# 16 Biais et variance d'un estimateur

## 1 Biais et variance pour l'estimateur d'un paramètre d'un modèle paramétrique

Nous notons $\hat{\mu}$ la moyenne ($\hat{\mu} = E[\mathbf{x}]$) et $\hat{\sigma}^2$ la variance ($\hat{\sigma}^2 = E[(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])]$) d'un modèle paramétrique $P(\mathbf{x}; \mu, \sigma)$ dont on fait l'hypothèse qu'il aurait généré des données observées $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$. Étant données les observations, nous nous intéressons à des estimateurs $\hat{\mu}_{est}$ et $\hat{\sigma}_{est}$ des paramètres $\hat{\mu}$ et $\hat{\sigma}$.

Un estimateur est non biaisé si sa moyenne sur l'ensemble de tous les jeux de données possibles est égale à la valeur du paramètre : $E[\hat{\mu}_{est}] = \hat{\mu}$, $E[\hat{\sigma}^2_{est}] = \hat{\sigma}^2$.

En plus d'un faible biais, un bon estimateur possède une faible variance : $Var(\hat{\mu}_{est}) = E[(\hat{\mu} - \hat{\mu}_{est})^2]$, $Var(\hat{\sigma}_{est}) = E[(\hat{\sigma} - \hat{\sigma}_{est})^2]$.

## 2 Estimateurs par maximum de vraissemblance

Etant donné un modèle paramétrique et un jeu de données supposé avoir été généré par ce modèle, l'estimateur par *maximum de vraissemblance* ("maximum likelihood") associe à un paramètre la valeur qui rend le jeu de données observé le plus probable.

$$\hat{\mu}_{ML} = \underset{\mu}{\operatorname{argmax}}\, P(\mathbf{x}; \mu, \sigma^2)$$
$$\Rightarrow \{\text{A l'endroit d'un extremum, la dérivée première s'annule}\}$$
$$\partial P(\mathbf{x}; \mu, \sigma^2)/\partial \mu = 0$$

$$\hat{\sigma}^2_{ML} = \underset{\sigma^2}{\operatorname{argmax}}\, P(\mathbf{x}; \mu, \sigma^2)$$
$$\Rightarrow \{\text{A l'endroit d'un extremum, la dérivée première s'annule}\}$$
$$\partial P(\mathbf{x}; \mu, \sigma^2)/\partial \sigma^2 = 0$$

### 2.1 Exemple d'une loi normale

Supposons que des échantillons scalaires $X = x_1, x_2, \ldots, x_n$ aient été générés selon une loi normale.

$$x_i \sim \mathcal{N}(\mu, \sigma^2)$$
$$P(x_i; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right)$$

$$
\begin{aligned}
& P(X;\mu,\sigma) \\
= & \{\text{Hypothèse : les échantillons sont indépendants}\} \\
& \prod_i P(x_i;\mu,\sigma) \\
= & \\
& (2\pi\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right]
\end{aligned}
$$

Supposons que la moyenne $\hat{\mu}$ du modèle soit connue. Calculons alors l'estimateur par maximum de vraissemblance de la variance.

$$
\begin{aligned}
& \hat{\sigma}^2_{ML} \\
= & \{\text{Par définition d'un estimateur par maximum de vraissemblance.}\} \\
& \underset{\sigma^2}{\text{argmax}}\, P(X;\hat{\mu},\sigma^2) \\
= & \{\text{Le logarithme est une fonction monotone qui ne change pas le lieu du maximum.}\} \\
& \underset{\sigma^2}{\text{argmax}}\, log\left[P(X;\hat{\mu},\sigma^2)\right] \\
\Rightarrow & \{\text{A l'endroit du max, la dérivée s'annule.}\} \\
& \partial log\left[P(X;\hat{\mu},\sigma^2)\right]/\partial\sigma^2 = 0 \\
= & \{\text{Notation : } s \triangleq \sigma^2\} \\
& \partial log\left[P(X;\hat{\mu},s)\right]/\partial s = 0 \\
= & \{\text{Définition de } P\} \\
& -(n/2)\partial log(s)/\partial s - \partial\left[(2s)^{-1}\sum_i (x_i-\hat{\mu})^2\right]\partial s = 0 \\
= & \{\text{Calcul des dérivées.}\} \\
& -(n/2)s^{-1} + 1/2s^{-2}\sum_i (x_i-\hat{\mu})^2 = 0 \\
= & \{\text{Factorisation pour faire apparaître } s\} \\
& (n/2)s^{-2}\left[\left(1/n\sum_i (x_i-\hat{\mu})^2\right) - s\right] = 0 \\
\Rightarrow & \{\text{Pour une variance finie, le second facteur doit s'annuler.}\} \\
& \hat{s}_{ML} \triangleq \hat{\sigma}^2_{ML} = 1/n\sum_i (x_i-\hat{\mu})^2
\end{aligned}
$$

Quel est le biais de cet estimateur ?

$$
\begin{aligned}
& E\left[\hat{\sigma^2_{ML}}\right] \\
= & \{\text{Voir dérivation ci-dessus.}\} \\
& E\left[n^{-1}\sum_i (x_i - \hat{\mu})^2\right] \\
= & \{\text{Linéarité de l'opérateur espérance E.}\} \\
& n^{-1}\sum_i E\left[x_i^2 - 2x_i\hat{\mu} + \hat{\mu}^2\right] \\
= & \{\text{Linéarité de l'opérateur espérance E.} \\
& \text{Chaque } x_i \text{est supposé avoir été généré par la même loi normale.}\} \\
& n^{-1}\left(nE[x_i^2] - 2n\hat{\mu}E[x_i] + n\hat{\mu}^2\right) \\
= & \{\text{Arithmétique.}\} \\
& E[x_i^2] - 2\hat{\mu}^2 + \hat{\mu}^2 \\
= & \{\hat{\sigma}^2 = E[x_i^2] - E[x_i]^2 = E[x_i^2] - \hat{\mu}^2\} \\
& \hat{\sigma}^2
\end{aligned}
$$

Cet estimateur est sans biais. On peut montrer qu'un estimateur non biaisé obtenu par l'approche du maximum de vraissemblance est également de variance minimale.

Considérons maintenant que la moyenne ne soit pas connue et calculons les estimateurs par maximum de vraissemblance de la maoyenne et de la variance.

$$
\begin{aligned}
& \hat{\mu}_{ML} \\
= & \{\text{Par définition d'un estimateur par maximum de vraissemblance.}\} \\
& \underset{\mu}{\text{argmax}}\, P(X; \mu, s) \\
\Rightarrow & \{\text{Le logarithme est une fonction monotone qui ne change pas le lieu du maximum.} \\
& \text{A l'endroit du max, la dérivée s'annule.}\} \\
& \partial/\partial\mu \; log\left[(2\pi s)^{-N/2} exp\left(-(2s)^{-1}\sum_i (x_i - \mu)^2\right)\right] = 0 \\
= & \{\text{Annulation des facteurs qui ne dépendent pas de } \mu.\} \\
& \partial/\partial\mu\left(-(2s)^{-1}\sum_i (x_i - \mu)^2\right) = 0 \\
= & \{\text{Calcul des dérivées.}\} \\
& -(2s)^{-1}\left(-2\sum_i (x_i - \mu)\right) = 0 \\
= & \{\text{Arithmétique}\} \\
& s^{-1}\left(\sum_i x_i \; - n\mu\right) = 0 \\
\Rightarrow & \{\text{Arithmétique}\} \\
& \hat{\mu}_{ML} = \frac{1}{n}\sum_i x_i
\end{aligned}
$$

Quel est le biais de cet estimateur ?

$$E[\hat{\mu}_{ML}] = E[1/n \sum_i x_i] = 1/n \sum_i E[x_i] = 1/n \times n \times E[x_i] = \hat{\mu}$$

Il s'agit don d'un estimateur non biaisé.

$$\begin{aligned}
& \hat{s}_{ML} \\
= & \{\text{Voir la précédente dérivation quand la moyenne était supposée connue.}\} \\
& n^{-1} \sum_i (x_i - \mu)^2 \\
= & \{\text{Utilisation de l'estimateur par maximum de vraissemblance de la moyenne.}\} \\
& n^{-1} \sum_i \left( x_i - n^{-1} \sum_i x_i \right)^2 \\
= & \{\text{Développement.}\} \\
& n^{-1} \sum_i x_i^2 \ - 2n^{-2} \sum_i \left( x_i \sum_i x_i \right) \ + \left( n^{-1} \sum_i x_i \right)^2 \\
= & \{\text{Arithmétique}\} \\
& n^{-1} \sum_i x_i^2 \ - 2 \left( n^{-1} \sum_i x_i \right)^2 \ + \left( n^{-1} \sum_i x_i \right)^2 \\
= & \{\text{Arithmétique}\} \\
& n^{-1} \sum_i x_i^2 \ - \left( n^{-1} \sum_i x_i \right)^2
\end{aligned}$$

Quel est le biais de cet estimateur de la variance ?

$$
\begin{aligned}
& E[\hat{s}_{ML}] \\
&= \{\text{Voir dérivation ci-dessus.}\} \\
& E\left[n^{-1}\sum_i x_i^2 \; - \left(n^{-1}\sum_i x_i\right)^2\right] \\
&= \{\text{Linéarité de l'opérateur E.}\} \\
& n^{-1}\sum_i E[x_i^2] \; - n^{-2}E\left[(\sum_i x_i)^2\right] \\
&= \{\hat{s} = E[x_i^2] - E[x_i]^2\} \\
& \hat{s} + \hat{\mu}^2 - n^{-2}E\left[\sum_i x_i^2 \; + \sum_i \sum_{j\neq i} x_i x_j\right] \\
&= \{\text{Linéarité de l'opérateur E.}\} \\
& \hat{s} + \hat{\mu}^2 - n^{-2}\left(n(\hat{s} + \hat{\mu}^2) \; + \sum_i \sum_{j\neq i} E[x_i]E[x_j]\right) \\
&= \{\text{Définition de } \mu.\} \\
& \hat{s} + \hat{\mu}^2 - n^{-2}\left(n(\hat{s} + \hat{\mu}^2) \; + n(n-1)\mu^2\right) \\
&= \{\text{Arithmétique}\} \\
& \hat{s} + \hat{\mu}^2 - n^{-1}\left(\hat{s} + \hat{\mu}^2 \; + n\hat{\mu}^2 - \hat{\mu}^2\right) \\
&= \{\text{Arithmétique}\} \\
& \hat{s} - n^{-1}\hat{s} \\
&= \{\text{Arithmétique}\} \\
& \frac{n-1}{n}\hat{s}
\end{aligned}
$$

L'estimateur de la variance $\hat{s}_{ML}$ est maintenant biaisé. Nous pouvons construire un estimateur sans biais :

$$s' = \left(\frac{n-1}{n}\right)^{-1} \hat{s}_{ML} = \frac{1}{n-1}\sum_i (x_i - \mu)^2$$

$s'$ est alors sans biais mais n'est plus de variance minimale (bien que, parmi les estimateurs non biaisés, il soit de variance minimale).

# 3 Analyse biais-variance pour la régression

Nous supposons que les données observées aient été générées par une fonction de la forme $y = f(\mathbf{x}) + \epsilon$ avec $\epsilon$ un bruit gaussien de moyenne nulle et de variance $\sigma^2$.

A partir d'un jeu de données $\{(\mathbf{x_i}, y_i)\}$, nous apprenons un modèle prédictif, par exemple un modèle linéaire $h(\mathbf{x}) = \boldsymbol{\beta}^T\mathbf{x} + \beta_0$, afin de minimiser l'erreur quadratique $\sum_i (y_i - h(\mathbf{x_i}))^2$.

Pour un nouveau point $\mathbf{x}^*$ qu'elle est l'espérance de l'erreur commise sur la prédiction de $y^* = f(\mathbf{x}^*) + \epsilon$, soit $E\left[(y^* - h(\mathbf{x}^*))^2\right]$. Il s'agit de la moyenne de l'erreur sur l'ensemble infini de tous les jeux de données d'entraînement possibles.

Notons $\overline{x} = E[x]$, la valeur moyenne de x. Rappelons un résultat utile :

$$
\begin{aligned}
& E[(x-\overline{x})^2] \\
= & \{\text{Arithmétique}\} \\
& E[x^2 - 2x\overline{x} + \overline{x}^2] \\
= & \{\text{Linéarité de } E.\} \\
& E[x^2] - 2\overline{x}E[x] + \overline{x}^2 \\
= & \{\text{Par définition de } \overline{x}.\} \\
& E[x^2] - 2\overline{x}^2 + \overline{x}^2 \\
= & \{\text{Arithmétique}\} \\
& E[x^2] - \overline{x}^2
\end{aligned}
$$

Nous décomposons l'espérance de l'erreur de prédiction ("Expected Prediction Error" ou EPE) d'un modèle de régression en biais, variance et bruit.

$$
\begin{aligned}
& E\left[(h(\mathbf{x}^*) - y^*)^2\right] \\
= & \{\text{Linéarité de E}\} \\
& E[h(\mathbf{x}^*)^2] - 2E[h(\mathbf{x}^*)]E[y^*] + E[y^{*2}] \\
= & \{E[z^2] = E[(z-\overline{z})^2] + \overline{z}^2 \\
& y = f(\mathbf{x}) + \epsilon \text{ avec } \epsilon \text{ un bruit gaussien de moyenne nulle. Donc } \overline{y}^* = f(\mathbf{x}^*)\} \\
& E\left[\left(h(\mathbf{x}^*) - \overline{h(\mathbf{x}^*)}\right)^2\right] + \overline{h(\mathbf{x}^*)}^2 - 2\overline{h(\mathbf{x}^*)}f(\mathbf{x}^*) + E\left[(y^* - f(\mathbf{x}^*))^2\right] + f(\mathbf{x}^*)^2 \\
= & \{E\left[(y^* - f(\mathbf{x}^*))^2\right] = E[\epsilon^2] = \sigma^2\} \\
& E\left[\left(h(\mathbf{x}^*) - \overline{h(\mathbf{x}^*)}\right)^2\right] + \left(\overline{h(\mathbf{x}^*)} - f(\mathbf{x}^*)\right)^2 + \sigma^2 \\
= & \{\text{Introduction des définition de la variance, du biais et du bruit.}\} \\
& \text{Variance } + \text{ Biais}^2 \text{ } + \text{ Bruit}^2
\end{aligned}
$$

- La variance mesure la variation de la prédiction $h(\mathbf{x}^*)$ d'un jeu de données d'entraînement à l'autre.
- Le biais mesure l'erreur moyenne de $h(\mathbf{x}^*)$.
- Le bruit mesure la variation de $y^*$ par rapport à $f(\mathbf{x}^*)$.