Master Science de la matière
École Normale Supérieure de Lyon
Université Claude Bernard Lyon I

Rapport de stage
Samuel Niang
M2 Physique - Concepts et applications

Calibration des calorimètres de CMS pour la reconstruction de flux de particules.

Résumé :
Pour reconstruire la trajectoire d'une particule, il est nécessaire de connaître son énergie. Cette énergie est estimée à partir d'un détecteur hadronique et d'un détecteur électromagnétiques, problématique de ce stage est : connaissant les dépôts d'énergie d'une particules dans ces deux calorimètres, quelle est son énergie ?
Pour répondre à ces questions, j'ai développé des algorithmes qui prennent en données d'entrainement des particules simulées et qui pour chaque couple d'énergie déposée (e_{cal}, h_{cal}) nous donne une énergie calibrée e_{calib}.
On peut alors ré-expliquer la problématique des algorithmes de la façon suivante : comment modéliser un nuage de points $(e_{cal}, h_{cal}, e_{true})$ par une surface $e_{calib} = f(h_{cal}, e_{true})$?

Mots clefs : *Calibration, Modélisation, Physique des particules*

Stage encadré par :
Colin Bernet `colin.bernet@cern.ch`
Bâtiment Paul Dirac
4, Rue Enrico Fermi
69622 Villeurbanne Cedex
Tél. : +33 (0) 4 72 44 84 57

18 juillet 2017

Table des matières

1 Introduction

Le but de ce stage est de trouver une méthode de calibration des calorimètres hadroniques et électromagnétiques de CMS, c'est à dire, pour une particules qui va laisser un dépôt d'énergie h_{cal}, e_{cal} dans chacun des calorimètres, comment approximer sa vraie énergie e_{true} ? Cette énergie dite énergie calibrée e_{calib} sera déterminée à l'aide de particules issues d'une simulation très précise (prenant en compte les défauts des calorimètres) qui serviront de données d'entrainement aux différents algorithme que j'ai développés durant mon stage.
Le but final de cette calibration sera d'améliorer la reconstruction des flux de particules (particules flow).

2 Méthodes de calibrations développées pendant le stage

2.1 Explications valables pour toutes les méthodes

2.1.1 Séparation des données

expliquer :
- séparation ecal = 0, ecal != 0
- limite ecal + hcal < 150

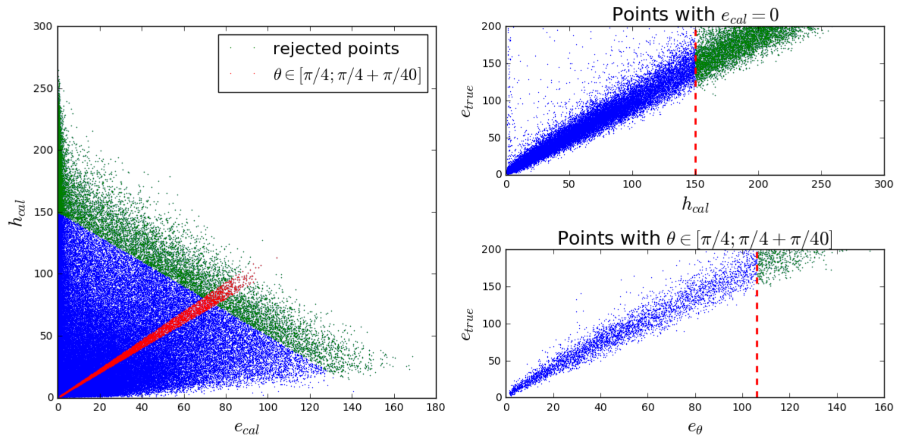


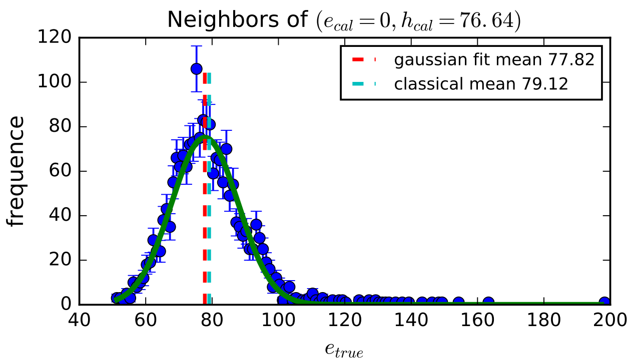
Figure 1 – On place une limite à $e_{cal} + h_{cal} = 150$. Á gauche, ..., en haut à droite, ..., en bas à doite, ...

2.1.2 Moyenne / moyenne de la gaussienne ajustée (”gaussian fit”, ”gaussienne fitée”) ?

A différents moment, nous aurons besoin de calculer des moyennes. Or souvent, la moyenne classique ne serait pas représentative de ce que tous souhaitons montrer car certains points sont ont des valeurs e_{cal}, h_{cal} mal estimées car la simulation prend en compte les défauts des calorimètres. Il serait donc alors incorrecte de les prendre en compte pour juger l'efficacité d'une calibration car ils sont complètement incohérents.

Pour résoudre ce problème, nous allons ajuster une gaussienne de la distribution des points à moyenner et choisir considérer que la moyenne à prendre en compte est la moyenne de la gaussienne. Ainsi, les points aberrants totalement écarté du centre de la distribution ne perturberont pas le calcul de la moyenne alors que dans le cas d'une moyenne classique, ils peuvent fortement attirer la moyenne vers eux.

Ces points aberrants peuvent également venir d'une particule qui se serait décomposée avant le calorimètre. Ainsi on trouve près de l'origine, des points à fort e_{true} et pour de faibles valeurs de e_{cal} et h_{cal}, et ces points ne sont pas du tout représentatif de l'efficacité d'une calibration.



Ici, on peut voir sur cet exemple que si nous prenons la moyenne classique de e_{true}, on obtient 79.12, or la moyenne de la gaussienne fitée est de 77.82, au vu de ce que nous avons dit précédemment, nous considèrerons que la seconde est plus judicieuse.

2.1.3 Comment est fait un fit ?

expliquer :
- barre d'erreur
- minimisation du chi2
- un bon chi2 réduit ?

2.2 Régression Linéaire

Pour s'entrainer à l'utilisation de *SciKit Learn*, j'ai d'abord utilisé la régression linéaire. Il s'agit alors de représenter les relations entre les énergies par :

$$e_{true} = a_1 e_{cal} + a_2 h_{cal} + b \tag{1}$$

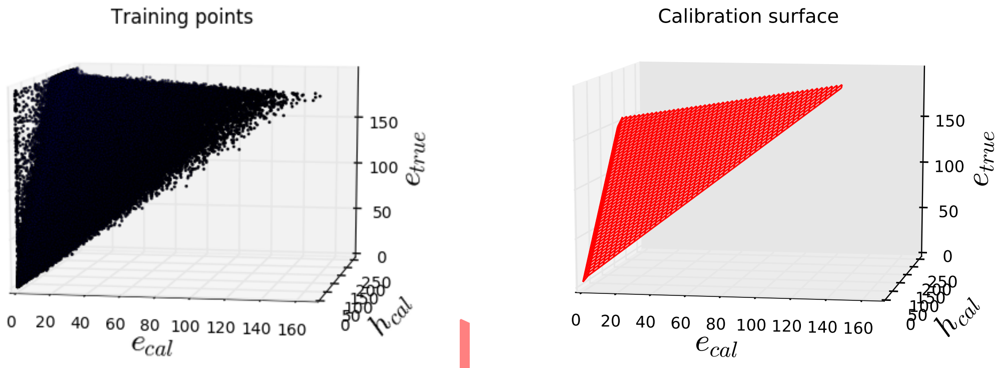


Figure 2 – Le nuage de points modélisé (à gauche) par un plan (à droite).

Nous avons ainsi modélisé le nuage de point par un plan, pour voir si cela était réaliste, nous allons d'abord regarder ce qui se passe dans le plan $e_{cal} = 0$:

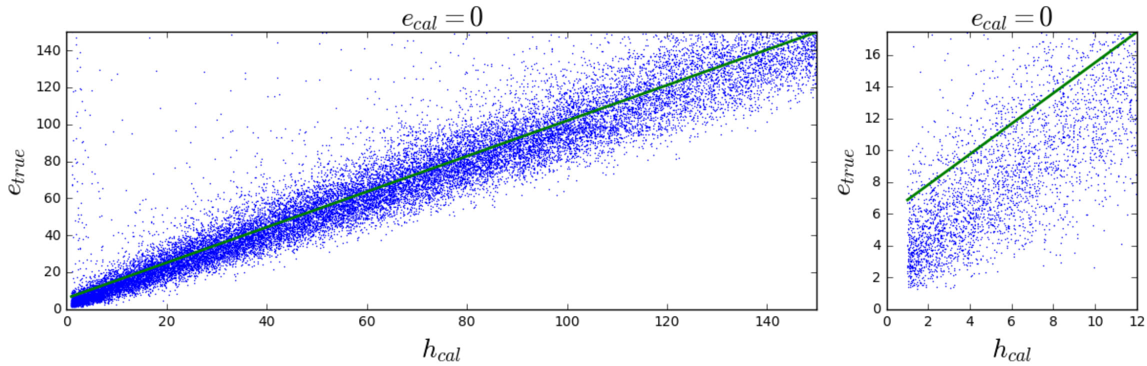


Figure 3 – Courbe de calibration pour $e_{cal} = 0$.

Nous constatons alors que la courbe ne passe pas par le le coeur du nuage de point à faible énergie. Pour avoir une vue d'ensemble, nous allons tracer e_{calib}/e_{true} qui doit être proche de 1 si la calibration est bonne.

En regardant la figure de droite, nous constatons que comme prévu la régression linéaire est mauvaise à faible énergie car en moyenne, e_{calib}/e_{true} n'est pas proche de 1. Plus intéressant, la figure de droite met en avant les non-linéarités du nuage de point.

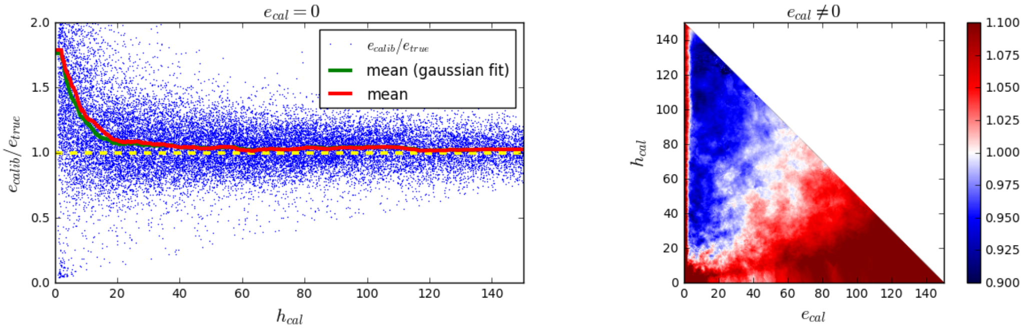


Figure 4 – e_{calib}/e_{true} en fonction de e_{cal} et h_{cal}.

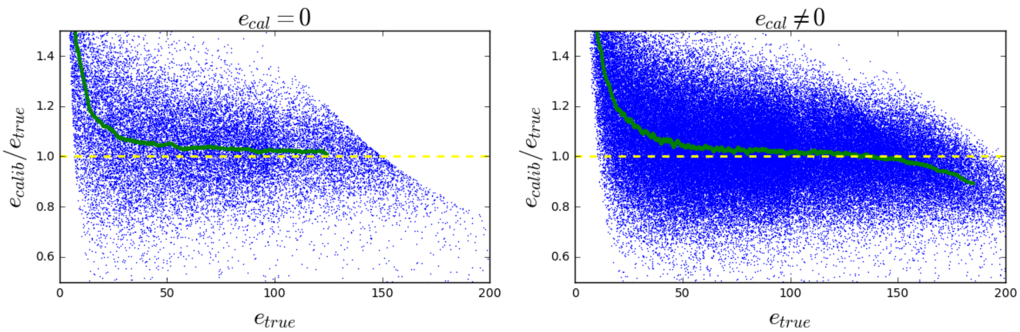


Figure 5 – e_{calib}/e_{true} en fonction de e_{true}.

Ici nous constatons que à faible et haut e_{true}, la calibration ne donne pas de bons résultats. En effet, la courbe de la moyenne (fit gaussien) s'écarte très fortement d'une constante égale à 1.

2.3 Méthode des "legos"

2.3.1 Principe général de l'algorithme

Comme nous l'avons vu précédemment, il faut une calibration qui prenne en compte les non-linéarité. Ici, l'idée est de découper le plan (e_{cal}, h_{cal}) en carré et de calculer la moyenne des e_{true} dans chaque carré qui sera la valeur e_{calib}.
Ainsi pour prédire une énergie de e^i_{calib} pour un point (e^i_{cal}, h^i_{cal}), nous allons regarder dans quel carré il se trouve et retourner la valeur d'énergie calibrée correspondante, faisant apparaître ainsi des "legos".

2.3.2 Résultat de la calibration

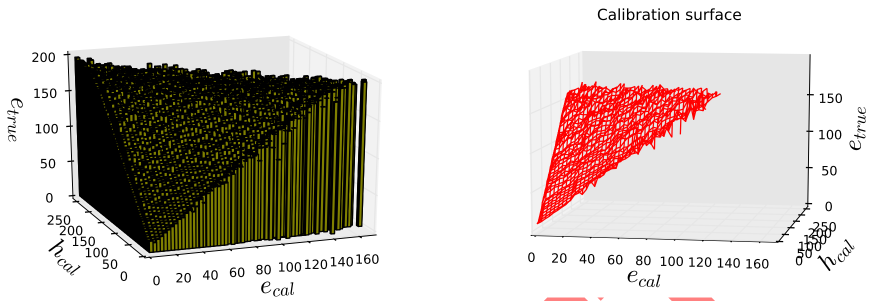


Figure 6 – Le nuage de points modélisé par des legos (à gauche) ainsi que la surface correspondante (à droite). 100 × 100 legos.

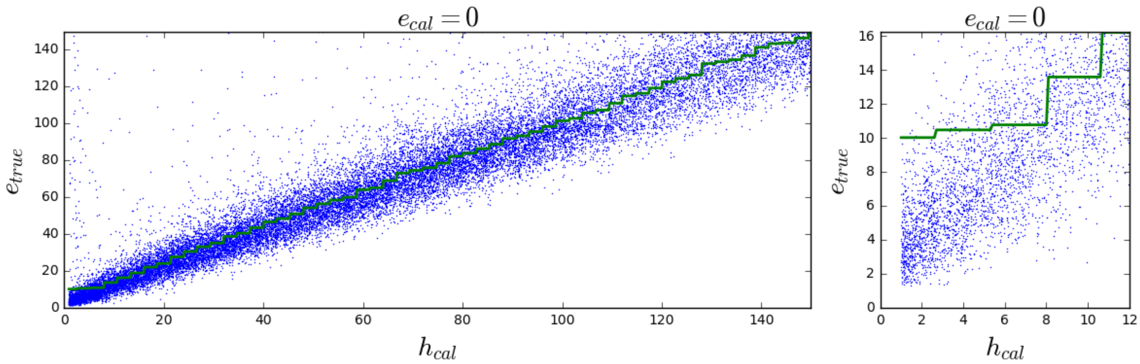


Figure 7 – Courbe de calibration pour $e_{cal} = 0$.

Bien que cette méthode prenne en compte les linéarité, nous pouvons voir sur les figures ci-dessus qu'il y a un effet de pas, ce qui n'est pas bon car beaucoup trop d'événements on le même e_{calib} et la courbe de calibration ne suit pas bien le coeur de distribution, surtout à faible énergie.

Cet effet de pas se retrouve également si l'on trace e_{calib}/e_{true} en fonction de e_{true} (Fig. 8) et nous y voyons alors un structure (des hyperboles) liées aux points qui ont la même énergie de calibration (contrairement à la régression linéaire Fig. 5).
Cet illustration montre à nouveau que nous sommes loin d'une répartition des points autour de $e_{calib}/e_{true} = 1$.

Ici nous constatons malgré tout que nous avons mieux pris en compte la non-linéarité, mais que en majorité, l'énergie calibrée est sur-estimée.

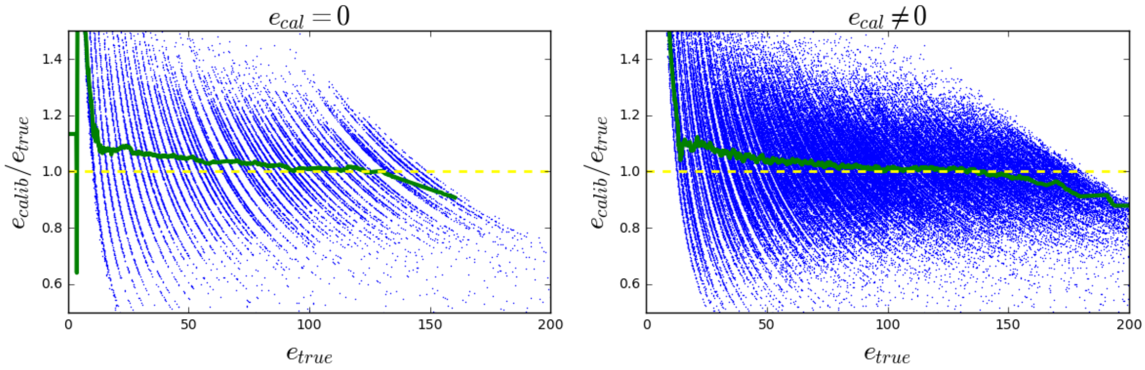


Figure 8 – e_{calib}/e_{true} en fonction de e_{true}.

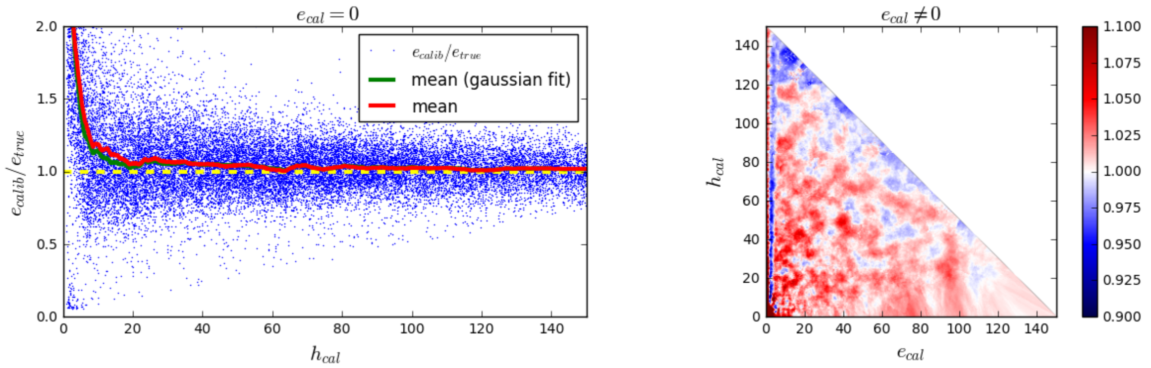


Figure 9 – e_{calib}/e_{true} en fonction de e_{cal} et h_{cal}.

2.4 Méthode des plus proches voisins (KNN)

2.4.1 Principe général de l'algorithme

Nous utilisons encore des données simulées pour effectuer une calibration, chaque particule simulée i est vue comme un point d'un espace tridimensionnel possédant des coordonnées $(e^i_{cal}, h^i_{cal}, e^i_{true})$, correspondant respectivement à l'énergie déposée dans le calorimètre électromagnétique, dans le calorimètre hadronique et l'énergie vraie.
Pour trouver l'énergie calibrée d'un point de coordonnées (e^0_{cal}, h^0_{cal}) :

— on recherche ses k plus proches voisins dans le plan $(e_{cal}, h_{cal}) \to (e^i_{cal}, h^i_{cal}), i \in [|1, ..., k|]$

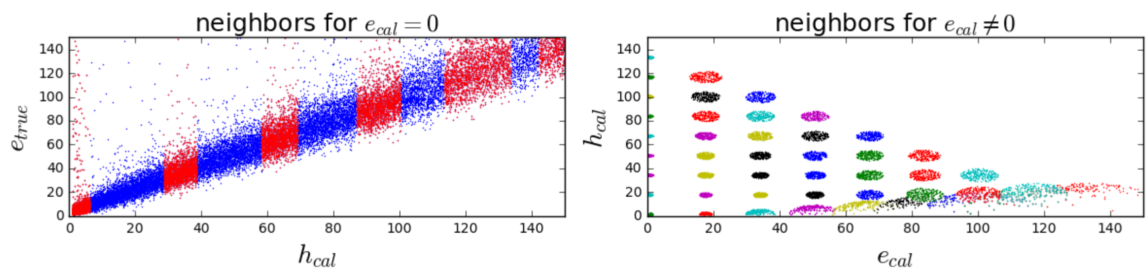


Figure 10 – $n_{voisins} = 2000$ pour $e_{cal} = 0$, $n_{voisins} = 250$ pour $e_{cal} \neq 0$

— on effectue une moyenne pondérée une moyenne pondérée de l'énergie vraie de ces plus proches voisins $\to e^0_{calib}$: l'énergie calibrée

La moyenne pondérée va donc s'exprimer ainsi :

$$e^0_{calib} = \frac{\sum_{i=1}^{k} g(e^i_{cal}, h^i_{cal}) \times e^i_{true}}{\sum_{i=1}^{k} g(e^i_{cal}, h^i_{cal})} \quad (2)$$

Dans notre cas nous avons pris pour g la distribution gaussienne $g(\vec{x}) = \exp -\frac{1}{2}(\frac{(\vec{x}-\vec{x}^0)^2}{\sigma^2})$, pour donner plus d'importance aux plus proches des k plus proches voisins.

2.4.2 Résultat de la calibration

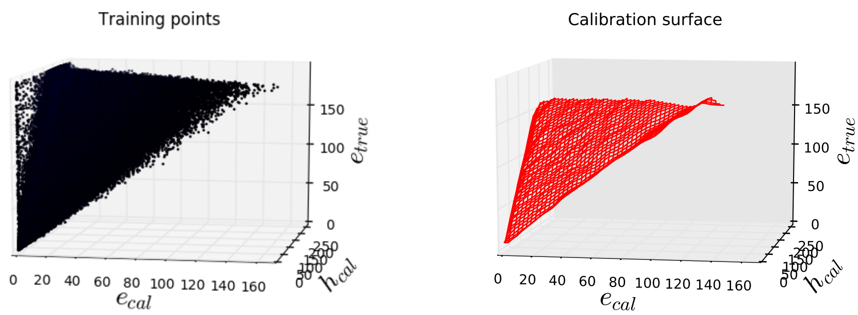


Figure 11 – Le nuage de points modélisé (à gauche) par une surface (à droite).

Nous constatons ici que la surface est beaucoup plus lisse que pour la méthode précédente, mais en regardant le cas particulier de $e_{cal} = 0$, nous constatons encore une fois que à faible énergie, la courbe de calibration ne passe pas par le coeur de la distribution. Cela vient du fait qu'il y a des points aberrants qui ont une forte énergie vraie mais qui ont une très faible énergie détectée par les calorimètres.
Il nous faut donc un moyen pour ne plus les prendre en compte pour avoir une courbe de calibration plus réaliste.

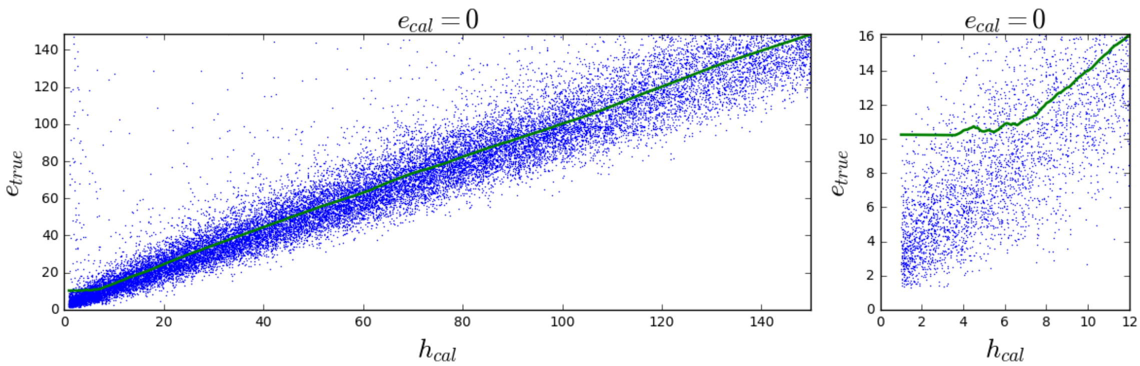


Figure 12 – Courbe de calibration pour $e_{cal} = 0$.

Malgré tout, en regardant la Fig. 13, nous constatons que les points sont mieux répartis autour de $e_{calib}/e_{true} = 1$.

Nous prenons également en compte les non-linéarité dans ce cas mais nous sur-estimons encore la valuer de l'énergie calibrée (encore une fois, à cause de ces points à fort e_{true}

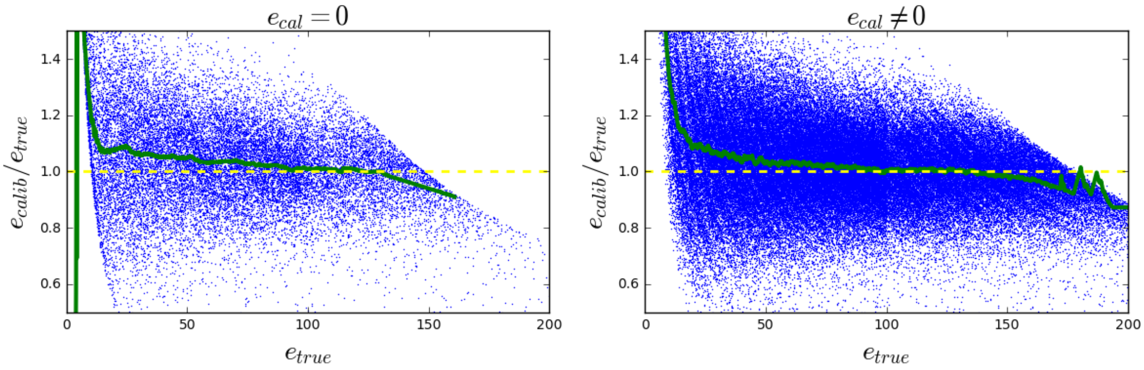


Figure 13 – e_{calib}/e_{true} en fonction de e_{true}.

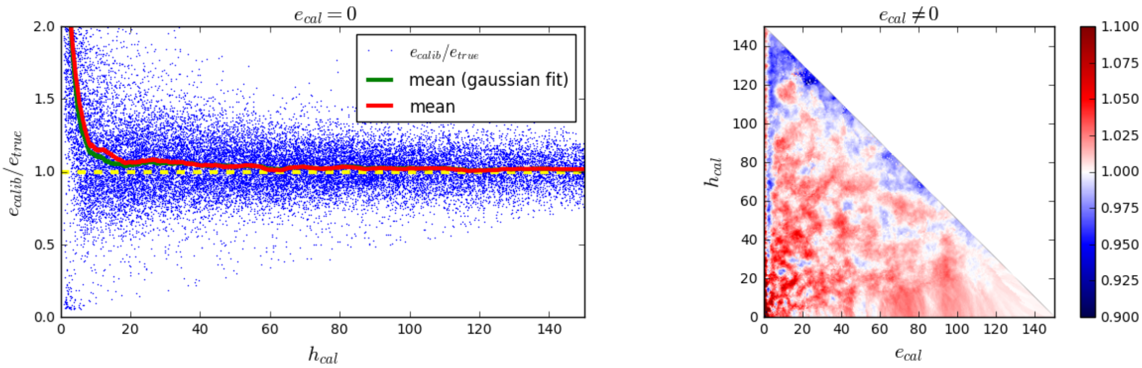


Figure 14 – e_{calib}/e_{true} en fonction de e_{cal} et h_{cal}.

2.5 KNN Gaussian Cleaning

2.5.1 Principe général de l'algorithme

Cette méthode est assez similaires à la précédente. Elle se base sur la constatation que la distribution en énergie vraie des paquets de plus proches voisins est un distribution gaussienne. Nous allons donc en utilisant la méthode des moindres carrés trouver les paramètres de la gaussienne en question et ne prendre en compte les plus proches voisins dont l'énergie vraie est $\mu - c\sigma \leq e^i_{true} \leq \mu + c\sigma$ (nous prenons par défaut $c = 2$), avec μ, σ la moyenne et l'écart type de la distribution gaussienne.
Principe de l'algorithme :

- on considère des points $(e^{0,j}_{cal}, h^{0,j}_{cal})$ où nous allons évaluer l'énergie calibrée.
- pour chaque $(e^{0,j}_{cal}, h^{0,j}_{cal})$:
  - on recherche ses k plus proches voisins dans le plan $(e_{cal}, h_{cal}) \rightarrow (e^i_{cal}, h^i_{cal}), i \in [|1, ..., k|]$
  - on trouve la gaussienne correspondante $\mu - c\sigma \leq e^i_{true} \leq \mu + c\sigma$
  - on ne conserve que les voisins dont : $\mu - c\sigma \leq e^i_{true} \leq \mu + c\sigma$
  - on effectue une moyenne pondérée une moyenne pondérée de l'énergie vraie de ces plus proches voisins $\rightarrow e^0_{calib}$: l'énergie calibrée
- on effectue une interpolation pour donner une valeur d'énergie calibrée quelque soit (e^0_{cal}, h^0_{cal})

2.5.2 Efficacité du fit

2.5.3 Résultat de la calibration

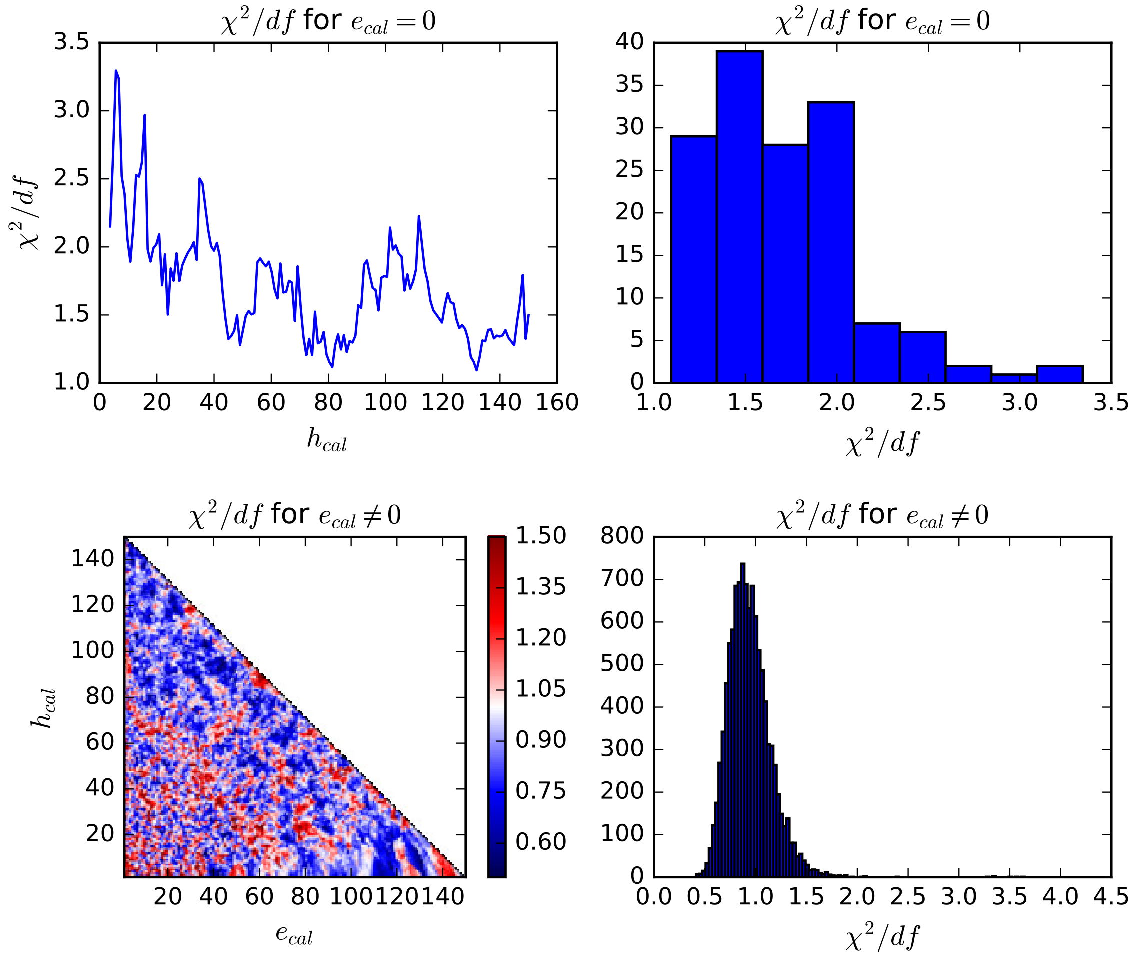


FIGURE 15 – Le χ^2 réduit pour chaque fit effectué.

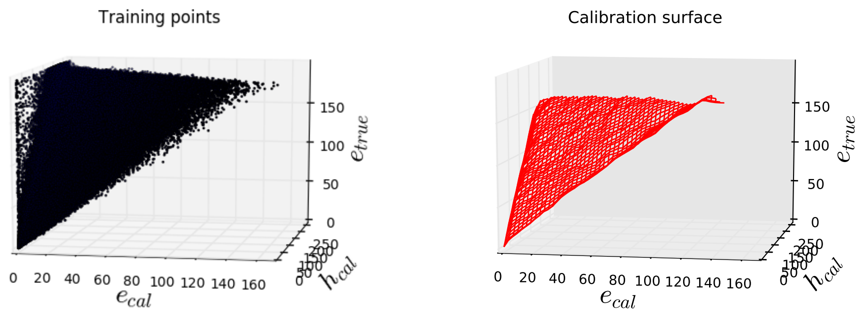


FIGURE 16 – Le nuage de points modélisé (à gauche) par une surface (à droite).

2.6 KNN Gaussian Fit

2.6.1 Principe général de l'algorithme

Ici, il s'agit du même principe que précédemment mais nous allons considérer que la valeur de e_{calib} est la moyenne de la gaussienne. Principe de l'algorithme :

- on considère des points $(e^{0,j}_{cal}, h^{0,j}_{cal})$ où nous allons évaluer l'énergie calibrée.
- pour chaque $(e^{0,j}_{cal}, h^{0,j}_{cal})$:
  - on recherche ses k plus proches voisins dans le plan $(e_{cal}, h_{cal}) \to (e^i_{cal}, h^i_{cal}), i \in [|1, ..., k|]$

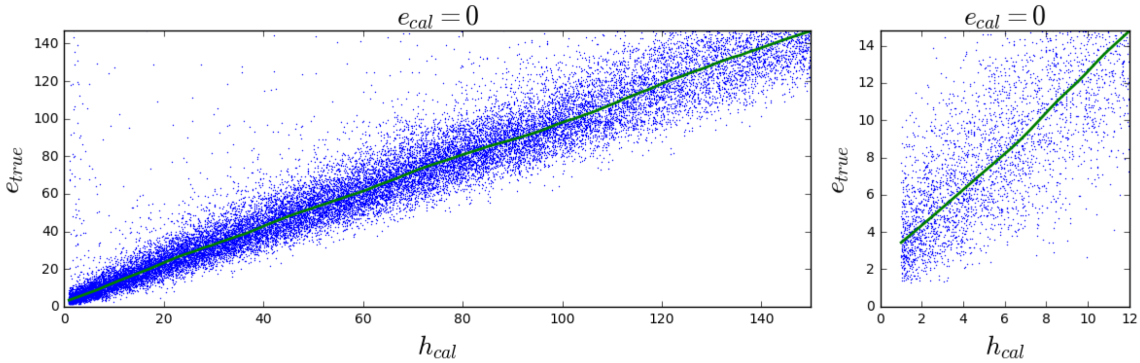


Figure 17 – Courbe de calibration pour $e_{cal} = 0$.

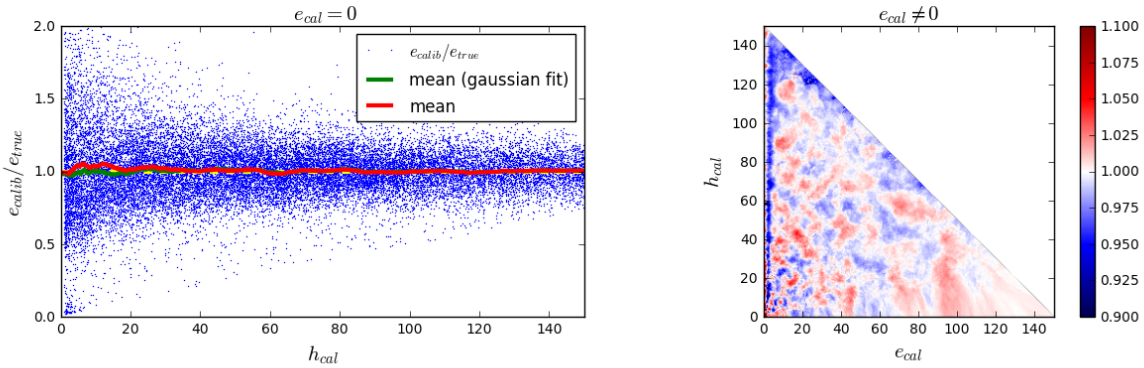


Figure 18 – e_{calib}/e_{true} en fonction de e_{cal} et h_{cal}.

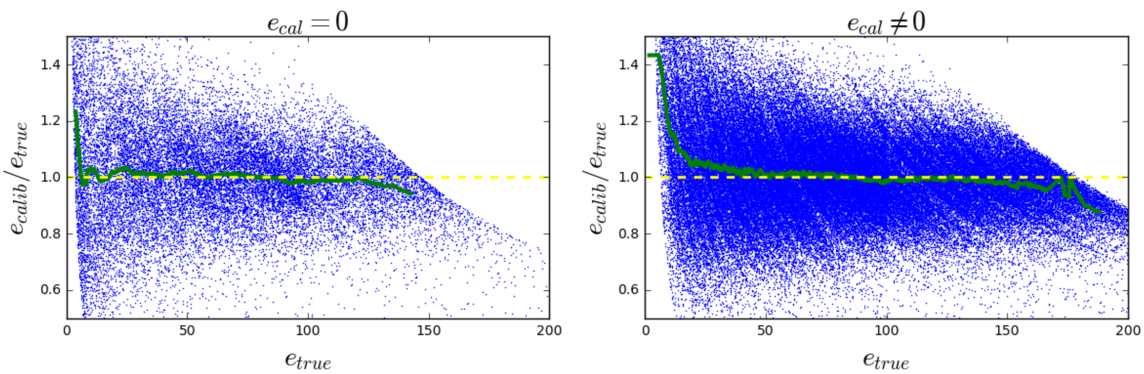


Figure 19 – e_{calib}/e_{true} en fonction de e_{true}.

  - on trouve la gaussienne correspondante $\rightarrow \sigma, \mu$
  - $\rightarrow e^0_{calib} = \mu$
- on effectue une interpolation pour donner une valeur d'énergie calibrée quelque soit (e^0_{cal}, h^0_{cal})

2.6.2 Résultat de la calibration

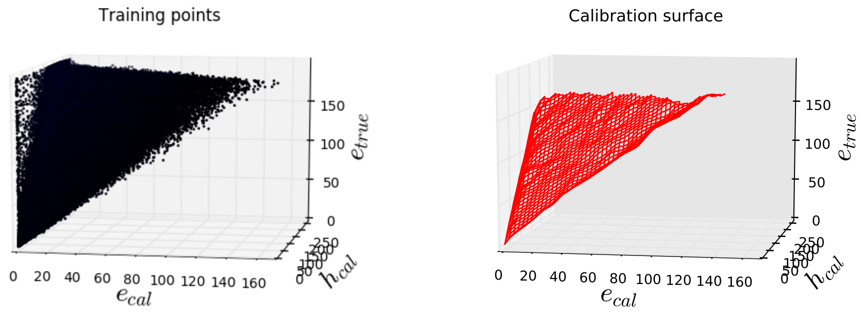


Figure 20 – Le nuage de points modélisé (à gauche) par une surface (à droite).

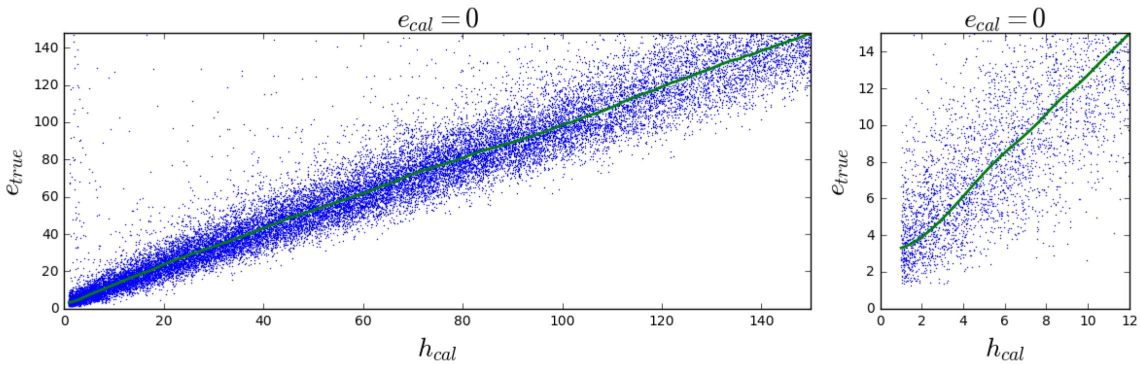


Figure 21 – Courbe de calibration pour $e_{cal} = 0$.

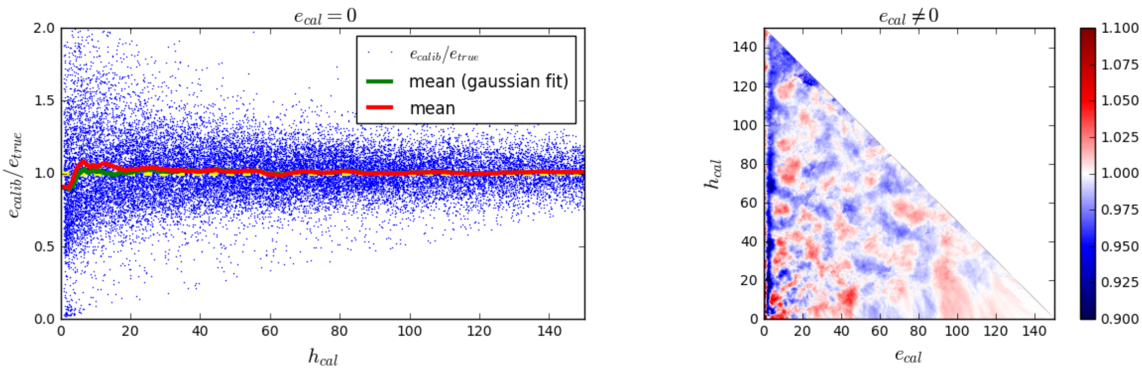


Figure 22 – e_{calib}/e_{true} en fonction de e_{cal} et h_{cal}.

3 Comparaison des méthodes

3.1 Méthodes basées sur KNN

3.2 Meilleure méthode

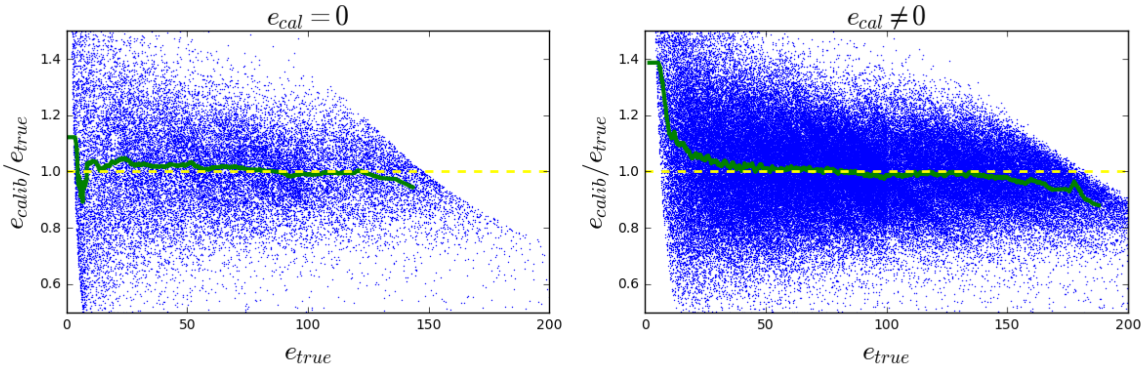


Figure 23 – e_{calib}/e_{true} en fonction de e_{true}.

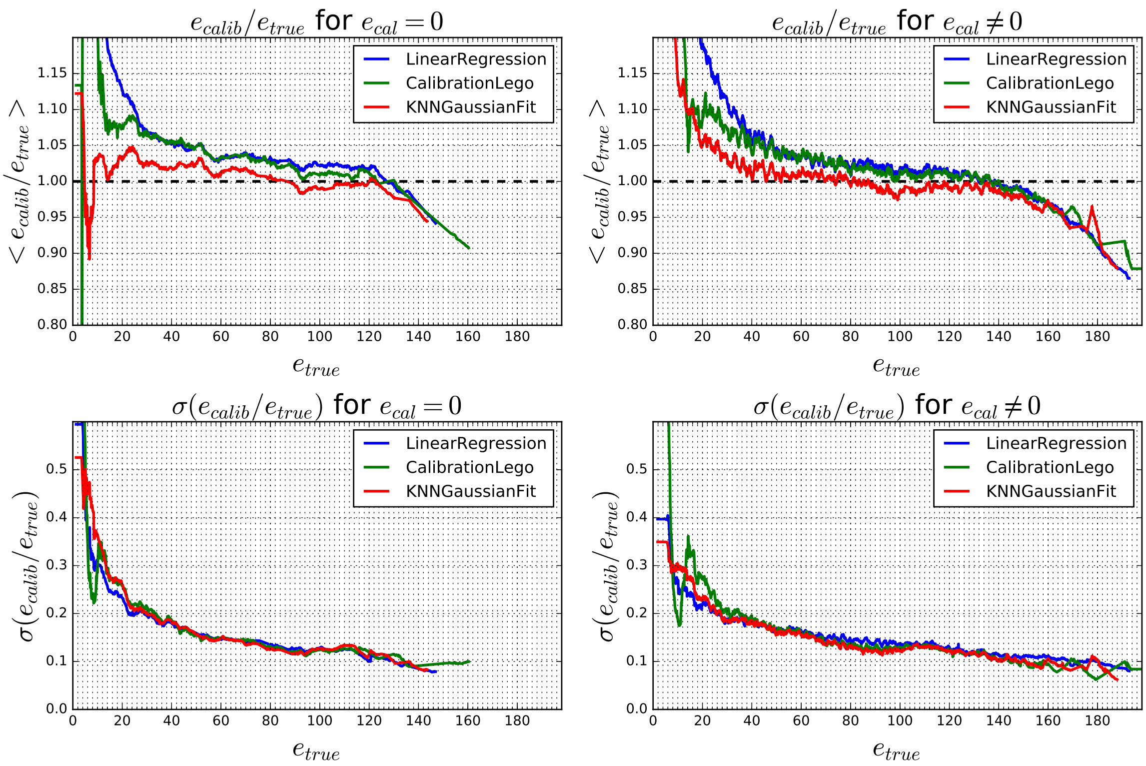


Figure 24 – e_{calib}/e_{true} en fonction de e_{true}.

4 Partage du programme

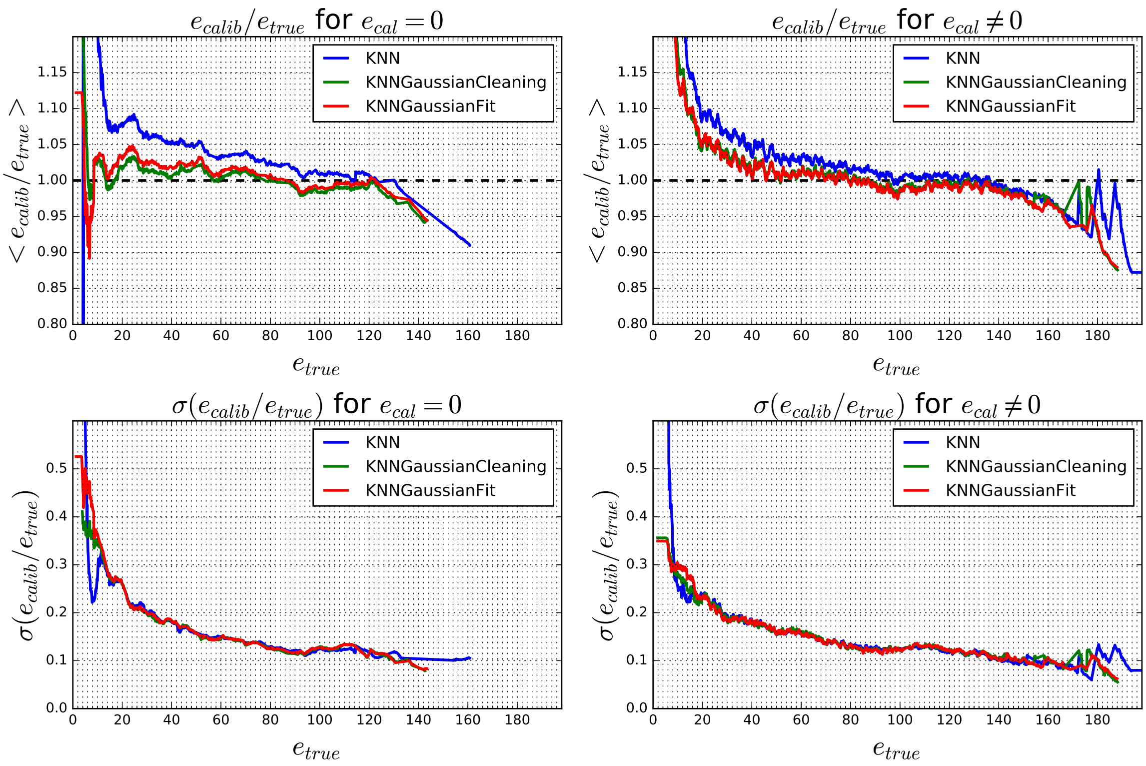


Figure 25 – e_{calib}/e_{true} en fonction de e_{true}.

5 Annexes

5.1 Comment créer une calibration ?

5.2 Fonctions utiles du programme