Biais/Variance

Cette section est basée sur James et al. (2021), chapitre 2.

Que souhaite-t-on faire ?

Supposons que l'on observe une variable réponse, notée Y, pouvant être quantitative, qualitative ou autre et p variables explicatives, notées $X_1, \dots, X_p$, pouvant aussi être quantitatives, qualitatives ou autres. On suppose aussi qu'il existe une certaine relation entre Y et $X = (X_1, \dots, X_p)$. Cette relation peut s'écrire de façon gérérale comme

$$Y = f(X) + \epsilon. \tag{1}$$

Ici, f est une fonction de $X_1, \dots, X_p$ que l'on souhaite estimer à l'aide des données et ε est un terme d'erreur. Dans le cadre de ce cours, on supposera que la variable aléatoire ε est indépendente de X et que sa moyenne est nulle et sa variance σ^2. Le modèle Équation 1 est un modèle général dans le sens où tout ce que l'on va faire dans le cadre de ce cours peut s'écrire sous cette forme, bien que l'on ne soit pas toujours capable de donner une équation pour f. La fonction f représente l'information systématique que X donne à propos de Y. La Figure 1 présente les différents éléments du modèle de façon visuel.

```
# Load required package
library(ggplot2)

# Set seed for reproducibility
set.seed(42)

# 1. Generate data
n <- 100
x <- sort(runif(n, 0, 2 * pi))
f_x <- sin(x)
epsilon <- rnorm(n, mean = 0, sd = 0.3)
y <- f_x + epsilon
```

```r
# 2. Create data frame
data <- data.frame(x = x, y = y, f_x = f_x)

# 3. Plot
ggplot(data, aes(x, y)) +
  # Vertical lines: error = Y - f(X)
  geom_segment(
    aes(x = x, xend = x, y = f_x, yend = y),
    color = "gray60", linetype = "dashed"
  ) +
  # Observed points
  geom_point(color = "black", alpha = 0.7, size = 2) +
  # True function
  geom_line(aes(x = x, y = f_x), color = "blue", size = 1.2) +
  labs(
    x = "X", y = "Y"
  ) +
  theme_minimal()
```

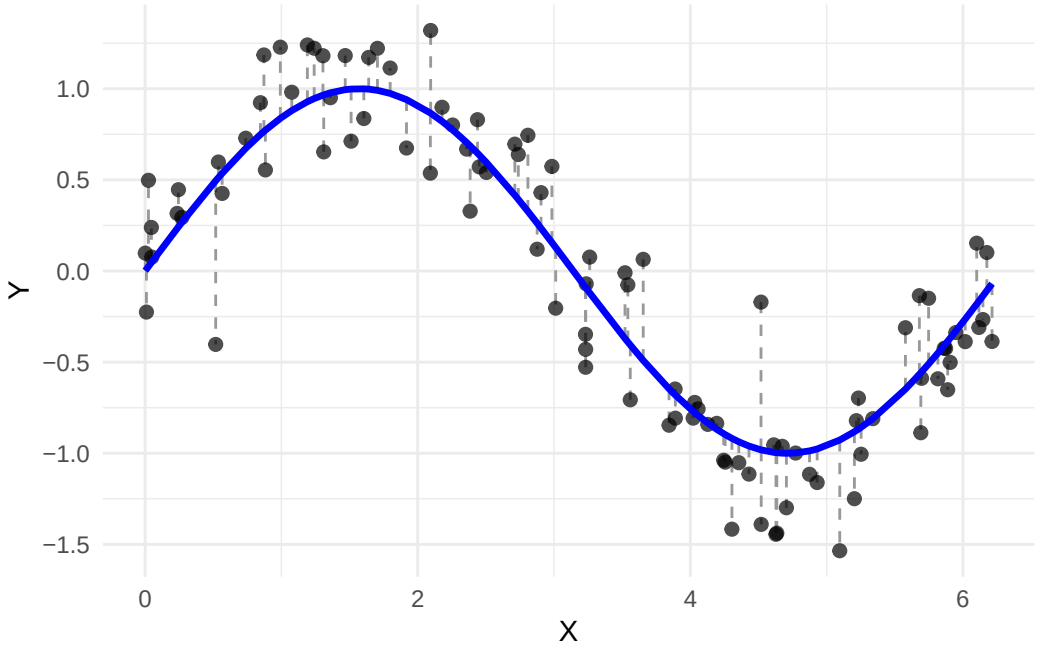

FIGURE 1 – Les différents éléments du modèle. Les points représentent les données observées (X, Y). La courbe bleue représente la fonction f et les lignes pointillées représentent l'erreur associée à chaque observation.

Dans la suite du cours, on verra quelques méthodes permettant d'estimer la fonction f. Cependant avant de voir comment estimer f, on va supposer que l'on a déjà un estimateur, noté $\hat{f}$, estimé à partir de n observations de (X, Y) et on va s'intéresser à la qualité de cet estimateur.

Exemple : la régression linéaire simple

Dans le cadre de la régression linéaire simple, on fait l'hypothèse que la fonction f est de la forme : $f(x) = ax+b$. Dans ce cas, l'estimation de la fonction f se résume à l'estimation des coefficients a et b.

Remarque : Compromis exactitude / interprétabilité

Dépendent de l'objectif de l'étude, on peut devoir faire un choix entre l'exactitude de nos prédictions et l'interprétabilité de notre modèle. En effet, si on restreint notre modèle à être linéaire, nos paramètres seront interprétables et même visualisable, mais ce sera peut-être au détriment du pouvoir prédictif du modèle, e.g. si la "vraie" relation entre X et Y n'est pas linéaire. Inversement, un modèle pouvant estimer des relations plus compliqué (et donc plus flexible) aura plus de paramètres et donc sera plus difficile à interpréter.

Remarque : *No free lunch in statistics*

On pourrait se demander pourquoi cette discussion sur différents estimateurs de f et la mesure de leur qualité d'ajustement. En effet, pourquoi ne pas juste consider un unique "meilleur" modèle ? La réponse est simple : il n'existe pas de méthode qui domine toutes les autres sur tous les jeux de données. Ainsi, une méthode peut très bien fonctionner sur un certain jeu de données et pas du tout sur d'autre jeu de données. Et même en ne considérant qu'un jeu de données, en fonction de l'objectif de l'étude (explication, prédiction, classification, …), le meilleur modèle peut être différent.

Mesurer la qualité de l'ajustement

Pour évaluer la qualité de notre estimateur $\hat{f}$, on a besoin d'une mesure nous indiquant si nos prédictions, $\widehat{Y} = \hat{f}(X)$, sont proches des données observées. Dit autrement, on cherche à quantifier si la réponse prédite pour chaque observation est proche de la vraie réponse pour cette observation.

Définition : Erreur quadratique moyenne

Dans le cas où Y est une variable quantitative, une mesure de la qualité de l'estimateur $\hat{f}$ est l'**erreur quadratique moyenne** (*mean square error*, MSE). Celle-ci est définie comme

$$MSE(Y, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2,$$

où $\hat{y}_i = \hat{f}(x_i)$ est la prédiction que $\hat{f}$ donne pour l'observation i.

L'idée est que la MSE sera petite si les réponses prédites sont proches des vraies réponses et grande si elles ne le sont pas. On peut aussi voir la MSE comme une mesure de la distance moyenne entre les vraies valeurs et les valeurs prédites. On cherche donc à avoir une distance moyenne petite.

L'erreur quadratique moyenne est une bonne mesure dans le cas de variables quantitatives, cependant, elle n'est pas utilisable avec des variables qualitatives. Dans ce cas, une mesure populaire est le taux d'erreur.

Définition : Taux d'erreur

Dans le cas où Y est une variable qualitative, une mesure de la qualité de l'estimateur $\hat{f}$ est le **taux d'erreur** (*error rate*, ER). Celle-ci est définie comme

$$ER(Y, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i \neq \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i \neq \hat{f}(x_i)).$$

où $\hat{y}_i = \hat{f}(x_i)$ est la prédiction que $\hat{f}$ donne pour l'observation i.

Dans la même idée que la MSE, l'ER est la proportion d'erreur qui sont faites par l'estimateur $\hat{f}$. C'est aussi une mesure de la distance moyenne entre les vraies valeurs et les valeurs prédites.

Compromis biais/variance

Lorsque l'on cherche à minimiser l'erreur quadratique moyenne (ou le taux d'erreur), on cherche à le faire sur les données observées, mais aussi sur les éventuelles futures observations de même processus. Il se trouve que, après avoir choisi un estimateur $\hat{f}$, on peut décomposer l'espérance de l'erreur en trois quantités : un terme de biais de $\hat{f}$ au carré, un terme de variance de $\hat{f}$ et un terme correspondant à la variance de l'erreur σ^2.

Compromis biais/variance

Ayant un estimateur $\hat{f}$ de f, l'espérance de l'erreur de l'estimateur peut se décomposer de la façon suivante :

$$\mathbb{E}\left[(Y-\widehat{Y})^2\right] = \mathbb{E}\left[(Y-\hat{f}(X))^2\right] = \mathrm{Var}(\hat{f}(X)) + \mathrm{Biais}(\hat{f}(X))^2 + \sigma^2.$$

Preuve

Tout d'abord, montrons que l'espérance de l'erreur de l'estimateur se décompose en une partie réductible et en une partie irréductible.

$$\begin{aligned}
\mathbb{E}\left[\left(Y-\widehat{Y}\right)^2\right] &= \mathbb{E}\left[\left(Y-\hat{f}(X)\right)^2\right] \\
&= \mathbb{E}\left[\left(f(X)+\varepsilon-\hat{f}(X)\right)^2\right] \\
&= \mathbb{E}\left[\left(f(X)-\hat{f}(X)\right)^2\right] + 2\mathbb{E}\left[\left(f(X)-\hat{f}(X)\right)\varepsilon\right] + \mathbb{E}[\varepsilon^2] \\
&= \mathbb{E}\left[\left(f(X)-\hat{f}(X)\right)^2\right] + 2\mathbb{E}\left[\left(f(X)-\hat{f}(X)\right)\right]\underbrace{\mathbb{E}[\varepsilon]}_{=0} + \sigma^2 \\
&= \underbrace{\mathbb{E}\left[\left(f(X)-\hat{f}(X)\right)^2\right]}_{\text{réductible}} + \underbrace{\sigma^2}_{\text{irréductible}}.
\end{aligned}$$

On utilise la linéarité de l'espérance et le fait que X et ε soient indépendants. On s'intéresse maintenant à la partie "réductible". L'astuce est de faire apparaître $\mathbb{E}\left[\hat{f}(X)\right]$.

$$\begin{aligned}
\mathbb{E}\left[\left(f(X)-\hat{f}(X)\right)^2\right] &= \mathbb{E}\left[\left(f(X)-\mathbb{E}\left[\hat{f}(X)\right]+\mathbb{E}\left[\hat{f}(X)\right]-\hat{f}(X)\right)^2\right] \\
&= \underbrace{\mathbb{E}\left[\left(f(X)-\mathbb{E}\left[\hat{f}(X)\right]\right)^2\right]}_{\text{A}} \\
&\quad - 2\underbrace{\mathbb{E}\left[\left(f(X)-\mathbb{E}\left[\hat{f}(X)\right]\right)\left(\hat{f}(X)-\mathbb{E}\left[\hat{f}(X)\right]\right)\right]}_{\text{B}} \\
&\quad + \underbrace{\mathbb{E}\left[\left(\hat{f}(X)-\mathbb{E}\left[\hat{f}(X)\right]\right)^2\right]}_{\text{C}}.
\end{aligned}$$

A. La fonction $f(X)$ n'étant pas aléatoire, on a $\mathbb{E}[f(X)] = f(X)$ et donc

$$\begin{aligned}\mathbb{E}\left[\left(f(X)-\mathbb{E}\left[\hat{f}(X)\right]\right)^2\right] &= \mathbb{E}\left[\left(\mathbb{E}\left[f(X)-\hat{f}(X)\right]\right)^2\right] \\ &= \mathbb{E}\left[f(X)-\hat{f}(X)\right]^2 \\ &= \text{Biais}(\hat{f}(X))^2.\end{aligned}$$

B. En développant l'expression et en utilisant l'indépendance des variables, on trouve que $B = 0$.
C. En utilisant la définition de la variance,

$$\mathbb{E}\left[\left(\hat{f}(X)-\mathbb{E}\left[\hat{f}(X)\right]\right)^2\right] = \text{Var}(\hat{f}).$$

Finalement, on a

$$\mathbb{E}\left[\left(f(X)-\hat{f}(X)\right)^2\right] = \text{Biais}(\hat{f}(X))^2 + \text{Var}(\hat{f}(X)).$$

D'où le résultat.

Remarque

Comme les trois quantités dans lesquelles on a décomposer l'espérance de l'erreur de l'estimateur sont non-négative, celle-ci est minoré par l'erreur irréductible. Elle ne peut donc jamais être inféireure à σ^2.

Qu'implique le compromis biais/variance ? Comme on cherche à minimiser l'espérance de l'erreur de l'estimateur, on voudrait une méthode qui nous donne à la fois un faible biais et une faible variance. Le biais est l'erreur d'approximation faite lorsque l'on approche un problème compliqué par un modèle simple. Par exemple, estimer un modèle linéaire lorsque que la relation ne l'est pas. La variance est la quantité par laquelle l'estimateur va changer si on apprend l'estimateur sur un jeu de données différents. Dit d'une autre façon, comme on estime f par $\hat{f}$ à l'aide d'un jeu de données particulier et l'estimer avec un autre jeu de données implique un estimateur $\hat{f}$ différents. Idéalement, l'estimateur $\hat{f}$ ne devrait pas trop changé en changeant le jeu de données d'entraînement. La Figure 2 présente un jeu de données et différents estimateurs $\hat{f}$. En faisant varier le paramètre λ, on obtient des modèles plus ou moins flexible. La Figure 3 montre la valeur du biais, de la variance et de la MSE pour les modèles estimés pour la Figure 2. On remarque que plus λ est petit, plus la variance est grande, mais le biais est petit (le modèle est flexible). Inversement, plus λ est grand, plus le biais est grand est la variance petite (le modèle est rigide). La courbe de MSE en fonction du paramètre est une courbe en U. Comme on cherche à minimiser la MSE, i.e. à faire un compromis entre le biais et la variance, on peut prendre $\lambda = 0.5$.

```r
# Load packages
library(ggplot2)
library(dplyr)
library(tidyr)

set.seed(42)

# 1. Simulate a single dataset
n <- 100
sigma <- 0.3
x <- sort(runif(n, 0, 1))
y <- 4 * x * (1 - x) * log(x) + 2 + rnorm(n, 0, sigma)
df <- data.frame(x = x, y = y)

# 2. Define grid and spans to compare
x_grid <- seq(0, 1, length.out = 300)
spans_to_plot <- c(0.15, 0.3, 0.5, 0.75, 1.0)

# 3. Compute loess fits for each span
fits <- lapply(spans_to_plot, function(s) {
  loess_model <- loess(y ~ x, data = df, span = s)
  y_hat <- predict(loess_model, newdata = data.frame(x = x_grid))
  data.frame(x = x_grid, y_hat = y_hat, span = paste0("  = ", s))
})

fit_df <- bind_rows(fits)

# 4. Plot
ggplot() +
  geom_point(data = df, aes(x, y), color = "black", alpha = 0.5, size = 2) +
  geom_line(data = fit_df, aes(x, y_hat, color = span), size = 1.1) +
  scale_color_viridis_d(option = "C") +
  labs(
    x = "Y", y = "Y",
    color = "Paramètre"
  ) +
  theme_minimal(base_size = 14)
```

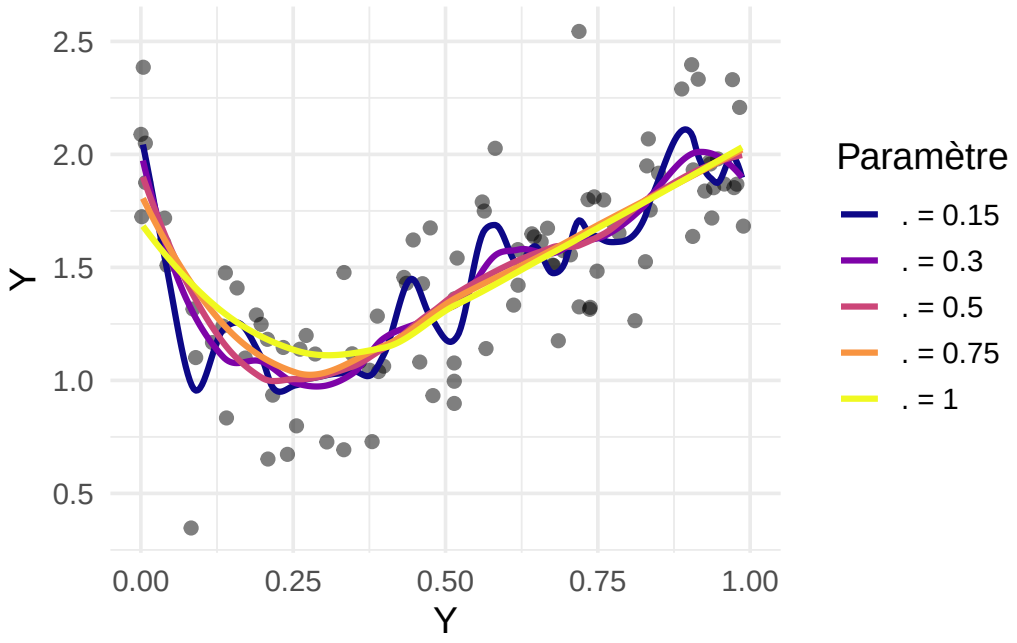

FIGURE 2 – Différents estimateurs de la fonction f.

```r
# Load packages
library(ggplot2)
library(dplyr)
library(tidyr)

set.seed(42)

# Parameters
n <- 100              # number of observations per dataset
n_sim <- 100          # number of simulated datasets
spans <- seq(0.1, 1, length.out = 15)  # LOESS smoothing parameters
sigma <- 0.1          # noise standard deviation
x_grid <- seq(0.1, 1, length.out = 200)
f_true <- 4 * x_grid * (1 - x_grid) * log(x_grid) + 2

# Storage for predictions
results <- list()

for (s in spans) {
  pred_matrix <- matrix(NA, nrow = length(x_grid), ncol = n_sim)
```

```r
  for (sim in 1:n_sim) {
    x <- sort(runif(n, 0.01, 1.1))
    y <- 4 * x * (1 - x) * log(x) + 2 + rnorm(n, 0, sigma)
    df <- data.frame(x = x, y = y)

    # Fit loess model with span = s
    model <- loess(y ~ x, data = df, span = s, degree = 2)
    pred <- predict(model, newdata = data.frame(x = x_grid), )

    pred_matrix[, sim] <- pred
  }

  # For each point in x_grid, compute bias², variance, MSE
  mean_pred <- rowMeans(pred_matrix, na.rm = TRUE)
  bias2 <- (mean_pred - f_true)^2
  var_pred <- apply(pred_matrix, 1, var, na.rm = TRUE)
  mse <- bias2 + var_pred

  results[[as.character(s)]] <- data.frame(
    span = s,
    Biais2 = mean(bias2),
    Variance = mean(var_pred),
    MSE = mean(mse)
  )
}

# Combine and reshape results
results_df <- bind_rows(results)
results_long <- pivot_longer(
  results_df,
  cols = c("Biais2", "Variance", "MSE"),
  names_to = "component", values_to = "value"
)

# Plot
ggplot(results_long, aes(x = span, y = value, color = component)) +
  geom_line(size = 1.2) +
  geom_point(size = 2) +
  scale_color_manual(
    values = c("Biais2" = "#0D0887FF", "Variance" = "#9C179EFF", "MSE" = "#ED7953FF")
  ) +
  labs(
```

```
    x = "Paramètre de lissage",
    y = "",
    color = ""
  ) +
  theme_minimal(base_size = 14)
```

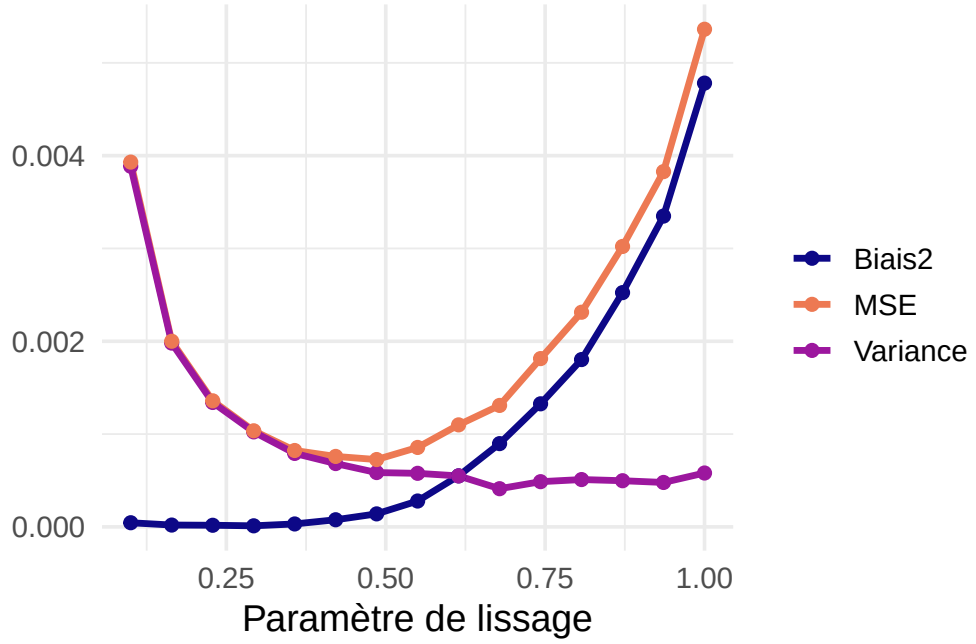

FIGURE 3 – Compromis biais/variance

De manière générale, plus le modèle est flexible, plus la variance va être grande et plus le biais va être faible et inversement. Le changement relatif entre le biais et la variance indique si l'espérance de l'erreur de l'estimateur augmente ou diminue. Lorsque l'on augmente la flexibilité de notre modèle, le biais va avoir tendence à diminuer plus vite que la variance augmente et donc la MSE diminue. À un moment, augmenter la flexibilité n'a plus d'impact sur le biais mais commence à impacter la variance de façon significative. Et par conséquent, le MSE augmente.

> Remarque : Pourquoi un compromis ?
>
> On l'appelle le compromis biais/variance parcequ'il est très facile d'avoir un modèle non biaisé (s'il passe par tous les points) ou avec une variance très faible (une ligne horizontale). Le compromis est donc de trouver un modèle avec la fois un modèle avec un faible biais et une faible variance.

James, Gareth, Daniela Witten, Trevor Hastie, et Robert Tibshirani. 2021. *An Introduction to Statistical Learning : With Applications in R*. Springer Texts in Statistics. New York, NY : Springer US. https://doi.org/10.1007/978-1-0716-1418-1.