Cher Monsieur,

J'ai parcouru le site des méthodes robustes sans toutefois essayer de comprendre les détails mathématiques : je suis "un lent" et je ne comprends vraiment les choses que si je les pratique "in vivo." Mais j'ai beaucoup "réagi" à ce qui y est dit et voici quelques aspects de ces "réactions."

Sur le fond ce qui y est dit est incontestable du moins pour qui a en effet travaillé à un vrai problème en vue d'une solution praticable. Vous êtes peut-être un peu injuste à l'égard de l'Académie car il y a eu des personnes dans ce monde qui ont eu le souci du réel : je pense par exemple à R.A. Fisher qui a développé un grand nombre de techniques statistiques dans le cadre d'une station agricole. Mais il est vrai que le grand nombre a peu d'intérêt pour les vrais problèmes (voir Annals of Statistics). J'ai également beaucoup apprécié vos références à Boileau : je crois en effet que la pratique de ses préceptes a de grandes vertus.

La méthode probabiliste que vous préconisez est fort intéressante. Elle fait penser, sinon dans la lettre, du moins dans son esprit, à du Bayes empirique : on donne une vraisemblance de trouver un paramètre ou une observation dans une région donnée et on le fait en s'appuyant sur les observations disponibles en les pondérant. Il reste néanmoins que la méthode sera plus ou moins pertinente et plus ou moins efficace selon le problème que l'on tente de résoudre. Ainsi le recours à l'entropie donne souvent comme résultat une loi Gaussienne (peut-être parce qu'il y a un log), loi que beaucoup décrient comme invraisemblable. Ma position est devenue la suivante : étant donné le problème P j'utilise la méthode M; si je travaille bien, j'aurai "in fine" une opinion quant à la validité des résultats que j'ai obtenus avec cette méthode M, et le bulletin final sera

résultats obtenus avec la méthode + opinion quant à la confiance que j'ai que ces résultats sont fiables.

L'ennui avec cette pratique c'est que le "client" souvent n'aime guère cette incertitude et préfère une "certitude incertaine !"

Une surprise : je n'ai pas trouvé mention dans les divers documents que j'ai examinés des notions de robustesse déjà sur le marché, ne serait-ce que pour dire qu'elles ne sont pas pertinentes. J'en connais trois que je vais brièvement commenter.

Le mouvement de la statistique robuste

Les trois noms qui me viennent à l'esprit sont ceux de Tukey, Huber et Hampel (THH). Au départ il y a le constat banal que les valeurs excentriques ont un effet délétère sur le calcul de la moyenne. Depuis la nuit des temps les statisticiens ont élagué les valeurs extrêmes avant de calculer la moyenne mais THH ont systématisé et étendu la technique au moins dans quatre directions. La première consiste à faire du minimax dans un voisinage du modèle retenu (Huber). La seconde à exiger la continuité de certaines fonctionnelles par rapport à la convergence en loi (Hampel). La troisième (Huber) concerne l'estimation et consiste à rechercher des estimateurs à partir de fonctions qui pondèrent à la baisse les valeurs excentriques, la justification étant que même dans les expériences les mieux contrôlées il y a un pourcentage d'erreurs qui n'est pas négligeable (de l'ordre de 5%). La dernière est le "jackknife" (Tukey) qui consiste à ajuster les modèles en retirant du corpus des données disponibles un pourcentage d'observations : l'ensemble des ajustements ainsi réalisés renseigne sur la stabilité du modèle.

Il fut un temps où la statistique robuste "was all the rage" mais la mayonnaise est un peu retombée. Il y a tout d'abord ceux qui pensent que l'on n'a pas le droit d'ignorer les valeurs excentriques. Il y a ainsi un jeux de données (voir par exemple le "textbook" de Pisani, Purves et al.) concernant la calibration des poids standard issu du Bureau of Standards américain : sur 100 observations il y en a quelques unes qui sont excentriques et les gens du Bureau disent qu'on ne peut les ignorer car elle reflètent les particularités du processus de mesure. Il y a ensuite le fait qu'il y a passablement d'arbitraire dans les divers choix qui doivent être faits lors de la mise en oeuvre des méthodes robustes et qu'on ne fait en fait que repousser le problème. Cela donne lieu à tout une "cuisine" manipulatoire dont il est très souvent difficile de tracer les effets. Et cela me mène à mon troisième argument : les méthodes robustes sont très hétérogènes. Quand on travaille à un "vrai" problème, on est amené à conduire de multiples analyses et plus la batterie d'outils que l'on utilise est hétérogène, plus il est difficile de parvenir à l' "opinion de fiabilité" que je mentionnais plus haut. Il m'est souvent apparu qu'il est plus efficace de travailler avec les méthodes classiques car elles sont très homogènes.

J'ai une fois eu l'occasion d'utiliser avec "succès" je crois une technique robuste : il s'agissait d'analyser des données fiscales relatives aux cantons suisses qui ont des configurations extrêmement hétérogènes. Une revue suisse de statistique et d'économie a refusé de publier le papier que j'avais écrit avec un fiscaliste parce que je me refusais à faire des tests statistiques jugeant qu'il n'avaient en l'état aucun sens.

La robustesse en chimie analytique

C'est là que j'ai rencontré des idées et des soucis proches de ceux qui semblent être les votres bien que les solutions auxquelles on y a recours soient plus proches de la statistique usuelle que des votres. Les chimistes que j'ai fréquentés ont le problème suivant. Il mettent au point des méthodes d'analyse de contenu chimique qui doivent être pratiquées "à grande échelle." Souvent l'analyse dépend de manière cruciale de ce que certains paramètres, par exemple la température, aient des valeurs très précises. Or ils savent que dans "grande échelle" il y a "grand incertitude." Ils essayent donc de prévoir et quantifier l'effet sur les résultats de l'analyse que vont avoir des valeurs des paramètres qui ne sont pas dans le protocole d'origine.

La robustesse dans les problèmes de sonar

C'est un domaine où l'on fait me semble-t-il ce que l'on peut plutôt que ce que l'on veut. Si l'on consulte un ouvrage "classique" (par exemple : M. Bouvet, Traitement des signaux pour les systèmes sonar, Masson) on constate que l'on ne s'éloigne guère des modèles linéaires, stationnaires et Gaussiens, ce qui est loin de représenter ce qui se passe dans l'océan. Mais même dans ce cadre il y a des problèmes de robustesse. En effet les signaux sonores enregistrés sont des réalisations de processus stochastiques et la statistique des ces signaux est une statistique de dimension infinie. Dans ce cadre l'existence d'une fonction de vraisemblance n'est pas garantie et dépend de relations fines entre les paramètres des objects stochastiques en jeu. Ainsi dans le cas Gaussien les covariances doivent satisfaire

$$R_2 = R_1^{\frac{1}{2}} \, (I + T) \, R_1^{\frac{1}{2}},$$

T auto-adjoint, Hilbert-Schmidt avec des valeurs propres strictement supérieures à -1. Comment peut-on "come to terms" avec de telles exigences quand on dispose d'observations en nombre fini même si elles sont très abondantes (de Brucq : *Ainsi, pour des signaux échantillonnés à* $\Delta t = 1,2510^{-4}$ *secondes, pour décrire les fréquences allant de 0 à 3000 Hertz, une heure d'observation fournit déjà plus de* 10^7 *valeurs numériques.*) ?

J'ai eu l'occasion de travailler à ce problème pour l'Office of Naval Research (USA). On a fait la chose suivante : on a obtenu la forme de la vraisemblance dans un cadre assez général qui fait très peu d'hypothèses sur le signal (qui est transient) et suppose que le bruit a la forme AG, G processus aléatoire Gaussien, A variable aléatoire positive indépendante de G (l'introduction de A permet d'obtenir une famille de lois bien plus grande que les lois Gaussiennes tout en permettant un traitement analytique). Ensuite cette forme très générale a été "calibrée" à partir de données réelles. Comme je n'avais pas de "security clearance" j'ignore le détail de ces opérations de calibration, mais mon collègue C.R. Baker (Statistics, UNC-Chapel Hill) m'a assuré que, pour les probabilités de fausse alarme très basses, la méthode donne dans certains cas des améliorations spectaculaires.

Il y a des cas où l'on peut contourner au moins partiellement la question de la robustesse. Ainsi dans le cadre de l'assurance-maladie il y a en Suisse un processus d'égalisation des risques entre caisses-maladie : celles qui ont beaucoup de "bons risques" payent une prime et celles qui ont beaucoup de "mauvais risques" reçoivent une prime. Les compensations se font sur la base de l'âge et du sexe, ce qui est évidemment insuffisant. Une programme national de recherche est arrivé à la conclusion que les coûts de la santé dépendent de l'état de santé et qu'il faut en tenir compte lors du processus d'égalisation ! L'analyse a été faite par des économètres et spécialistes du secteur de la santé à l'aide de régressions logistiques fort complexes et de manipulations de données plutôt opaques. Je ne pourrais accepter les conclusions de l'étude que si j'étais à même d'en reconstruire les détails, ce que je ne crois pas possible (la plupart des analystes ne tiennent pas le journal détaillé de leurs activités et de ce qui motive leurs choix). Il est bien plus simple dans ce cas d'examiner l'historique des remboursements dont a bénéficié chaque assuré : on ne fait aucune hypothèse et les faits "sautent aux yeux."

Un dernier mot concerne la modélisation dans le secteur de la banque et finance : certains du moins des acteurs ont en général conscience qu'ils utilisent des modèles parfaitement artificiels. Mais ils savent aussi que le fait que tous utilisent ces modèles suffit. En effet, comme me l'a dit un gestionnaire de fonds, il est très difficile de convaincre un client de baser un investissement à partir de méthodes que nul autre utilise !

Voilà donc quelques unes des remarques qui me sont venues à l'esprit en parcourant le site "méthodes robustes." J'espère qu'elle ont un intérêt. Je me réjouis de consulter l'ouvrage que vous préparez.

Bonnes salutations. Gualtierotti