Régression barycentrique pour un nuage de point

Objectif :

La régression barycentrique permet de définir une fonction mathématique qui, à un N-uplet associe un P-uplet, en respectant un certain nombre fini de contraintes ponctuelles.

Applications :

Via ce procédé, on peut construire une fonction mathématique de x passant par toutes les valeurs expérimentales Yi associés aux Xi. Ainsi, à partir d'un lot de résultats expérimentaux, on pourra extrapoler les résultats ponctuels pour les généraliser sur des intervalles.

Formulation mathématique de l'objectif:

Soit U = {(x1, y1, z1, …), (x2, y2, …), …, (xk, yk, …)} un ensemble de K n-uplets deux à deux distincts, et V un ensemble de K p-uplets. On a donc :

$$card(U) = card(V)$$

On va chercher une fonction continue, dérivable, qui associera, au i-eme N-uplet de l'ensemble U le i-eme P-uplet de l'ensemble V. f la fonction définie par :

$$f_{j\in\mathbb{N}\geq 1} : x\in\mathbb{R}^n \rightarrow f_j(x) = \frac{\sum_{i=1}^{i=K} d(U_i, x, j)\times V_i}{\sum_{i=1}^{i=K} d(U_i, x, j)}$$

$$d : (a, b, j)\in(\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}) \rightarrow \left(\sum_{i=1}^{n}\left((a_i - b_i)^2\right)\right)^{-j}$$

f est donc la moyenne des éléments de V, pondérés par d(Ui, x, j), l'inverse de la distance carrée entre le point de calcul et le point Ui, élevée à la puissance j.

Etudions déjà sommairement la fonction d:

$$d\in(C^\infty)_m$$

$$a\rightarrow b \Rightarrow \left(\sum_{i=1}^{n}\left((a_i - b_i)^2\right)\right)\rightarrow 0$$
$$\Rightarrow d(a, b, j\geq 1)\rightarrow +\infty$$

Vérifions que la fonction f colle aux hypothèses fixées:

La fonction f doit associer, à tout n-uplet de U, le p-uplet correspondant de V. Vérifions:

$$f_j(u_i \in U) = v_i \in V \, ?$$

$$f_j(u \in U) \sim \frac{\sum_{i=1}^{i=K} d(u_i, u, j) \times V_i}{\sum_{i=1}^{i=K} d(u_i, u, j)} \Rightarrow f_j(u \in U) \sim \frac{\left(\sum_{i=1, u_i \neq u}^{i=K} d(u_i, u, j) \times V_i \right) + d(u, u, j) \times v}{\left(\sum_{i=1, u_i \neq u}^{i=K} d(u_i, u, j) \right) + d(u, u, j)}$$

$$\Rightarrow f_j(u \in U) \sim \frac{\vec{x} \in \mathbb{R}^p + d(u, u, j) \times v}{x \in \mathbb{R} + d(u, u, j)}, d(u, u, j) \to +\infty$$

$$\Rightarrow f_j(u \in U) \sim \frac{\vec{x} \in \mathbb{R}^p}{x \in \mathbb{R} + d(u, u, j)} + \frac{d(u, u, j) \times v}{x \in \mathbb{R} + d(u, u, j)}$$

$$\Rightarrow f_j(u \in U) \sim \vec{0} + \frac{v}{\frac{x \in \mathbb{R}}{d(u, u, j)} + 1}$$

$$\Rightarrow f_j(u \in U) \sim v$$

La fonction f, à tout n-uplet de U, associe donc bien le p-uplet de V correspondant. On peut démontrer que f est également continue, dérivable. Prenons le cas de la dimension 1 (n=p=1):

$$f_j(x) = \frac{\sum_{i=1}^{i=K} d(u_i, x, j) \times \vec{V}_i}{\sum_{i=1}^{i=K} d(u_i, x, j)}, d(u_i, x, j) = \left(\sum_{i=1}^{n} \left((a_i - b_i)^2 \right) \right)^{-j}$$

$$n = p = 1 \Rightarrow f_j(x) = \frac{\sum_{i=1}^{i=K} (u_i - x)^{-2j} \times V_i}{\sum_{i=1}^{i=K} (u_i - x)^{-2j}}, d(u_i, x, j) = (u_i - x)^{-2j}$$

D'où une dérivée de f:

$$\frac{2j \left(-\left(\sum_{i=1}^{k} \left(-(u_i - x)^{-2j-1} v_i \right) \right) \left(\sum_{i=1}^{k} (u_i - x)^{-2j} \right) + \left(\sum_{i=1}^{k} (u_i - x)^{-2j} v_i \right) \left(\sum_{i=1}^{k} \left(-(u_i - x)^{-2j-1} \right) \right) \right)}{\left(\sum_{i=1}^{k} (u_i - x)^{-2j} \right)^2}$$

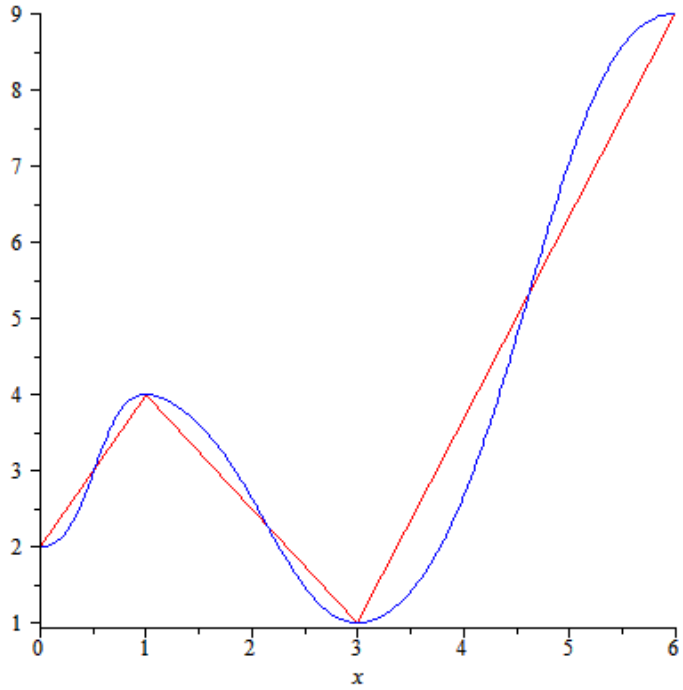
Il est à noter que pour j >= 1, la dérivée de f en tout x de U est nulle. Il en va de même pour de dimensions n et p supérieur à 1. On constate même que toutes dérivées de f jusqu'à la j-eme incluse sont nulles pour tout n-uplet de U. Quels graphs éclaireront surement tout cela:

Cas à 1 dimension:

Pour n=p=1, avec j=1, on trace, en rouge, la ligne brisée qui relie les couples (u, v) pour toute paire d'élément u et v en regard de UxV, et en bleu, la fonction f précédemment définie, avec

U = {0, 1, 3, 6}

V = {2, 4, 1, 9}

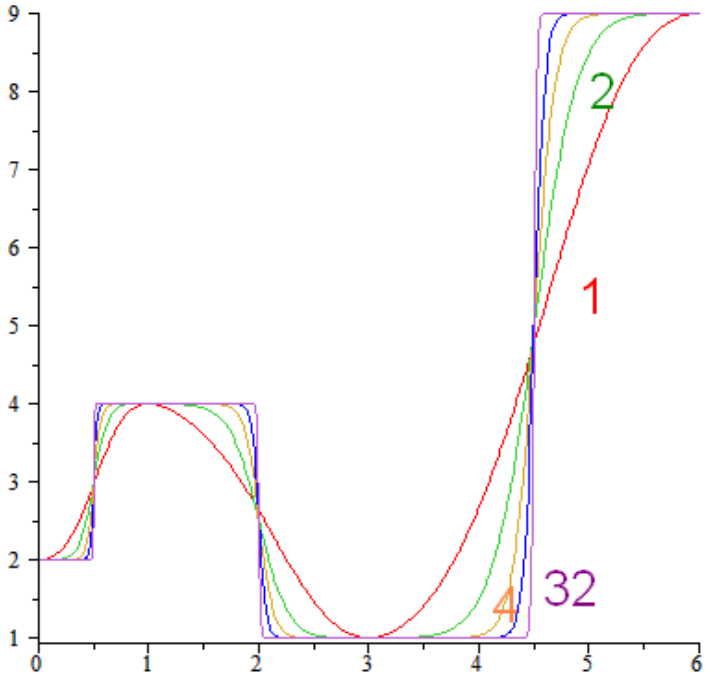


On remarque que:

- La fonction f passe par les valeurs [0, 2], [1, 4], [3, 1] et [6, 9] issus de UxV
- La fonction f a une dérivée nule en ces points
- La fonction f vaut, à peu près, la moyenne des deux points les plus proches lorsqu'on se situe entre deux points (pour x=4,5, f(x) vaut environ 5,15 qui est proche de (9+1)/2)
- La fonction f ne présente pas de valeur aberrante (toute valeur f(x) dans l'intervalle est correcte)

Influence de j:

Avec les mêmes points, en trace les fonctions f1, f2, f4, f8 et f32 (avec f* la fonction f pour une valeur de j fixée à *):

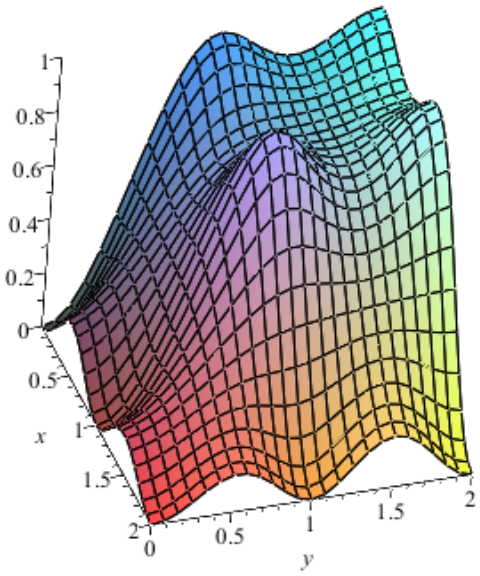


Lorsque j augmente, les valeurs prises par la fonction génèrent des paliers au niveau des valeurs expérimentales (les valeurs de UxV). Une bonne valeur de j sera donc de 1, pour éviter ces plateaux (sauf si on désire les garder).

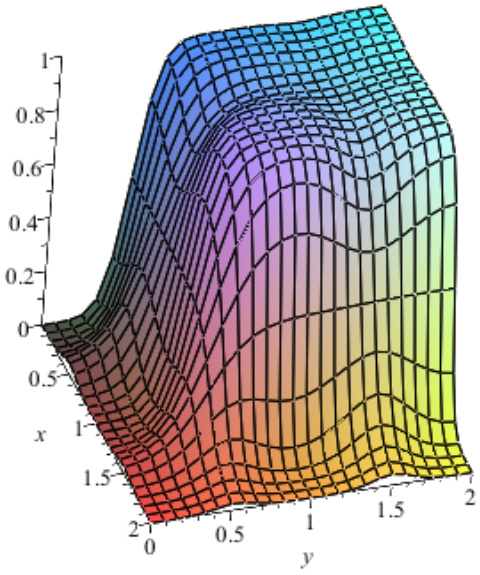
Cas 2D:

A un couple (x,y), on veut associer un scalaire z (ou un vecteur n). Les valeurs prises sont:

U = {[0,0],[0,1],[0,2],[1,0],[1,1],[1,2],[2,0],[2,1],[2,2]} et V={0,1,1,0,1,1,0,0,0} et j=1



La fonction f passe toujours par tous les points, mais des "vallées" apparaisent lorsqu'on se situe entre deux points. En effet, en dimension 1, si on choisit une valeur de x, alors on ne trouve que deux valeurs, dans U, qui soient proches de x (pour x=3, si U={1,2,4,5,6}, on a deux valeurs seulement, 2 et 4, qui sont proches de x). A l'inverse, en ajoutant une nouvelle dimension, si on choisit un point x=[0.4, 1.2], on trouve plusieurs points de l'ensemble U qui lui sont proches ([0,1], [0,2], [1,1], [1,2]). Ces points "parasitent" la valeur de la fonction, et sont à l'origine de ces vallées. Or, on a vu qu'en augmentant la valeur de j, on génère des plateaux. Essayons j=2:



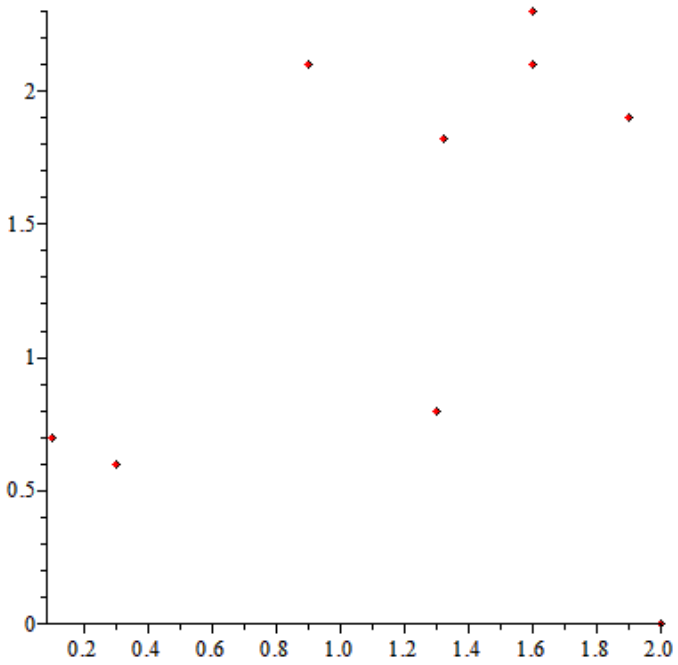
On observe que les vallées issues des points parasite de la dimension ajoutée viennent compenser les plateaux créés par l'incrémentation de j (j=1 → j=2).Donc, il est conseillé d'utiliser, comme valeur pour j:

$$j = dim(U) = n$$

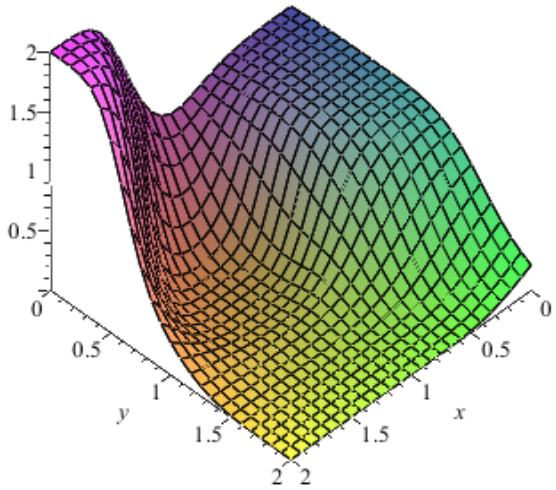
Grille U quelconque:

Dans le cas précédent, on a utilisé un ensemble U particulier, puisque les points étaient régulièrement répartis. Que se passe-t-il avec une grille inégale ?

Les valeurs de U utilisées sont :



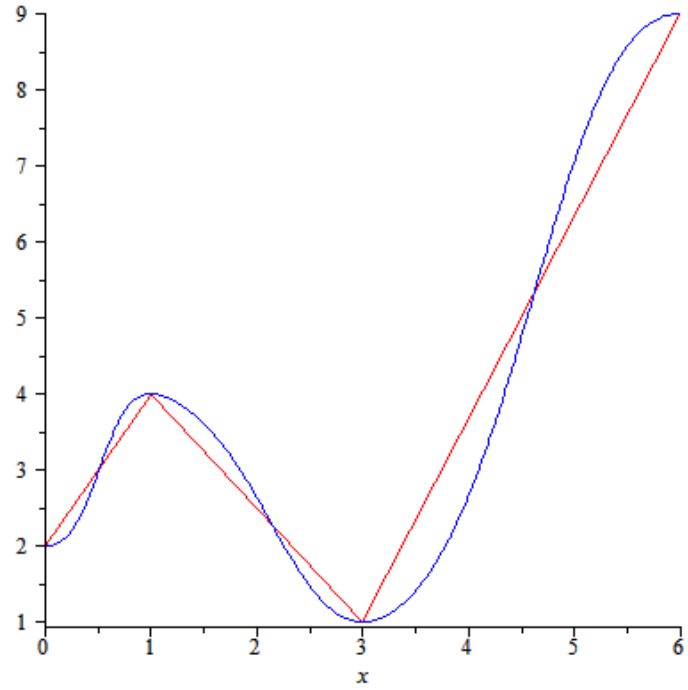
Aux points [<1,<1] (les deux en bas à gauche), on associe la valeur v=1. Au point [2,0] on associe la valeur v=2. A tous les autres points, on associe v=0. Résultat :



La fonction, comme démontré, passe par tous les points issus de UxV, sans plateau ni vallée parasite. La méthode accepte donc l'utilisation d'une grille de points non-homogènement répartis.

Hors cadre :

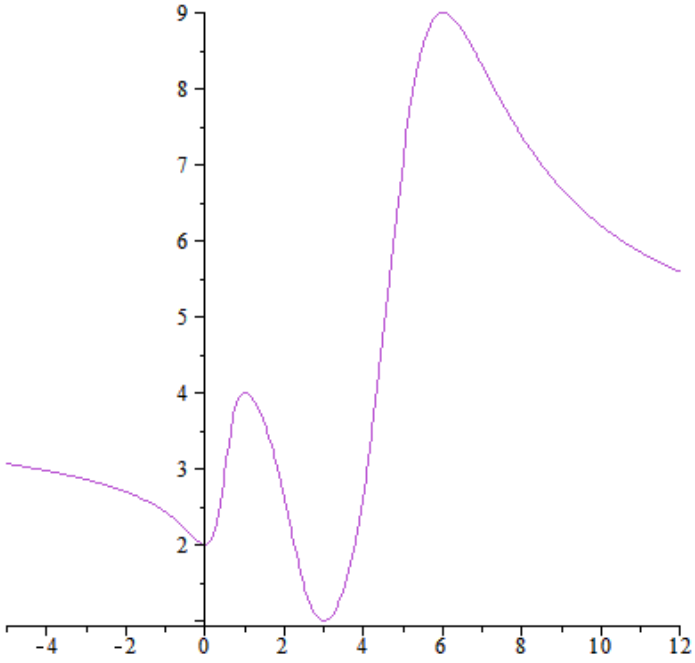
Reprenons le cas 1D, plus simple à comprendre ici :



U = {0, 1, 3, 6}, inclus dans l'intervalle [0,6] et V={2, 4, 1, 9}

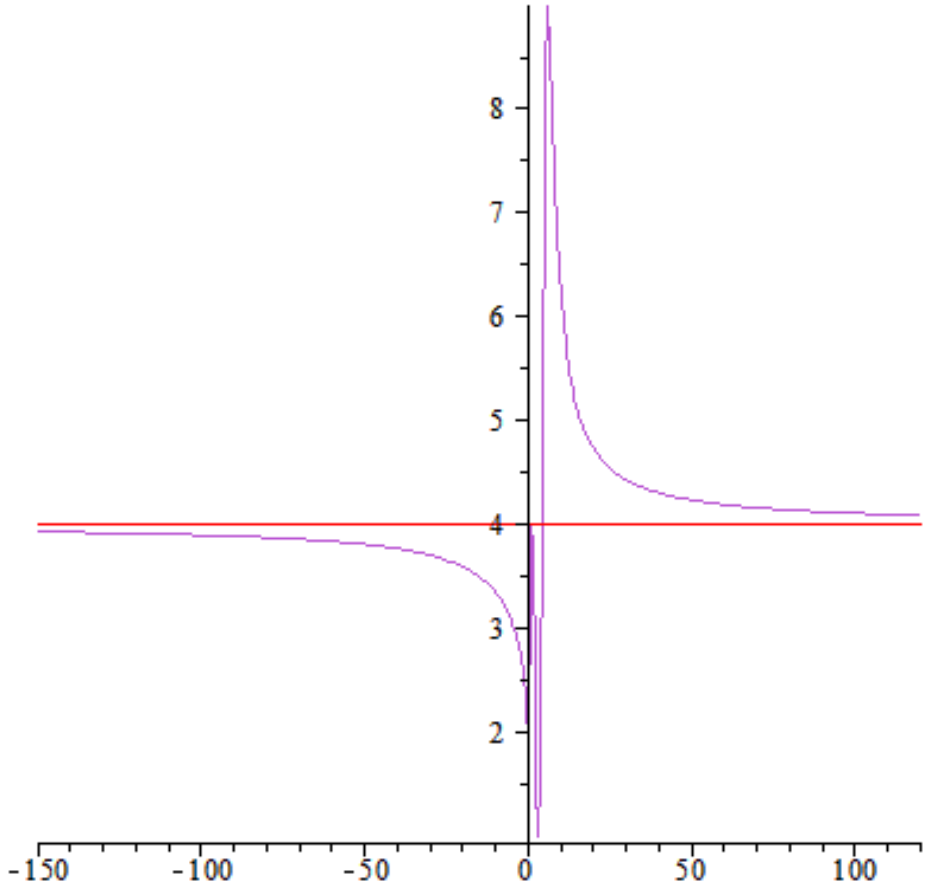
Que se passe-t-il si on demande la valeur de la fonction f en 8 ou en -2, valeurs qui sont en dehors de l'intervalle dans lequel U s'inscrit ([0,6]) ?

Prolongeons le tracé :



On s'aperçoit que les valeurs en dehors de l'intervalle ne conservent pas la tendance de l'ensemble des points de UxV. En d'autres termes, entre x=3 et x=6, la valeur de V augmente (le tracé rouge monte). On s'attend donc à ce qu'après x=6, la courbe monte toujours. Or, le tracé de la fonction f descend. La fonction f est donc très souvent invalide en dehors de l'intervalle de travail (en dehors de [0,6]). Néanmoins, f ne présente pas de valeur aberrante. On peut montrer, par le calcul, que f

tend vers la moyenne de V (la moyenne des valeurs mesurées) lorsqu'on s'éloigne de l'ensemble U des points de mesure. Prolongeons encore le graphe :



La fonction tend vers 4, valeur moyenne de V (V={2,4,1,9}, moyenne 4).

On peut le montrer rapidement en rappellant que f est une moyenne des valeurs de V, pondérée par l'inverse de la distance entre le point où on demande la mesure (ici, en 1D, x, en 2D, (x,y)). On, si on se place loin des points U, la distance entre le point où on demande la mesure et n'importe quel point de U est quasiment la même (la distance entre Paris et n'importe quel quartier de Lyon est quasiment constante, et égale à la distance Paris-Lyon). Finalement, hors du cadre des U, f est une moyenne des V, pondérés par l'inverse d'une valeur quasi constante (la distance). Donc, f est une moyenne des V (faire la moyenne d'un ensemble de nombre en multipliant tous ces nombres par une constante c, puis en divisant la moyenne par c ne change rien et est égale à la moyenne de l'ensemble de nombres, sans faire intervenir c).

Donc, hors cadre, la fonction n'est pas aberrante, mais elle sera sûrement erronée.

Limites de la méthode :

Régression:

Si on observe la formulation de f, on s'aperçoit que le nombre de calculs requis pour évaluer f en un point dépend du nombre de points expérimentaux (le nombre de valeurs dans U et V). Or, une régression n'est pas sensée dépendre du nombre de points utilisés (une régression linéaire, de la forme y=a.x + b , n'a pas une complexité qui dépend du nombre de points utilisés pour faire la régression). Le cas présent n'est donc pas, à proprement parler, une régression. Néanmoins, elle permet de construire une fonction qui va avoir un résultat semblable à une régression et ce quelques soient les points utilisés.

Limite de calculabilité :

Si on considère U formé de K n-uplets et V, formé de k p-uplets. Alors, d'après la définition de la fonction de régression, f aura une complexité de la forme :

$$O(k\times(4n+p))$$

La complexité est donc linéairement dépendant du nombre de points évalués. Souvent, on utilise une grille homogène, faite de a^n points (en 1D, on découpe l'intervalle d'étude en a points régulièrement répartis ; en 2D, on découpe en a points pour l'abscisse et en a points pour l'ordonnée, on a donc a² points). Alors, la complexité, pour une grille homogène de cette forme, est :

$$O((4n+p)a^n)$$

Avec cette méthode, la complexité explose très vite avec le nombre de points, surtout en dimensions 3 ou 4 (2 est déjà parfois limite). Point de vue informatique, la complexité reste parfaitement raisonnable et la fonction est calculable sans trop de difficultés. En revanche, un tableur aura beaucoup plus de mal à gérer une telle formule, longue à écrire (mais qui peut être générée automatiquement, de sorte que l'utilisateur n'ait qu'à la copier/coller). Cette méthode peut donc requérir une macro pour fonctionner.

Hors cadre:

Lorsque l'on s'éloigne clairement des points de mesure, autrement dit, lorsque tous les points de mesure se trouvent dans un même demi-espace de frontière passant par le poit en lequel on demande le calcul (pour le cas 1D, si U contient des valeurs incluses dans [0,4], alors tout ce qui est hors de [0,4] est considéré comme étant "hors cadre"; dans le cas 2D, tout ce qui n'est pas dans le plus petit carré qui contient l'ensemble des points de U est hors-cadre), alors le résultat du calcul en ce point tend à s'approcher de la moyenne des valeurs de V à mesure que l'on s'éloigne des points de U. Le modèle n'est donc pas abberrant, mais il ne conserve pas la "tendance" en dehors des limites: supposons qu'en 1D, dans l'intervalle [0,4], les points sont situés quasiment sur la droite d'équation y=x, alors, en 6 (qui est hors de [0,4] donc hors cadre), la fonction renverra un nombre plus proche de 2 (la moyenne de y=x sur [0,4]) que de 6 (la valeur de y en x=6).

MONIER Vincent

29/07/2011