De la pratique à la théorie

Quelques exemples vécus,

en particulier les problèmes du calage à poids bornés

Jean-Claude DEVILLE (ex-Crest/Ensai)

Plan de l’exposé:

Liminaire : Remerciements, fierté ... et perplexité !

I- De la pratique à la théorie

De quoi on parle et comment ça marche (les ‘outils’)

Exemple : ma première publication !!

La suite (de la démographie aux sondages)

II- Début aux sondages : la pratique !

L’échantillonnage équilibré et le recensement

Pondération des enquêtes et calage : état des lieux et surprise !

CALMAR ou le calage devenu opérationnel à l’INSEE : comment, pourquoi !!

La confrontation avec la pratique

III- Les problèmes spécifiques au calage ‘serré’

Une pratique naturelle et naïve…mais !

Un cas qui marche à peu près bien : le calage ‘métrique’

Pour le calage généralisé à poids bornés un gros problème de fond que je ne sais pas bien contourner

Remerciements au comité Waksberg

Fierté : sorte de prix Nobel en statistique d'enquête

Des prédécesseurs qui m'ont tant appris et fait réfléchir !

Francophonie : pas moyen de refuser

Perplexité ! :

cette conférence

Une recherche de retraité sans outils :

Informatique et calcul ?

Littérature et contacts scientifiques ?

Environnement…

D’où ce plan : ‘raconter ma vie’ avec ce fil directeur des outils, puis ce que j’**essaie** de faire maintenant

I-De la pratique à la théorie : on dit l'inverse quand on débute

La pratique= le quotidien des 'praticiens' qui 'produisent' avec des outils et dans un environnement

Bonne pratique = ce qui 'marche' (ce dont 'on' est satisfait)

Respect aux praticiens

Et la théorie alors ?

Un discours sur une ‘réalité’, en particulier sur des pratiques

Dans le meilleur des cas ça permet de comprendre et de rebondir

Dans le pire c’est un verbiage creux

La pratique de la théorie

Pitié pour les ‘théoriciens’

PRATIQUE ←-------→ *OUTILS* ←-----→ *THEORIE*

Sans les outils deux univers déconnectés (passer de l'un à l'autre, fabriquer des outils)

L'évolution des outils liée

aux 'envies' de la pratique

aux 'fantaisies' de la théorie

Exemple : ma première publication avec G.Calot consacrée aux résultats principaux de l'enquête 'familles' de …1962 !!!

Nuptialité et fécondité selon le milieu socio culturel, Economie et Statistique n°27,1971

Il y avait des régressions !!

$$Prop\ de\ familles \text{ à } k\ enfants = a\ Nombre\ moyen\ final + b$$

$$k = 0,1,\ldots,5,6\ et\ +$$

sur la 'population' des 25 CSP

OUTIL : Calculatrices électriques 4 opérations

Comment ? assez indescriptible !

C’était une pratique qui appliquait la théorie

Mais les choses étaient en train de changer

J'étais passé de la 'division démographie au 'département recherche' (1976) pour travailler à ma façon sur le calendrier de constitution des familles soit l'aspect *temporel* des données :

= analyse factorielle dans L^2 (covariance d'un processus du second ordre)

1974 Méthodes Statistiques et Numériques de l'Analyse harmonique (Annales de l'INSEE n° 15)

OUTIL: Besoin : pas mal de calcul et diagonalisation d'une matrice symétrique positive de grosse taille pour l'époque

Apprendre à programmer efficacement (un peu comme tout le monde !)

Aller porter son paquet de cartes au 'terminal'

Perforer des cartes à trous…etc

Accessoirement des maths pas tout à fait élémentaires

La suite (de la démographie aux sondages, 1976 -1983)

Les terminaux en libre service (machine à écrire puis écrans)

Les logiciels généraux (SAS en particulier)

La socialisation des savoirs et des savoir-faire dans les salles de terminaux

(les calculettes programmables stage O.S.)

La suite des outils : ordinateurs pour tous, capacités de calculs et de stockage, délocalisation des ressources…big data et cloud !

Production méthodo :

1982:Analyse de données qualitatives: comment analyser des calendriers ? (annales de l'INSEE n° 45) (femmes mariées 3 où 4 fois)

1983:Correspondance analysis, with an extension toward nominal Time-Series (Journal of Econometrics 22) avec G. SAPORTA

II-**Début aux sondages** : Je n'y connaissais rien !

Ni à la théorie (Hajek 1981, puis un peu tous les articles de Sarndal, Rao, Fuller et bien d'autres)

Ni la pratique : fabriquer le nouvel échantillon maître (83-85), puis 'tirer' des échantillons pour les autres !

L'échantillonnage équilibré et le recensement (86-87)

Efficient Sampling algorithmes and balanced sample (avec Jean Marie GROSBRAS et Nicole ROTH) Compsat 1988 Copenhague

Deux papiers que j’aime bien dans cette période

-Une théorie des enquêtes par quota/ A theory of quota Survey Survey Methodology/Techniques d'Enquête (Décembre 1991)

-Constrained Samples, Conditional inference, Weighting : Three aspects of the utilisation of Auxiliary Information .Proceding of the Workshop "Auxiliary Information in Surveys" Statistics Sweden, ÖREBRO

Pondération des enquêtes et calage : état des lieux et surprise !

« ***Redre*** » méthode de pondération (+programme) de Y.Lemel

Minimiser la variance des poids sous contrainte de somme 1 et de respecter certaines proportions et moyennes (pour avoir un 'bon' échantillon)

Remettre les choses sur pied :

Estimer des totaux ; certains sont connus par des 'sources externes'

Utiliser l'aléatoire du sondage et donc Horvitz-Thompson et ses POIDS

Minimiser une distance entre de nouveaux poids et HT

Regarder ce que ça fait !

Miracle : avec les moindres carrés on trouve l'estimateur par régression

$$\hat{t}_{reg} = \hat{t} + \widehat{B'}(\hat{X} - X) = \sum_s w_k(s) y_k$$

qui apparait comme un estimateur pondéré commode, sans modèle (!)

et dont on sait estimer la variance!

1988 Estimation linéaire et redressement sur l'information auxiliaire dans les enquêtes par sondage dans 'Essais en l'honneur d'Edmond MALINVAUD ' pp 915 à 927 Economica

L'équipe « Sondages » se renforce (Pascal Ardilly puis Olivier Sautory)

Rencontre avec Carl Sarndal et visite de Statistique Canada (87-88)

Une trouvaille : minimiser une distance ‘additive’ entre les poids

Ca règle la question du ‘raking ratio’ avec une distance de type entropie

Ca permet (en principe !) un contrôle des variations de poids (distance bornée)

Pour passer à la pratique un outil et des expérimentations

‘Calmar‘ dû à Olivier et de la propagande

1991 : Calibration Estimators (avec C.E. SÄRNDAL) Journées de Méthodologie Statistiques des 13-14 Mars 1991 INSEE Méthodes n° 29 - 30 - 31 (décembre 1992)

Le calage devient opérationnel à l’INSEE : comment, pourquoi !!

La confrontation avec la pratique :

-Choix des variables de calage :‘explicatives’ des *y* d’intérêt

-Et pour corriger des non-réponses (substitut du RR ou des cellules) ? → La forme des poids $F(q_k x'_k \lambda)$ rappelle un modèle linéaire généralisé→introduire les facteurs de non-réponse dans les *x*

-Pour plus de souplesse on passe au calage généralisé , où les poids ont la forme $g_k = F(z'_k \lambda)$ avec des z_k ‘quelconques’

Remarque : les q_k furent introduit pour justifier l’estimateur ratio….avant qu’on s’aperçoive que c’est une variante de calage généralisé !!!

Mais les *z* ne sont pas une information auxiliaire (total ?)

Problèmes non ou mal résolus :

Choix des variables (dim x=dim z)

Existence d'une solution ?

Statut du λ=effet fixe de la NR+ alea d'échantillonnage

Un peu de biblio narcissique

Calibration Estimators in Survey Sampling (avec C.E. SÄRNDAL) Journal of the American Statistical Association (JASA), June 1992, Vol 87, n° 418, pp 376-382

Generalized Raking Procedures in Survey Sampling (avec C.E. SÄRNDAL et O. SAUTORY) Journal of the American Statistical Association (JASA), september 1993,Vol. 88, N° 423, pp. 1013-1020.

1998 La correction de la non-réponse par calage ou par échantillonnage équilibré.In *Recueil de la Section des méthodes d'enquêtes, 26ème congrès de la Société de Statistique du Canada*, pages 103-110,Sherbrooke.

2000 Generalized calibration and application to weighting for non-response. In *Compstat - Proceedings in Computational Statistics:14th Symposium Held in Utrecht, The Netherlands*, pages 65-76, New York.Springer-Verlag.

2003- Calage,Calage Géneralisé et Hypercalage,*53eme session de l'Institut International de Statistique*,Berlin

Les méthodes à poids bornés: favorite des ‘praticiens’ :
Sentiment de sécurité sur la stabilité des estimations (robustesse?)

Comment on fait ? On cherche :

-de nouveaux poids g_k compris entre deux valeurs m et M

- obtenir l’égalité $t_X = \sum_s d_k x_k \, g_k = Xg$ (calage) .

Calage classique : on minimise une ‘ distance’ additive $\sum_s G(g_k\,, 1)$ où G doit avoir de bonne propriétés (convexité stricte et dérivabilité) en plus des limites voulues.

La situation est simple :

soit le problème n'admet pas de solution,

soit admet une solution unique donnée par la résolution des équations de calage $t = \sum_s d_k x_k \;\; F(q_k x'_k \lambda)$ où est F une fonction définie sur $\mathbb{R}$ déduite de la distance , croissante et 'régulière' (dérivable, de limites m et M les limites étant atteintes soit asymptotiquement soit à distance finie). Si la distance est définie sur $\mathbb{R}$ tout entier on a toujours une solution et il en va naturellement de même quand m et M sont suffisamment proches de $\pm\infty$. A l'inverse si M-m est faible on n'aura généralement pas de solution. Si M-m=0, la seule solution possible est un estimateur par ratio !

La pratique : chercher par tâtonnements des calages où *m* et *M* seront respectivement le plus grand et le plus petit possible.

On constate les choses suivantes :

- ***M* et *m* ne dépendent pas de la distance utilisée**
- **La plupart des poids convergent vers *M* ou *m***

Pourquoi ?

En fait, on a un peu changé le problème, et on étudie la position relative de la variété linéaire $Xg=t$ (de dimension $n-p$) avec le n-cube centré en $c=(M+m)\mathbf{1}/2$ et de côté $D=M-m$.

Supposons c fixé pour commencer ;

La procédure de tâtonnement produit (en général) un g sur une $p-1$-face ($g+H$ n'a qu'un point commun avec le cube)

On varie alors c sur la droite (0, **1)** pour pour minimiser le max des g_k . Le vecteur de poids trouvé est sur une $p-2$-face du cube : $n-p+2$ coordonnée valent m ou M , les autres fixent sa place dans la $p-2$-face.

Le tâtonnement consiste à rechercher un couple (c,D) et un vecteur ‘calant’ g tels que $max_k\ (|g_k - c|)$ soit minimum.

Pour une étude plus complète :
Calage serrè des poids d'enquête Monique Graf (colloque sondage,Dijon 2014)°

Utilise une technique de programmation linéaire qui conduit au même résultat que le tâtonnement sans référence à une distance

Commentaires :

On a en fait estimé 2 paramètres de plus m et M . Quel sont les conséquences sur les justifications (asymptotiques !) du calage ?

Sans doute rien de dramatique, mais va t'en savoir…

On minimise une distance de type l_∞ . Approximable par l_p pour p grand ?

On pourrait généraliser à des situations plus complexes , par exemple avec des m_k et M_k (robustesse, domaines…).

Il est plus parlant de regarder ce qui se passe dans $\mathbb{R}^p$ espace image de *X* . On étudie donc la position relative de *t* et de $K_{c,D}$ image par X du cube de centre *c* et de coté *D*.

C'est un polytope de dimension *p*, fermé convexe et symétrique de centre *Xc*. Il y a calage quand *t* est intérieur à $K_{c,D}$.

A *c* fixé , c'est toujours possible pour $D \geq D_c$:

-Si < on minimise la distance (calage 'habituel')

-si = on obtient un point $t_g = Xg$ de la frontière de K_{c,D_c} résultat du tâtonnement. t_g est dans une *p-1* (ou moins !) face de K_{c,D_c} qui est l'image d'une *p-1* face du *n*-cube.

C’est aussi la limite de $F(q_k x'_k \lambda)$ pour λ infini.

On remarque maintenant que Xc (le centre) est autorisé à varier dans la droite qui contient 0 et $\hat{t} = X\mathbf{1}$ **.** La minimisation de D_c achève le tâtonnement.

Il en résulte que *n-p+2* des g_k valent *m* ou *M* (‘équations’ d’une *p*—2 face) , les *p-2* restant servant à définir notre point dans sa *p-2*-face.

Comme annoncé par la pratique, ce résultat explique le groupement de presque tous les poids aux deux bornes, indépendamment de la distance utilisée pour faire les calculs !
La recherche de *c* est en fait l’estimation d’un paramètre supplémentaire qui amène notre point au bord de sa *p*-1-face, sur une *p-2* face avec un g_k de plus à la borne.

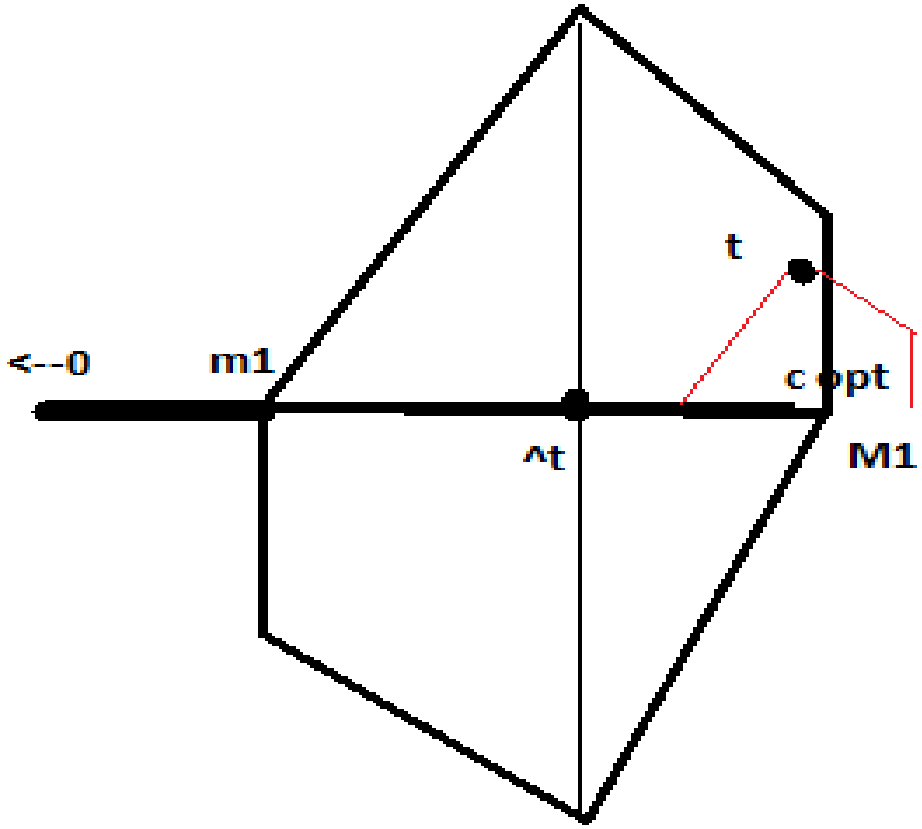


Illustration du bazar dans l'espace image

Où les choses se compliquent:Où on en est ??

1- Si t intérieur à $K = X(Cn)$ $t = Xg$ avec $g_k = F(q_k x'_k \lambda)$ et $\lambda \in \mathbb{R}^p$

2-Si t à la frontière de K presque tout les poids valent m ou M et correspondent à un $\lambda = \pm\infty$ (pour au moins une composante) si on utilise une distance.

En fait les g admissibles constituent une variété V de dimension p telle que $V+H$ recouvre le cube C_n . Tout point du cube se projette le long de H en un point unique de V.

3-S_K= ensemble des sommets de *K* est l'image par *X* d'une partie C_{ap} des sommets de C_n , qui définissent (enveloppe convexe)le 'contour apparent'. Les autres se 'projettent' à l'intérieur de *K*. (ex :un 3-cube)

$g \in C_{\text{ap}}$ si $g + Ker(X) \cap C_n$ est réduit à un point

Cela vaut pour le calage 'métrique' où *X* détermine, avec la métrique, la variété *V.*

Le contour apparent, lui, ne dépend que de *Ker(X)* !!

Est-ce encore vrai dans le cas du calage généralisé , où *V* est de la forme $g_k = F(z'_k \lambda)$ ($\lambda \in \mathbb{R}^p$) avec des z_k 'quelconques' ?

NON !

Un exemple minimaliste :

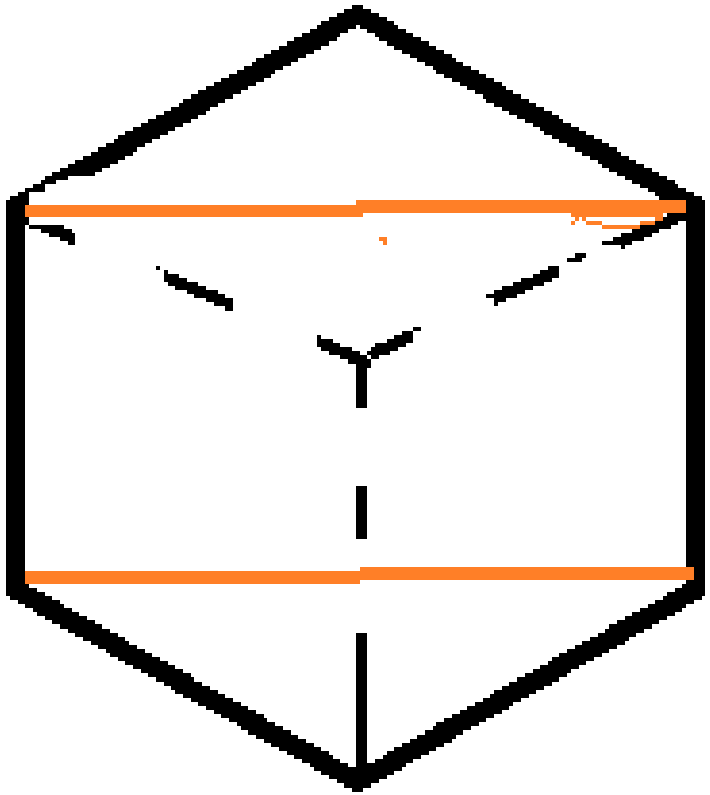
$$X = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \; et \; Z = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

$$Ker(X)=(1 \;\; 1 \; -1)'$$

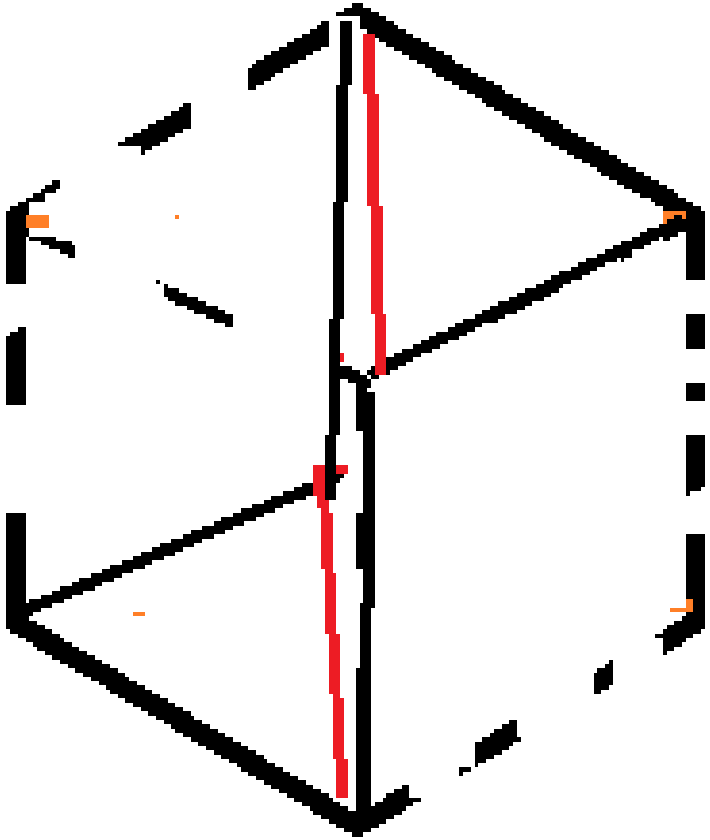
c'est du calage ordinaire et tout va bien, avec n'importe quelle fonction F!

Si on fait $X = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix}$ c'est la catastrophe !

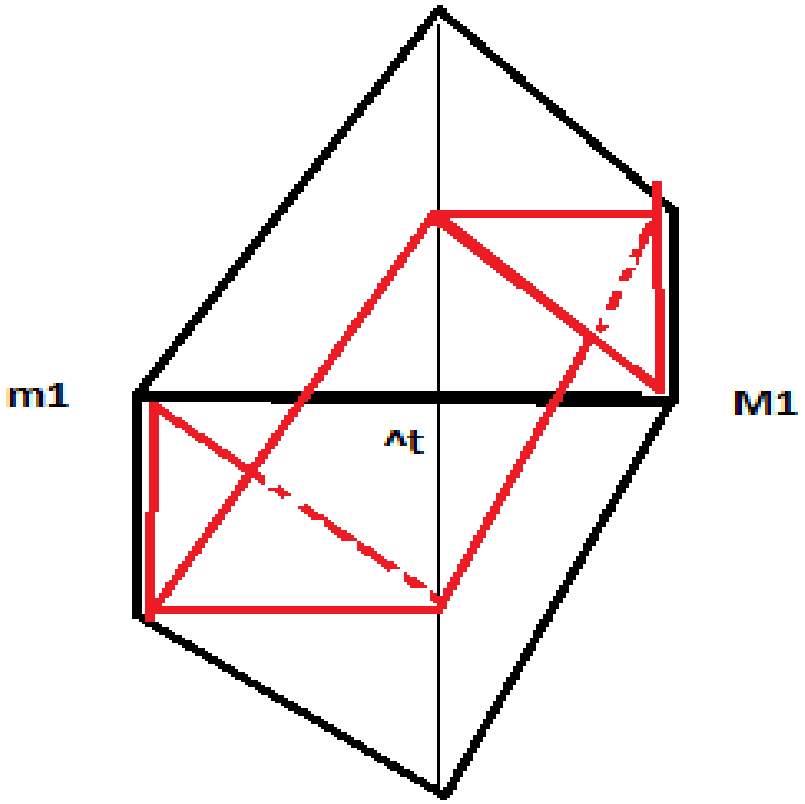
$$Ker(X)=(1 \; -1 \;\; -1)'$$



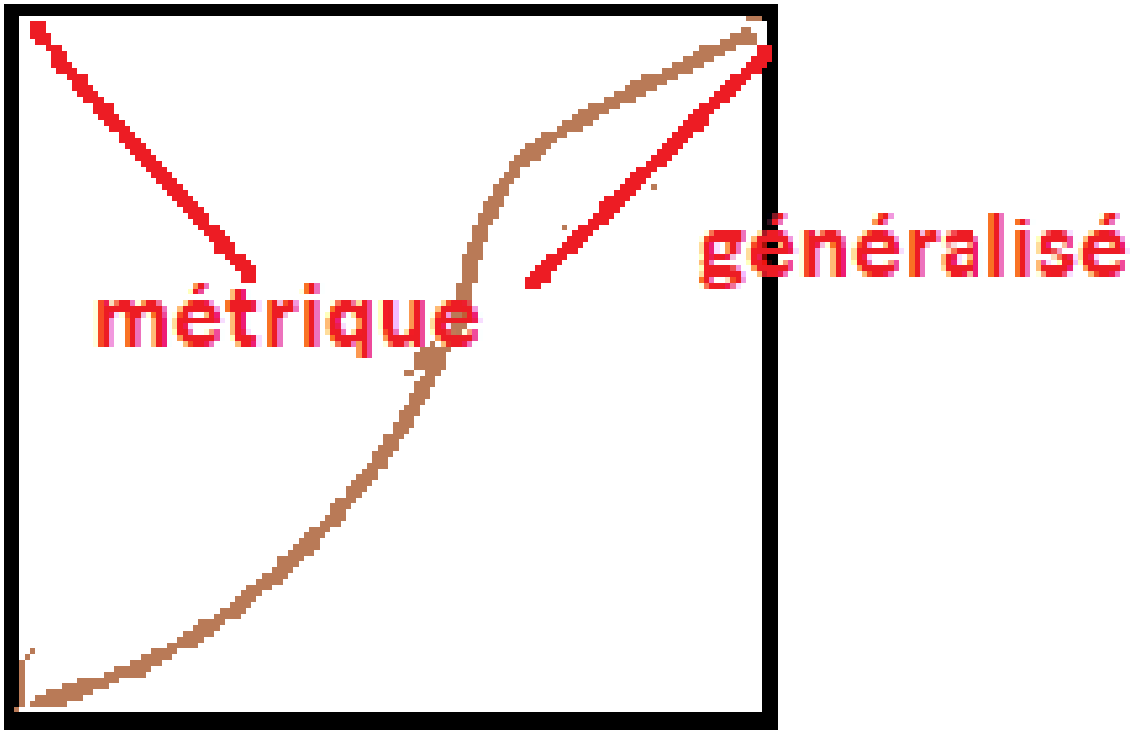
calage ordinaire



mauvais calage généralisé



Les sommets (1 1 0) et (0 0 1) du cube ne sont pas atteints…et on trouve des cas où le calage a deux solutions distinctes (ou plus)


métrique
généralisé

Une condition nécessaire évidente pour que ‘ça’ marche encore pour le calage généralisé :

V (défini uniquement par Z) atteint tous les sommets du contour apparent de C_n (défini uniquement par X).

Cette condition est également suffisante dans de nombreux cas.

Elle est elle-même entrainée par $z_k = q_k x_k$ où les q_k sont des matrices diagonales inversibles. Autrement dit pour chaque k les coordonnées des z_k et des x_k ont le même signe.(même orthant)

En particulier si toutes les variables du problème sont positives le calage généralisé à poids bornés fonctionne comme un calage métrique.

Ouf, c’est fini !

Merci de votre présence, et de votre attention