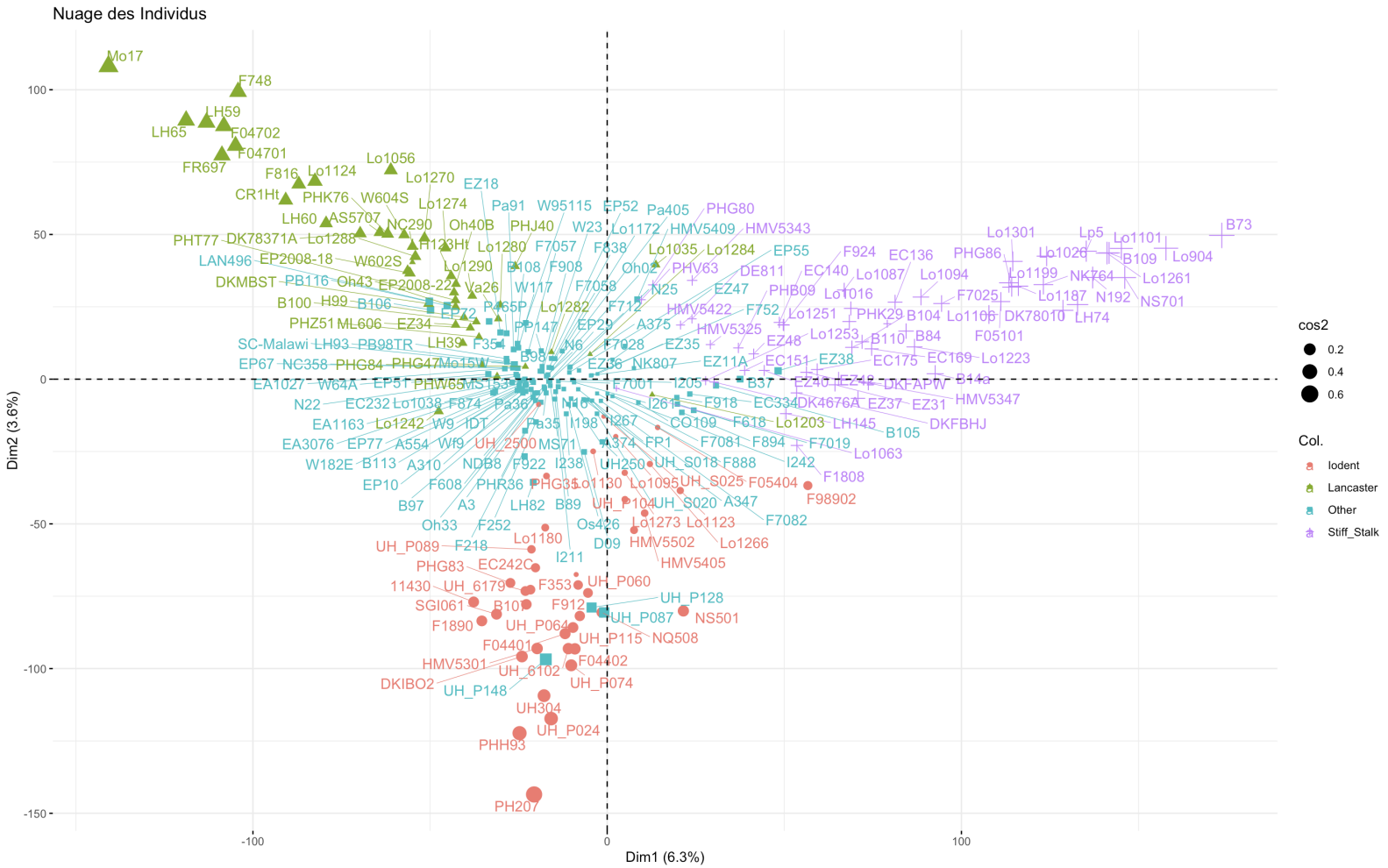


Figure: ACP des données génotypiques

Plan

Notations

- $n_I \in \mathbb{N}$: nombre d'individus ;
- $n_E \in \mathbb{N}$: nombre d'environnements ;
- $n_{Ri,e} \in \mathbb{N}$: nombre de répétitions par individu et par environnement ;
- Y variable aléatoire représentant le rendement de l'individu i dans l'environnement e et pour la répétition r, $Y = (Y_{i,e,r})_{i \in [1,n_I], e \in [1,n_E], r \in [1,n_{Ri,e}]}$.

Définition du modèle

$$Y_{i,e,r} = \mu$$

Définition du modèle

$$Y_{i,e,r} = \mu + \alpha_e$$

Définition du modèle

$$Y_{i,e,r} = \underbrace{\mu + \alpha_e}_{:=\mu_e}$$

Définition du modèle

$$Y_{i,e,r} = \mu_e + U_{i,e}$$

Définition du modèle

$$U \sim N(0, \sigma_G^2 A)$$

Définition du modèle

$$\cancel{U \sim N(0, \sigma_G^2 A)} \Longrightarrow U_{i,e} \sim N(\mu_{G_i}, \sigma_{G_i}^2),\ IND;$$

Définition du modèle

$$Y_{i,e,r} = \mu_e + U_{i,e}$$

Avec,

$$U_{i,e} \sim N(\mu_{G\,i}, \sigma_{G\,i}^2),\ IND;$$
$$\mu_{G\,i} := \sum_j x_{i,j}\beta_j;$$
$$\sigma_{G\,i}^2 := e^{\sum_j x_{i,j}\beta'_j + \beta'_0}.$$

$$x_{i,j} = \begin{cases} 2 \text{ si le marqueur est présent} \\ 0 \text{ si le marqueur n'est pas présent} \end{cases}$$

Définition du modèle

$$Y_{i,e,r} = \mu_e + U_{i,e}$$

Avec,

$$U_{i,e} \sim N(\mu_{G\,i}, \sigma_{G\,i}^2),\ IND;$$
$$\mu_{G\,i} := \sum_j x_{i,j}\beta_j;$$
$$\sigma_{G\,i}^2 := e^{\sum_j x_{i,j}\beta'_j + \beta'_0}.$$

$$x_{i,j} = \begin{cases} 2 \text{ si le marqueur est présent} \\ 0 \text{ si le marqueur n'est pas présent} \end{cases}$$

Définition du modèle

$$Y_{i,e,r} = \mu_e + U_{i,e} + E_{i,e,r}$$

Avec,

$$E_{i,e,r} \sim N(0, \sigma_{E_e}^2), \ IND$$

Définition du modèle

Propriété (Modèle final)

$$\boxed{Y_{i,e,r} = \mu_e + U_{i,e} + E_{i,e,r}}$$

$$U_{i,e} \sim N(\mu_{G\,i}, \sigma_{G\,i}^2),\ IND;$$

$$E_{i,e,r} \sim N(0, \sigma_{E\,e}^2),\ IND$$

Plan

Calcul de vraisemblance

$$\mathcal{L}(Y) = -\frac{\log(2\pi)}{2}\sum_{i,e} n_{Ri,e} - \frac{1}{2}\sum_{i,e}\log|\Sigma_{i,e}| - \sum_{i,e}\frac{\tilde{Y}_{i,e}^{T}\cdot\Sigma_{i,e}^{-1}\cdot\tilde{Y}_{i,e}}{2}$$

Calcul de vraisemblance

$$\mathcal{L}(Y) = \sum_{i,e} -\frac{1}{2\sigma^2_{E_e}}\Big(\sum_{r=1}^{n_{Ri,e}} (Y_{i,e,r} - (\mu_e + \mu_{Gi}))^2$$

$$+ \frac{\sigma^2_{G_i}}{\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i}} \left(\sum_{r=1}^{n_{Ri,e}} Y_{i,e,r} - (\mu_e + \mu_{Gi})\right)^2 \Big)$$

$$- \frac{1}{2}(n_{Ri,e} - 1)\log(\sigma^2_{E_e}) - \frac{1}{2}\log\left[(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})\right] - \frac{\log(2\pi)}{2} n_{Ri,e}$$

Plan

Première génération des données

$$\boxed{Y_{i,e,r} = \mu_e + U_{i,e} + E_{i,e,r}}$$

$$U_{i,e} \sim N(\mu_{G_i}, \sigma_{G_i}^2),\ IND \quad \text{où} \quad \mu_{G_i} = \sum_j x_{i,j}\beta_j, \sigma_{G_i}^2 = e^{\sum_j x_{i,j}\beta'_j + \beta'_0};$$

$$E_{i,e,r} \sim N(0, \sigma_{E_e}^2),\ IND$$

- On garde les $x_{i,j}$ de nos données
- $\beta_j \sim \mathcal{N}(0,1)$
- $\beta'_j \sim \mathcal{N}(0,1)$
- $E_{i,e,r} \sim \mathcal{N}(0,1)$
- μ_e ?

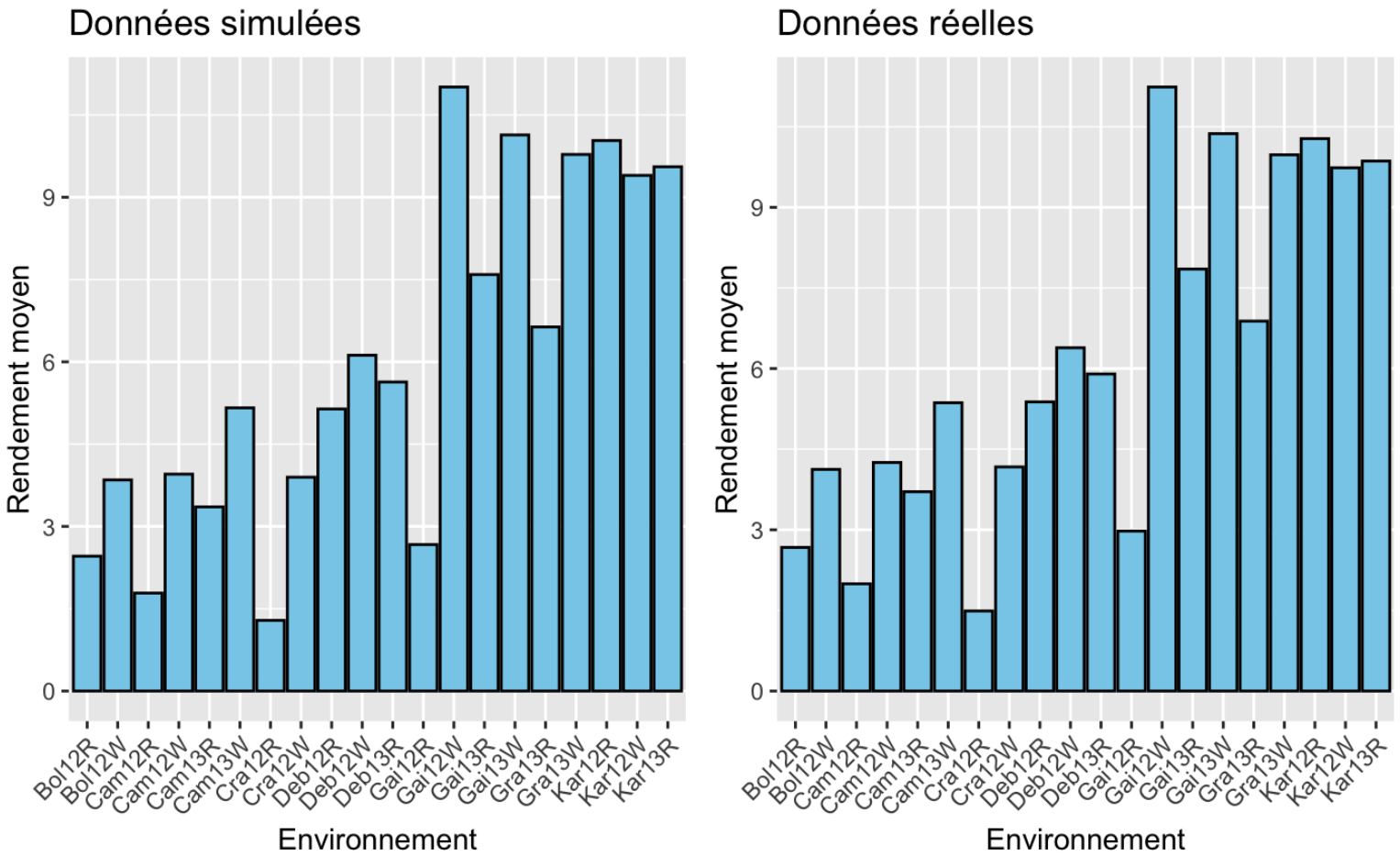
Inférence à partir d'un modèle simplifié

$$Y = \mu_e + U + E$$

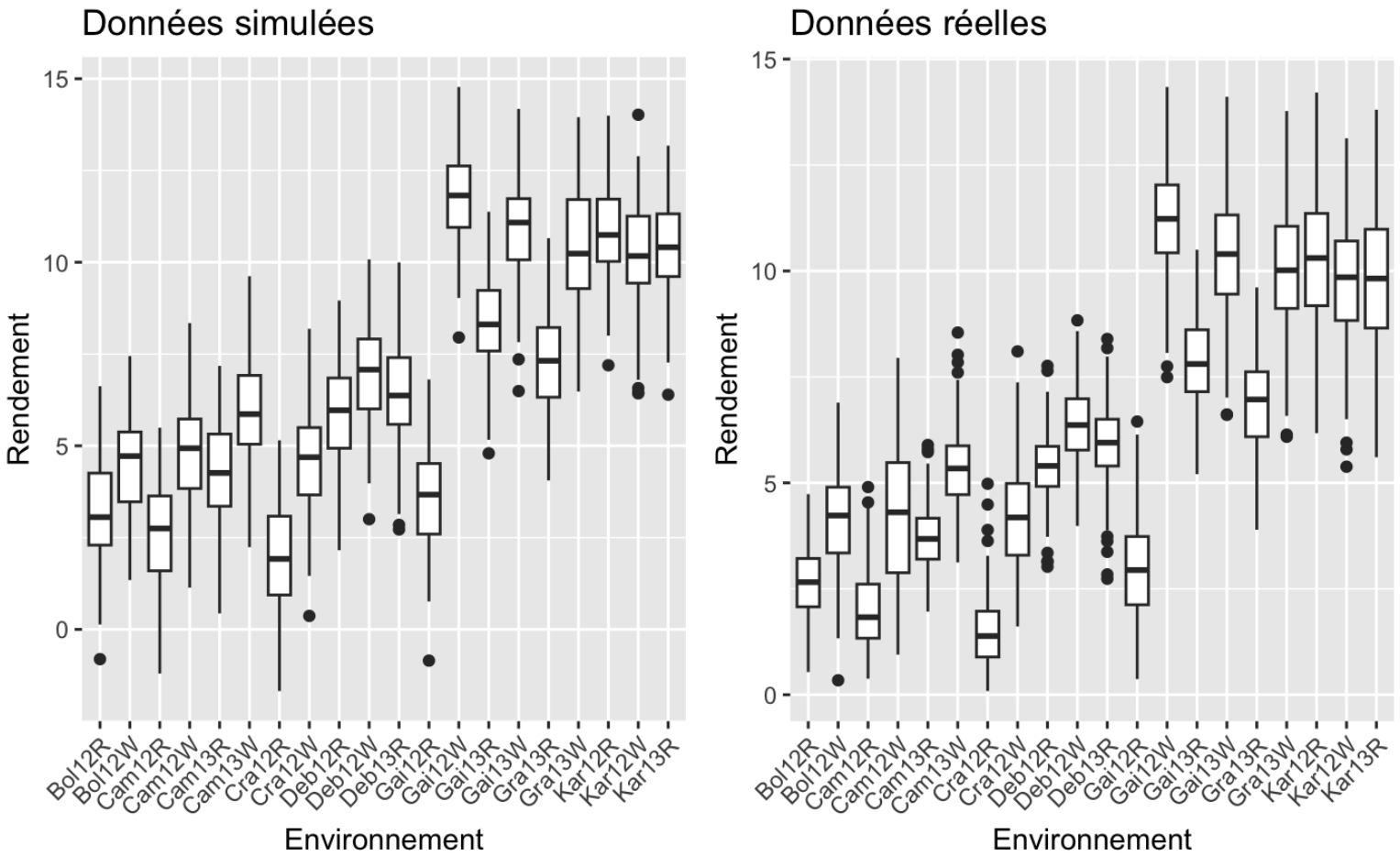
avec $U \sim N(0, \sigma^2_G A)$ avec σ^2_G variance génétique, A matrice de parenté (Kinship) et $E \sim N(0, \sigma^2_{E_e})$.

On récupère les valeurs de μ_e, et on renormalise les autres paramètres pour avoir de meilleurs ordres de grandeurs.

Génération des données - Résultats (1/2)



Génération des données - Résultats (2/2)



Plan

Expression de la vraisemblance pour $n_{Ri,e} = 1$

$$\begin{aligned}\hat{\theta} &= \text{argmin}_\theta \, -\mathcal{L}(Y; \mu_e, \mu_{Gi}, \sigma_{Gi}, \sigma_{E_e}) \\ &= \text{argmin}_\theta \Big(\sum_{i,e} \log(\sigma_{Gi}^2 + \sigma_{E_e}^2) + \sum_{i,e} \frac{(Y_{i,e} - (\mu_e + \mu_{Gi}))^2}{\sigma_{Gi}^2 + \sigma_{E_e}^2}\Big)\end{aligned}$$

Algorithme d'optimisation utilisant la méthode Rprop

```r
# Initialisation de theta_current avec des valeurs de 1
theta_current <- lapply(theta, function(x) torch_tensor(rep(1, length(x)), requires_grad =
  ↪  TRUE))

# Initialisation de l'optimizer
theta_optimizer <- optim_rprop(theta_current)

# Parametres
num_iterations <- 100
loss_vector <- vector("numeric", length = num_iterations)

# Iterations
for (i in 1:num_iterations) {
  # Derivees a 0
  theta_optimizer$zero_grad()
  ## Forward
  loss <- calcul_vraisemblance(theta_current,x_values,df_Y)
  ## Backward
  loss$backward()
  ## Mise a jour des parametres
  theta_optimizer$step()
  ## Stockage de la perte actuelle pour l'affichage graphique
  loss_vector[i] <- loss %>% as.numeric()
}
```

Optimisation

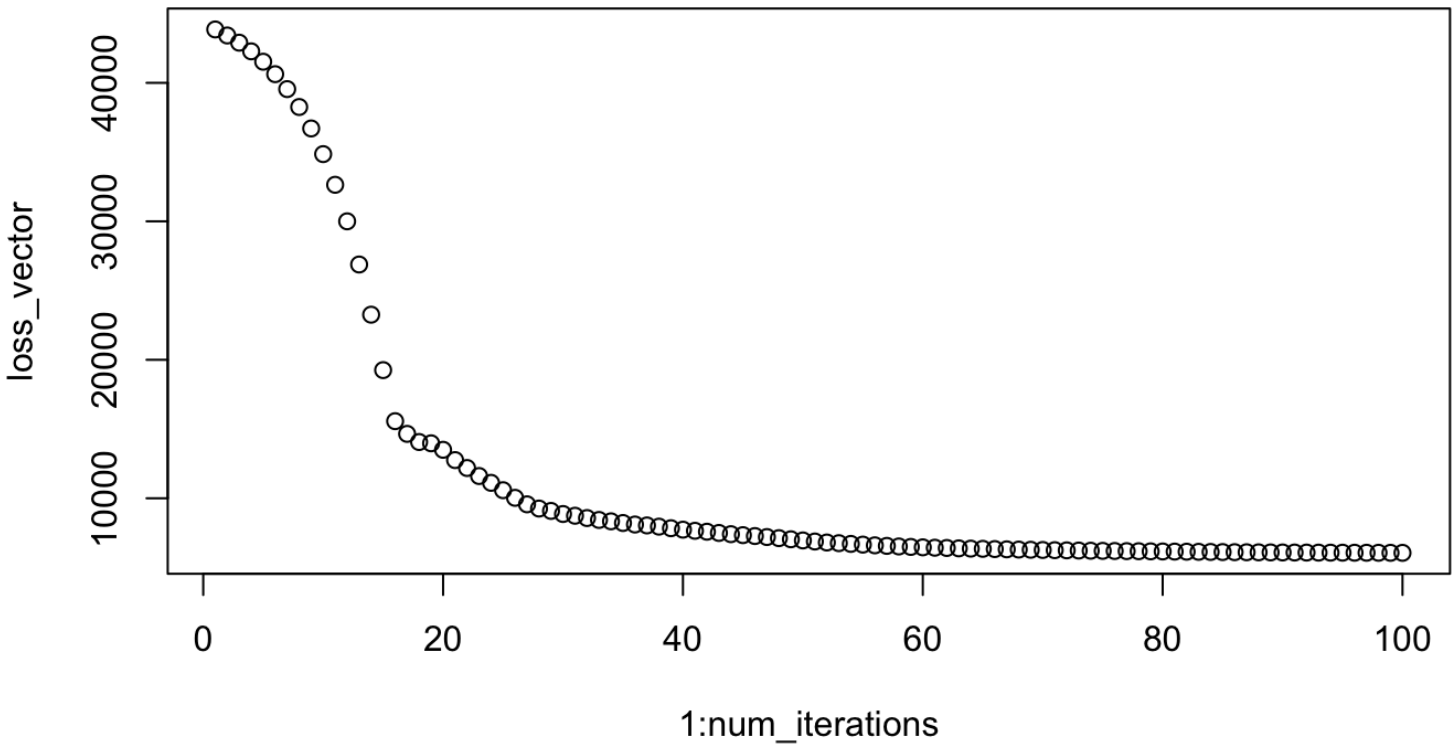
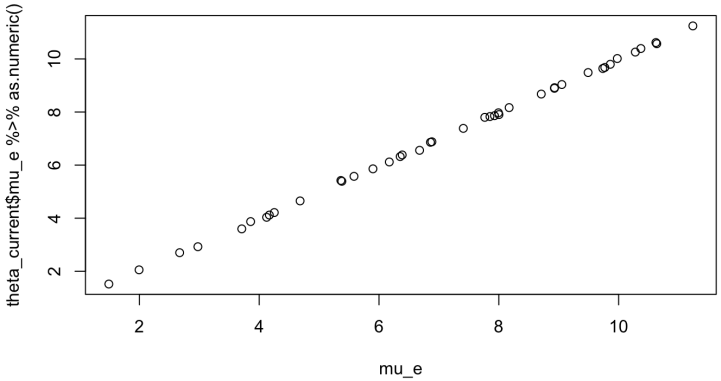
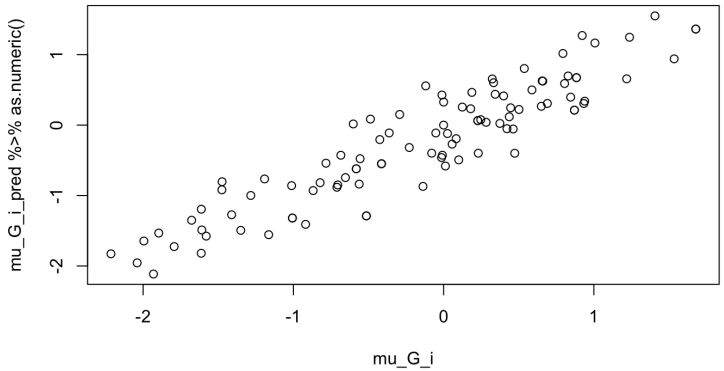
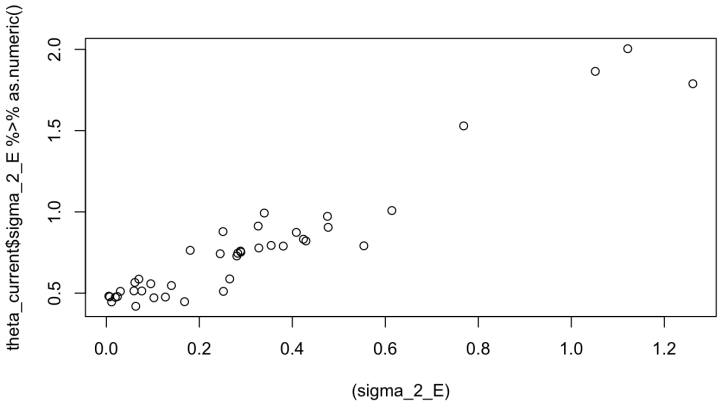


Figure: Vecteur de perte en fonction des itérations pour $n_{Ri,e} = 1$

Résultats obtenus







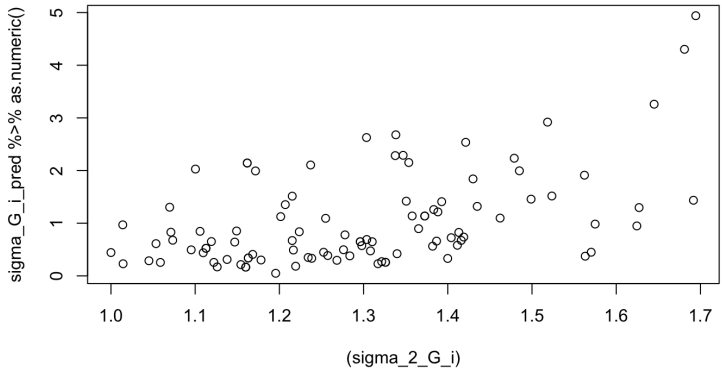


Figure: Paramètres inférés pour $n_{Ri,e} = 1$

Discussion

Problème d'identifiabilité ?

Pour $n_{Ri,e} = 1$:

$$\begin{aligned}\hat{\theta} &= \text{argmin}_\theta \, -\mathcal{L}(Y; \mu_e, \mu_{G_i}, \sigma_{G_i}, \sigma_{E_e}) \\ &= \text{argmin}_\theta \Big(\sum_{i,e} \log(\sigma_{G_i}^2 + \sigma_{E_e}^2) + \sum_{i,e} \frac{(Y_{i,e} - (\mu_e + \mu_{G_i}))^2}{\sigma_{G_i}^2 + \sigma_{E_e}^2}\Big)\end{aligned}$$

Cas général ($n_{Ri,e} \geq 1$):

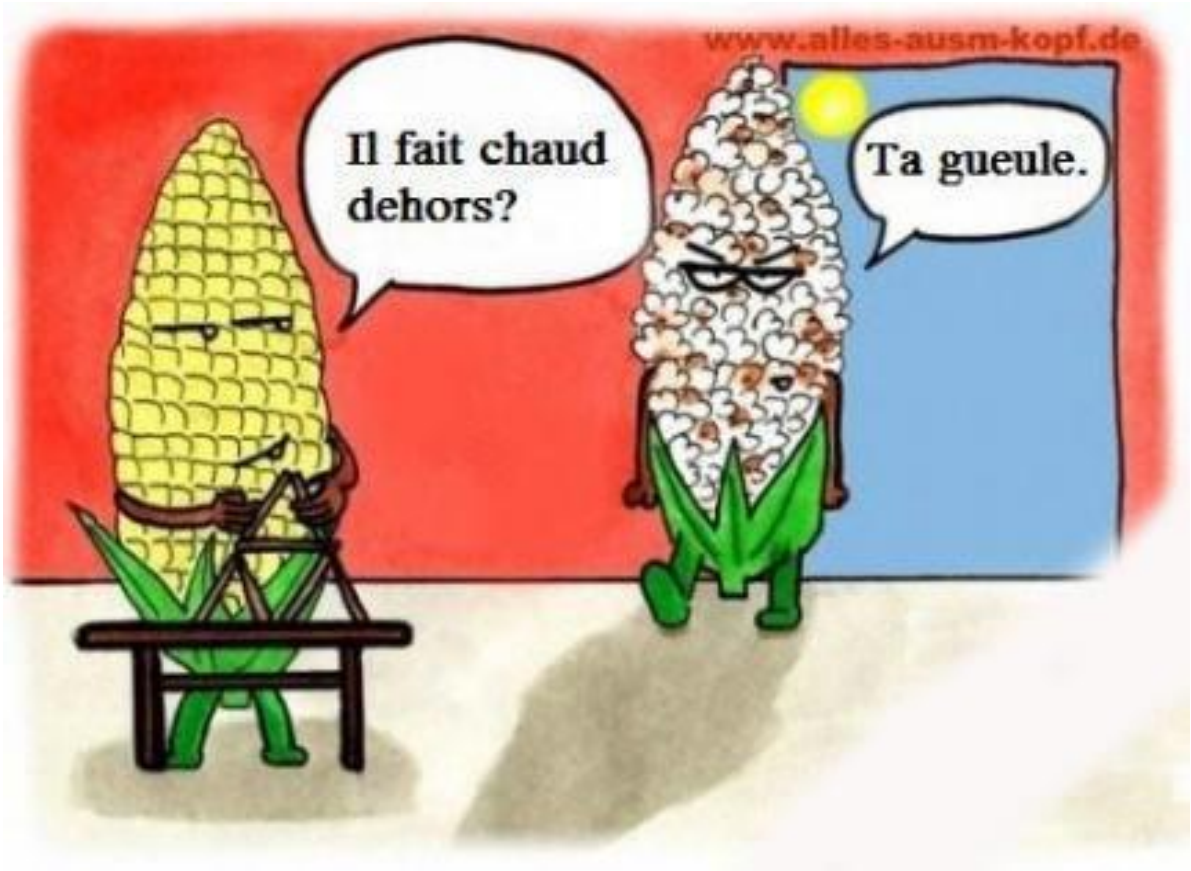
$$\begin{aligned}\mathcal{L}(Y) = \sum_{i,e} -\frac{1}{2\sigma_{E_e}^2}\Big(&\sum_{r=1}^{n_{Ri,e}} (Y_{i,e,r} - (\mu_e + \mu_{Gi}))^2 \\ &+ \frac{\sigma_{G_i}^2}{\sigma_{E_e}^2 + n_{Ri,e}\sigma_{G_i}^2} \left(\sum_{r=1}^{n_{Ri,e}} Y_{i,e,r} - (\mu_e + \mu_{Gi})\right)^2\Big) \\ &- \frac{1}{2}(n_{Ri,e} - 1)\log(\sigma_{E_e}^2) - \frac{1}{2}\log\left[(\sigma_{E_e}^2 + n_{Ri,e}\sigma_{G_i}^2)\right] - \frac{\log(2\pi)}{2} n_{Ri,e}\end{aligned}$$

Plan

Conclusion et perspectives

Pour conclure ...

Merci pour votre attention !



Bibilographie

- Martin W Ganal et al. "A large maize (Zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome" (2011)
- Emilie J Millet et al. "Genome-wide analysis of yield in Europe: allelic effects vary with drought and heat scenarios" (2016)
- Miguel A Raffo et al. "Genomic prediction for grain yield and micro-environmental sensitivity in winter wheat" (2023)
- Renaud Rincent et al. "Recovering power in association mapping panels with variable levels of linkage disequilibrium" (2014)
- Paul M VanRaden. "Efficient methods to compute genomic predictions" (2008)