Mathématiques pour les Sciences du Vivant

Rapport Final

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

Auteurs:
Salma Guennouni Assimi, Adrien Sardi

Encadrés par:
Tristan Mary-Huard (INRAE-AgroParisTech) et Laurence Moreau (INRAE, UMR GQE-Moulon)

Novembre 2023 - Mars 2024

Contents

1 Données **3**
- 1.1 Données génotypiques . . . 3
- 1.2 Données phénotypiques . . . 4

2 Modèle **5**
- 2.1 Espérance de Y . . . 5
- 2.2 Variance de Y . . . 5
- 2.3 Modèle final . . . 6
- 2.4 Calcul de la log-vraisemblance . . . 8

3 Génération des données **10**
- 3.1 Première génération des données . . . 10
- 3.2 Génération des paramètres de renormalisation grâce à un modèle linéaire mixte simplifié . . . 10
- 3.3 Adaptation des données générées à nos données réelles . . . 11

4 Inférence des paramètres **13**
- 4.1 Maximum de vraisemblance et paramètre à inférer . . . 13
- 4.2 Implémentation du maximum de vraisemblance . . . 13
- 4.3 Mini-batch . . . 14

5 Résultats et discussion **15**
- 5.1 Résultats . . . 15
- 5.2 Discussion . . . 17

6 Conclusion **19**

A Calcul de la Covariance **21**

B Calculs autour de la matrice $\Sigma_{i,e}$ **21**
- B.1 Inverse de la matrice $\Sigma_{i,e}$ . . . 21
- B.2 Déterminant de la matrice $\Sigma_{i,e}$ . . . 22

C Calcul de la vraisemblance **22**

Introduction

Dans le contexte actuel de changement climatique, un des enjeux majeurs de l'amélioration des plantes est de développer des variétés performantes mais aussi robustes face aux aléas climatiques (sécheresse, stress thermique ou encore gelées tardives) de plus en plus forts et de plus en plus fréquents. Leurs effets sont également accentués par l'utilisation de pratiques culturales plus respectueuses de l'environnement (réduction de l'irrigation, des engrais, des pesticides...).

Dans les programmes de sélection actuels, les variétés sont essentiellement sélectionnées pour leur performance moyenne observée dans un réseau de plusieurs essais représentatifs de la zone de culture, sans prendre en compte de façon explicite leur stabilité. Il est important de changer de paradigme en sélectionnant sur la stabilité des variétés au-delà de leur performance moyenne. L'objectif de nos travaux est donc d'évaluer la possibilité de changer de critère pour sélectionner sur la base de la distribution des performances et d'évaluer la possibilité de prédire ce critère sur la base du contenu génétique d'une variété.

Afin de réaliser cet objectif nous avons travaillé pendant plusieurs mois avec deux chercheurs spécialistes du domaine. Nous avions accès à une base de données comprenant plusieurs expériences menées sur des variétés de maïs. Pour chaque expérience nous avions des informations sur le phénotype des plantes comme le rendement, et pour chaque variété nous avions le génotype. Les expériences nous ont ainsi permis d'évaluer les variétés dans des conditions contrastées.

Dans ce rapport, nous proposons un modèle permettant la prédiction de la distribution de la performance d'un individu sur la base de son génotype. Nous implémentons ensuite en langage `R` la génération de données simulées, ainsi que l'inférence des paramètres du modèle par maximum de vraisemblance. En particulier, nous utilisons les outils de différentiation automatique disponibles dans le package `Torch` pour l'optimisation. Grâce aux données simulées, nous évaluons la performance de l'approche proposée.

1 Données

Nous disposons d'un jeu de données public [2] correspondant à un panel de 244 lignées de maïs génotypées pour 602 356 marqueurs SNP et évaluées pour leur performance hybride dans 22 environnements contrastés (combinaisons emplacement × année x conditions d'essais) caractérisés par des variables environnementales.

En d'autres terme nous avons à notre disposition deux types de données pour chacun de nos individus : des données génotypiques (les lignées de maïs pour 602 356 marqueurs SNP) et des données phénotypiques. Ces données phénotypiques correspondent à la réponse d'un individu (i.e. une plante) pour un environnement donné (un environnement correspondant à une combinaison emplacement × année x conditions d'essais).

Dans un premier temps, et afin de mieux comprendre nos données, nous allons les analyser. Nous commençons d'abord par les données génotypiques puis nous regarderons les données phénotypiques.

1.1 Données génotypiques

Dans un premier temps, nous avons choisi de restreindre notre analyse à un sous-ensemble de données en nous focalisant sur les données provenant d'une puce de génotypage 50k SNP, largement utilisée dans la communauté du maïs [1].

Afin de donner du sens à nos données, nous pouvons effectuer une Analyse en Composantes Principales (ACP). Cela permet une visualisation et une interprétation simplifiée des relations entre les variétés.

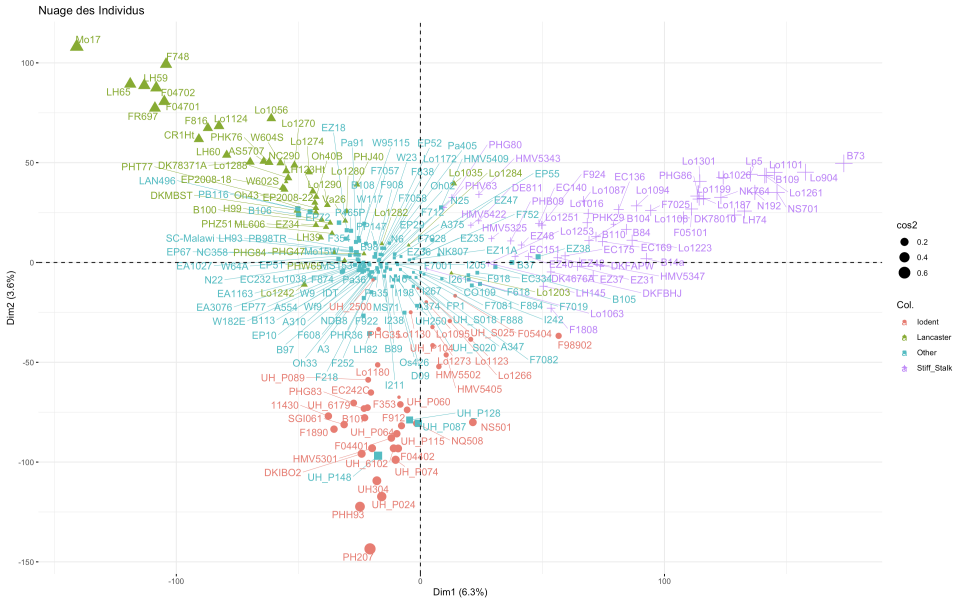

Figure 1: ACP des données génotypiques

La première chose que l'on remarque est la tendance de nos données et la structuration en groupes. On remarque de plus que partant du centre l'étalement se fait dans 3 directions qui mènent à 3 variétés particulières qui se démarquent: B73, Mo17 et PH207. La structuration en trois groupes découle de deux facteurs : Premièrement, la diversité du maïs au sein du groupe génétique étudié se présente naturellement sous forme de trois groupes distincts, avec B73, Mo17 et PH207 en tant que variétés emblématiques. Deuxièmement,

un biais de sélection des marqueurs sur la puce est observé, certains d'entre eux étant spécifiquement choisis pour leurs variations entre B73 (la première variété de maïs séquencée) et Mo17. En conséquence, les données de la puce amplifient la divergence génétique entre B73 et Mo17.

Concernant les marqueurs étudiés, il y a plusieurs points d'attention. Parmi les 50 000 marqueurs que l'on étudie, certains ont un nom qui débute par "SYN" et d'autres par "PZ". Ceux débutant par "SYN" présentent des biais dans leur détection et sont donc retirés de l'analyse pour éviter des conclusions erronées ou des distorsions dans les résultats. Par ailleurs, pour sélectionner des marqueurs ayant de bonnes propriétés, nous cherchons les marqueurs dont la fréquence allélique soit à peu près équilibrée.

1.2 Données phénotypiques

Comme expliqué précédemment, les données phénotypiques nous donnent la réponse des individus étudiés (i.e. les plantes) pour différents environnements. Chaque environnement correspond à une combinaison emplacement × année × conditions d'essais. Par exemple : la combinaison (Moulon, 2015, Pas d'irrigation) nous donne un environnement. L'ensemble des individus est exposé aux environnements, ce qui nous permet d'obtenir pour chaque individu sa réponse dans un environnement donnée.

Lorsqu'on parle de réponse, on peut considérer plusieurs caractéristiques : taille des graines, taille des plantes, nombre de graines, rendement,... Afin de mener notre étude, nous nous sommes intéressés à la donnée la plus pertinente : le rendement.

Nous étudions donc pour chaque individu son rendement pour un environnement donné. Notre objectif devient donc : **Prédire la variation du rendement d'un individu sur un ensemble d'environnements, à partir de son génotype**. C'est une question intéressante et pertinente à se poser car, comme on peut le voir en Figure 2, il est difficile de quantifier l'apport des individus et celui de l'environnement lorsqu'on s'intéresse au rendement.

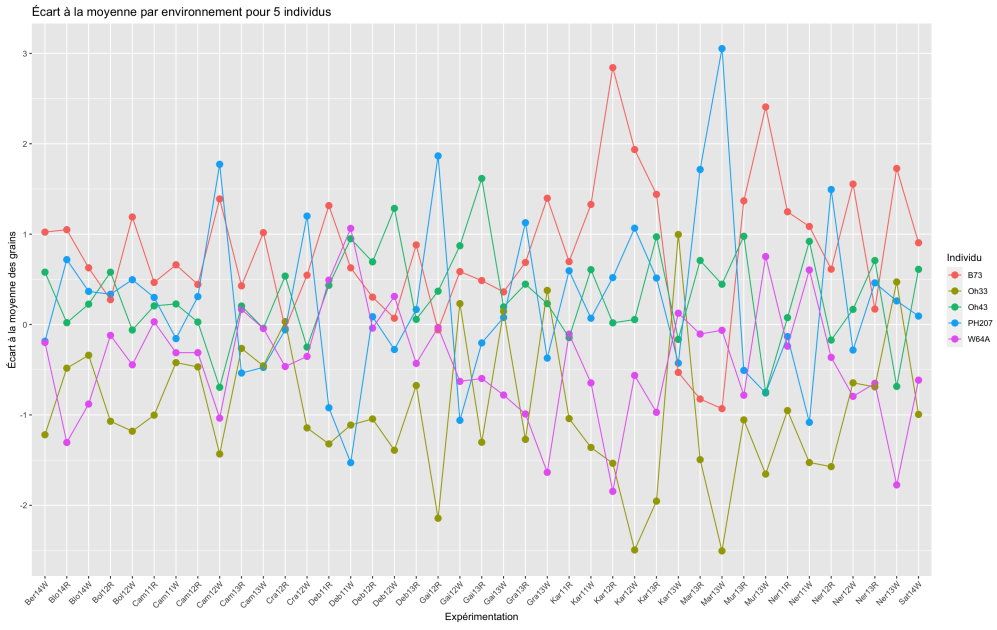

Figure 2: Écart à la moyenne à l'environnement pour différents individus

2 Modèle

L'objectif de cette partie est de proposer un modèle permettant de faire le lien entre l'information génétique des individus, l'environnement dans lequel on les observe et la distribution de leur rendement. En nous inspirant du travail de [3], nous allons construire un modèle linéaire mixte qui permet de décrire au mieux le rendement mais surtout de mettre en avant certains paramètres qui nous intéressent. Commençons par définir plusieurs grandeurs.

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

On peut remarquer ici que le nombre de répétitions est dépendant de i et de e. En effet, pour des données réelles le nombre de répétitions dépend du choix et des contraintes des expérimentateurs, et varie donc pour chaque individu dans chaque environnement.

2.1 Espérance de Y

Tout d'abord commençons par écrire l'espérance de Y. Comme vu dans la partie précédente, le rendement moyen dépend de l'environnement que l'on considère. Cela est logique car, par exemple, un environnement plus humide devrait entraîner un rendement plus important. Nous avons donc un terme de moyenne qui dépend de l'environnement. D'autre part, la moyenne dépend aussi de l'individu que l'on considère. Pour aller même plus loin, on souhaite pouvoir modéliser le fait que pour un environnement donné deux individus n'ont pas les mêmes performances. Et inversement, pour deux environnements différents, un individu n'a pas les mêmes performances. Finalement on souhaite :

$$\mathbb{E}[Y_{i,e,r}] = \mu + \alpha_e + \mu_{G_i}$$

Remarquons que le terme μ est une constante arbitraire pour le moment.

2.2 Variance de Y

Maintenant que l'on connaît les propriétés de la moyenne que l'on souhaite, on peut s'intéresser à la variance du rendement. Déjà, comme tout modèle, nous devons introduire un terme d'erreur pour rendre des effets que nous ne considérons pas ou que nous ne pouvons prévoir lors des expérience. Nous choisissons de plus que le terme d'erreur est de moyenne nulle. On a ainsi une variable d'erreur :

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

où les erreurs sont indépendantes d'un individu à un autre et d'un environnement à un autre, et indépendantes du génotype. On suppose ici que la variance d'erreur n'est pas la même dans tous les environnements.

Finalement, un dernier paramètre nous intéresse. Celui de la variance génétique, que nous pouvons aussi appeler robustesse. C'est même le paramètre qui nous intéresse le plus car c'est véritablement ce critère que l'on cherche à inférer.

Afin de proposer une expression pour le terme nous donnant la variance, nous pourrions nous inspirer de ce qui a été fait dans l'article [3] : $U \sim N(0, \sigma_G^2 A)$ où A est la matrice d'apparentement (dite aussi matrice de Kinship). Cependant, nous souhaitons prendre en compte le critère de robustesse propre à chaque individu i. On introduit donc :

$$U_{i,e} \sim N(\mu_{Gi}, {\sigma_G}_i^2),\ IND$$

Ainsi U est une variable aléatoire qui représente l'effet du génome. Les $U_{i,e}$ sont tirés de manière indépendante d'un environnement à un autre, et sont indépendants de l'erreur. En introduisant ce paramètre, nous souhaitons prédire plus précisément les caractéristiques de robustesse des individus que l'on étudie.

On note par ailleurs $x_{i,j}$ le trait de l'individu i au marqueur j. On fait l'hypothèse que que l'individu fait partie d'une lignée et que les $x_{i,j}$ que l'ont considère sont ceux qui ont un effet sur le génotype de l'individu.

Par ailleurs, nous pouvons préciser la forme que nous attendons pour nos paramètres :

- $\mu_{Gi} = f(g_i)$. Nous choisissons ici de considérer $f(g_i) = \sum_j x_{i,j}\beta_j$ où $x_{i,j}$ prend ses valeurs dans $\{0, 1\}$ et β_j représente l'effet du marqueur j sur la moyenne μ_{Gi} pour l'individu i. Ainsi, la ressemblance génétique entre les individus est contenue dans la moyenne $\mu_{G\,i}$.
- ${\sigma_G}_i^2 = f'(g_i) = e^{\sum_j x_{i,j}\beta'_j + \beta'_0}$ où la fonction exponentielle est utilisée pour imposer des valeurs positives pour la variance et où β'_j représente l'effet du marqueur j sur la variance $\sigma^2_{G\,i}$ pour l'individu i. Finalement, nous introduisons un terme β'_0 dans l'expression de la variance qui nous permet de modéliser une renormalisation de nos données.

2.3 Modèle final

Finalement, nous obtenons le modèle suivant :

$$\boxed{Y_{i,e,r} = (\mu + \alpha_e) + U_{i,e} + E_{i,e,r}}$$

où Y représente le phénotype de l'individu i dans l'environnement e pour la répétition r, α_e l'effet fixe de l'environnement, U l'effet du génotype, et E une erreur aléatoire associée aux mesures.

Par soucis d'identifiabilité, on regroupe la somme $\mu + \alpha_e$ en un seul terme que l'on note μ_e. Ainsi

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

$$\begin{aligned}
&U_{i,e} \sim N(\mu_{Gi}, {\sigma_G}_i^2),\ IND;\\
&\mu_{Gi} = \sum_j x_{i,j}\beta_j;\\
&{\sigma_G}_i^2 = e^{\sum_j x_{i,j}\beta'_j + \beta'_0}.
\end{aligned}$$

Enfin, pour le modèle d'erreur, on pose :

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

où les erreurs sont indépendantes d'un individu à un autre et d'un environnement à un autre, et indépendantes de U. On a bien $U \perp E$.

Le modèle ainsi défini a plusieurs propriétés :

$$\mathbb{E}[Y_{i,e,r}] = \mu + \alpha_e + \mu_{Gi} = \mu_e + \mu_{Gi}$$

et

$$\mathbb{V}[Y_{i,e,r}] = {\sigma_G}_i^2 + {\sigma_E}_e^2$$

Le premier résultat était attendu et même désiré. Il confirme que les effets de moyenne sont bien modélisés dans notre modèle. Ensuite, pour la variance, l'expression découle directement de la somme des deux lois normales. On voit apparaître le paramètre de robustesse, qui, comme nous l'avons dit, est le paramètre qui nous intéresse le plus et que l'ont cherche à inférer.

Pour le calcul de la covariance nous devons travailler un peu plus. En effet, l'introduction des répétitions (qui n'ont pas d'impact direct sur la définition de nos données) entraîne l'apparition d'un terme de covariance non nul. Ainsi, pour (i, e) différent de (i', e') on a

$$\mathrm{Cov}[Y_{i,e,r}, Y_{i',e',r'}] = {\sigma_G}_i^2 \delta_{i,i'} \delta_{e,e'}$$

car les $E_{i,e,r}$ sont indépendants, les $U_{i,e}$ également, et $U \perp E$. Le détail des calcul est en annexe.

Ainsi, la matrice de covariance de $Y_{i,e}$ (vecteur de taille $n_{Ri,e}$) pour i et e fixés s'écrit :

$$\Sigma_{i,e} = \begin{pmatrix} ({\sigma_G}_i^2 + \sigma_{E_e}^2) & \sigma_{G_i}^2 & \dots & \sigma_{Gi}^2 \\ \sigma_{Gi}^2 & ({\sigma_G}_i^2 + \sigma_{E_e}^2) & \dots & \sigma_{Gi}^2 \\ \dots & \dots & \dots & \dots \\ \sigma_{Gi}^2 & \sigma_{Gi}^2 & \dots & ({\sigma_G}_i^2 + \sigma_{E_e}^2) \end{pmatrix}$$

Ici, $\Sigma_{i,e}$ est une matrice de taille $n_{Ri,e} \times n_{Ri,e}$, $n_{Ri,e}$ étant un nombre qui peut varier pour chaque individu i et chaque expérience e. Il dépend complètement des données et des expériences.

Le calcul de la Covariance nous renseigne sur l'indépendance des $Y_{i,e,r}$ et des $Y_{i',e',r}$. Notamment on remarque qu'il n'y a de la dépendance qu'au sein des répétitions. Finalement, on peut réécrire la matrice de covariance sous la forme:

$$\Sigma_{i,e} = \sigma_{E_e}^2 I + \sigma_{Gi}^2 J$$

où I est la matrice identité et J la matrice pleine de 1.

2.4 Calcul de la log-vraisemblance

Comme les $Y_{i,e,r}$ suivent des lois normales dont nous avons l'expression des espérances et de la matrice de covariance, la vraisemblance peut maintenant être calculée. Cela va nous permettre d'appliquer une méthode de maximum de vraisemblance, que l'on implémentera à travers l'utilisation d'outils de différenciation automatique (cf Partie 4).

Calculons donc la log-vraisemblance :

$$\begin{aligned}\mathcal{L}(Y;\mu_e,\mu_{Gi},\sigma_{Gi},\sigma_{E_e}) &= \sum_{i,e}\mathcal{L}(Y_{i,e};\mu,\alpha_e,\mu_{Gi},\sigma_{Gi},\sigma_{E_e})_{|(i,e)} \\ &= \sum_{i,e}\left(-\log\sqrt{(2\pi)^{n_{Ri,e}}|\Sigma_{i,e}|} - \frac{\tilde{Y}_{i,e}^T\cdot\Sigma_{i,e}^{-1}\cdot\tilde{Y}_{i,e}}{2}\right)\end{aligned}$$

où $\tilde{Y}_{i,e}$ désigne le vecteur de taille $n_{Ri,e}$ recentré en 0. C'est à dire que pour tout $r\in[1,n_{Ri,e}]$

$$\tilde{Y}_{i,e,r} \coloneqq Y_{i,e,r} - \mathbb{E}[Y_{i,e,r}] = Y_{i,e,r} - (\mu_e+\mu_{Gi}).$$

Finalement, on obtient la propriété suivante :

Propriété 1

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

où n_I est le nombre d'individus et n_E est le nombre d'environnements, comme définis précédemment.

Nous remarquons que nous avons besoin d'inverser la matrice $\Sigma_{i,e}$. Ainsi, nous allons effectuer quelques calculs algébriques pour déterminer cet inverse. Après des calculs (tous détaillés en annexe), nous obtenons :

$$\Sigma_{i,e}^{-1} = \frac{1}{\sigma_{E_e}^2}I - \frac{\sigma_{Gi}^2}{\sigma_{E_e}^2(\sigma_{E_e}^2+n_{Ri,e}\cdot\sigma_{G_i}^2)}J$$

Nous pouvons également calculer le déterminant de la matrice $\Sigma_{i,e}$ (cf calculs en annexe), et on trouve

$$|\Sigma_{i,e}| = (\sigma_{E_e}^2+n_{Ri,e}\cdot\sigma_{Gi}^2)(\sigma_{E_e}^2)^{n_{Ri,e}-1}$$

Nous pouvons alors réécrire l'expression de la vraisemblance.

Propriété 2 *(Expression de la vraisemblance)*

$$\mathcal{L}(Y;\mu,\alpha_e,\mu_{Gi},\sigma_{Gi},\sigma_{E_e}) = \sum_{i,e}\frac{1}{2\sigma^2_{E_e}}\left\{-\sum_{r=1}^{n_{Ri,e}}\tilde{Y}^2_{i,e,r}+\frac{\sigma^2_{G_i}}{\sigma^2_{E_e}+n_{Ri,e}\sigma^2_{G_i}}\left(\sum_{r=1}^{n_{Ri,e}}\tilde{Y}_{i,e,r}\right)^2\right\}$$
$$-\frac{1}{2}\sum_{i,e}(n_{Ri,e}-1)\log(\sigma^2_{E_e})-\frac{1}{2}\sum_{i,e}\log\left[(\sigma^2_{E_e}+n_{Ri,e}\sigma^2_{G_i})\right]-\frac{\log(2\pi)}{2}\sum_{i,e}n_{Ri,e}$$

Que l'on peut aussi écrire

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

Nous avons ainsi une expression de la vraisemblance qui dépend uniquement des paramètres que l'on cherche à inférer ou bien de données déjà connues. Nous pourrons ainsi utiliser cette expression pour inférer les paramètres du modèle.

3 Génération des données

Afin de tester et de vérifier le modèle d'inférence via maximum de vraisemblance que nous construisons et détaillons en Partie 4, nous avons l'intention de générer des données simulées. Pour ce faire, nous allons suivre trois étapes distinctes :

- Tout d'abord, nous allons générer certaines données en utilisant leurs lois respectives.
- Ensuite, nous procéderons à la génération de paramètres de renormalisation en effectuant une inférence à l'aide d'un modèle linéaire mixte simplifié.
- Enfin, nous utiliserons les paramètres inférés pour renormaliser les premiers paramètres générés.

3.1 Première génération des données

Nous devons générer des données de rendement à partir des paramètres suivants : μ_e, μ_{G_i}, σ_{G_i}, σ_{E_e}.

La première étape consiste à fixer les grandeurs utilisées : nombre d'individus, nombre de marqueurs, nombre de répétitions et nombre d'environnements. Ici, nous effectuons nos simulations en considérant un nombre d'individus égal à 100 (parmi les 230 individus), un nombre de marqueurs égal à 10 (parmi les 50 000 marqueurs), et en utilisant tous les environnements à notre disposition.

Afin de générer μ_{G_i} et σ_{G_i} nous souhaitons passer par la génération des $x_{i,j}$, β_j et β'_j. Concernant les $x_{i,j}$, il est compliqué de les générer efficacement par un tirage prenant en compte la complexité génétique des plantes et la dépendance entre les individus. En effet, cette dépendance est déterminée par la liaison physique des marqueurs le long des chromosomes mais également par l'histoire évolutive du panel d'individus de maïs étudiés. C'est pourquoi nous préférons utiliser les données génétiques que nous possédons, i.e. nous conservons la matrice des $x_{i,j}$ pour générer les données.

Concernant les β_j et β'_j, ceux-ci sont générés en tirant une loi normale centrée réduite dans un premier temps. Nous faisons de même pour les variances d'erreurs $\sigma^2_{E_e}$, qui suivent également une loi normale centrée réduite.

Pour que nos données soient réalistes, nous allons ensuite procéder à une renormalisation de ces paramètres. Enfin, concernant les μ_e, nous allons également nous baser sur nos données réelles pour la génération des données. Pour cela, nous allons inférer les valeurs de plusieurs paramètres à partir d'un modèle simplifié.

3.2 Génération des paramètres de renormalisation grâce à un modèle linéaire mixte simplifié

Nous souhaitons obtenir des valeurs cohérentes avec nos données. Pour cela nous allons inférer les paramètres du modèle linéaire mixte suivant :

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

où cette fois $U \sim N(0, \sigma^2_G A)$ avec σ^2_G qui représente la variance génétique et A qui est la matrice de parenté (Kinship) et $E \sim N(0, \sigma^2_{E_e})$. Nous allons inférer μ_e et σ_{E_e} pour chaque environnement et cela nous donnera les valeurs que l'on cherche.

La première étape consiste à construire les données Y et x dont on a besoin. En particulier on construit Y_e et x_e pour un environnement et pour une expérience données.

Un fois cela fait, nous calculons la matrice dite de Kinship, i.e. la matrice de parenté entre les individus. Ici la matrice est calculée selon la méthode de Van Raden [5]. Ce calcul commence par un centrage des données génotypiques, en soustrayant la moyenne de chaque colonne. Ensuite, un produit matriciel est effectué entre les données centrées et leur transposée. Enfin, cette matrice est normalisée en multipliant chaque élément par le nombre de colonnes de la matrice et en divisant par la somme de sa diagonale. Cette normalisation assure notamment que la diagonale de la matrice de parenté est égale à 1, fournissant ainsi une mesure de la similarité génétique entre chaque paire d'individus. La Figure 3 présente la matrice de Kinship obtenue.

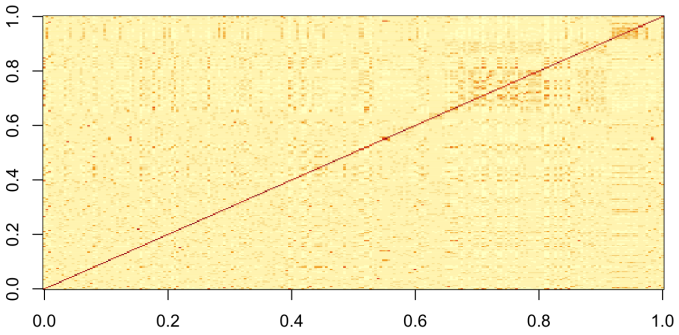

Figure 3: Matrice de Kinship

Nous passons ensuite à l'inférence des paramètres. Pour cela, nous utilisons la bibliothèque MM4LMM [4], développée notamment par Tristan Mary-Huard. Nous récupérons ainsi les valeurs des paramètres recherchés.

3.3 Adaptation des données générées à nos données réelles

Une fois que cela est fait nous pouvons renormaliser nos données. Les effets de l'environnement, les erreurs et les effets génétiques pour chaque individu, sont renormalisés grâce aux valeurs des paramètres inférés grâce au modèle simplifié, pour garantir que leur espérance et variance soit cohérente avec l'espérance et la variance génétique attendues.

Finalement, nous obtenons nos données. Une bonne façon de procéder pour vérifier la cohérence de nos données et de vérifier si la moyenne et la variance du Y_{gen} généré est proche des données réelles. Pour cela calculons la moyenne et la variances de nos données pour plusieurs expériences.

Nous obtenons en Figure 4 la moyenne de notre rendement généré et nous pouvons le comparer à notre rendement réel. Nous sommes assez satisfaits des résultats car nous remarquons que les tendances ainsi que les valeurs sont proches des données réelles, ce qui est cohérent car nous avons utilisé les moyennes issues des données réelles pour générer nos données. Cela nous donne une première façon de vérifier que notre génération de données se passe correctement.

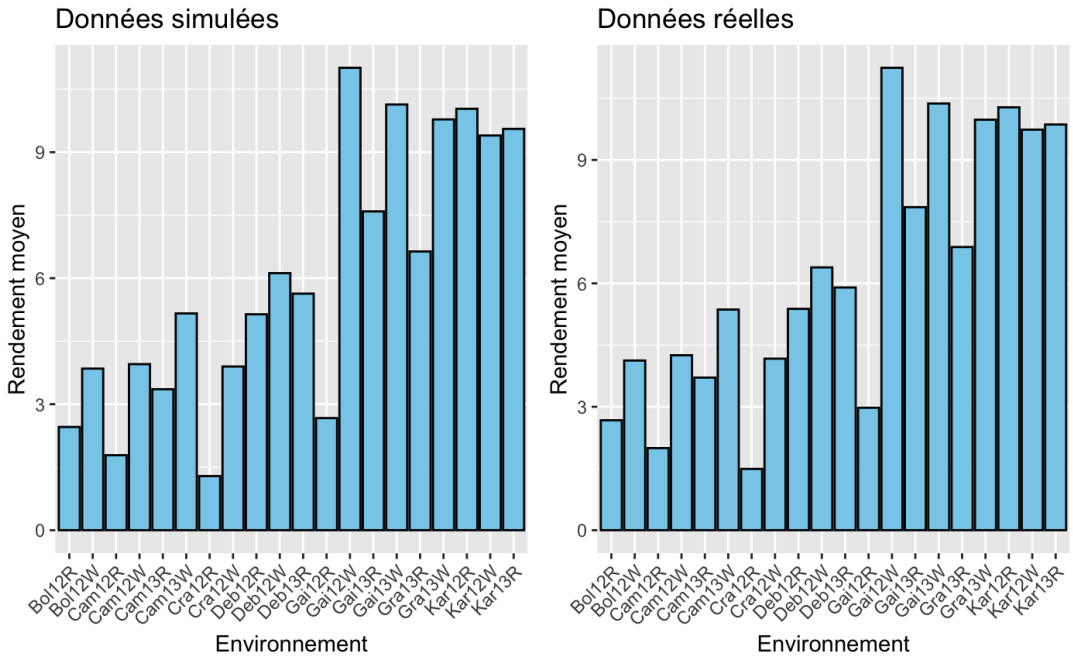

Figure 4: Comparaison du rendement moyen généré avec les données réelles

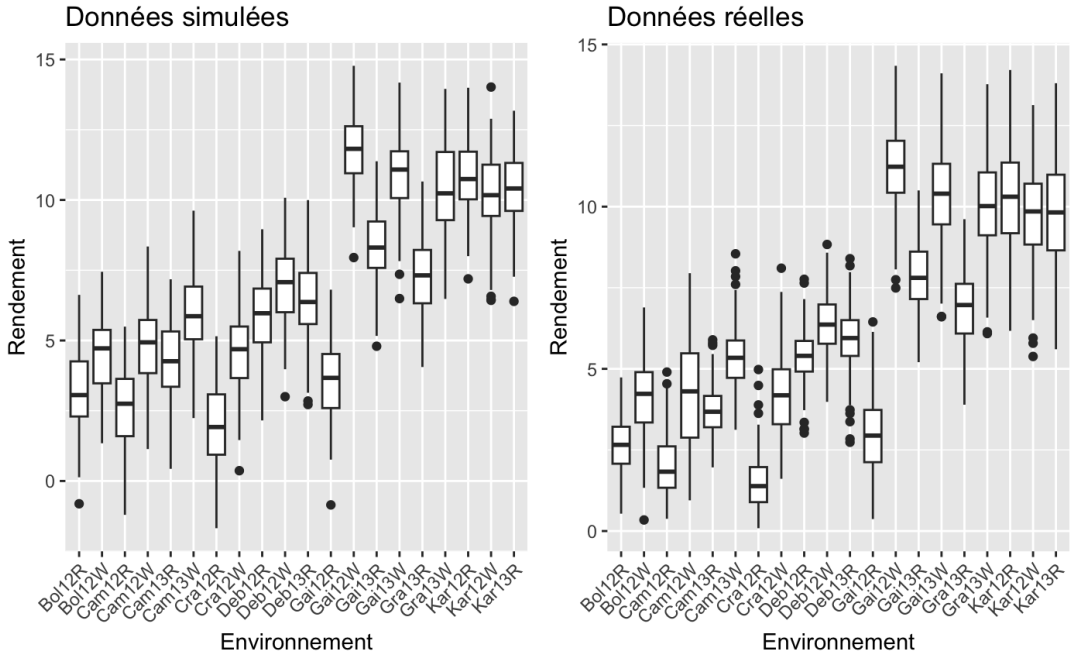

Figure 5: Comparaison de la distribution statistique des données générées avec les données réelles

La Figure 5 nous montre la distribution statistique pour les données générées. Une fois encore nous sommes satisfaits de nos résultats car nous remarquons qu'en tendance et en valeur nos résultats sont très proches des résultats réels. Nous allons pouvoir utiliser nos données générées pour tester notre modèle.

4 Inférence des paramètres

4.1 Maximum de vraisemblance et paramètre à inférer

Repartons de la Propriété 2 qui nous donne une expression de la vraisemblance. Posons $\theta = (\mu_e, \mu_{Gi}, \sigma_{Gi}, \sigma_{E_e})$. Maximiser la vraisemblance revient à minimiser l'opposé de la vraisemblance. On cherche donc $\hat{\theta}$ comme étant

$$\hat{\theta} \coloneqq \text{argmin}_\theta \, -\mathcal{L}(Y; \mu_e, \mu_{Gi}, \sigma_{Gi}, \sigma_{E_e})$$

En utilisant la Propriété 2 nous en déduisons :

Propriété 3

$$\begin{aligned}
\hat{\theta} &= \text{argmin}_\theta \, -\mathcal{L}(Y; \mu_e, \mu_{Gi}, \sigma_{Gi}, \sigma_{E_e}) \\
&= \text{argmin}_\theta \sum_{i,e} \frac{1}{\sigma^2_{E_e}} \left\{ \sum_{r=1}^{n_{Ri,e}} \left(Y_{i,e,r} - (\mu_e + \mu_{Gi})\right)^2 - \frac{\sigma^2_{G_i}}{\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i}} \left(\sum_{r=1}^{n_{Ri,e}} Y_{i,e,r} - (\mu_e + \mu_{Gi}) \right)^2 \right\} \\
&\quad + \sum_{i,e} (n_{Ri,e} - 1) \log(\sigma^2_{E_e}) + \sum_{i,e} \log\left[(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i}) \right]
\end{aligned}$$

En ayant retiré les termes n'influençant pas la maximisation.

Dans notre cas nous pouvons même réexprimer cette valeur en fonction de β et β' comme définis précédemment. Cela nous donne alors

$$\begin{aligned}
\hat{\theta} = \text{argmin}_\theta \sum_{i,e} \frac{1}{\sigma^2_{E_e}} &\left\{ \sum_{r=1}^{n_{Ri,e}} \left(Y_{i,e,r} - (\mu_e + \sum_j x_{i,j}\beta_j) \right)^2 - \frac{e^{\beta'_0 + \sum_j x_{i,j}\beta'_j}}{\sigma^2_{E_e} + n_{Ri,e} \cdot e^{\beta'_0 + \sum_j x_{i,j}\beta'_j}} \left(\sum_{r=1}^{n_{Ri,e}} Y_{i,e,r} - (\mu_e + \sum_j x_{i,j}\beta_j) \right)^2 \right\} \\
&+ \sum_{i,e} (n_{Ri,e} - 1) \log(\sigma^2_{E_e}) + \sum_{i,e} \log\left[(\sigma^2_{E_e} + n_{Ri,e} \cdot e^{\beta'_0 + \sum_j x_{i,j}\beta'_j}) \right]
\end{aligned}$$

Nous pouvons alors utiliser cette expression pour construire en `R` une fonction qui réalise l'inférence. Pour cela nous utilisons la librairie `Torch` afin de trouver un minimiseur de notre fonction.

4.2 Implémentation du maximum de vraisemblance

Afin de réaliser le maximum de vraisemblance, nous réalisons en réalité un minimum de -logvraisemblance. Nous écrivons donc une fonction en utilisant la librairie `Torch` qui calcule l'opposé de la logvraisemblance $-\mathcal{L}$, et nous utilisons une méthode d'optimisation fournie par la bibliothèque `Torch` pour trouver les valeurs des paramètres du modèle qui minimisent $-\mathcal{L}$. Pour cela, nous utilisons une méthode de descente de gradient, et ici plus précisément l'algorithme Rprop (Resilient Backpropagation). Nous spécifions alors le nombre

d'itérations pour l'optimisation et à chaque itération, nous réinitialisons les gradients, calculons $-\mathcal{L}$, effectuons la rétropropagation et mettons à jour les paramètres. Quand le nombre maximal d'itérations est atteint, on a alors accès aux valeurs des paramètres estimés $\hat{\theta} = \text{argmin}_\theta -\mathcal{L}(Y; \mu_e, \mu_{Gi}, \sigma_{Gi}, \sigma_{E_e})$.

```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 <- 500
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()
}
```

Listing 1: Algorithme d'optimisation utilisant la méthode Rprop

4.3 Mini-batch

Lors de l'optimisation des paramètres de notre modèle, et surtout car on dispose d'un grand nombre de données, il est avantageux d'utiliser des mini-batch de données au lieu d'utiliser l'ensemble complet des données à chaque itération. On définit alors un nombre d'epochs qui décrit le nombre de fois où le modèle voit l'ensemble complet des données d'entraînement. Ainsi, lors d'une epoch, le jeu de données d'entraînement est divisé en plusieurs batches et chaque batch est utilisé pour calculer les gradients et mettre à jour les paramètres du modèle. Après avoir parcouru tous les batches de l'ensemble de données, une epoch est terminée. L'apprentissage se poursuit donc sur plusieurs epochs, où chaque epoch consiste à parcourir l'ensemble complet des données d'entraînement. Cela permet un entraînement efficace sur de grands ensembles de données tout en réduisant notre temps de calcul et en améliorant la stabilité de notre optimisation.

5 Résultats et discussion

5.1 Résultats

Nous effectuons l'optimisation de nos paramètres selon le code 1. La figure 6 illustre la progression de la minimisation de l'opposé de la logvraisemblance, au fil des itérations.

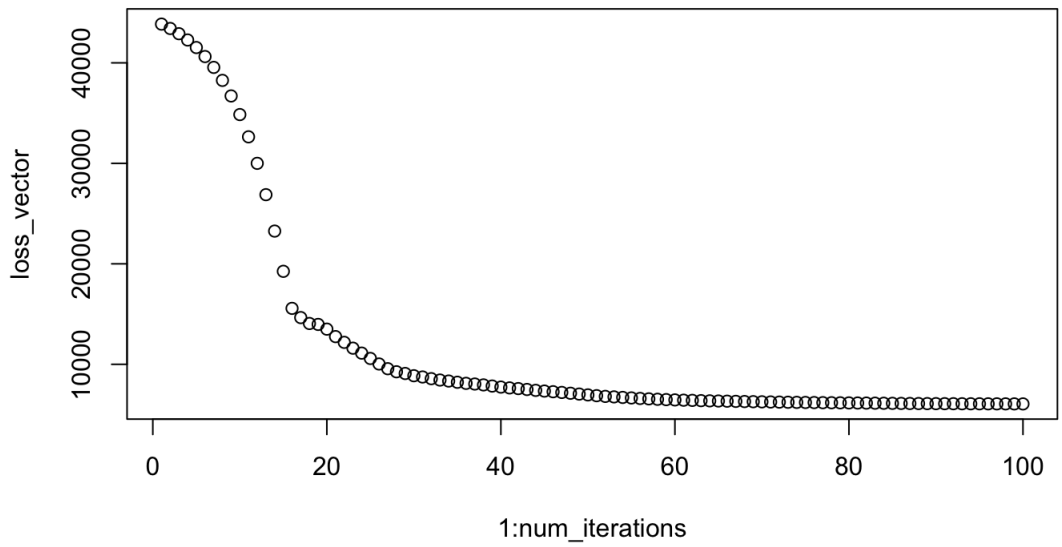

Figure 6: Optimisation au fil des itérations

Nous observons une diminution claire de la fonction de perte au fur et à mesure des itérations, ce qui indique une amélioration continue des estimations de nos paramètres. Cette tendance confirme l'efficacité de notre algorithme d'optimisation dans la recherche des valeurs optimales.

Ensuite, une fois l'optimisation terminée, nous pouvons accéder aux valeurs estimées des paramètres du modèle. Nous comparons ici ces valeurs estimées avec les valeurs réelles pour évaluer la performance de notre modèle. Pour ce faire, nous traçons des graphiques de dispersion où chaque point représente une paire de valeurs : une valeur réelle et une valeur prédite par notre modèle.
Ces graphiques de dispersion nous permettent d'observer visuellement l'ajustement entre les valeurs réelles et les valeurs prédites. Une bonne adéquation serait représentée par des points alignés sur une ligne diagonale, ce qui signifierait une bonne correspondance entre les valeurs réelles et prédites. Une telle correspondance refléterait la capacité de notre modèle à prédire efficacement les valeurs en fonction des données disponibles, alors que des écarts significatifs entre les valeurs réelles et prédites pourraient indiquer des lacunes dans notre modèle ou des biais dans nos données.

Premièrement, nous comparons les valeurs réelles et prédites pour le paramètre μ_e, qui représente les effets environnementaux. Ensuite, nous faisons de même pour le paramètre $\sigma^2_{E_e}$, qui représente la variance des erreurs. Enfin, nous comparons les valeurs réelles et prédites pour les paramètres μ_{G_i} et $\sigma^2_{G_i}$, qui représentent respectivement les moyennes génétiques et leurs variances pour chaque individu. Nous utilisons également des graphiques de dispersion pour cette comparaison.

Les résultats obtenus sont les suivants :

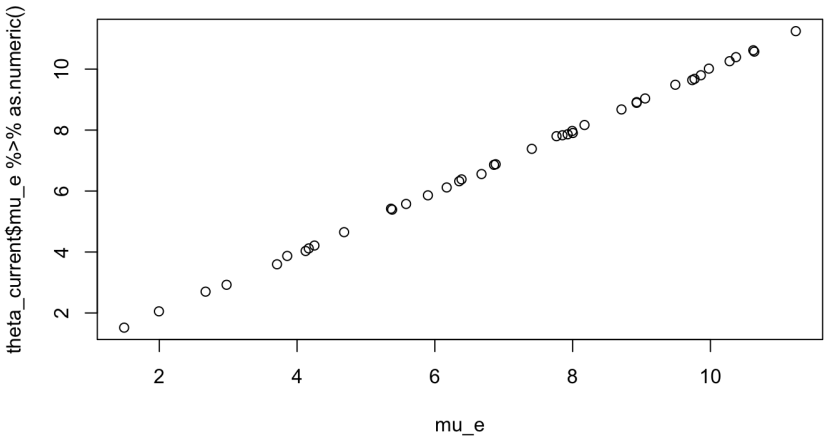

Figure 7: Comparaison des valeurs réelles et prédites pour μ_e.

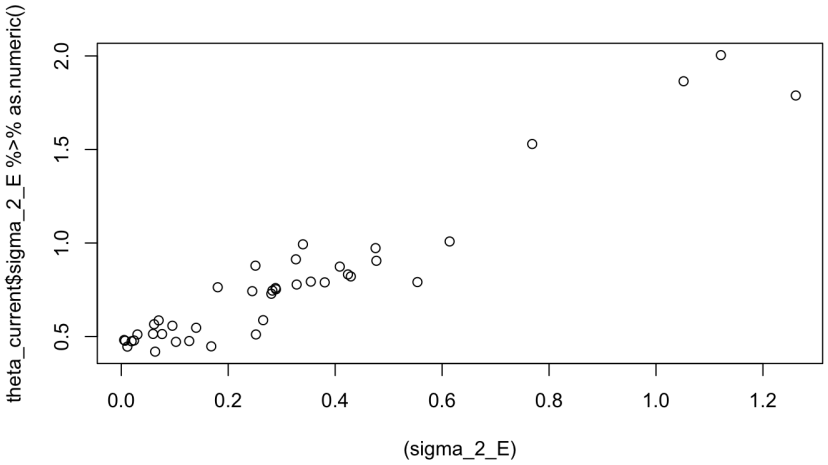

Figure 8: Comparaison des valeurs réelles et prédites pour $\sigma^2_{E_e}$.

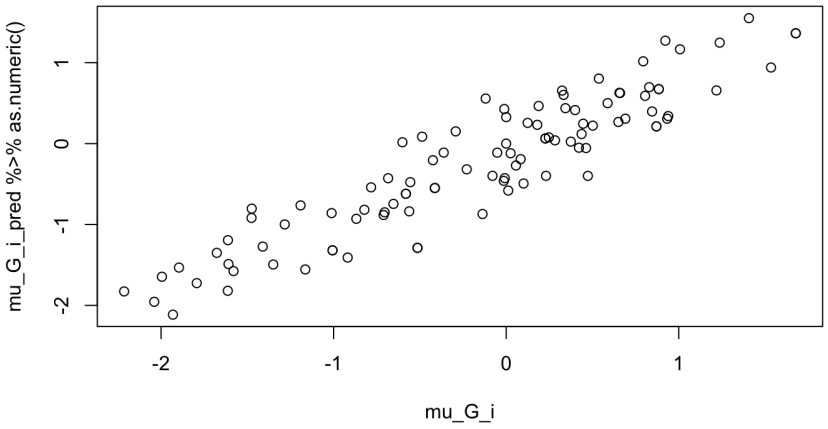

Figure 9: Comparaison des valeurs réelles et prédites pour μ_{G_i}.

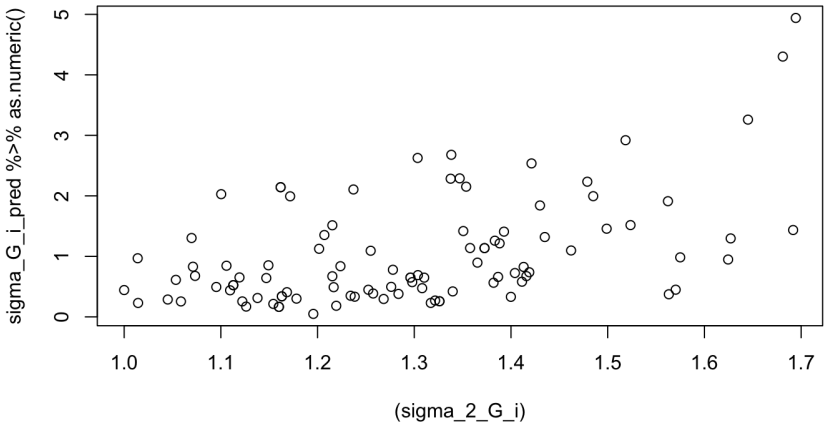

Figure 10: Comparaison des valeurs réelles et prédites pour $\sigma^2_{G_i}$.

5.2 Discussion

Nous remarquons que les valeurs de μ_e sont quasi-parfaitement prédites, comme en témoigne la fonction linéaire que nous obtenons lors du tracé du graphique de dispersion dans la Figure 7.
Pour $\sigma^2_{E_e}$ et μ_{G_i}, bien que la correspondance ne soit pas parfaite, nous observons néanmoins une concordance raisonnable entre les valeurs réelles et prédites, indiquant une certaine capacité du modèle à estimer ces paramètres, bien que beaucoup moins précisément que pour μ_e.
En revanche, pour $\sigma^2_{G_i}$, les variances des moyennes génétiques, nous observons des écarts bien plus significatifs entre les valeurs réelles et prédites. Une première piste d'explication de ces résultats peut venir de l'absence de répétition dans les tests que nous avons menés. En effet, comme détaillé plus haut, la variance du rendement est égale à la somme de la variance génétique et de la variance de l'erreur.

$$\mathbb{V}[Y_{i,e,r}] = \sigma^2_{Gi} + \sigma^2_{E_e}$$

En l'absence de répétition, le modèle n'a pas de réel moyen de distinguer quelle composante de la variance vient d'un effet génétique et quelle composante vient d'un effet d'erreur. En inférant nos paramètres sur des données comportant des répétitions, nous permettrons à l'algorithme de distinguer quelle partie de la variance vient d'un effet d'erreur et quelle partie vient d'un effet génétique. En effet, la variable d'erreur $E_{e,r}$ est tirée pour chaque répétition alors que le terme $U_{i,e}$ lui est le même entre les répétitions.
Nous n'avons pas eu suffisamment de temps pour explorer davantage et résoudre les écarts observés dans la prédiction des valeurs de $\sigma^2_{G_i}$. Cela nécessitera une analyse plus approfondie et une exploration plus détaillée des facteurs sous-jacents.

6 Conclusion

Dans ce projet, nous avons proposé un nouveau modèle statistique pour étudier les rendements de variétés de maïs sur la base de la distribution de sa valeur génétique. Ce modèle est particulièrement novateur car il prend compte de la robustesse d'un individu, ce qui, à notre connaissance, n'avait jamais été fait auparavant. Nous avons ensuite mis en œuvre ce modèle en utilisant le langage de programmation `R`, en générant des données simulées et en réalisant l'inférence des paramètres par maximum de vraisemblance grâce aux outils de différentiation automatique disponibles dans le package `Torch`.

Grâce à nos données simulées, nous avons pu évaluer la performance de notre approche. Les résultats obtenus ont montré que notre modèle était capable de prédire avec succès certains des paramètres étudiés. Cependant, des défis subsistent, notamment en ce qui concerne la prédiction du paramètre de robustesse. Il est clair que des travaux supplémentaires sont nécessaires pour améliorer la capacité de notre modèle à prédire ce paramètre crucial.

Notre travail constitue donc une première étape prometteuse dans le développement d'un modèle prédictif de la performance des individus en fonction de leur génotype. Nos résultats fournissent une base solide pour des recherches futures visant à affiner et à améliorer notre approche, en vue d'une application pratique dans le domaine de la génétique quantitative.

Ce projet a été extrêmement enrichissant car il nous a permis tout d'abord de nous familiariser davantage avec les enjeux et les concepts de génétique quantitative, notamment grâce à l'expertise de Mme. Laurence Moreau. De plus, éclairés par l'expertise de M. Tristan Mary-Huard, nous avons également pu mettre en pratique les outils de modélisation introduits dans les différents cours du Master, et nous avons pu approfondir nos connaissances sur les outils d'optimisation modernes couramment utilisés en intelligence artificielle, ouvrant ainsi de nouvelles perspectives de recherche et d'application dans ce domaine en constante évolution.

Nous remercions chaleureusement nos encadrants qui ont été d'une grande aide tout au long de ce projet.

References

[1] 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". In: *PloS one* 6.12 (2011), e28334.

[2] Emilie J Millet et al. "Genome-wide analysis of yield in Europe: allelic effects vary with drought and heat scenarios". In: *Plant Physiology* 172.2 (2016), pp. 749–764.

[3] Miguel A Raffo et al. "Genomic prediction for grain yield and micro-environmental sensitivity in winter wheat". In: *Frontiers in Plant Science* 13 (2023), p. 1075077.

[4] Renaud Rincent et al. "Recovering power in association mapping panels with variable levels of linkage disequilibrium". In: *Genetics* 197.1 (2014), pp. 375–387.

[5] Paul M VanRaden. "Efficient methods to compute genomic predictions". In: *Journal of dairy science* 91.11 (2008), pp. 4414–4423.

A Calcul de la Covariance

Reprenons le calcul de la covariance de Y.

$$\begin{aligned}
\mathrm{Cov}[Y_{i,e,r}, Y_{i',e',r'}] &= \mathrm{Cov}[(\mu+\alpha_e) + U_{i,e} + E_{i,e,r}, (\mu+\alpha_{e'}) + U_{i',e'} + E_{i',e',r'}] \\
&= \mathrm{Cov}[U_{i,e} + E_{i,e,r}, U_{i',e'} + E_{i',e',r'}] \\
&= \mathrm{Cov}[U_{i,e}, U_{i',e'}] + \mathrm{Cov}[E_{i,e,r}, E_{i',e',r'}] + \mathrm{Cov}[U_{i',e'}, E_{i,e,r}] + \mathrm{Cov}[U_{i,e}, E_{i',e',r'}] \\
&= \mathrm{Cov}[U_{i,e}, U_{i',e'}] \\
&= \mathrm{V}[U_{i,e}] \cdot \delta_{i,i'} \cdot \delta_{e,e'} \\
&= \sigma^2_{G_i} \delta_{i,i'} \delta_{e,e'}
\end{aligned}$$

Finalement on obtient bien,

$$\mathrm{Cov}[Y_{i,e,r}, Y_{i',e',r'}] = \sigma^2_{G_i} \delta_{i,i'} \delta_{e,e'}$$

B Calculs autour de la matrice $\Sigma_{i,e}$

B.1 Inverse de la matrice $\Sigma_{i,e}$

Fixons un couple $(i,e) \in [1,n_i] \times [1,n_E]$. Dans le but d'alléger les calculs nous appellerons $\Sigma \coloneqq \Sigma_{i,e}$ et nous noterons $n_{Ri,e} = n_{Ri,e}$. Pour rappel, $\Sigma_{i,e}$ est une matrice de taille $n_{Ri,e}$ et $\Sigma = \sigma^2_{E_e} I + \sigma^2_{G_i} J$.

$$\Sigma = \sigma^2_{E_e} I + \sigma^2_{G_i} J \implies \Sigma^2 = \sigma^4_{E_e} I + 2\sigma^2_{E_e}\sigma^2_{G_i} J + \sigma^4_{G_i} J^2$$

Or J est la matrice pleine de 1, de taille $n_{Ri,e}$. Ainsi, on remarque facilement que $J^2 = n_{Ri,e} J$
Dès lors,

$$\Sigma^2 = \sigma^4_{E_e} I + (2\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})\sigma^2_{G_i} J$$

On obtient donc

$$\Sigma^2 = (2\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})(\sigma^2_{E_e} I + \sigma^2_{G_i} J) - (2\sigma^4_{E_e} + n_{Ri,e}\sigma^2_{E_e}\sigma^2_{G_i} - \sigma^4_{E_e}) I = (2\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})\Sigma - \sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i}) I$$

Nous avons donc déterminé un polynôme annulateur de la matrice de covariance $\Sigma_{i,e}$, et celui-ci défini par:

$$P(X) = X^2 - (2\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i}) X + \sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i}) I$$

Dès lors, nous pouvons déterminer l'inverse de $\Sigma_{i,e}$:

$$P(\Sigma) = 0 \iff \Sigma^2 - (2\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})\Sigma + \sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i}) I = 0 \iff \Sigma[\frac{-1}{\sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})}(\Sigma - (2\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i}) I)] = I$$

Or $\Sigma - (2\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i}) I = \sigma^2_{E_e} I + \sigma^2_{G_i} J - (2\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i}) I = -(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i}) I + \sigma^2_{G_i} J$
D'où $\Sigma^{-1} = \frac{-1}{\sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})}(-(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i}) I + \sigma^2_{G_i} J) = \frac{1}{\sigma^2_{E_e}} I - \frac{\sigma^2_{G_i}}{\sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})} J$
On a donc:

$$\Sigma^{-1} = \begin{pmatrix}
\frac{(n_{Ri,e}-1)\sigma^2_{G_i} + \sigma^2_{E_e}}{\sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})} & -\frac{\sigma^2_{G_i}}{\sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})} & \cdots & -\frac{\sigma^2_{G_i}}{\sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})} \\
-\frac{\sigma^2_{G_i}}{\sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})} & \frac{(n_{Ri,e}-1)\sigma^2_{G_i} + \sigma^2_{E_e}}{\sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})} & \cdots & -\frac{\sigma^2_{G_i}}{\sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})} \\
\cdots & \cdots & \cdots & \cdots \\
-\frac{\sigma^2_{G_i}}{\sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})} & -\frac{\sigma^2_{G_i}}{\sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})} & \cdots & \frac{(n_{Ri,e}-1)\sigma^2_{G_i} + \sigma^2_{E_e}}{\sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})}
\end{pmatrix}$$

soit en factorisant :

$$\Sigma^{-1} = \frac{1}{\sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})} \begin{pmatrix} (n_{Ri,e}-1)\sigma^2_{G_i} + \sigma^2_{E_e} & -\sigma^2_{G_i} & ... & -\sigma^2_{G_i} \\ -\sigma^2_{G_i} & (n_{Ri,e}-1)\sigma^2_{G_i} + \sigma^2_{E_e} & ... & -\sigma^2_{G_i} \\ ... & ... & ... & ... \\ -\sigma^2_{G_i} & -\sigma^2_{G_i} & ... & (n_{Ri,e}-1)\sigma^2_{G_i} + \sigma^2_{E_e} \end{pmatrix}$$

Dans la suite nous préférerons utiliser la forme compacte :

$$\begin{aligned} \Sigma^{-1} &= \frac{1}{\sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})} \left((n_{Ri,e}\sigma^2_{G_i} + \sigma^2_{G_e})I - \sigma^2_{G_i}J\right) \\ &= \frac{1}{\sigma^2_{E_e}} I - \frac{\sigma^2_{G_i}}{\sigma^2_{E_e}(\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})} J \end{aligned}$$

B.2 Déterminant de la matrice $\Sigma_{i,e}$

Nous pouvons également réaliser le calcul du déterminant de la matrice $\Sigma_{i,e}$. Comme $\Sigma_{i,e}$ s'écrit sous la forme $\sigma^2_{E_e} I + \sigma^2_{G_i} J$, alors v est vecteur propre de $\Sigma = \sigma^2_{E_e} I + \sigma^2_{G_i} J$ de valeur propre λ ssi v est vecteur propre de J de valeur propre $\frac{\lambda - \sigma^2_{E_e}}{\sigma^2_{G_i}}$. En effet, $\Sigma_{i,e}$ est une matrice symétrique réelle, donc diagonalisable dans une base orthonormale constituée de vecteurs propres. Autrement dit, $\Sigma = \sigma^2_{E_e} I + \sigma^2_{G_i} J = PDP^{-1}$. Dès lors:

$$\sigma^2_{G_i} J = PDP^{-1} - P(\sigma^2_{E_e} I)P^{-1}$$

i.e.

$$J = P[\frac{1}{\sigma^2_{G_i}}(D - \sigma^2_{E_e} I)]P^{-1}$$

.
Par ailleurs, on a $J^2 = n_{Ri,e}J$ i.e. $J(J - n_{Ri,e}I) = 0$, donc les seules valeurs propres possibles de J sont 0 et $n_{Ri,e}$.
On remarque de plus que que le vecteur de taille 1 ne comportant que des 1 est un vecteur propre de J de valeur propre $n_{Ri,e}$, et que l'on peut trouver $n_{Ri,e} - 1$ vecteurs propres linéairement indépendants de valeurs propres 0 car J est une matrice pleine de 1 et de taille $n_{Ri,e} \times n_{Ri,e}$ avec $n_{Ri,e} - 1$ vecteurs colonnes liés.
Par conséquent, on sait que $\Sigma_{i,e}$ a pour valeurs propres $\sigma^2_{E_e} + r\sigma^2_{G_i}$ avec multiplicité 1 et $\sigma^2_{E_e}$ avec multiplicité $n_{Ri,e} - 1$. En particulier, son déterminant vaut

$$|\Sigma| = (\sigma^2_{E_e} + n_{Ri,e}\sigma^2_{G_i})(\sigma^2_{E_e})^{n_{Ri,e}-1}$$

C Calcul de la vraisemblance

Nous repartons de la Propriété 1. Nous avons la formule suivante

$$\mathcal{L}(Y; \mu_e, \mu_{Gi}, \sigma_{Gi}, \sigma_{E_e}) = -\frac{\log(2\pi)}{2} \sum_{i,e} n_{Ri,e} - \frac{1}{2} \sum_{i,e} \log|\Sigma_{i,e}| - \sum_{i,e} \frac{\tilde{Y}^T_{i,e} \cdot \Sigma^{-1}_{i,e} \cdot \tilde{Y}_{i,e}}{2}$$

Nous pouvons utiliser les résultats des sections précédentes pour calculer la formule de la vraisemblance.

Pour commencer, pour i et e fixés on a :

$$\begin{aligned}\tilde{Y}_{i,e}^T \cdot \Sigma_{i,e}^{-1} \cdot \tilde{Y}_{i,e} &= \tilde{Y}_{i,e}^T \cdot \left(\frac{1}{\sigma_{E_e}^2} I - \frac{\sigma_{G_i}^2}{\sigma_{E_e}^2(\sigma_{E_e}^2 + n_{Ri,e}\sigma_{G_i}^2)} J \right) \cdot \tilde{Y}_{i,e} \\ &= \frac{1}{\sigma_{E_e}^2} \sum_{r=1}^{n_{Ri,e}} \tilde{Y}_{i,e,r}^2 - \frac{\sigma_{G_i}^2}{\sigma_{E_e}^2(\sigma_{E_e}^2 + n_{Ri,e}\sigma_{G_i}^2)} \tilde{Y}_{i,e}^T \cdot J \cdot \tilde{Y}_{i,e} \\ &= \frac{1}{\sigma_{E_e}^2} \sum_{r=1}^{n_{Ri,e}} \tilde{Y}_{i,e,r}^2 - \frac{\sigma_{G_i}^2}{\sigma_{E_e}^2(\sigma_{E_e}^2 + n_{Ri,e}\sigma_{G_i}^2)} \left(\sum_{r=1}^{n_{Ri,e}} \tilde{Y}_{i,e,r} \right)^2\end{aligned}$$

Ainsi on obtient

$$\sum_{i,e} \frac{\tilde{Y}_{i,e}^T \cdot \Sigma_{i,e}^{-1} \cdot \tilde{Y}_{i,e}}{2} = \sum_{i,e} \left\{ \frac{1}{2\sigma_{E_e}^2} \sum_{r=1}^{n_{Ri,e}} \tilde{Y}_{i,e,r}^2 - \frac{\sigma_{G_i}^2}{2\sigma_{E_e}^2(\sigma_{E_e}^2 + n_{Ri,e}\sigma_{G_i}^2)} \left(\sum_{r=1}^{n_{Ri,e}} \tilde{Y}_{i,e,r} \right)^2 \right\}$$

D'autre part nous pouvons réutiliser l'expression du déterminant que nous avons obtenue

$$\begin{aligned}\frac{1}{2} \sum_{i,e} \log |\Sigma_{i,e}| &= \frac{1}{2} \sum_{i,e} \log \left[(\sigma_{E_e}^2 + n_{Ri,e}\sigma_{G_i}^2)(\sigma_{E_e}^2)^{n_{Ri,e}-1} \right] \\ &= \frac{1}{2} \sum_{i,e} (n_{Ri,e} - 1) \log(\sigma_{E_e}^2) + \frac{1}{2} \sum_{i,e} \log \left[(\sigma_{E_e}^2 + n_{Ri,e}\sigma_{G_i}^2) \right]\end{aligned}$$

On obtient finalement

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

Que l'on peut aussi écrire

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