Apprentissage Statistique & Data mining

PHILIPPE BESSE

Version Octobre 2006

Institut de Mathématiques de Toulouse
Laboratoire de Statistique et Probabilités — UMR CNRS C5583
Institut National des Sciences Appliquées de Toulouse — 31077 – Toulouse cedex 4.

Chapitre 1

Introduction

1 Objectif

L'objet de ce cours est d'introduire, sous une forme homogène et synthétique, les techniques de modélisation statistique et issues de la théorise de l'apprentissage utilisées le plus couramment en *fouille de données* ou *data mining* dans des champs d'applications très divers : industriels, marketing, ou encore en relation avec des thématiques de recherche en Biologie, Épidémiologie...

La première partie ou premier objectif d'une telle démarche : l'*exploration statistique* et la recherche de classes est développée dans un autre document (Baccini et Besse 2000). Ce cours se focalise sur le deuxième objectif de la fouille de données qui est la recherche d'informations pertinentes (de pépites d'information) pour l'aide à la décision et la prévision.

La section 2 suivante de ce chapitre introduit à la *fouille de données* tandis que la section 3 reprend ces objectifs dans le cadre général de la modélisation afin d'en élargir les champs d'application. La section 4 décrit la stratégie très généralement mise en place pour optimiser choix de méthodes et choix de modèles ; la section 5 décrit brièvement quelques exemples d'application et notamment ceux utilisés pour illustrer ce cours. Enfin, la section 6 liste rapidement les méthodes qui sont abordées et les raisons qui ont conduit à ce choix.

2 Motivations du *data mining*

2.1 Origine

Le développement des moyens informatiques et de calcul permet le stockage (bases de données), le traitement et l'analyse d'ensembles de données très volumineux. Plus récemment, le perfectionnement des logiciels et de leurs interfaces offrent aux utilisateurs, statisticiens ou non, des possibilités de mise en œuvre très simples de ces méthodes. Cette évolution, ainsi que la popularisation de nouvelles techniques algorithmiques (réseaux de neurones, support vector machine...) et outils graphiques, conduit au développement et à la commercialisation de logiciels (Enterprise miner, Clementine, Insightfull miner...) intégrant un sous-ensemble de méthodes statistiques et algorithmiques utilisées sous la terminologie de *Data Mining* généralement traduit par *fouille de données* (voir Tufféry 2007 pour un exposé plus complet et détaillé). Cette approche, dont la présentation est principalement issue du marketing spécialisé dans la gestion de la relation client (GRC) (*client relation management* ou CRM), trouve également des développements et applications industrielles en contrôle de qualité ou même dans certaines disciplines scientifiques dès lors que les ingénieurs et chercheurs sont confrontés à un volume de données important. L'accroche publicitaire souvent citée par les éditeurs de logiciels (SAS) est :

> *Comment trouver un diamant dans un tas de charbon sans se salir les mains.*

Nous proposons d'évaluer et d'expérimenter la réalité de cette annonce qui s'adresse à un marché en pleine expansion. Les entreprises sont en effet très motivées pour tirer parti et amortir, par une aide à la décision quantifiée, les coûts de stockage des teras octets que leur service informatique s'emploie à administrer.

2.2 Environnement

Le contexte informationnel de la fouille de données est celui des *data wharehouses*. Un entrepôt de données, dont la mise en place est assuré par un gestionnaire de données (data manager) est un ensemble de bases relationnelles extraites des données brutes de l'entreprise et relatives à une problématique :

- gestion des stocks (flux tendu), des ventes d'un groupe afin de prévoir et anticiper au mieux les tendances du marché,
- suivi des fichiers clients d'une banque, d'une assurance, associés à des données socio-économiques (INSEE), à l'annuaire, en vue de la constitution d'une segmentation (typologie) pour cibler des opérations de marketing ou des attributions de crédit. La *gestion de la relation client* (GRC ou CRM) vise à une individualisation ou personnalisation de la production et de la communication afin d'évacuer la notion de *client moyen*.
- recherche, spécification puis ciblage de *niches* de marché les plus profitables (banque) ou au contraire les plus risquées (assurance) ;
- suivi en ligne des paramètres de production (traçabilité) en contrôle de qualité pour détecter au plus vite l'origine d'une défaillance ;
- prospection textuelle (*text mining*) et veille technologique ;
- *web mining* et comportement des internautes ;
- ...

Cet environnement se caractérise par

- une informatique hétérogène faisant intervenir des sites distants (Unix, Dos, NT, VM...) à travers le réseau de l'entreprise (intranet) ou même des accès extérieurs (internet). Des contraintes d'efficacité, de fiabilité ou de sécurité conduisent à répartir, stocker l'information à la source plutôt qu'à la dupliquer systématiquement ou à la centraliser.
- L'incompatibilité logique des informations observées sur des échantillons différents ne présentant pas les mêmes strates, les mêmes codifications.
- Des volumes et flux considérables de données issues de saisies automatisées et chiffrés en téra-octets.
- Contrairement à une démarche statistique traditionnelle (planification de l'expérience), les données analysées sont stockées à d'autres fins (comptabilité, contrôle de qualité...) et sont donc *préalables* à l'analyse.
- La nécessité de ne pas exclure *a priori* un traitement *exhaustif* des données afin de ne pas laisser échapper, à travers le crible d'un *sondage*, des groupes de faibles effectifs mais à fort impact économique.

3 Apprentissage statistique

Un peu de recul permet d'inscrire la démarche de la fouille de données dans un contexte plus large et donc potentiellement plus propice à d'autres domaines d'application.

3.1 Objectif général

Dès qu'un phénomène, qu'il soit physique, biologique ou autre, est trop complexe ou encore trop bruité pour accéder à une description analytique débouchant sur une modélisation déterministe, un ensemble d'approches ont été élaborées afin d'en décrire au mieux le comportement à partir d'une série d'observations. Citons la reconnaissance de la parole ou de caractères manuscrits, l'imagerie médicale ou satellitaire, la prévision d'une grandeur climatique ou économique, du comportement d'un client...la plupart des disciplines scientifiques sont concernées. Historiquement, la Statistique s'est beaucoup développée autour de ce type de problèmes et a proposé des *modèles* incorporant d'une part des *variables explicatives ou prédictives* et, d'autre part, une composante aléatoire ou *bruit*. Il s'agit alors d'*estimer* les *paramètres* du modèle à partir des observations en contrôlant au mieux les propriétés et donc le comportement de de la partie aléatoire. Dans la même situation, la communauté informatique parle plutôt d'*apprentissage* visant le même objectif. Apprentissage machine (ou *machine learning*), reconnaissance de forme (pattern recognition) en sont les principaux mots-clefs.

3.2 Problématiques

Supervisé vs. *non-supervisé*

Distinguons deux types de problèmes : la présence ou non d'une variable à *expliquer* Y ou d'une *forme* à reconnaître qui a été, conjointement avec X, observée sur les mêmes objets. Dans le premier cas il s'agit bien d'un problème de modélisation ou *apprentissage supervisé* : trouver une fonction ϕ susceptible, au mieux selon un critère à définir, de reproduire Y ayant observé X.

$$Y = \phi(X) + \varepsilon$$

où ε symbolise le bruit ou erreur de mesure avec le parti pris le plus commun que cette erreur est additive. En cas d'erreur multiplicative, une transformation logarithmique ramène au problème précédent.

Dans le cas contraire, en l'absence d'une variable à expliquer, il s'agit alors d'apprentissage dit *non-supervisé*. L'objectif généralement poursuivi est la recherche d'une typologie ou taxinomie des observations : comment regrouper celles-ci en classes homogènes mais les plus dissemblables entre elles. C'est un problème de classification (*clustering*).

Attention, l'anglais *classification* se traduit plutôt en français par discrimination ou classement (apprentissage supervisé) tandis que la recherche de classes (*clustering*) (apprentissage non-supervisé) fait appel à des méthodes de classification ascendante hiérarchique ou à des algorithmes de réallocation dynamique (k-means) ou de cartes auto-organisatrices (Kohonen). Ces méthodes de classification ou *clustering* ne sont pas abordées ici, elles ont été regroupées avec les techniques exploratoires (Baccini et Besse 2000).

Modélisation vs. *apprentissage*

Tout au long de ce document, les termes de *modélisation* et d'*apprentissage* sont utilisées comme des synonymes ce qui est abusif tant que les objectifs d'une étude n'ont pas été clairement explicités. Dans la tradition statistique, la notion de *modèle* est centrale surtout avec une finalité *explicative*. Il s'agit alors d'approcher la réalité, le *vrai* modèle, supposé exister, éventuellement basé sur une théorie physique, économique... sous-jacente. Le choix du modèle (cf. ci-dessous) est alors guidé par des critères d'*ajustement* et les décisions de validité, de présence d'effets, basées sur des *tests* reposant eux-mêmes sur des hypothèses probabilistes. L'interprétation du rôle de chaque variable explicative est prépondérante dans la démarche.

En revanche, si l'objectif est essentiellment *prédictif*, il apparaît que le meilleur modèle n'est pas nécessairement celui qui ajusterait le mieux le vrai modèle. La théorie de l'*apprentissage* (Vapnik, 1999) montre alors que le cadre théorique est différent et les majorations d'erreur requièrent une autre approche. Les choix sont basés sur des critères de qualité de *prévision* visant à la recherche de *modèles parcimonieux*, c'est-à-dire de complexité (nombre de paramètres ou flexibilité limitée) dont l'interprétabilité passe au deuxième plan. La deuxième devise des Shadoks (voir figure devshad) n'est pas une référence à suivre en Statistique !

Discrimination vs. *régression*

Le type des variables statistiques considérées diffèrent selon l'espace dans lequel elles prennent leurs valeur. Elles peuvent être qualitatives à valeurs dans un ensemble de cardinal fini ou quantitatives à valeurs réelles voire fonctionnelles (Besse et Cardot, 2003). Ce dernier cas est introduit en annexe par le chapitre **??**. Certaines méthodes d'apprentissage ou de modélisation s'adaptent à tout type de variables explicatives tandis que d'autres sont spécialisées. Enfin, si Y à expliquer est qualitative, on parle de discrimination, classement ou reconnaissance de forme tandis que si Y est quantitative on parle, par habitude, d'un problème de régression. Dans ce cas encore, certaines méthodes sont spécifiques (régression linéaire, analyse discriminante) tandis que d'autres s'adaptent sans modification profonde remettant en cause leur principe (réseaux de neurones, arbres de décision...).

Statistique, informatique et taille des données

Lorsque des hypothèses relatives au modèle (linéarité) et aux distributions sont vérifiées c'est-à-dire, le plus souvent, lorsque l'échantillon ou les résidus sont supposés suivre des lois se mettant sous la forme d'une famille exponentielle (gaussienne, binomiale, poisson...), les techniques statistiques de modélisation tirées du modèle linéaire général sont optimales (maximum de vraisemblance) et, surtout dans le cas d'échantillons de taille restreinte, il semble difficile de faire beaucoup mieux.

En revanche, dès que les hypothèses distributionnelles ne sont pas vérifiées, dès que les relations sup-

FIG. 1.1 – *Shadok : devise numéro 2*

posées entre les variables ne sont pas linéaires ou encore dès que le volume des données est important, d'autre méthodes viennent concurrencer l'approche statistique classique.

Prenons un exemple simple : expliquer une variable quantitative Y par un ensemble $\{X^1, \ldots, X^p\}$ de variables également quantitatives :

$$Y = \phi(X^1, \ldots, X^p) + \varepsilon.$$

observées sur un échantillon $(y_i, \boldsymbol{x}_i); i = 1, \ldots, n$ de taille n Si ϕ est supposée linéaire et p petit, de l'ordre d'une dizaine ; le problème est bien connu et largement débattu dans la littérature. Dans le cas où ϕ n'est pas franchement linéaire et n grand, il est possible d'estimer précisément un nombre plus important de paramètres et donc d'envisager des modèles plus sophistiqués. Si on s'en tient au modèle gaussien usuel, même le cas le plus simple d'un modèle polynômial devient vite problématique. En effet, lorsque ϕ est linéaire, prenons $p = 10$, la procédure de choix de modèle est confrontée à un ensemble de 2^{10} modèles possibles et des algorithmes astucieux permettent encore de s'en sortir. En revanche, considérer pour ϕ un simple polynôme du deuxième voire troisième degré avec toutes ses interactions, amène à considérer un nombre considérable de paramètres et donc, par explosion combinatoire, un nombre astronomique de modèles possibles. D'autres méthodes doivent alors être considérées en prenant en compte nécessairement la complexité algorithmique des calculs. Ceci explique l'implication d'une autre discipline, l'informatique, dans cette problématique. Le souci de calculabilité l'emporte sur la définition mathématique du problème qui se ramène à l'optimisation d'un critère d'ajustement de ϕ sur un ensemble de solutions plus ou moins riche. Ces méthodes ont souvent été développées dans un autre environnement disciplinaire : informatique, intelligence artificielle… ; k plus proches voisins, réseaux de neurones, arbres de décisions, *support vector machine* deviennent des alternatives crédibles dès lors que le nombre d'observations est suffisant ou le nombre de variables très important.

3.3 Stratégies de choix

Choix de méthode

Avec l'avènement du *data mining*, de très nombreux articles comparent et opposent les techniques sur des jeux de données publics et proposent des améliorations incrémentales de certains algorithmes. Après une période fiévreuse où chacun tentait d'afficher la suprématie de sa méthode, un consensus s'est établi autour de l'idée qu'il n'y a pas de "meilleure méthode". Chacune est plus ou moins bien adaptée au problème posé, à la nature des données ou encore aux propriétés de la fonction ϕ à approcher ou estimer. Sur le plan méthodologique, il est alors important de savoir comparer des méthodes afin de choisir la plus pertinente. Cette comparaison repose sur une estimation d'erreur (de régression ou de classement) qu'il est nécessaire

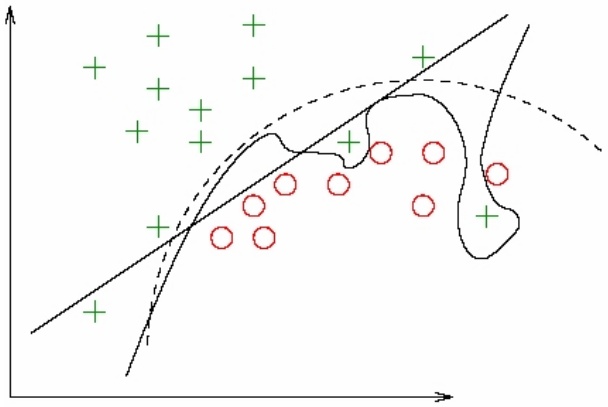

FIG. 1.2 – *Sous-ajustement linéaire et sur-ajustement local (proches voisins) d'un modèle quadratique.*

de conduire avec soin. Un chapitre (5) est consacré à ce point.

Choix de modèle : équilibre biais-variance

Tous les auteurs s'accordent pour souligner l'importance qu'il y a à construire des modèles *parcimonieux* quelque soit la méthode utilisée. Toutes les méthodes sont concernées : nombre de variables explicatives, de feuilles dans un arbre ou de neurones dans une couche cachée.... Seuls les algorithmes de combinaison de modèles (bagging, boosting) contournent cette étape au prix d'un accroissement sensible du volume des calculs et surtout de l'interprétabilité des résultats obtenus.

L'alternative est claire, plus un modèle est complexe et donc plus il intègre de paramètres et plus il est flexible donc capable de s'ajuster aux données engendrant ainsi une erreur faible d'ajustement. En revanche, un tel modèle peut s'avérer défaillant lorsqu'il s'agira de prévoir ou généraliser, c'est-à-dire de s'appliquer à des données qui n'ont pas participé à son estimation.

L'exemple élémentaire de la figure 10.1 illustre ce point fondamental dans le cas d'un problème de discrimination dans $\mathbb{R}^2$. Une frontière dolnt le modèle "vrai" est quadratique est, à cause d'"erreurs de mesure" sous-ajustée par une régression linéaire mais surajustée par un polynôme de degré plus élevé ou l'algorithme local des k plus proches voisins.

Ce problème s'illustre aussi facilement en régression classique. Ajouter des variables explicatives dans un modèle ne peut que réduire l'erreur d'ajustement (le R^2) et réduit le biais si le "vrai" modèle est un modèle plus complet. Mais, ajouter des variables fait rédhibitoirement croître la variance des estimateurs et donc celle des prévisions qui se dégradent, voire explose, avec la multicolinéarité des variables explicatives. Un risque pour le modèle, ou erreur quadratique de prévision, s'exprimant comme le carré du biais plus la variance, il est important d'optimiser le dosage entre biais et variance en contrôlant le nombre de variables dans le modèle (sa complexité) afin de minimiser le risque. Ces remarques conduisent à la définition de critères de choix de modèle dont le C_p de Mallows fut un précurseur en régression suivi par d'autres propositions : Akaïke (AIC), Schwartz (BIC)...

Parfois plus que celui de la méthode, le choix du bon modèle dans une classe ou ensemble de modèles pour une méthode donnée est primordial. En conséquence, les problèmes d'optimisation considérés doivent mettre en œuvre un critère qui prend en compte la *complexité du modèle*, c'est-à-dire la complexité de l'espace ou de la classe dans lequel la solution est recherchée.

Choix de modèle : sélection vs. *régularisation*

Selon la méthode considérée, la complexité du modèle s'exprime de différentes façons. Simple lors d'une sélection de variable en régression linéaire, la complexité est directement liée à la dimension de l'espace engendré et donc au nombre de variables. Les choses se compliquent pour les modèles non-linéaires lorsque, à dimension fixée, c'est la plus ou moins grande flexibilité des solutions qui doit être pénalisée.

C'est typiquement le cas en régression non-paramétrique ou fonctionnelle. Une pénalisation faisant intervenir la norme carrée de la dérivée seconde contrôle la flexibilité d'un lissage spline. La "largeur de fenêtre" du noyau contrôle également la régularité de la solution. En régression linéaire, si le nombre et les variables sont déterminés, la version "ridge" de la régression pénalise la norme carrée du vecteur des paramètres et restreint ainsi, par *régularisation*, l'espace des solutions pour limiter l'effet de la multicolinéarité.

Enfin, pour aborder en toute généralité les situations les plus compliquées, Vapnik (1999) a formalisé la théorie de l'apprentissage en introduisant une notion particulière de dimension pour toute famille de modèles.

4 Stratégie du *data mining*

4.1 Les données

Dans la majorité des problèmes rencontrés, des caractéristiques ou variables $X = (X^1, \ldots, X^p)$ dites explicatives ou prédictives ont été observées sur un ensemble de n objets, individus ou unités statistiques. Un premier travail, souvent fastidieux mais incontournable, consiste à mener une exploration statistique de ces données : allure des distributions, présence de données atypiques, corrélations et cohérence, transformations éventuelles des données, description multidimensionnelle, réduction de dimension, classification. C'est l'objet d'un cours distinct d'exploration statistique (Baccini et Besse 2000). La deuxième partie décrit les outils de modélisation statistique ou encore d'apprentissage utilisables pour la modélisation à fin de prévision d'une variable *cible* Y par les variables explicatives X^j.

L'enchaînement, éventuellement itératif, de ces étapes (exploration puis apprentissage) constitue le fondement de la fouille de données.

Pour comprendre la structure et bien appréhender le contenu de ce cours, il est important d'intégrer rapidement ce qu'est la stratégie à mettre en œuvre pour aboutir au bon *apprentissage* ou encore au bon *modèle prédictif* recherché à partir des données observées.

Attention, contrairement à une démarche statistique traditionnelle dans laquelle l'observation des données est intégrée à la méthodologie (planification de l'expérience), les données sont ici *préalables* à l'analyse. Néanmoins il est clair que les préoccupations liées à leur analyse et à son objectif doivent intervenir le plus en amont possible pour s'assurer quelques chances de succès.

4.2 Les étapes de l'apprentissage

Les traitements s'enchaînent de façon assez systématique selon le schéma suivant et quelque soit le domaine d'application :

i. Extraction des données avec ou sans échantillonnage faisant référence à des techniques de sondage appliquées ou applicables à des bases de données.

ii. Exploration des données pour la détection de valeurs aberrantes ou seulement atypiques, d'incohérences, pour l'étude des distributions des structures de corrélation, recherche de typologies, pour des transformations des données…

iii. Partition aléatoire de l'échantillon (apprentissage, validation, test) en fonction de sa taille et des techniques qui seront utilisées pour estimer une erreur de prévision en vue des étapes de choix de modèle, puis de choix et certification de méthode.

iv. Pour chacune des méthodes considérées : modèle linéaire général (gaussien, binomial ou poissonien), discrimination paramétrique (linéaire ou quadratique) ou non paramétrique, k plus proches voisins, arbre, réseau de neurones (perceptron), support vecteur machine, combinaison de modèles (bagging, boosting).
   - estimer le modèle pour une valeur donnée d'un paramètre de *complexité* : nombre de variables, de voisins, de feuilles, de neurones, durée de l'apprentissage, largeur de fenêtre… ;
   - optimiser ce paramètre (sauf pour les combinaisons de modèles affranchies des problèmes de surapprentissage) en fonction de la technique d'estimation de l'erreur retenue : échantillon de validation, validation croisée, approximation par pénalisation de l'erreur d'ajustement (critères C_p,

AIC).

v. Comparaison des modèles optimaux obtenus (un par méthode) par estimation de l'erreur de prévision sur l'échantillon test ou, si la présence d'un échantillon test est impossible, sur le critère de pénalisation de l'erreur (AIC d'Akaîke par exemple) s'il en existe une version pour chacune des méthodes considérées.

vi. Itération éventuelle de la démarche précédente (validation croisée), si l'échantillon test est trop réduit, depuis (iii). Partitions aléatoires successives de l'échantillon pour moyenner sur plusieurs cas l'estimation finale de l'erreur de prévision et s'assurer de la robustesse du modèle obtenu.

vii. Choix de la méthode retenue en fonction de ses capacités de prévision, de sa robustesse mais aussi, éventuellement, de l'interprétabillité du modèle obtenu.

viii. Ré-estimation du modèle avec la méthode, le modèles et sa complexitée optimisés à l'étape précédente sur l'ensemble des données.

ix. exploitation du modèle sue la base.

5 Exemples et jeux de données

En plus des exemples "pédagogiques" permettant d'illustre simplement les différentes méthodes étudiées, d'autres exemples en "vraie grandeur" permettent d'en évaluer réellement la pertinence mais aussi toute la complexité de mise en œuvre. D'autres exemples sont encore plus concrètement proposés en travaux dirigés avec leur traitement informatique.

5.1 Banque, finance, assurance : Marketing

L'objectif est une communication personnalisée et adaptée au mieux à chaque client. L'application la plus courante est la recherche d'un *score* estimé sur un échantillon de clientèle pour l'apprentissage puis extrapolé à l'ensemble en vu d'un objectif commercial :

- *Appétence* pour un nouveau produit financier : modélisation de la probabilité de posséder un bien (contrat d'assurance...) puis application à l'ensemble de la base. Les clients, pour lesquels le modèle prédit la possession de ce bien alors que ce n'est pas le cas, sont démarchés (télé marketing, publipostage ou mailing, phoning,...) prioritairement.
- *Attrition* ; même chose pour évaluer les risques de départ (churn) des clients par exemple chez un opérateur de téléphonie. Les clients pour lesquels le risque prédit est le plus important reçoivent des incitations à rester.
- *Risque* pour l'attribution d'un crédit bancaire ou l'ouverture de certains contrats d'assurance.
- ...

L'exemple traité reprend les données bancaires de Baccini et Besse 2000. Après la phase exploratoire, il s'agit de construire un score d'appétence de la carte Visa Premier dans l'idée de fidéliser les meilleurs clients. La variable à prédire est binaire : possession ou non de cette carte en fonction des avoirs et comportements bancaires décrits par 32 variables sur un millier de clients.

5.2 Environnement : pic d'ozone

L'objectif est de prévoir pour le lendemain les risques de dépassement de seuils de concentration d'ozone dans les agglomérations à partir de données observées : concentrations en 03, NO3, NO2... du jour, et d'autres prédites par Météo-France : température, vent... Encore une fois, le modèle apprend sur les dépassements observés afin de prédire ceux à venir.

Il s'agit d'un problème de régression : la variable à prédire est une concentration mais elle peut aussi être considérée comme binaire : dépassement ou non d'un seuil. Il y a 8 variables explicatives dont une est déjà une prévision de concentration d'ozone mais obtenue par un modèle déterministe de mécanique des fluides (équation de Navier et Stockes). L'approche statistique vient améliorer cette prévision en modélisant les erreurs et en tenant compte d'observations de concentration d'oxyde et dioxyde d'azote, de vapeur d'eau, de la prévision de la température ainsi que de la force du vent.

5.3 Santé : aide au diagnostic

Les outils statistiques sont largement utilisés dans le domaine de la santé. Ils le sont systématiquement lors des essais cliniques dans un cadre législatif stricte mais aussi lors d'études épidémiologiques pour la recherche de facteurs de risques dans des grandes bases de données ou encore pour l'aide au diagnostic. L'exemple étudié illustre ce dernier point : il s'agit de prévoir un diagnostic à partir de tests biologiques et d'examens élémentaires. Bien entendu, la variable à prédire, dont l'évaluation nécessite souvent une analyse très coûteuse voire une intervention chirurgicale, est connue sur l'échantillon nécessaire à l'estimation des modèles.

Dans l'exemple étudié (breast cancer), il s'agit de prévoir le type de la tumeur (bénigne, maligne) lors d'un cancer du sein à l'aide de 9 variables explicatives biologiques.

5.4 Biologie : sélection de gènes

Les techniques de microbiologie permettent de mesurer simultanément l'expression (la quantité d'ARN messager produite) de milliers de gènes dans des situations expérimentales différentes, par exemple entre des tissus sains et d'autres cancéreux. L'objectif est donc de déterminer quels gènes sont les plus susceptibles de participer aux réseaux de régulation mis en cause dans la pathologie ou autre phénomène étudié. Le problème s'énonce simplement mais révèle un redoutable niveau de complexité et pose de nouveaux défis au statisticien. En effet, contrairement aux cas précédents pour lesquels des centaines voire des milliers d'individus peuvent être observés et participer à l'apprentissage, dans le cas des biopuces, seuls quelques dizaines de tissus son analysés à cause essentiellement du prix et de la complexité d'une telle expérience. Compte tenu du nombres de gènes ou variables, le problème de discrimination est sévèrement indéterminé. D'autes approches, d'autres techniques sont nécessaires pour pallier à l'insuffisance des méthodes classiques de discrimination.

L'exemple reprend les données de Baccini et Besse (2000) concernant les différences d'expression des gènes en croisant deux facteurs lors d'une expérience de régime alimentaire (5 régimes) chez des souris (2 génotypes). La suite de l'étude conduit donc à rechercher les gènes expliquant au mieux les distinctions entre génotypes et aussi entre régimes.

5.5 Exemples industriels

Les exemples ci-dessous sont cités à titre illustratif mais leur complexité, inhérente à beaucoup de problèmes industriels, ne permet pas de les détailler à des fins pédagogiques.

Motorola : Détection de défaillance

Un procédé de fabrication de microprocesseurs comporte des centaines d'étapes (photogravures, dépôts, cuissons, polissages, lavages...) dont tous les paramètres, équipement et mesures physiques (températures, pressions...), sont enregistrés dans une grande base de données permettant la traçabilité des produits manufacturés. Le test électrique de chaque microprocesseur ne peut se faire qu'en fin de fabrication lorsque ceux-ci sont achevés. Il est évidement important de pouvoir déterminer, lors de l'apparition d'une baisse du rendement et en utilisant les données de la base, l'équipement ou la fourniture responsable de la défaillance afin d'y remédier le plus rapidement possible.

Airbus : Aide au pilotage

Les graphes de la figure 1.3 tracent les enregistrements des commandes et positions d'un avion en vol. Ceux-ci mettent en évidence un phénomène de résonance entre l'appareil et le comportement du pilote qui est très dangereux pour la sécurité. L'objectif est de construire un modèle susceptible, en temps réel, de détecter une telle situation afin d'y remédier par exemple en durcissant les commandes de vol électriques. Le problème est très spécifique car les données, ou signaux, sont mesurées en temps réel et constituent des discrétisations de courbes.

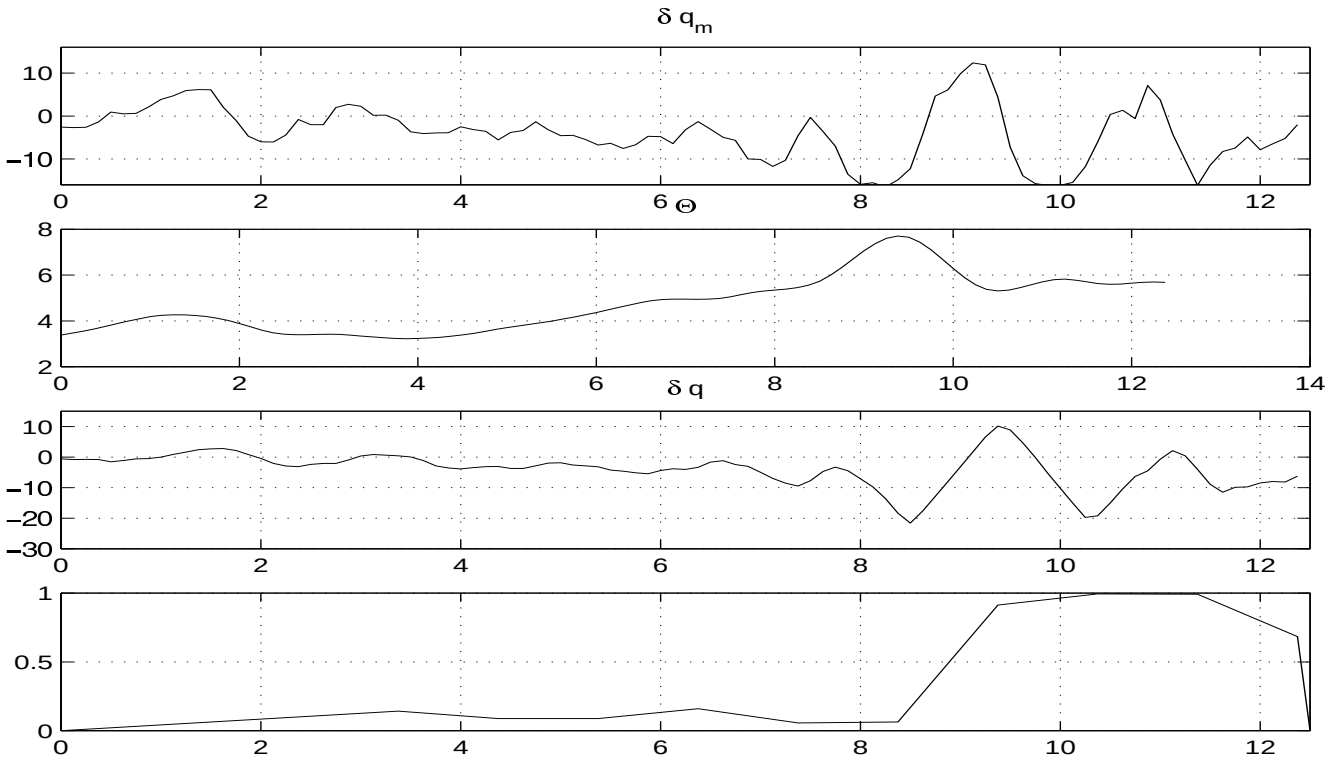

FIG. 1.3 – Airbus : Pompage piloté révélé par l'observation des paramètres en temps réel. De (haut en bas) : manche, assiette, gouverne comparer avec la prévision qu'en fait un réseau de neurones.

6 Contenu

Il a fallu faire des choix dans l'ensemble des techniques proposées et leurs nombreux avatars. La forme et le contenu sont guidés par les besoins exprimés lors des stages réalisées par les étudiants du département Génie Mathématique de l'INSA, du Master professionnel de Statistique & Économétrie ou encore par les thèmes des collaborations industrielles et scientifiques du laboratoire de Statistique et Probabilités[1]. Le lecteur peut se faire une idée du nombre très important de méthodes et variantes concernées par l'apprentissage supervisée ou non supervisé en consultant une boîte à outil Mathlab de classification[2]. Remarquons que les principaux logiciels commerciaux (SAS, Splus, SPSS, Matlab. . .) ou gratuits (R), performants et s'imposant par des interfaces très conviviales (Enterprise Miner, Insightfull Miner, Clementine), contribuent largement à la diffusion, voire la pénétration, de méthodes très sophistiquées dans des milieux imperméables à une conceptualisation mathématique trop abstraite.

Chaque méthode ou famille de méthodes de modélisation et d'apprentissage parmi les plus répandues, est présentée de façon plus ou moins succincte dans un chapitre distinct avec un objectif prédictif. La régression linéaire classique en statistique prend une place particulière à titre pédagogique. Très antérieure aux autres, elle donne lieu a une bibliographie abondante. Conceptuellement plus simple, elle permet d'introduire plus facilement les problématiques rencontrées comme celle du choix d'un modèle par ses deux approches types : la sélection de variable ou la régularisation (*ridge*). Pour une meilleure compréhension des logiciels qui y font largement référence, une introduction (annexe) au *modèle linéaire général* fournit le cadre théorique nécessaire à l'unification des régressions linéaire, loglinéaire et logistique ; cette dernière reste toujours très utilisée en scoring. La présentation de l'analyse discriminante décisionnelle, paramétrique ou non paramétrique, les k plus proches voisins, permet d'introduire également des notions de théorie bayésienne de la décision. Un chapitre incontournable est consacré aux techniques d'estimation d'une erreur de prévision sur lesquelles reposent les choix opérationnels décisifs : de modèle, de méthode mais aussi l'évaluation de la précision des résultats escomptés. Les chapitres suivants sont consacrées aux techniques algorithmiques : arbres binaires de décision (*classification and regression trees* ou CART) et à celles plus directement issues de la théorie de l'apprentissage machine (*machine learning*) : réseau de neurones et perceptron, agrégation de modèles (*boosting, random forest*), support vector machine (SVM). Enfin un chapitre conclusif propose une comparaison systématique des méthodes sur les différents jeux de données. Des annexes apportent des compléments théoriques ou méthodologiques : modélisation de données fonctionnelles, introduction au modèle linéaire général, bootstrap.

Le choix a été fait de conserver et expliciter, dans la mesure du possible, les concepts originaux de chaque méthode dans son cadre disciplinaire tout en tâchant d'homogénéiser notations et terminologies. L'objectif principal est de faciliter la compréhension et l'interprétation des techniques des principaux logiciels pour en faciliter une *utilisation pertinente et réfléchie*. Ce cours ne peut être dissocié de séances de travaux dirigés sur ordinateur à l'aide de logiciels (SAS, R...) pour traiter des données en vraie grandeur dans toute leur complexité.

[1] http ://www.lsp.ups-tlse.fr
[2] http ://tiger.technion.ac.il/ eladyt/classification/

Chapitre 2

Régression linéaire

1 Introduction

Ce chapitre ne propose qu'une introduction au modèle gaussien, à sa définition et à son estimation en privilégiant l'objectif de prévision. Il s'attarde donc sur le problème délicat du choix de modèle afin, principalement, d'en introduire et d'en illustrer les grands principes dans le cas relativement simple d'un modèle linéaire. Une section introduit le modèle d'analyse de covariance mais de nombreux aspects : colinéarité, points influents, tests, analyse de variance, modèle multinomial ou poissonien (modèle log-linéaire)... sont négligés et à rechercher dans la bibliographie de même qu'une présentation globale du *modèle linéaire général* incluant toutes ces approches et seulement résumée en annexe. Les statistiques des tests élémetaires sont explicitées afin de faciliter la lectures et l'interprétation des résultats issus des logiciels.

Le but premier de ce chapitre est donc l'explication ou plutôt, la modélisation dans un but prédictif, d'une variable *quantitative* par plusieurs variables quantitatives (régression linéaire multiple) ou par un mélange de variables quantitatives et qualitatives (analyse de covariance).

2 Modèle

Le modèle de régression linéaire multiple est l'outil statistique le plus habituellement mis en œuvre pour l'étude de données multidimensionnelles. Cas particulier de modèle linéaire, il constitue la généralisation naturelle de la régression simple.

Une variable quantitative Y dite *à expliquer* (ou encore, réponse, exogène, dépendante) est mise en relation avec p variables quantitatives $X^1, \ldots, X^p$ dites *explicatives* (ou encore de contrôle, endogènes, indépendantes, régresseurs).

Les données sont supposées provenir de l'observation d'un échantillon statistique de taille n ($n > p+1$) de $\mathbb{R}^{(p+1)}$:

$$(x_i^1, \ldots, x_i^j, \ldots, x_i^p, y_i) \quad i = 1, \ldots, n.$$

L'écriture du *modèle linéaire* dans cette situation conduit à supposer que l'espérance de Y appartient au sous-espace de $\mathbb{R}^n$ engendré par $\{\mathbf{1}, X^1, \ldots, X^p\}$ où $\mathbf{1}$ désigne le vecteur de $\mathbb{R}^n$ constitué de "1" . C'est-à-dire que les $(p+1)$ variables aléatoires vérifient :

$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i \quad i = 1, 2, \ldots, n$$

avec les hypothèses suivantes :

i. Les ε_i sont des termes d'erreur indépendants et identiquement distribués ; $E(\varepsilon_i) = 0, Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$.

ii. Les termes x^j sont supposés déterministes (facteurs contrôlés) **ou bien** l'erreur $\boldsymbol{\varepsilon}$ est indépendante de la distribution conjointe de $X^1, \ldots, X^p$. On écrit dans ce dernier cas que :

$$E(Y|X^1, \ldots, X^p) = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \cdots + \beta_p X^p \text{ et } Var(Y|X^1, \ldots, X^p) = \sigma^2.$$

iii. Les paramètres inconnus $\beta_0, \ldots, \beta_p$ sont supposés constants.

iv. En option, pour l'étude spécifique des lois des estimateurs, une quatrième hypothèse considère la normalité de la variable d'erreur $\boldsymbol{\varepsilon}$ ($\mathcal{N}(0, \sigma^2\mathbf{I})$). Les ε_i sont alors i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

Les données sont rangées dans une matrice $\mathbf{X}(n \times (p+1))$ de terme général x_i^j, dont la première colonne contient le vecteur $\mathbf{1}$ ($x_0^i = 1$), et dans un vecteur $\mathbf{Y}$ de terme général y_i. En notant les vecteurs $\boldsymbol{\varepsilon} = [\varepsilon_1 \cdots \varepsilon_p]'$ et $\boldsymbol{\beta} = [\beta_0 \beta_1 \cdots \beta_p]'$, le modèle s'écrit matriciellement :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

3 Estimation

Conditionnellement à la connaissance des valeurs des X^j, les paramètres inconnus du modèle : le vecteur β et σ^2 (paramètre de nuisance), sont estimés par minimisation des carrés des écarts (M.C.) ou encore, en supposant (iv), par maximisation de la vraisemblance (M.V.). Les estimateurs ont alors les mêmes expressions, l'hypothèse de normalité et l'utilisation de la vraisemblance conférant à ces derniers des propriétés complémentaires.

3.1 Estimation par M.C.

L'expression à minimiser sur $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ s'écrit :

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i^1 - \beta_2 x_i^2 - \cdots - \beta_p x_i^p)^2 &= \|\mathbf{y} - \mathbf{X}\beta\|^2 \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.\end{aligned}$$

Par dérivation matricielle de la dernière équation on obtient les *"équations normales"* :

$$\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = 0$$

dont la solution correspond bien à un minimum car la matrice hessienne $2\mathbf{X}'\mathbf{X}$ est semi définie-positive.

Nous faisons l'hypothèse supplémentaire que la matrice $\mathbf{X}'\mathbf{X}$ est inversible, c'est-à-dire que la matrice $\mathbf{X}$ est de rang $(p+1)$ et donc qu'il n'existe pas de colinéarité entre ses colonnes. En pratique, si cette hypothèse n'est pas vérifiée, il suffit de supprimer des colonnes de $\mathbf{X}$ et donc des variables du modèle. Des diagnostics de colinéarité et des critères aident au choix des variables.

Alors, l'estimation des paramètres β_j est donnée par :

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

et les valeurs ajustées (ou estimées, prédites) de $\mathbf{y}$ ont pour expression :

$$\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

où $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ est appelée "*hat matrix*" ; elle met un chapeau à $\mathbf{y}$. Géométriquement, c'est la matrice de projection orthogonale dans $\mathbb{R}^n$ sur le sous-espace Vect($\mathbf{X}$) engendré par les vecteurs colonnes de $\mathbf{X}$.

On note

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

le vecteur des résidus ; c'est la projection de $\mathbf{y}$ sur le sous-espace orthogonal de Vect($\mathbf{X}$) dans $\mathbb{R}^n$.

3.2 Propriétés

Les estimateurs des M.C. $b_0, b_1, \ldots, b_p$ sont des estimateurs sans biais : $E(\mathbf{b}) = \boldsymbol{\beta}$, et, parmi les estimateurs sans biais fonctions linéaires des y_i, ils sont de variance minimum (théorème de Gauss-Markov) ; ils sont donc "BLUE" : *best linear unbiaised estimators*. Sous hypothèse de normalité, les estimateurs du M.V. sont uniformément meilleurs (efficaces) et coïncident avec ceux des M.C.

On montre que la matrice de covariance des estimateurs se met sous la forme

$$E[(\mathbf{b}-\boldsymbol{\beta})(\mathbf{b}-\boldsymbol{\beta})'] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1},$$

celle des prédicteurs est

$$E[(\widehat{\mathbf{y}}-\mathbf{X}\boldsymbol{\beta})(\widehat{\mathbf{y}}-\mathbf{X}\boldsymbol{\beta})'] = \sigma^2\mathbf{H}$$

et celle des estimateurs des résidus est

$$E[(\mathbf{e}-\mathbf{u})((\mathbf{e}-\mathbf{u}))'] = \sigma^2(\mathbf{I}-\mathbf{H})$$

tandis qu'un estimateur sans biais de σ^2 est fourni par :

$$s^2 = \frac{\|\mathbf{e}\|^2}{n-p-1} = \frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2}{n-p-1} = \frac{\text{SSE}}{n-p-1}.$$

Ainsi, les termes $s^2 h_i^i$ sont des estimations des variances des prédicteurs $\widehat{y}_i$.

3.3 Sommes des carrés

SSE est la somme des carrés des résidus (*sum of squared errors*),

$$\text{SSE} = \|\mathbf{y}-\widehat{\mathbf{y}}\|^2 = \|\mathbf{e}\|^2.$$

On définit également la somme totale des carrés (*total sum of squares*) par

$$\text{SST} = \|\mathbf{y}-\bar{y}\mathbf{1}\|^2 = \mathbf{y}'\mathbf{y} - n\bar{y}^2$$

et la somme des carrés de la régression (*regression sum of squares*) par

$$\text{SSR} = \|\widehat{\mathbf{y}}-\bar{y}\mathbf{1}\|^2 = \widehat{\mathbf{y}}'\widehat{\mathbf{y}} - n\bar{y}^2 = \mathbf{y}'\mathbf{H}\mathbf{y} - n\bar{y}^2 = \mathbf{b}'\mathbf{X}'\mathbf{y} - n\bar{y}^2.$$

On vérifie alors : SST = SSR + SSE.

3.4 Coefficient de détermination

On appelle *coefficient de détermination* le rapport

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

qui est donc la part de variation de Y expliquée par le modèle de régression. Géométriquement, c'est un rapport de carrés de longueur de deux vecteurs. C'est donc le cosinus carré de l'angle entre ces vecteurs : $\mathbf{y}$ et sa projection $\widehat{\mathbf{y}}$ sur Vect$(\mathbf{X})$.

Attention, dans le cas extrême où $n = (p+1)$, c'est-à-dire si le nombre de variables explicatives est grand comparativement au nombre d'observations, $R^2 = 1$. Ou encore, il est géométriquement facile de voir que l'ajout de variables explicatives ne peut que faire croître le coefficient de détermination.

La quantité R est appelée *coefficient de corrélation multiple* entre Y et les variables explicatives, c'est le coefficient de corrélation usuel entre $\mathbf{y}$ et sa prévision (ou projection) $\widehat{\mathbf{y}}$.

4 Inférences dans le cas gaussien

En principe, l'hypothèse optionnelle (iv) de normalité des erreurs est nécessaire pour cette section. En pratique, des résultats asymptotiques, donc valides pour de grands échantillons, ainsi que des études de simulation, montrent que cette hypothèse n'est pas celle dont la violation est la plus pénalisante pour la fiabilité des modèles.

4.1 Inférence sur les coefficients

Pour chaque coefficient β_j on montre que la statistique

$$\frac{b_j - \beta_j}{\sigma_{b_j}}$$

où $\sigma^2_{b_j}$, variance de b_j est le $j$ème terme diagonal de la matrice $s^2(\mathbf{X}'\mathbf{X})^{-1}$, suit une loi de Student à $(n-p-1)$ degrés de liberté. Cette statistique est donc utilisée pour tester une hypothèse H_0 : $\beta_j = a$ ou pour construire un intervalle de confiance de niveau $100(1-\alpha)\%$:

$$b_j \pm t_{\alpha/2;(n-p-1)}\sigma_{b_j}.$$

Attention, cette statistique concerne un coefficient et ne permet pas d'inférer conjointement (cf. §3.4) sur d'autres coefficients car ils sont corrélés entre eux ; de plus elle dépend des absences ou présences des autres variables X^k dans le modèle. Par exemple, dans le cas particulier de deux variables X^1 et X^2 très corrélées, chaque variable, en l'absence de l'autre, peut apparaître avec un coefficient significativement différent de 0 ; mais, si les deux sont présentes dans le modèle, elles peuvent chacune apparaître avec des coefficients insignifiants.

De façon plus générale, si $\mathbf{c}$ désigne un vecteur non nul de $(p+1)$ constantes réelles, il est possible de tester la valeur d'une combinaison linéaire $\mathbf{c}'\mathbf{b}$ des paramètres en considérant l'hypothèse nulle H_0 : $\mathbf{c}'\mathbf{b} = a$; a connu. Sous H_0, la statistique

$$\frac{\mathbf{c}'\mathbf{b} - a}{(s^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c})^{1/2}}$$

suit une loi de Student à $(n-p-1)$ degrés de liberté.

4.2 Inférence sur le modèle

Le modèle peut être testé globalement. Sous l'hypothèse nulle H_0 : $\beta_1 = \beta_2 = \ldots = \beta_p = 0$, la statistique

$$\frac{\text{SSR}/p}{\text{SSE}/(n-p-1)} = \frac{\text{MSR}}{\text{MSE}}$$

suit une loi de Fisher avec p et $(n-p-1)$ degrés de liberté. Les résultats sont habituellement présentés dans un tableau *"d'analyse de la variance"* sous la forme suivante :

Source de variation	*d.d.l.*	*Somme des carrés*	*Variance*	*F*
Régression	p	SSR	MSR=SSR/p	MSR/MSE
Erreur	$n-p-1$	SSE	MSE=SSE/$(n-p-1)$	
Total	$n-1$	SST		

4.3 Inférence sur un modèle réduit

Le test précédent amène à rejeter H_0 dès que l'une des variables X^j est liée à Y. Il est donc d'un intérêt limité. Il est souvent plus utile de tester un modèle réduit c'est-à-dire dans lequel certains coefficients, à l'exception de la constante, sont nuls contre le modèle complet avec toutes les variables. En ayant éventuellement réordonné les variables, on considère l'hypothèse nulle H_0 : $\beta_1 = \beta_2 = \ldots = \beta_q = 0, q < p$.

Notons respectivement SSR_q, SSE_q, R^2_q les sommes de carrés et le coefficient de détermination du modèle réduit à $(p-q)$ variables. Sous H_0, la statistique

$$\frac{(\text{SSR} - \text{SSR}_q)/q}{\text{SSE}/(n-p-1)} = \frac{(R^2 - R^2_q)/q}{(1-R^2)/(n-p-1)}$$

suit une loi de Fisher à q et $(n-p-1)$ degrés de liberté.

Dans le cas particulier où $q = 1$ ($\beta_j = 0$), la F-statistique est alors le carré de la t-statistique de l'inférence sur un paramètre et conduit donc au même test.

4.4 Prévision

Connaissant les valeurs des variables X^j pour une nouvelle observation : $\mathbf{x}_0' = [x_0^1, x_0^2, \ldots, x_0^p]$ appartenant au domaine dans lequel l'hypothèse de linéarité reste valide, une prévision, notée $\widehat{y}_0$ de Y ou $E(Y)$ est donnée par :

$$\widehat{y}_0 = b_0 + b_1 x_0^1 + \cdots + b_p x_0^p.$$

Les intervalles de confiance des prévisions de Y et $E(Y)$, pour *une* valeur $\mathbf{x}_0 \in \mathbb{R}^p$ et en posant $\mathbf{v}_0 = (1|\mathbf{x}_0')' \in \mathbb{R}^{p+1}$, sont respectivement

$$\begin{aligned} \widehat{y}_0 &\pm t_{\alpha/2;(n-p-1)}\, s(1 + \mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2}, \\ \widehat{y}_0 &\pm t_{\alpha/2;(n-p-1)}\, s(\mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2}. \end{aligned}$$

4.5 Exemple

Le modèle de régression linéaire n'est pas adapté à l'explication d'une variable binaire comme dans le cas des données bancaires. Ceci est abordé dans le chapitre suivant en utilisant la régression logistique tandis que d'autres exemples de données sont utilisées dans ce chapitre. Les premières sont extraites de Jobson (1991) et décrivent les résultats comptables de 40 entreprises du Royaume Uni.

RETCAP	Return on capital employed
WCFTDT	Ratio of working capital flow to total debt
LOGSALE	Log to base 10 of total sales
LOGASST	Log to base 10 of total assets
CURRAT	Current ratio
QUIKRAT	Quick ratio
NFATAST	Ratio of net fixed assets to total assets
FATTOT	Gross sixed assets to total assets
PAYOUT	Payout ratio
WCFTCL	Ratio of working capital flow to total current liabilities
GEARRAT	Gearing ratio (debt-equity ratio)
CAPINT	Capital intensity (ratio of total sales to total assets)
INVTAST	Ratio of total inventories to total assets

Modèle complet

La procédure SAS/REG est utilisée dans le programme suivant. Beaucoup d'options sont actives afin de fournir la plupart des résultats même si certains sont redondants ou peu utiles.

```
options linesize=110 pagesize=30 nodate nonumber;
title;
proc reg data=sasuser.ukcomp1 all;
 model RETCAP = WCFTCL   WCFTDT   GEARRAT   LOGSALE   LOGASST
       NFATAST  CAPINT   FATTOT   INVTAST   PAYOUT   QUIKRAT   CURRAT
       /dw covb  Influence cli clm tol vif collin R P;
output out=resout h=lev p=pred r=res student=resstu ;
run;
```

```
Analysis of Variance
                          Sum of          Mean
Source          DF       Squares        Square         F Value       Prob>F
                (1)
Model           12      0.55868 (2)     0.04656 (5)    8.408 (7)     0.0001 (8)
Error           27      0.14951 (3)     0.00554 (6)
```

```
C Total          39      0.70820 (4)
  Root MSE        0.07441 (9)    R-square      0.7889 (12)
  Dep Mean        0.14275 (10)    Adj R-sq     0.6951 (13)
  C.V.           52.12940 (11)
```

(1)	degrés de liberté de la loi de Fisher du test global
(2)	SSR
(3)	SSE ou déviance
(4)	SST=SSE+SSR
(5)	SSR/DF
(6)	s^2 =MSE=SSE/DF est l'estimation de σ^2
(7)	Statistique F du test de Fisher du modèle global
(8)	$P(f_{p;n-p-1} > F)$; H_0 est rejetée au niveau α si $P < \alpha$
(9)	s =racine de MSE
(10)	moyenne empirique de la variable à expliquée
(11)	Coefficient de variation $100\times$ (9)/(10)
(12)	Coefficient de détermination R^2
(13)	Coefficient de détermination ajusté R'^2

```
Parameter Estimates
                 Parameter      Standard    T for H0:                          Variance
Variable  DF      Estimate         Error   Parameter=0 Prob>|T|   Tolerance    Inflation
                     (1)           (2)         (3)        (4)          (5)          (6)
INTERCEP   1      0.188072    0.13391661       1.404     0.1716    .            0.00000000
WCFTCL     1      0.215130    0.19788455       1.087     0.2866    0.03734409  26.77799793
WCFTDT     1      0.305557    0.29736579       1.028     0.3133    0.02187972  45.70441500
GEARRAT    1     -0.040436    0.07677092      -0.527     0.6027    0.45778579   2.18442778
LOGSALE    1      0.118440    0.03611612       3.279     0.0029    0.10629382   9.40788501
LOGASST    1     -0.076960    0.04517414      -1.704     0.0999    0.21200778   4.71680805
...
```

(1)	estimations des paramètres (b_j)
(2)	écarts-types de ces estimations (s_{b_j})
(3)	statistique T du test de Student de H_0 : $b_j = 0$
(4)	$P(t_{n-p-1} > T)$; H_0 est rejetée au niveau α si $P < \alpha$
(5)	$1 - R^2_{(j)}$
(6)	VIF=$1/(1 - R^2_{(j)})$

Ces résultats soulignent les problèmes de colinéarités. De grands "VIF" sont associés à de grands écart-types des estimations des paramètres. D'autre part les nombreux tests de Student non significatifs montrent que trop de variables sont présentes dans le modèle. Cette idée est renforcée par le calcul de l'indice de conditionnement (explicité dans la section suivante : 8.76623/0.00125).

5 Choix de modèle

De façon un peu schématique, on peut associer la pratique de la modélisation statistique à trois objectifs qui peuvent éventuellement être poursuivis en complémentarité.

Descriptif : Il vise à rechercher de façon exploratoire les liaisons entre Y et d'autres variables, potentiellement explicatives, X^j qui peuvent être nombreuses afin, par exemple d'en sélectionner un sous-ensemble. À cette stratégie, à laquelle peuvent contribuer des Analyses en Composantes Principales, correspond des algorithmes de recherche (pas à pas) moins performants mais économiques en temps de calcul si p est grand.

Attention, si n est petit, et la recherche suffisamment longue avec beaucoup de variables explicatives, il sera toujours possible de trouver un "bon" modèle expliquant y ; c'est l'effet *data mining* dans les modèles économétriques appelé maintenant *data snooping*.

Explicatif : Le deuxième objectif est sous-tendu par une connaissance *a priori* du domaine concerné et dont des résultats théoriques peuvent vouloir être confirmés, infirmés ou précisés par l'estimation des paramètres. Dans ce cas, les résultats inférentiels précédents permettent de construire *le* bon test

conduisant à la prise de décision recherchée. Utilisées hors de ce contexte, les statistiques de test n'ont plus alors qu'une valeur indicative au même titre que d'autres critères plus empiriques.

Prédictif : Dans le troisième cas, l'accent est mis sur la qualité des estimateurs et des prédicteurs qui doivent, par exemple, minimiser une erreur quadratique moyenne. C'est la situation rencontrée en *apprentissage*. Ceci conduit à rechercher des modèles *parcimonieux* c'est-à-dire avec un nombre volontairement restreint de variables explicatives. Le "meilleur" modèle ainsi obtenu peut donner des estimateurs légèrement biaisés au profit d'un compromis pour une variance plus faible. Un bon modèle n'est donc plus celui qui explique le mieux les données au sens d'une déviance (SSE) minimale (ou d'un R^2 max) au prix d'un nombre important de variables pouvant introduire des colinéarités. Le bon modèle est celui qui conduit aux prévisions les plus fiables.

Certes, le théorème de Gauss-Markov indique que, parmi les estimateurs sans biais, celui des moindres carrés est de variance minimum. Néanmoins, il peut être important de préférer un estimateur légèrement biaisé si le gain en variance est lui plus significatif. C'est tout le problème de trouver un bon équilibre entre biais et variance afin de minimiser un risque quadratique de prévision. Il y a principalement deux façons de "biaiser" un modèle dans le but de restreindre la variance :

- en réduisant le nombre de variables explicatives et donc en simplifiant le modèle,
- en contraignant les paramètres du modèle, en les rétrécissant (*schrinkage*), en régression *ridge* qui opère une régularisation.

Commençons par décrire les procédures de sélection.

5.1 Critères

De nombreux critères de choix de modèle sont présentés dans la littérature sur la régression linéaire multiple. Citons le critère d'information d'Akaïke (`AIC`), celui bayésien de Sawa (`BIC`)... (cf. chapitre 5). Ils sont équivalents lorsque le nombre de variables à sélectionner, ou niveau du modèle, est fixé. Le choix du critère est déterminant lorsqu'il s'agit de comparer des modèles de niveaux différents. Certains critères se ramènent, dans le cas gaussien, à l'utilisation d'une expression pénalisée de la fonction de vraisemblance afin de favoriser des modèles parcimonieux. En pratique, les plus utilisés ou ceux généralement fournis par les logiciels sont les suivants.

Statistique du F de Fisher

Ce critère, justifié dans le cas explicatif car basé sur une qualité d'ajustement est aussi utilisé à titre indicatif pour comparer des séquences de modèles emboîtés. La statistique partielle de Fisher est

$$\frac{(\text{SSR} - \text{SSR}_q)/s}{\text{SSE}/(n-p-1)} = \frac{(R^2 - R_q^2)}{1-R^2)} \frac{n-p-1}{q}$$

dans laquelle l'indice q désigne les expressions concernant le modèle réduit avec $(p-q)$ variables explicatives. On considère alors que si l'accroissement $(R^2 - R_q^2)$ est suffisamment grand :

$$R^2 - R_R^2 > \frac{q}{(n-p-1)} F_{\alpha;q,(n-p-1)},$$

l'ajout des q variables au modèle est justifié.

R^2 et R^2 ajusté

Le coefficient de détermination $R^2 = 1 - \text{SSE/SST}$, directement lié à la déviance (SSE) est aussi un indice de qualité mais qui a la propriété d'être monotone croissant en fonction du nombre de variables. Il ne peut donc servir qu'à comparer deux modèles de même niveau c'est-à-dire avec le même nombre de variables.

En revanche, le R^2 *ajusté* :

$$R'^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}.$$

dans lequel le rapport SSE/SST est remplacé par un rapport des estimations sans biais des quantités σ^2 et σ_y^2 introduit une pénalisation liée au nombre de paramètres à estimer.

Ce coefficient s'exprime encore par

$$1 - \frac{(n-1)\text{MSE}}{\text{SST}}$$

ainsi dans la comparaison de deux modèles partageant la même SST, on observe que $R'^2 > R'^2_j$ si et seulement si MSE<MSE$_j$; MSE et MSE$_j$ désignant respectivement l'erreur quadratique moyenne du modèle complet et celle d'un modèle à j variables explicatives. Maximiser le R^2 ajusté revient donc à minimiser l'erreur quadratique moyenne.

C_p de Mallows

Cet indicateur est une estimation de l'erreur quadratique moyenne de prévision qui s'écrit aussi comme la somme d'une variance et du carré d'un biais. L'erreur quadratique moyenne de prévision sécrit ainsi : :

$$\text{MSE}(\widehat{y}_i) = \text{Var}(\widehat{y}_i) + [\text{Biais}(\widehat{y}_i)]^2$$

puis après sommation et réduction :

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\text{MSE}(\widehat{y}_i) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\text{Var}(\widehat{y}_i) + \frac{1}{\sigma^2}\sum_{i=1}^{n}[\text{Biais}(\widehat{y}_i)]^2.$$

En supposant que les estimations du modèle complet sont sans biais et en utilisant des estimateurs de $Var(\widehat{y}_i)$ et σ^2, l'expression de l'erreur quadratique moyenne totale standardisée (ou réduite) pour un modèle à j variables explicatives s'écrit :

$$C_p = (n-q-1)\frac{\text{MSE}_j}{\text{MSE}} - [n-2(q+1)]$$

et définit la valeur du C_p de Mallows pour les q variables considérées. Il est alors d'usage de rechercher un modèle qui minimise le C_p tout en fournissant une valeur inférieure et proche de $(q+1)$. Ceci revient à considérer que le "vrai" modèle complet est moins fiable qu'un modèle réduit donc biaisé mais d'estimation plus précise.

Akaïke's Information criterion *(AIC)*

A compléter

PRESS de Allen

Il s'agit l'introduction historique de la validation croisée. On désigne par $\widehat{y}_{(i)}$ la prévision de y_i calculée sans tenir compte de la $i$ème observation $(y_i, x_i^1, \ldots, x_i^p)$, la somme des erreurs quadratiques de prévision (PRESS) est définie par

$$\text{PRESS} = \sum_{i=1}^{n}(y_i - \widehat{y}_{(i)})^2$$

et permet de comparer les capacités prédictives de deux modèles. Le chapitre 5 donne plus de détails sur ce type d'estimation.

5.2 Algorithmes de sélection

Lorsque p est grand, il n'est pas raisonnable de penser explorer les 2^p modèles possibles afin de sélectionner le "meilleur" au sens de l'un des critères ci-dessus. Différentes stratégies sont donc proposées qui doivent être choisies en fonction de l'objectif recherché et des moyens de calcul disponibles ! Trois types d'algorithmes sont résumés ci-dessous par ordre croissant de temps de calcul nécessaire c'est-à-dire par nombre croissant de modèles considérés parmi les 2^p et donc par capacité croissante d'optimalité. On donne pour chaque algorithme l'option `selection` à utiliser dans la procédure `REG` de SAS.

Pas à pas

Sélection (`forward`) À chaque pas, une variable est ajoutée au modèle. C'est celle dont la valeur p ("prob value")associée à la statistique partielle du test de Fisher qui compare les deux modèles est minimum.

La procédure s'arrête lorsque toutes les variables sont introduites ou lorsque p reste plus grande qu'une valeur seuil fixée par défaut à $0,50$.

Élimination (`backward`) L'algorithme démarre cette fois du modèle complet. À chaque étape, la variable associée à la plus grande valeur p est éliminée du modèle. La procédure s'arrête lorsque les variables restant dans le modèle ont des valeurs p plus petites qu'un seuil fixé par défaut à $0,10$.

Mixte (`stepwise`) Cet algorithme introduit une étape d'élimination de variable après chaque étape de sélection afin de retirer du modèle d'éventuels variables qui seraient devenues moins indispensables du fait de la présence de celles nouvellement introduites.

Global

L'algorithme de Furnival et Wilson est utilisé pour comparer tous les modèles possibles en cherchant à optimiser l'un des critères : R^2, R^2 ajusté, ou C_p de Mallows (`rsquare, adjrsq, cp`) choisi par l'utilisateur. Par souci d'économie, cet algorithme évite de considérer des modèles de certaines sous-branches de l'arborescence dont on peut savoir a priori qu'ils ne sont pas compétitifs. En général les logiciels exécutant cet algorithme affichent le (`best=1`) ou les meilleurs modèles de chaque niveau.

5.3 Exemple

Parmi les trois types d'algorithmes et les différents critères de choix, une des façons les plus efficaces consistent à choisir les options du programme ci-dessous. Tous les modèles (parmi les plus intéressants selon l'algorithme de Furnival et Wilson) sont considérés. Seul le meilleur pour chaque niveau, c'est-à-dire pour chaque valeur p du nombre de variables explicatives sont donnés. Il est alors facile de choisir celui minimisant l'un des critères globaux (C_p ou BIC ou ...).

```
options linesize=110 pagesize=30 nodate nonumber;
title;
proc reg data=sasuser.ukcomp2 ;
model RETCAP = WCFTCL   WCFTDT   GEARRAT   LOGSALE   LOGASST
       NFATAST  CAPINT   FATTOT   INVTAST   PAYOUT   QUIKRAT   CURRAT
       / selection=rsquare cp rsquare bic best=1;
run;
```

```
   N = 40     Regression Models for Dependent Variable: RETCAP
  R-sq. Adjust. C(p)    BIC    Variables in Model
In       R-sq
 1 0.105 0.081 78.393 -163.2 WCFTCL
 2 0.340 0.305 50.323 -173.7 WCFTDT QUIKRAT
 3 0.615 0.583 17.181 -191.1 WCFTCL NFATAST CURRAT
 4 0.720 0.688  5.714 -199.2 WCFTDT LOGSALE NFATAST CURRAT
 5 0.731 0.692  6.304 -198.0 WCFTDT LOGSALE NFATAST QUIKRAT CURRAT
 6 0.748 0.702  6.187 -197.2 WCFTDT LOGSALE NFATAST INVTAST QUIKRAT CURRAT
 7 0.760 0.707  6.691 -195.7 WCFTDT LOGSALE LOGASST NFATAST FATTOT QUIKRAT CURRAT
 8 0.769 0.709  7.507 -193.8 WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST QUIKRAT CURRAT
 9 0.776 0.708  8.641 -191.5 WCFTCL WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST QUIKRAT
                             CURRAT
10 0.783 0.708  9.744 -189.1 WCFTCL WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST PAYOUT
                             QUIKRAT CURRAT
11 0.786 0.702 11.277 -186.4 WCFTCL WCFTDT LOGSALE LOGASST NFATAST CAPINT FATTOT INVTAST
                             PAYOUT QUIKRAT CURRAT
12 0.788 0.695 13.000 -183.5 WCFTCL WCFTDT GEARRAT LOGSALE LOGASST NFATAST CAPINT FATTOT
                             INVTAST PAYOUT QUIKRAT CURRAT
```

Dans cet example, C_p et BIC se comportent de la même façon. Avec peu de variables, le modèle est trop biaisé. Ils atteignent un minimum pour un modèle à 4 variables explicatives puis croissent de nouveau selon la première bissectrice. La maximisation du R^2 ajusté conduirait à une solution beaucoup moins parcimonieuse. On note par ailleurs que l'algorithme remplace WCFTCL par WCFTDT. Un algorithme par sélection ne peut pas aboutir à la solution optimale retenue.

5.4 Choix de modèle par régularisation

L'autre stratégie qui cherche à conserver l'ensemble ou tout du moins la plupart des variables explicatives pose un problème de *multicolinéarité*. Il est résolu par une procédure de régularisation.

Problème

L'estimation des paramètres ainsi que celle de leur écart-type (standard error) nécessite le calcul explicite de la matrice $(\mathbf{X}'\mathbf{X})^{-1}$. Dans le cas dit *mal conditionné* où le déterminant de la matrice $\mathbf{X}'\mathbf{X}$ n'est que légèrement différent de 0, les résultats conduiront à des estimateurs de variances importantes et même, éventuellement, à des problèmes de précision numérique. Il s'agit donc de diagnostiquer ces situations critiques puis d'y remédier. Dans les cas descriptif ou prédictif on supprime des variables à l'aide des procédures de choix de modèle mais, pour un objectif explicatif nécessitant toutes les variables, d'autres solutions doivent être envisagées : algorithme de résolution des équations normales par transformations orthogonales (procédure `orthoreg` de SAS) sans calcul explicite de l'inverse pour limiter les problèmes numériques, régression biaisée (ridge), régression sur composantes principales.

VIF

La plupart des logiciels proposent des diagnostics de colinéarité. Le plus classique est le *facteur d'inflation de la variance* (VIF)

$$V_j = \frac{1}{1 - R_j^2}$$

où R_j^2 désigne le coefficient de détermination de la régression de la variable X^j sur les autres variables explicatives ; R_j est alors un coefficient de corrélation multiple, c'est le cosinus de l'angle dans $\mathbb{R}^n$ entre X^j et le sous-espace vectoriel engendré par les variables $\{X^1, \ldots, X^{j-1}, X^{j+1}, \ldots, X^p\}$. Plus X^j est "linéairement" proche de ces variables et plus R_j est proche de 1 ; on montre alors que la variance de l'estimateur de β_j est d'autant plus élevée. Évidemment, cette variance est minimum lorsque X^j est orthogonal au sous-espace engendré par les autres variables.

Conditionnement

De façon classique, les qualités numériques de l'inversion d'une matrice sont quantifiées par son *indice de conditionnement*. On note $\lambda_1, \ldots, \lambda_p$ les valeurs propres de la matrice des corrélations $\mathbf{R}$ rangées par ordre décroissant. Le déterminant de $\mathbf{R}$ est égal au produit des valeurs propres. Ainsi, des problèmes numériques, ou de variances excessives apparaissent dès que les dernières valeurs propres sont relativement trop petites. L'*indice de conditionnement* est le rapport

$$\kappa = \lambda_1 / \lambda_p$$

de la plus grande sur la plus petite valeur propre.

En pratique, si $\kappa < 100$ on considère qu'il n'y a pas de problème. Celui-ci devient sévère pour $\kappa > 1000$. Cet indice de conditionnement donne un aperçu global des problèmes de colinéarité tandis que les VIF, les tolérances ou encore l'étude des vecteurs propres associés au plus petites valeurs propres permettent d'identifier les variables les plus problématiques.

Régression ridge

Ayant diagnostiqué un problème mal conditionné mais désirant conserver toutes les variables, il est possible d'améliorer les propriétés numériques et la variance des estimations en considérant un estimateur légèrement biaisé des paramètres. L'estimateur "*ridge*" est donné par

$$\mathbf{b}_R = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y},$$

qui a pour effet de décaler de la valeur k toutes les valeurs propres de la matrice à inverser et, plus particulièrement, les plus petites qui reflètent la colinéarité. On montre que cela revient encore à estimer le modèle par les moindres carrés sous la contrainte que la norme du vecteur[1] $\boldsymbol{\beta}$ des paramètres ne soit pas

[1] En pratique, la contrainte ne s'applique pas au terme constant β_0 mais seulement aux coefficients du modèle.

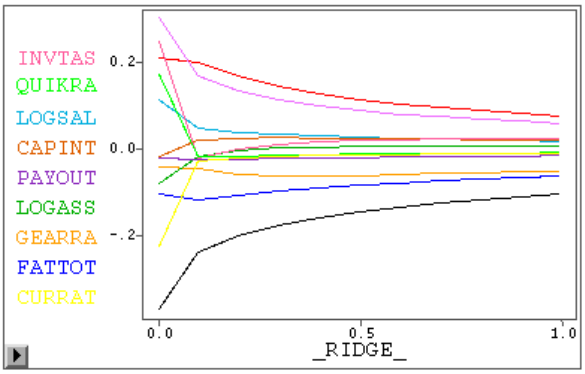

FIG. 2.1 – *Retour sur capital : Evolution des paramètres de la régression* ridge *en fonction du paramètre de régularisation.*

trop grande :

$$\mathbf{b}_R = \arg\min_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \ ; \ \|\boldsymbol{\beta}\|^2 < c \right\}.$$

C'est encore, en introduisant un multiplicateur de Lagrange dans le problème de minimisation, un problème de moindres carrés pénalisés :

$$\mathbf{b}_R = \arg\min_{\beta} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \}.$$

Cela revient à pénaliser la norme de l'estimateur pour empêcher les coefficients d'exploser et donc pour limiter la variance. On parle aussi d'estimateur à rétrécisseur (*shrinkage*). Comme dans tout problème de régularisation, il est nécessaire de fixer la valeur du paramètre λ ; la validation croisée peut être utilisée à cette fin mais la lecture du graphique (cf. figure 2.1) montrant l'évolution des paramètres en fonction du coefficient ridge est souvent suffisante. La valeur est choisie au point où la décroissance des paramètres devient faible et quasi-linéaire. Une autre version (*lasso*) de régression biaisée est obtenue en utilisant la norme en valeur absolue pour définir la contrainte sur les paramètres.

Régression sur composantes principales

L'Analyse en Composantes Principales est, entre autres, la recherche de p variables dites principales qui sont des combinaisons linéaires des variables initiales de variance maximale sous une contrainte d'orthogonalité (cf. Baccini et Besse (2000) pour des détails). En désignant par $\mathbf{V}$ la matrice des vecteurs propres de la matrice des corrélations $\mathbf{R}$ rangés dans l'ordre décroissant des valeurs propres, les valeurs prises par ces variables principales sont obtenues dans la matrice des composantes principales

$$\mathbf{C} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{V}.$$

Elles ont chacune pour variance la valeur propre λ_j associée. Le sous-espace engendré par ces variables principales est le même que celui engendré par les variables initiales. Il est donc géométriquement équivalent de régresser Y sur les colonnes de $\mathbf{C}$ que sur celles de $\mathbf{X}$. Les problèmes de colinéarité sont alors résolus en supprimant les variables principales de plus faibles variances c'est-à-dire associées aux plus petites valeurs propres ou encore en exécutant un algorithme de choix de modèle sur les composantes.

La solution obtenue présente ainsi de meilleures qualités prédictives mais, les coefficients de la régression s'appliquant aux composantes principales, un calcul complémentaire est nécessaire afin d'évaluer et d'interpréter les effets de chacune des variables initiales.

Régression PLS

Une dernière approche est largement utilisée, par exemple en chimiométrie, afin de pourvoir traiter les situations présentant une forte multicolinéarité et même, lorsque le nombre d'observations est inférieur au nombre de prédicteurs. Il s'agit de la régression PLS (*partial least square*).

Comme pour la régression sur composantes principales, le principe est de rechercher un modèle de régression linéaire sur un ensemble de composantes orthogonales contruites à partir de combinaisons linéaires des variables explicatives centrées. Dans le cas de la PLS, la construction des composantes est optimisée pour que celles-ci soient le plus liées à la variable Y à prédire au sens de la covariance empirique, alors que les composantes principales ne visent qu'à extraire une part de variance maximale sans tenir compte d'une variable cible.

Soit $\mathbf{X}(n \times p)$ la matrice des prédicteurs centrés avec n pouvant être inférieur à p. On cherche une matrice $\mathbf{W}$ de coefficients ou pondérations définissant les q composantes T_k par combinaisons linéaires des variables X_j :

$$\mathbf{T} = \mathbf{XW}.$$

La matrice $\mathbf{W}$ est solution du problème suivant :

$$\begin{aligned} \text{Pour} \quad k = 1, \ldots, q, \quad \mathbf{w}_k &= \arg\max_{\mathbf{w}} \mathrm{Cov}(Y, T_k)^2 \\ &= \arg\max_{\mathbf{w}} \mathbf{w}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w} \\ \text{Avec} \quad \mathbf{w}_k'\mathbf{w_k} = 1 \quad \text{et} \quad \mathbf{t}_k'\mathbf{t}_k &= \mathbf{w}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w} = 0, \quad \text{pour} \quad \ell = 1 \ldots, k-1. \end{aligned}$$

La matrice $\mathbf{W}$ est obtenue par la démarche itérative de l'algorithme 1 ; il suffit ensuite de calculer la régression de Y sur les q variables T_k centrées ainsi construites. Le choix du nombre de composanteq q est optimisé par validation croisée.

Cet algorithme se généralise directement à une variable explicative multidimensionnelle (SIMPLS). Le critère à optimiser devient une somme des carrés des covariances entre une composante et chacune des variables réponse. Plusieurs variantes de la régression PLS multidimensionnelles ont été proposés (NIPALS, Kernel-PLS...) ; le même critère est optimisé mais sous des contraintes différentes.

Algorithm 1 régression PLS

$\mathbf{X}$ matrice des variables explicatives centrées,
Calcul de $\mathbf{W}$ matrice des coefficients.
Pour $k = 1$ à q **Faire**
  $\mathbf{w}_k = \frac{\mathbf{X'Y}}{\|\boldsymbol{X'Y}\|}$,
  $\mathbf{t}_k = \mathbf{X}\mathbf{w}_k$
  Déflation de $\mathbf{X}$: $\mathbf{X} = \mathbf{X} - \mathbf{t}_k\mathbf{t}_k'\mathbf{X}$
Fin Pour

6 Compléments

6.1 Modèles polynomiaux

En cas d'invalidation de l'hypothèse de linéarité, il peut être intéressant de considérer des modèles polynômiaux, très classiques pour décrire des phénomènes physiques, de la forme

$$Y = \beta_0 + \cdots + \beta_j X^j + \cdots + \gamma_{kl} X^k X^l + \cdots + \delta_j X^{j2}$$

qui sont encore appelés *surfaces de réponse* en plannification expérimentale. Ces modèles sont faciles à étudier dans le cadre linéaire, il suffit d'ajouter des nouvelles variables constituées des produits ou des carrés des variables explicatives initiales. Les choix : présence ou non d'une interaction entre deux variables, présence ou non d'un terme quadratique se traitent alors avec les mêmes outils que ceux des choix de variable mais en intégrant une contrainte lors de la lecture des résultats : ne pas considérer des modèles incluant des termes quadratiques dont les composants linéaires auraient été exclus ou encore, ne pas supprimer d'un modèle une variable d'un effet linéaire si elle intervient dans un terme quadratique.

La procédure `rsreg` de SAS est plus particulièrement adaptée aux modèles quadratiques. Elle ne comporte pas de procédure de choix de modèle mais fournit des aides et diagnostics sur l'ajustement de la surface ainsi que sur la recherche des points optimaux.

Attention : Ce type de modèle accroît considérablement les risques de colinéarité, il est peu recommandé de considérer des termes cubiques.

6.2 Influence, résidus, validation

Avant toute tentative de modélisation complexe, il est impératif d'avoir conduit des analyses uni et bivariées afin d'identifier des problèmes sur les distributions de chacune des variables : dissymétrie, valeurs atypiques (outliers) ou sur les liaisons des variables prises deux par deux : non-linéarité. Ces préliminaires acquis, des aides ou diagnostics associés à la régression linéaire multiple permettent de détecter des violations d'hypothèses (homoscédasticité, linéarité) ou des points influents dans ce contexte multidimensionnel (cf. figure 2.2).

Points influents

Comme toute méthode quadratique, l'estimation des paramètres est très sensible à la présence de points extrêmes susceptibles de perturber gravement les résultats. Une observation est *influente* sur les paramètres d'une régression si, à la fois,

- elle est éloignée du barycentre, et ce dans la direction d'un vecteur propre associé à une petite valeur propre (effet levier),
- elle provoque un grand résidu.

L'observation de la diagonale de la matrice $\mathbf{H}$ (hat matrix) révèle un effet levier potentiel tandis que l'analyse des résidus *studentisés* pointe ceux susceptibles de poser des problèmes (valeur absolue plus grande que 2).

Les deux diagnostics précédents sont combinés dans des mesures synthétiques proposées par différents auteurs. La plus utilisée est la distance de *Cook*

$$D_i = \frac{1}{s^2(p+1)}(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(i)})'(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(i)}) = \left[\frac{h_i^i}{1-h_i^i}\right]\frac{r_i^2}{(p+1)}$$

qui quantifie l'influence de la i-ème observation sur l'écart entre le prédicteur $\widehat{\mathbf{y}}$ et le prédicteur $\widehat{\mathbf{y}}_{(i)}$ calculé sans cette $i$ème observation. On conclut à une influence de l'observation i lorsque la valeur de D_i dépasse 1.

Tous ces critères sont illustrés dans les graphiques de la figure 2.2. Les tableaux ci-dessous fournis pas SAS illustrent ces quantités sur l'exemple des données comptables.

Obs	Dep Var RETCAP (1)	Predict Value (2)	Std Err Predict (3)	Lower95 Mean (4)	Upper95 Mean (5)	Lower95 Predict (6)	Upper95 Predict (7)	Residual (8)	Std Err Residual (9)	Student Residual (10)
1	0.2600	0.2716	0.053	0.1625	0.3808	0.0839	0.4593	-0.0116	0.052	-0.223
2	0.5700	0.3690	0.039	0.2882	0.4497	0.1962	0.5417	0.2010	0.063	3.183
3	0.0900	0.00897	0.063	-0.1205	0.1385	-0.1912	0.2092	0.0810	0.039	2.055
4	0.3200	0.2335	0.021	0.1903	0.2768	0.0748	0.3922	0.0865	0.071	1.212
5	0.1700	0.1164	0.046	0.0215	0.2113	-0.0634	0.2961	0.0536	0.058	0.920

...

Obs	-2-1-0 1 2 (11)	Cook's D (12)	Rstudent (13)	Hat Diag H (14)	Cov Ratio (15)	Dffits (15)	INTERCEP Dfbetas (15)	WCFTCL Dfbetas (15)	WCFTDT Dfbetas (15)
1	| |	0.004	-0.2194	0.5109	3.2603	-0.2242	0.0299	0.0632	-0.0911
2	| |******	0.302	3.9515	0.2795	0.0050	2.4611	0.9316	-0.3621	0.3705
3	| |****	0.832	2.1955	0.7192	0.6375	3.5134	0.5543	2.1916	-2.0241
4	| |**	0.010	1.2228	0.0803	0.8585	0.3613	-0.0132	-0.0835	0.1207
5	| |*	0.041	0.9175	0.3864	1.7591	0.7280	-0.0386	0.0906	0.0060

...

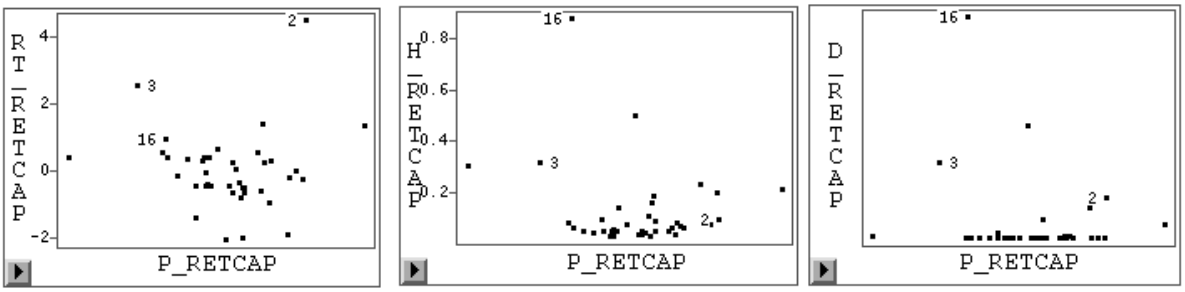

FIG. 2.2 – *Retour sur capital : Graphe des résidus studentisés, de la diagonale de la matrice* $\mathbf{H}$ *et de la distance de Cook en fonction des valeurs prédites.*

(1)	variable à expliquer y_i
(2)	valeur ajustée $\widehat{y}_i$
(3)	écart-type de cette estimation$s_{\widehat{y}_i}$
(4)et (5)	Intervalle de confiance pour l'estimation de $E(y_i)$
(6) et (7)	Intervalle de confiance pour l'estimation de y_i
(8)	résidus calculés e_i
(9)	écarts-types de ces estimations
(10)	résidus standardisés (ou studentisés internes) r_i
(11)	repérage graphique des résidus standardisés : $* = 0.5$.
(12)	Distance de Cook
(13)	résidus studentisés (externes) t_i
(14)	Termes diagonaux de la matrice chapeau $\mathbf{H}$
(15)	autres indicateurs d'influence

```
Sum of Residuals                     0
Sum of Squared Residuals        0.1495   (SSE)
Predicted Resid SS (Press)      1.0190   (PRESS)
```

Régression partielle

Un modèle de régression multiple est une technique *linéaire*. Il est raisonnable de s'interroger sur la pertinence du caractère linéaire de la contribution d'une variable explicative à l'ajustement du modèle. Ceci peut être réalisé en considérant une *régression partielle*.

On calcule alors deux régressions :

- la régression de Y sur les variables $X^1, \ldots, X^{j-1}, X^{j+1}, \ldots, X^p$, dans laquelle la $j$ème variable est omise, soit $\mathbf{r}_{y(j)}$ le vecteur des résidus obtenus.
- La régression de X^j sur les variables $X^1, \ldots, X^{j-1}, X^{j+1}, \ldots, X^p$. Soit $\mathbf{r}_{x(j)}$ le vecteur des résidus obtenus.

La comparaison des résidus par un graphe (nuage de points $\mathbf{r}_{y(j)} \times \mathbf{r}_{x(j)}$) permet alors de représenter la nature de la liaison entre X^j et Y *conditionnellement* aux autres variables explicatives du modèle.

Graphes

Différents graphiques permettent finalement de contrôler le bien fondé des hypothèses de linéarité, d'homoscédasticité, éventuellement de normalité des résidus.

- Le premier considère le nuage de points des résidus studentisés croisés avec les valeurs prédites. Les points doivent être uniformément répartis entre les bornes -2 et $+2$ et ne pas présenter de formes suspectes (cf. figure 2.2).
- Le deuxième croise les valeurs observées de Y avec les valeurs prédites. Il illustre le coefficient de détermination R qui est aussi la corrélation linéaire simple entre $\widehat{\mathbf{y}}$ et $\mathbf{y}$. Les points doivent s'aligner

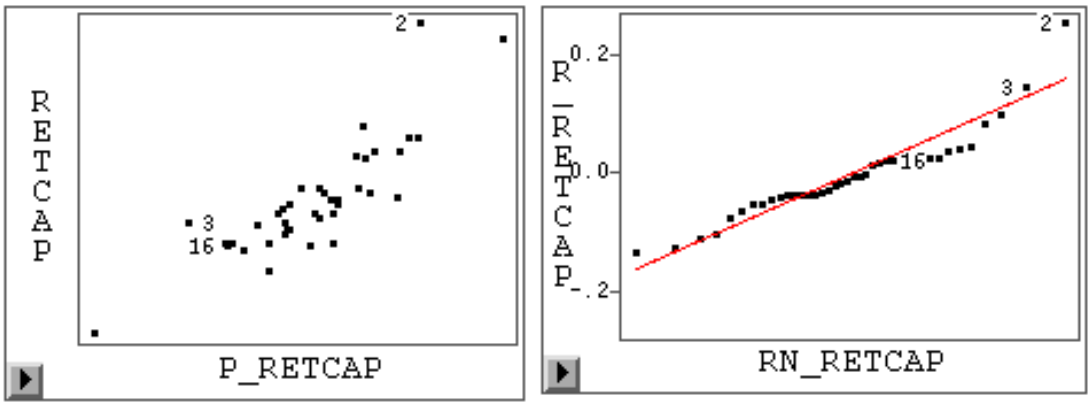

FIG. 2.3 – *Retour sur capital : Graphe des valeurs observées en fonction des valeurs prédites et droite de Henri des résidus (normal qq-plot).*

autour de la première bissectrice. Il peut être complété par l'intervalle de confiance des y_i ou celui de leurs moyennes. (cf. figure 2.3).

- La qualité, en terme de linéarité, de l'apport de chaque variable est étudiée par des régressions partielles. Chaque graphe de résidus peut être complété par une estimation fonctionnelle ou régression non-paramétrique (loess, noyau, spline) afin d'en facilité la lecture.
- Le dernier trace la droite de Henri (Normal QQplot) des résidus dont le caractère linéaire de la représentation donne une idée de la normalité de la distribution. (cf. figure 2.3)

7 Analyse de variance à un facteur

7.1 Introduction

Les techniques dites d'*analyse de variance* sont des outils entrant dans le cadre général du modèle linéaire et où une variable quantitative est expliquée par une ou plusieurs variables qualitatives. L'objectif essentiel est alors de comparer les moyennes empiriques de la variable quantitative observées pour différentes catégories d'unités statistiques. Ces catégories sont définies par l'observation des variables qualitatives ou *facteurs* prenant différentes modalités ou encore de variables quantitatives découpées en classes ou *niveaux*. Une combinaison de niveaux définit une *cellule*, *groupe* ou *traitement*.

Il s'agit donc de savoir si un facteur ou une combinaison de facteurs (*interaction*) a un *effet* sur la variable quantitative en vue, par exemple, de déterminer des conditions optimales de production ou de fabrication, une dose optimale de médicaments.... Ces techniques apparaissent aussi comme des cas particuliers de la régression linéaire multiple en associant à chaque modalité une *variable indicatrice* (dummy variable) et en cherchant à expliquer une variable quantitative par ces variables indicatrices. L'appellation "analyse de variance" vient de ce que les tests statistiques sont bâtis sur des comparaisons de sommes de carrés de variations.

L'analyse de variance est souvent utilisée pour analyser des données issue d'une *planification expérimentale* au cours de laquelle l'expérimentateur a la possibilité de contrôler *a priori* les niveaux des facteurs avec pour objectif d'obtenir le maximum de précision au moindre coût. Ceci conduit en particulier à construire des facteurs orthogonaux deux à deux (variables explicatives non linéairement corrélées) afin de minimiser la variance des estimateurs. On distingue le cas particulier important où les cellules ont le même effectif, on parle alors de *plan orthogonal* ou *équirépété* ou *équilibré* (balanced), qui conduit à des simplifications importantes de l'analyse de variance associée. On appelle plan *complet* un dispositif dans lequel toutes les combinaisons de niveaux ont été expérimentées. On distingue entre des modèles fixes, aléatoires ou mixtes selon le caractère déterministe (contrôlé) ou non des facteurs par exemple si les modalités résultent d'un

choix aléatoire parmi un grand nombre de possibles. Dans cette courte introduction seuls le modèle fixe à un facteur est considéré.

L'analyse de variance à un facteur est un cas particulier d'étude de relations entre deux variables statistiques : une quantitative Y admettant une densité et une qualitative X ou facteur qui engendre une partition ou classification de l'échantillon en J groupes, cellules ou classes indicées par j. L'objectif est de comparer les distributions de Y pour chacune des classes en particulier les valeurs des moyennes et variances. Un préalable descriptif consiste à réaliser un graphique constitué de diagrammes boites parallèles : une pour chaque modalité. Cette représentation donne une première appréciation de la comparaison des distributions (moyenne, variance) internes à chaque groupe. Les spécificités de la planification d'expérience ne sont pas abordées dans ce cours axé sur la fouille de données pour laquelle les données sont justement préalablement fournies. Les plans d'expérience sont surtout utilisés en milieu industriel : contrôle de qualité, optimisation des processus de production, ou en agronomie pour la sélection de variétés, la comparaison d'engrais, d'insecticides.... La bibliographie est abondante sur ce sujet.

7.2 Modèle

Pour chaque niveau j de X, on observe n_j valeurs $y_{1j}, \ldots, y_{n_j j}$ de la variable Y et où $n = \sum_{j=1}^{J} n_j$ ($n > J$) est la taille de l'échantillon. On suppose qu'à l'intérieur de chaque cellule, les observations sont indépendantes équidistribuées de moyenne μ_j et de variance *homogène* $\sigma_j^2 = \sigma^2$. Ceci s'écrit :

$$y_{ij} = \mu_j + \varepsilon_{ij}$$

où les ε_{ij} sont i.i.d. suivant une loi centrée de variance σ^2 qui sera supposée $\mathcal{N}(0, \sigma^2)$ pour la construction des tests. Cette dernière hypothèse n'étant pas la plus sensible. Les espérances μ_j ainsi que le paramètre de nuisance σ^2 sont les paramètres inconnus à estimer.

On note respectivement :

$$\begin{aligned}
\bar{y}_{.j} &= \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}, \\
s_j^2 &= \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2, \\
\bar{y}_{..} &= \frac{1}{n} \sum_{i=1}^{n_j} \sum_{j=1}^{J} y_{ij},
\end{aligned}$$

les moyennes et variances empiriques de chaque cellule, la moyenne générale de l'échantillon.

Les paramètres μ_j sont estimés sans biais par les moyennes $\bar{y}_{.j}$ et comme le modèle s'écrit alors :

$$y_{ij} = \bar{y}_{.j} + (y_{ij} - \bar{y}_{.j}),$$

l'estimation des erreurs est $e_{ij} = (y_{ij} - \bar{y}_{.j})$ tandis que les valeurs prédites sont $\widehat{y}_{ij} = \bar{y}_{.j}$.

Sous l'hypothèse d'homogénéité des variances, la meilleure estimation sans biais de σ^2 est

$$s^2 = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2}{n - J} = \frac{1}{n - J}[(n-1)s_1^2 + \cdots + (n_J - 1)s_J^2]$$

qui s'écrit donc comme une moyenne pondérée des variances empiriques de chaque groupe.

Notons $\mathbf{y}$ le vecteur des observations $[y_{ij} | i = 1, n_j; j = 1, J]'$ mis en colonne, $\boldsymbol{\varepsilon} = [\varepsilon_{ij} | i = 1, n_j; j = 1, J]'$ le vecteur des erreurs, $\mathbf{1}_j$ les variables indicatrices des niveaux et $\mathbf{1}$ la colonne de 1s. Le $i$ème élément d'une variable indicatrice (dummy variable) $\mathbf{1}_j$ prend la valeur 1 si la $i$ème observation y_i est associée au $j$ème et 0 sinon.

Comme dans le cas de la régression linéaire multiple, le modèle consiste à écrire que l'espérance de la variable Y appartient au sous-espace linéaire engendré par les variables explicatives, ici les variables indicatrices :

$$\mathbf{y} = \beta_0 \mathbf{1} + \beta_1 \mathbf{1}_1 + \cdots + \beta_J \mathbf{1}_J + \boldsymbol{\varepsilon}.$$

La matrice $\mathbf{X}$ alors construite n'est pas de plein rang $p+1$ mais de rang p. La matrice $\mathbf{X}'\mathbf{X}$ n'est pas inversible et le modèle admet une infinité de solutions. Nous disons que les paramètres β_j ne sont pas *estimables* ou identifiables. En revanche, certaines fonctions (combinaisons linéaires) de ces paramètres sont estimables et appelées *contrastes*.

Dans le cas du modèle d'analyse de variance à *un* facteur, la solution la plus simple adoptée consiste à considérer un sous-ensemble des indicatrices ou de combinaisons des indicatrices engendrant le même sous-espace de façon à aboutir à une matrice inversible. Ceci conduit à considérer différents modèles associés à différentes *paramétrisation*. *Attention*, les paramètres β_j ainsi que la matrice $\mathbf{X}$ prennent à chaque fois des significations différentes.

Un premier modèle (cell means model) s'écrit comme celui d'une régression linéaire multiple sans terme constant avec $\boldsymbol{\beta} = [\mu_1, \ldots, \mu_J]'$ le vecteur des paramètres :

$$\begin{aligned} \mathbf{y} &= \beta_1 \mathbf{1}_1 + \cdots + \beta_J \mathbf{1}_J + \boldsymbol{\varepsilon} \\ \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \end{aligned}$$

Les calculs se présentent simplement mais les tests découlant de ce modèle conduiraient à étudier la nullité des paramètres alors que nous sommes intéressés par tester l'égalité des moyennes.

Une autre paramétrisation, considérant cette fois le vecteur $\boldsymbol{\beta} = [\mu_J, \mu_1 - \mu_J, \ldots, \mu_{J-1} - \mu_J]'$ conduit à écrire le modèle (base cell model) de régression avec terme constant :

$$\mathbf{y} = \beta_0 \mathbf{1} + \beta_1 \mathbf{1}_1 + \cdots + \beta_{J-1} \mathbf{1}_{J-1} + \boldsymbol{\varepsilon}.$$

C'est celle de SAS alors que d'autres logiciels considèrent des paramètres d'effet différentiel $\mu_j - \mu_.$ par rapport à l'effet moyen $\mu_. = 1/J \sum_{j=1}^{J} \mu_j$. Ce dernier est encore un modèle (group effect model) de régression linéaire avec terme constant mais dont les variables explicatives sont des différences d'indicatrices et avec $\boldsymbol{\beta} = [\mu_., \mu_1 - \mu_., \ldots, \mu_{J-1} - \mu_.]'$:

$$\mathbf{y} = \beta_0 \mathbf{1} + \beta_1 (\mathbf{1}_1 - \mathbf{1}_J) + \cdots + \beta_{J-1} (\mathbf{1}_{J-1} - \mathbf{1}_J) + \boldsymbol{\varepsilon}.$$

7.3 Test

On désigne les différentes sommes des carrés des variations par :

$$\begin{aligned} \text{SST} &= \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{..})^2 = \sum_{j=1}^{J} \sum_{i=1}^{n_j} y_{ij}^2 - n\bar{y}_{..}^2, \\ \text{SSW} &= \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2 = \sum_{j=1}^{J} \sum_{i=1}^{n_j} y_{ij}^2 - \sum_{j=1}^{J} n_j \bar{y}_{.j}^2, \\ \text{SSB} &= \sum_{j=1}^{J} n_j (\bar{y}_{.j} - \bar{y}_{..})^2 = \sum_{j=1}^{J} n_j \bar{y}_{.j}^2 - n\bar{y}_{..}^2, \end{aligned}$$

où "T" signifie totale, "W" (within) intra ou résiduelle, "B" (between) inter ou expliquée par la partition. Il est facile de vérifier que SST=SSB+SSW.

On considère alors l'hypothèse

$$H_0 \ : \mu_1 = \cdots = \mu_J,$$

qui revient à dire que la moyenne est indépendante du niveau ou encore que le facteur n'a pas d'effet, contre l'hypothèse

$$H_1 \ : \exists (j,k) \text{ tel que } \mu_j \neq \mu_k$$

qui revient à reconnaître un effet ou une influence du facteur sur la variable Y.

Dans les modèles précédents, l'étude de cette hypothèse revient à comparer par un test de Fisher un modèle complet (les moyennes sont différentes) avec un modèle réduit supposant la nullité des paramètres β_j et donc l'égalité des moyennes à celle de la dernière cellule ou à la moyenne générale.

Les résultats nécessaires à la construction du test qui en découle sont résumés dans la table d'analyse de la variance :

Source de variation	*d.d.l.*	*Somme des carrés*	*Variance*	F
Modèle (inter)	$J-1$	SSB	MSB=SSB/$(J-1)$	MSB/MSW
Erreur (intra)	$n-J$	SSW	MSW=SSW/$(n-J)$	
Total	$n-1$	SST		

Pratiquement, un programme de régression usuel permet de construire estimation et test de la nullité des β_j sauf pour le premier modèle qui doit tester l'égalité au lieu de la nullité des paramètres.

Dans le cas de deux classes $(J = 2)$ on retrouve un test équivalent au test de Student de comparaison des moyennes de deux échantillons indépendants. Si l'hypothèse nulle est rejetée, la question suivante consiste à rechercher quelles sont les groupes ou cellules qui possèdent des moyennes significativement différentes. De nombreux tests et procédures ont été proposés dans la littérature pour répondre à cette question. Enfin, l'hypothèse importante du modèle induit par l'analyse de variance est l'homogénéité des variances de chaque groupe. Conjointement à l'estimation du modèle et en supposant la normalité, il peut être instructif de contrôler cette homogénéité par un test.

8 Analyse de covariance

L'*analyse de covariance* se situe encore dans le cadre général du modèle linéaire et où une variable quantitative est expliquée par plusieurs variables à la fois quantitatives et qualitatives. Dans les cas les plus complexes, ont peut avoir plusieurs facteurs (variables qualitatives) avec une structure croisée ou hiérarchique ainsi que plusieurs variables quantitatives intervenant de manière linéaire ou polynômiale. Le principe général, dans un but explicatif ou décisionnel, est toujours d'estimer des modèles "*intra-groupes*" et de faire apparaître (tester) des effets différentiels "*inter-groupes*" des paramètres des régressions. Ainsi, dans le cas plus simple où seulement une variable parmi les explicatives est quantitative, nous sommes amenés à tester l'hétérogénéité des constantes et celle des pentes (interaction) entre différents modèles de régression linéaire.

Ce type de modèle permet donc, toujours avec un objectif prédictif, de s'intéresser à la modélisation d'une variable quantitative par un ensemble de variables explicatives à la fois quantitatives et qualitatives. La possible prise en compte d'interactions complique singulièrement la procédure de sélection de variables.

8.1 Modèle

Le modèle est explicité dans le cas élémentaire où une variable quantitative Y est expliquée par une variable qualitative T à J niveaux et une variable quantitative, appelée encore covariable, X. Pour chaque niveau j de T, on observe n_j valeurs $x_{1j}, \ldots, x_{n_j j}$ de X et n_j valeurs $y_{1j}, \ldots, y_{n_j j}$ de Y ; $n = \sum_{j=1}^{J} n_j$ est la taille de l'échantillon.

En pratique, avant de lancer une procédure de modélisation et tests, une démarche exploratoire s'appuyant sur une représentation en couleur (une par modalité j de T) du nuage de points croisant Y et X et associant les droites de régression permet de se faire une idée sur les effets respectifs des variables : parallélisme des droites, étirement, imbrication des sous-nuages.

On suppose que les moyennes conditionnelles $E[Y|T]$, c'est-à-dire calculées à l'intérieur de chaque cellule, sont dans le sous-espace vectoriel engendré par les variables explicatives quantitatives, ici X. Ceci s'écrit :

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij} \; ; \quad j = 1, \ldots, J; \quad i = 1, \cdots, n_j$$

où les ε_{ij} sont i.i.d. suivant une loi centrée de variance σ^2 qui sera supposée $\mathcal{N}(0, \sigma^2)$ pour la construction des tests.

Notons $\mathbf{y}$ le vecteur des observations $[y_{ij}|i = 1, n_j; j = 1, J]'$ mis en colonne, $\mathbf{x}$ le vecteur $[x_{ij}|i = 1, n_j; j = 1, J]'$, $\boldsymbol{\varepsilon} = [\varepsilon_{ij}|i = 1, n_j; j = 1, J]'$ le vecteur des erreurs, $\mathbf{1}_j$ les variables indicatrices des niveaux et $\mathbf{1}$ la colonne de 1s. On note encore $\mathbf{x}.\mathbf{1}_j$ le produit terme à terme des deux vecteurs, c'est-à-dire le vecteur contenant les observations de $\mathbf{X}$ sur les individus prenant le niveau j de T et des zéros ailleurs.

La résolution simultanée des J modèles de régression est alors obtenue en considérant globalement le

modèle :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

dans lequel $\mathbf{X}$ est la matrice $n \times 2J$ constituée des blocs $[\mathbf{1}_j | \mathbf{x}.\mathbf{1}_j]$; $j = 1, \ldots, J$. L'estimation de ce modèle global conduit, par bloc, à estimer les modèles de régression dans chacune des cellules.

Comme pour l'analyse de variance, les logiciels opèrent une reparamétrisation faisant apparaître des effets différentiels par rapport au dernier niveau (SAS/GLM, SAS/INSIGHT) ou par rapport à un effet moyen (Systat), afin d'obtenir directement les bonnes hypothèses dans les tests. Ainsi, dans le premier cas, on considère la matrice de même rang (sans la Jème indicatrice)

$$\mathbf{X} = [\mathbf{1} | \mathbf{x} | \mathbf{1}_1 | \cdots | \mathbf{1}_{J-1} | \mathbf{x}.\mathbf{1}_1 | \cdots | \mathbf{x}.\mathbf{1}_{J-1}]$$

associée aux modèles :

$$y_{ij} = \beta_{0J} + (\beta_{0j} - \beta_{0J}) + \beta_{1J} x_{ij} + (\beta_{1j} - \beta_{1J}) x_{ij} + \varepsilon_{ij} \; ; \quad j = 1, \ldots, J-1; i = 1, \ldots, n_j.$$

8.2 Tests

Différentes hypothèses sont alors testées en comparant le modèle complet

$$\begin{aligned} \mathbf{y} = \beta_{0J}\mathbf{1} \quad &+ \quad (\beta_{01} - \beta_{0J})\mathbf{1}_1 + \cdots + (\beta_{0J-1} - \beta_{0J})\mathbf{1}_{J-1} + \beta_{1J}\mathbf{x} + \\ &+ \quad (\beta_{11} - \beta_{1J})\mathbf{x}.\mathbf{1}_1 + \cdots + (\beta_{1J-1} - \beta_{1J})\mathbf{x}.\mathbf{1}_{J-1} + \boldsymbol{\varepsilon} \end{aligned}$$

à chacun des modèles réduits :

$$\begin{aligned} (i) \qquad & \mathbf{y} = \beta_{0J}\mathbf{1} + (\beta_{01} - \beta_{0J})\mathbf{1}_1 + \cdots + (\beta_{0J-1} - \beta_{0J})\mathbf{1}_{J-1} + \beta_{1J}\mathbf{x} + \boldsymbol{\varepsilon} \\ (ii) \qquad & \mathbf{y} = \beta_{0J}\mathbf{1} + (\beta_{01} - \beta_{0J})\mathbf{1}_1 + \cdots + (\beta_{0J-1} - \beta_{0J})\mathbf{1}_{J-1} + \\ & + (\beta_{1j} - \beta_{1J})\mathbf{x}.\mathbf{1}_1 + \cdots + (\beta_{1J-1} - \beta_{1J})\mathbf{x}.\mathbf{1}_{J-1} + \boldsymbol{\varepsilon} \\ (iii) \qquad & \mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + (\beta_{1j} - \beta_{1J})\mathbf{x}.\mathbf{1}_1 + \cdots + (\beta_{1J-1} - \beta_{1J})\mathbf{x}.\mathbf{1}_{J-1} + \boldsymbol{\varepsilon} \end{aligned}$$

par un test de Fisher. Ceci revient à considérer les hypothèses suivantes :

- H_0^i : pas d'interaction, $\beta_{11} = \cdots = \beta_{1J}$, les droites partagent la même pente β_{1J},
- H_0^{ii} : β_{1J}=0,
- H_0^{iii} :$\beta_{01} = \cdots = \beta_{0J}$, les droites partagent la même constante à l'origine β_{0J}.

On commence donc par évaluer i), si le test n'est pas significatif, on regarde ii) qui, s'il n'est pas non plus significatif, conduit à l'absence d'effet de la variable X. De même, toujours si i) n'est pas significatif, on s'intéresse à iii) pour juger de l'effet du facteur T.

8.3 Choix de modèle

Ce cadre théorique et les outils informatiques (SAS/GLM) permettent de considérer des modèles beaucoup plus complexes incluant plusieurs facteurs, plusieurs variables quantitatives, voire des polynômes de celles-ci, ainsi que les diverses interactions entre qualitatives et quantitatives. Le choix du "bon" modèle devient vite complexe d'autant que la stratégie dépend, comme pour la régression linéaire multiple, de l'objectif visé :

descriptif : des outils multidimensionnels descriptifs (ACP, AFD, AFCM...) s'avèrent souvent plus efficaces pour sélectionner, en première approche, un sous-ensemble de variables explicatives avant d'opérer une modélisation,

explicatif : de la prudence est requise d'autant que les hypothèses ne peuvent être évaluées de façon indépendante surtout si, en plus, des cellules sont déséquilibrées ou vides,

prédictif : la recherche d'un modèle efficace, donc parcimonieux, peut conduire à négliger des interactions ou effets principaux lorsqu'une faible amélioration du R^2 le justifie et même si le test correspondant apparaît comme significatif. L'utilisation du C_p est théoriquement possible mais en général ce critère n'est pas calculé et d'utilisation délicate car nécessite la considération d'un "vrai" modèle de référence ou tout du moins d'un modèle de faible biais pour obtenir une estimation raisonnable de la variance de l'erreur. En revanche AIC et PRESS donnent des indications plus pertinentes. L'algorithme de recherche descendant est le plus couramment utilisé avec la contrainte suivante : *un effet principal n'est supprimé qu'à la condition qu'il n'apparaisse plus dans une interaction*.

8.4 Exemple

Les données, extraites de Jobson (1991), sont issues d'une étude marketing visant à étudier l'impact de différentes campagnes publicitaires sur les ventes de différents aliments. Un échantillon ou "panel" de familles a été constitué en tenant compte du lieu d'habitation ainsi que de la constitution de la famille. Chaque semaine, chacune de ces familles ont rempli un questionnaire décrivant les achats réalisés. Nous nous limitons ici à l'étude de l'impact sur la *consommation de lait* de quatre campagnes diffusées sur des chaînes locales de télévision. Quatre villes, une par campagne publicitaire, ont été choisies dans cinq différentes régions géographiques. Les consommations en lait par chacune des six familles par ville alors été mesurées (en dollars) après deux mois de campagne.

Les données se présentent sous la forme d'un tableau à 6 variables : la région géographique, les 4 consommations pour chacune des villes ou campagnes publicitaires diffusées, la taille de la famille. Cette situation est celle classique d'un modèle d'analyse de variance. Nous choisissons ici de conserver quantitative la variable taille de la famille et donc de modéliser la consommation de lait par un modèle d'analyse de covariance plus *économique* en degrés de liberté moins de paramètres sont à estimer.

On s'intéresse à différents modèles de régression visant à expliquer la consommation en fonction de la taille de la famille conditionnellement au type de campagne publicitaire.

```
proc glm data=sasuser.milk;
class pub;
model consom=pub taille pub*taille;
run;
```

Les résultats ci-dessous conduiraient à conclure à une forte influence de la taille mais à l'absence d'influence du type de campagne. Les droites de régression ne semblent pas significativement différentes.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PUB	3	227.1807	75.7269	0.57	0.6377 (1)
TAILLE	1	40926.0157	40926.0157	306.57	0.0001 (2)
TAILLE*PUB	3	309.8451	103.2817	0.77	0.5111 (3)

(1) Test de la significativité des différences des termes constants.
(2) Test de l'influence du facteur quantitatif.
(3) Test de la significativité des différences des pentes (interaction).

Néanmoins, pris d'un doute, le même calcul est effectué séparément pour chaque région :

```
proc glm data=sasuser.milk;
by region;
class pub;
model consom=pub taille pub*taille;
run;
```

Région	Source	DF	Type III SS	Mean Square	F Value	Pr > F
	PUB	3	72.02974	24.00991	4.62	0.0164
1	TAILLE	1	7178.32142	7178.32142	1380.25	0.0001
	TAILLE*PUB	3	217.37048	72.45683	13.93	0.0001
	PUB	3	231.73422	77.24474	30.36	0.0001
2	TAILLE	1	8655.25201	8655.25201	3402.34	0.0001
	TAILLE*PUB	3	50.15069	16.71690	6.57	0.0042
	PUB	3	79.54688	26.51563	6.01	0.0061
3	TAILLE	1	6993.30160	6993.30160	1585.35	0.0001
	TAILLE*PUB	3	173.19305	57.73102	13.09	0.0001
	PUB	3	415.66664	138.55555	15.23	0.0001
4	TAILLE	1	9743.37830	9743.37830	1071.32	0.0001
	TAILLE*PUB	3	361.39556	120.46519	13.25	0.0001

```
        PUB            3       15.35494       5.11831      0.79     0.5168
5       TAILLE         1     8513.28516    8513.28516   1314.71     0.0001
        TAILLE*PUB     3       52.75119      17.58373      2.72     0.0793
```

Il apparaît alors qu'à l'intérieur de chaque région (sauf région 5), les campagnes de publicité ont un effet tant sur la constante que sur la pente.

Ceci incite donc à se méfier des *interactions* (l'effet région compense l'effet publicité) et encourage à toujours conserver le facteur bloc (ici la région) dans une analyse de variance. Une approche complète, considérant *a priori* toutes les variables (3 facteurs), est ici nécessaire (cf. TP).

9 Exemple : Prévision de la concentration d'ozone

9.1 Les données

Les données proviennent des services de Météo-France et s'intéresse à la prévision de la concentration en Ozone dans 5 stations de mesure ; ces sites ont été retenus pour le nombre impoprtant de pics de pollution qui ont été détectés dans les périodes considérées (étés 2002, 2003, 2005). Un pic de pollution est défini ici par une concentration dépassant le seuil de $150\mu g/m^3$. Météo-France dispose déjà d'une prédvision (MOCAGE), à partir d'um modèle physique basé sur les équations du comportement dynamique de l'atmosphère (Navier et Stockes). Cette prévision fait partie du dispositif d'alerte des pouvoirs publics et prévoit donc une concentration de pollution à 17h locale pour le lendemain. L'objet du travail est d'en faire une évaluation statistique puis de l'améliorer en tenant compte d'autres variables ou plutôt d'autres prévisions faites par Météo-France. Il s'agit donc d'intégrer ces informations dans un modèle statistique global.

Les variables

Certaines variables de concentration ont été transformées afin de rendre symétrique (plus gaussienne) leur distribution.

O3-o Concentration d'ozone effectivement observée ou variable à prédire,

03-pr prévision "mocage" qui sert de variable explicative ;

Tempe Température prévue pour le lendemain,

vmodule Force du vent prévue pour le lendemain,

lno Logarithme de la concentration observée en monoxyde d'azote,

lno2 Logarithme de la concentration observée en dioxyde d'azote,

rmh20 Racine de la concentration en vapeur d'eau,

Jour Variable à deux modalités pour distinguer les jours "ouvrables" (0) des jours "fériés-WE" (1).

Station Une variable qualitative indique la station concernée : Aix-en-Provence, Rambouillet, Munchhausen, Cadarache, et Plan de Cuques.

Modèle physique

Les graphiques de la figure 2.4 représente la première prévision de la concentration d'ozone observée, ainsi que ses résidus, c'est-à-dire celle obtenue par le modèle physique MOCAGE. Ces graphes témoignent de la mauvaise qualité de ce modèle : les résidus ne sont pas répartis de façon symétrique et les deux nuages pésentent une légère forme de "banane" signifiant que des composantes non linéaires du modèle n'ont pas été prises en compte. D'autre part, la forme d'entonnoir des résidus montrent une forte hétéroscédasticité. Cela signifie que la variance des résidus et donc des prévisions croît avec la valeur. En d'autre terme, la qualité de la prévision se dégrade pour les concentrations élevées justgement dans la zone "sensible".

Modèle sans interaction

Un premier modèle est estimé avec R :

```
fit.lm=lm(O3-o~O3-pr+vmodule+lno2+lno+s-rmh2o+jour+station+TEMPE,data=donne)
```

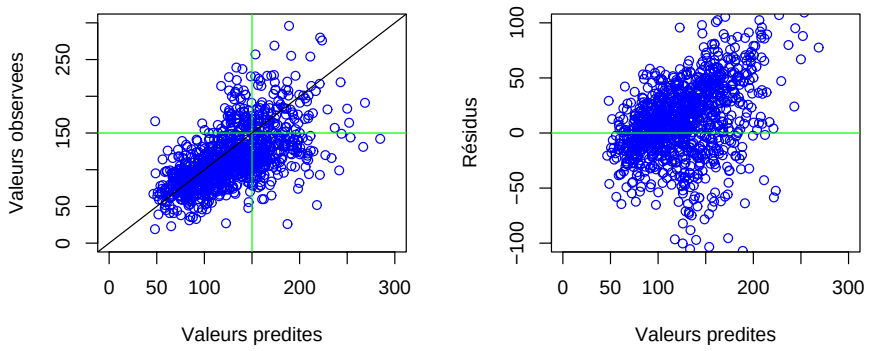

FIG. 2.4 – Ozone : prévision et résidus du modèle MOCAGE de Météo-France pour 5 stations.

Il introduit l'ensemble des variables explicatives mais sans interaction. Les résultats numériques sont fournis ci-dessous.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.99738    7.87028  -0.635  0.52559
O3_pr         0.62039    0.05255  11.805  < 2e-16 ***
vmodule      -1.73179    0.35411  -4.891 1.17e-06 ***
lno2        -48.17248    6.19632  -7.774 1.83e-14 ***
lno          50.95171    5.98541   8.513  < 2e-16 ***
s_rmh2o     135.88280   50.69567   2.680  0.00747 **
jour1        -0.34561    1.85389  -0.186  0.85215
stationAls    9.06874    3.37517   2.687  0.00733 **
stationCad   14.31603    3.07893   4.650 3.76e-06 ***
stationPla   21.54765    3.74155   5.759 1.12e-08 ***
stationRam    6.86130    3.05338   2.247  0.02484 *
TEMPE         4.65120    0.23170  20.074  < 2e-16 ***

Residual standard error: 27.29 on 1028 degrees of freedom
Multiple R-Squared: 0.5616,     Adjusted R-squared: 0.5569
F-statistic: 119.7 on 11 and 1028 DF,  p-value: < 2.2e-16
```

A l'exception de la variable indiquant la nature du jour, l'ensemble des coefficients sont jugés significativement différent de zéro mais la qualité de l'ajustement est faible (R^2).

Modèle avec interaction

La qualité d'ajustement du modèle précédent n'étant pas très bonne, un autre modèle est considéré en prenant en compte les interactions d'ordre 2 entgre les variables. Compte tenu de la complexité du modèle qui un découle, un choix automatique est lancé par élimination successive des termes non significatifs (algotithme backward). Le critère optimisé est celui (AIC) d'Akaïke. Plusieurs interactions ont été éliminées au cours de la procédure mais beaucoup subsistent dans le modèle. Attention, les effets principaux `lno2`, `vmodule` ne peuvent être retirés car ces variables apparaissent dans une interaction. En revanche on peut s'interroger sur l'opportunité de conserver celle entre la force du vent et la concentration de dioxyde d'azote.

```
           Df Deviance Resid. Df Resid. Dev        F    Pr(>F)
NULL                        1039    1745605
O3_pr       1   611680      1038    1133925 969.9171 < 2.2e-16 ***
station     4    39250      1034    1094674  15.5594 2.339e-12 ***
vmodule     1     1151      1033    1093523   1.8252 0.1769957
lno2        1      945      1032    1092578   1.4992 0.2210886
s_rmh2o     1    24248      1031    1068330  38.4485 8.200e-10 ***
```

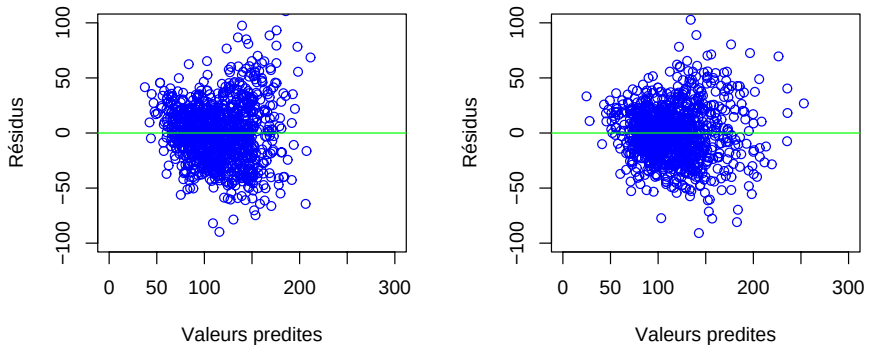

FIG. 2.5 – Ozone : Résidus des modèles linéaire et quadratique.

```
TEMPE              1   248891      1030     819439 394.6568 < 2.2e-16 ***
O3_pr:station      4    16911      1026     802528   6.7038 2.520e-05 ***
O3_pr:vmodule      1     8554      1025     793974  13.5642 0.0002428 ***
O3_pr:TEMPE        1    41129      1024     752845  65.2160 1.912e-15 ***
station:vmodule    4     7693      1020     745152   3.0497 0.0163595 *
station:lno2       4    12780      1016     732372   5.0660 0.0004811 ***
station:s_rmh2o    4    19865      1012     712508   7.8746 2.997e-06 ***
station:TEMPE      4    27612      1008     684896  10.9458 1.086e-08 ***
vmodule:lno2       1     1615      1007     683280   2.5616 0.1098033
vmodule:s_rmh2o    1     2407      1006     680873   3.8163 0.0510351 .
lno2:TEMPE         1     4717      1005     676156   7.4794 0.0063507 **
s_rmh2o:TEMPE      1    42982      1004     633175  68.1543 4.725e-16 ***
```

Ce sont surtout les graphes de la figure 2.5 qui renseignent sur l'adéquation des modèles. Le modèle quadratique fournit une forme plus "linéaire" des résidus et un meilleur ajustement avec un R^2 de 0,64 mais l'hétéroscédasticité reste présente, d'autres approches s'avèrent nécessaires afin de réduire la variance liée à la prévision des concentrations élevées.

9.2 Autres exemples

Les autres jeux de données étudiés dans ce cours ne se prètent pas à une modèle de régression multiple classique ; soit la variable à prédire est qualitative binaire et correspondent donc à un modèle de régression logistique (cancer et données bancaires), soit la situation est plus complexe car fait appel à un modèle mixte ou à effet aléatoire (régime des souris).

Chapitre 3

Régression logistique

1 Introduction

Dans ce chapitre, nous définissons le contexte pratique de la *régression logistique* qui s'intéressent plus particulièrement à la description ou l'explication d'observations constitués d'effectifs comme, par exemple, le nombre de succès d'une variable de Bernouilli lors d'une séquence d'essais. Contrairement aux modèles du chapitre précédent basés sur l'hypothèse de normalité des observations, les lois concernées sont discrètes et associées à des dénombrements : binomiale, multinomiale. Néanmoins, ce modèle appartient à la famille du *modèle linéaire général* (annexe) et partagent à ce titre beaucoup d'aspects (estimation par maximum de vraisemblance, tests, diagnostics) et dont la stratégie de mise en œuvre, similaire au cas gaussien, n'est pas reprise.

Une première section définit quelques notions relatives à l'étude de la liaison entre variables qualitatives. Elles sont couramment utilisées dans l'interprétation des modèles de régression logistique.

2 Odds et odds ratio

Une variable

Soit Y une variable qualitative à J modalités. On désigne la chance (ou *odds*[1] de voir se réaliser la $j$ème modalité plutôt que la $k$ème par le rapport

$$\Omega_{jk} = \frac{\pi_j}{\pi_k}$$

où π_j est la probabilité d'apparition de la $j$ème modalité. Cette quantité est estimée par le rapport n_j/n_k des effectifs observés sur un échantillon. Lorsque la variable est binaire et suit une loi de Bernouilli de paramètre π, l'odds est le rapport $\pi/(1-\pi)$ qui exprime une cote ou chance de gain.

Par exemple, si la probabilité d'un succès est 0.8, celle d'un échec est 0.2. L'odds du succès est 0.8/0.2=4 tandis que l'odds de l'échec est 0.2/0.8=0.25. On dit encore que la chance de succès est de 4 contre 1 tandis que celle d'échec est de 1 contre 4.

Table de contingence

On considère maintenant une table de contingence 2×2 croisant deux variables qualitatives binaires X^1 et X^2. les paramètres de la loi conjointe se mettent dans une matrice :

$$\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix}$$

où $\pi_{ij} = P[\{X^1 = i\} \text{ et } \{X^2 = j\}]$ est la probabilité d'occurence de chaque combinaison.

- Dans la ligne 1, l'odds que la colonne 1 soit prise plutôt que la colonne 2 est :

$$\Omega_1 = \frac{\pi_{11}}{\pi_{12}}.$$

[1] Il n'existe pas, même en Québécois, de traduction consensuelle de "odds" qui utilise néanmoins souvent le terme "cote".

- Dans la ligne 2, l'odds que la colonne 1 soit prise plutôt que la colonne 2 est :

$$\Omega_2 = \frac{\pi_{21}}{\pi_{22}}.$$

On appelle *odds ratio* (rapport de cote) le rapport

$$\Theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

Ce rapport prend la valeur 1 si les variables sont indépendantes, il est supérieur à 1 si les sujets de la ligne 1 ont plus de chances de prendre la première colonne que les sujets de la ligne 2 et inférieur à 1 sinon.

Exemple : supposons qu'à l'entrée dans une école d'ingénieurs, 7 garçons sur 10 sont reçus tandis que seulement 4 filles sur 10 le sont. L'odds des garçons est alors de 0.7/0.3=2.33 tandis que celle des filles est de 0.4/0.6=0.67. L'odds ratio est de 2.33/0.67=3.5. La chance d'être reçu est 3.5 plus grande pour les garçons que pour les filles.

L'odds ratio est également défini pour deux lignes (a, b) et deux colonnes (c, d) quelconques d'une table de contingence croisant deux variables à J et K modalités. L'odds ratio est le rapport

$$\Theta_{abcd} = \frac{\Omega_a}{\Omega_b} = \frac{\pi_{ac}\pi_{bd}}{\pi_{ad}\pi_{bc}} \quad \text{estimé par l'odds ratio empirique} \quad \widehat{\Theta}_{abcd} = \frac{n_{ac}n_{bd}}{n_{ad}n_{bc}}.$$

3 Régression logistique

3.1 Type de données

Cette section décrit la modélisation d'une variable qualitative Z à 2 modalités : 1 ou 0, succès ou échec, présence ou absence de maladie, panne d'un équipement, faillite d'une entreprise, bon ou mauvais client.... Les modèles de régression précédents adaptés à l'explication d'une variable quantitative ne s'appliquent plus directement car le régresseur linéaire usuel $\mathbf{X}\boldsymbol{\beta}$ ne prend pas des valeurs simplement binaires. L'objectif est adapté à cette situation en cherchant à expliquer les probabilités

$$\pi = P(Z = 1) \quad \text{ou} \quad 1 - \pi = P(Z = 0),$$

ou plutôt une transformation de celles-ci, par l'observation conjointe des variables explicatives. L'idée est en effet de faire intervenir une fonction réelle monotone g opérant de $[0, 1]$ dans $\mathbb{R}$ et donc de chercher un modèle linéaire de la forme :

$$g(\pi_i) = \mathbf{x}_i'\beta.$$

Il existe de nombreuses fonctions, dont le graphe présente une forme sigmoïdale et qui sont candidates pour remplir ce rôle, trois sont pratiquement disponibles dans les logiciels :

probit : g est alors la fonction inverse de la fonction de répartition d'une loi normale, mais son expression n'est pas explicite.

log-log avec g définie par

$$g(\pi) = \ln[-\ln(1 - \pi)]$$

mais cette fonction est dissymétrique.

logit est définie par

$$g(\pi) = \mathit{logit}(\pi) = \ln\frac{\pi}{1 - \pi} \quad \text{avec} \quad g^{-1}(x) = \frac{e^x}{1 + e^x}.$$

Plusieurs raisons, tant théoriques que pratiques, font préférer cette dernière solution. Le rapport $\pi/(1 - \pi)$, qui exprime une "cote", est l'*odds* et la *régression logistique* s'interprète donc comme la recherche d'une modélisation linéaire du "log odds" tandis que les coefficients de certains modèles expriment des "odds ratio" c'est-à-dire l'influence d'un facteur qualitatif sur le risque (ou la chance) d'un échec (d'un succès) de Z.

Cette section se limite à la description de l'usage élémentaire de la régression logistique. Des compléments concernant l'explication d'une variable qualitative ordinale (plusieurs modalités), l'intervention de variables explicatives avec effet aléatoire, l'utilisation de mesures répétées donc dépendantes, sont à rechercher dans la bibliographie.

3.2 Modèle binomial

On considère, pour $i = 1, \ldots, I$, différentes valeurs *fixées* $x_i^1, \ldots, x_i^q$ des variables explicatives $X^1, \ldots, X^q$. Ces dernières pouvant être des variables quantitatives ou encore des variables qualitatives, c'est-à-dire des facteurs issus d'une planification expérimentale.

Pour chaque groupe, c'est-à-dire pour chacune des combinaisons de valeurs ou facteurs, on réalise n_i observations ($n = \sum_{i=1}^{I} n_i$) de la variable Z qui se mettent sous la forme $y_1/n_1, \ldots, y_I/n_I$ où y_i désigne le nombre de "succès" observés lors des n_i essais. On suppose que toutes les observations sont indépendantes et qu'à l'intérieur d'un même groupe, la probabilité π_i de succès est constante. Alors, la variable Y_i sachant n_i et d'espérance $E(Y_i) = n_i\pi_i$ suit une loi *binomiale* $\mathcal{B}(n_i, \pi_i)$ dont la fonction de densité s'écrit :

$$P(Y = y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{(n_i - y_i)}.$$

On suppose que le vecteur des fonctions *logit* des probabilités π_i appartient au sous-espace vect$\{X^1, \ldots, X^q\}$ engendré par les variables explicatives :

$$\mathit{logit}(\pi_i) = \mathbf{x}_i'\boldsymbol{\beta} \quad i = 1, \ldots, I$$

ce qui s'écrit encore

$$\pi_i = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}} \quad i = 1, \ldots, I.$$

Le vecteur des paramètres est estimé par maximisation de la log-vraisemblance. Il n'y a pas de solution analytique, celle-ci est obtenue par des méthodes numériques itératives (par exemple Newton Raphson) dont certaines reviennent à itérer des estimations de modèles de régression par moindres carrés généralisés avec des poids et des métriques adaptés à chaque itération.

L'optimisation fournit une estimation $\mathbf{b}$ de $\boldsymbol{\beta}$, il est alors facile d'en déduire les estimations ou prévisions des probabilités π_i :

$$\widehat{\pi}_i = \frac{e^{\mathbf{x}_i'\mathbf{b}}}{1 + e^{\mathbf{x}_i'\mathbf{b}}}$$

et ainsi celles des effectifs

$$\widehat{y}_i = n_i\widehat{\pi}_i.$$

Remarques

i. La matrice $\mathbf{X}$ issue de la planification expérimentale est construite avec les mêmes règles que celles utilisées dans le cadre de l'analyse de covariance mixant variables explicatives quantitatives et qualitatives. Ainsi, les logiciels gèrent avec plus ou moins de clarté le choix des variables indicatrices et donc des paramètres estimables ou contrastes associés.

ii. La situation décrite précédemment correspond à l'observation de données *groupées*. Dans de nombreuses situations concrètes et souvent dès qu'il y a des variables explicatives quantitatives, les observations $\mathbf{x}_i$ sont toutes distinctes. Ceci revient donc à fixer $n_i = 1; i = 1, \ldots, I$ dans les expressions précédentes et la loi de Bernouilli remplace la loi binomiale. Certaines méthodes ne sont alors plus applicables et les comportements asymptotiques des distributions des statistiques de test ne sont plus valides, le nombre de paramètres tendant vers l'infini.

iii. Dans le cas d'une variable explicative X dichotomique, un logiciel comme SAS fournit, en plus de l'estimation d'un paramètre b, celle des odds ratios ; b est alors le log odds ratio ou encore, e^b est l'odds ratio. Ceci s'interprète en disant que Y a e^b fois plus de chance de succès (ou de maladie comme par un exemple un cancer du poumon) quand $X = 1$ (par exemple pour un fumeur).

3.3 Régressions logistiques polytomique et ordinale

La régression logistique adaptée à la modélisation d'une variable dichotomique se généralise au cas d'une variable Y à plusieurs modalités ou polytomique. Si ces modalités sont ordonnés, on dit que la

variable est qualitative ordinale. Ces types de modélisation sont très souvent utilisés en épidémiologie et permettent d'évaluer ou comparer des risques par exemples sanitaires. Des estimations d'odds ratio ou rapports de cotes sont ainsi utilisés pour évaluer et interpréter les facteurs de risques associés à différents types (régression polytomique) ou seuils de gravité (régression ordinale) d'une maladie ou, en marketing, cela s'applique à l'explication, par exemple, d'un niveau de satisfaction d'un client. Il s'agit de comparer entre elles des estimations de fonctions logit.

Dans une situation de *data mining* ou fouille de données, ce type d'approche se trouve lourdement pénalisé lorsque, à l'intérieur d'un même modèle polytomique ou ordinal, plusieurs types de modèles sont en concurrence pour chaque fonction logit associée à différentes modalités. Différents choix de variables, différents niveaux d'interaction rendent trop complexe et inefficace cette approche. Elle est à privilégier uniquement dans le cas d'un nombre restreint de variables explicatives avec un objectif explicatif ou interprétatif.

À titre illustratif, explicitons le cas simple d'une variable Y à k modalités ordonnées expliquée par une seule variable dichotomique X. Notons $\pi_j(X) = P(Y = j|X)$ avec $\sum_{j=1}^{k} \pi_j(X) = 1$. Pour une variable Y à k modalités, il faut, en toute rigueur, estimer $k-1$ prédicteurs linéaires :

$$g_j(X) = \alpha_j + \beta_j X \quad \text{pour} j = 1, \ldots, k-1$$

et, dans le cas d'une variable ordinale, la fonction lien logit utilisée doit tenir compte de cette situation particulière.

Dans la littérature, trois types de fonction sont considérées dépendant de l'échelle des rapports de cote adoptée :

- échelle basée sur la comparaison des catégories adjacentes deux à deux,
- sur la comparaison des catégories adjacentes supérieures cumulées,
- et enfin sur la comparaison des catégories adjacentes cumulées.

Pour $k = 2$, on retrouve les trois situations se ramènent à la même d'une variable dichotomique. C'est le dernier cas qui est le plus souvent adopté ; il conduit à définir les fonctions des "logits cumulatifs" de la forme :

$$\log \frac{\pi_{j+1} + \cdots + \pi_k}{\pi_1 + \cdots + \pi_j} \quad \text{pour} j = 1, \ldots, k-1.$$

Pour un seuil donné sur Y, les catégories inférieures à ce seuil, cumulées, sont comparées aux catégories supérieures cumulées. Les fonctions logit définies sur cette échelle dépendent chacune de tous les effectifs, ce qui peut conduire à une plus grande stabilité des mesures qui en découlent.

Si les variables indépendantes sont nombreuses dans le modèle ou si la variable réponse Y comporte un nombre élevé de niveaux, la description des fonctions logit devient fastidieuse. La pratique consiste plutôt à déterminer un coefficient global b (mesure d'effet) qui soit la somme pondérée des coefficients b_j. Ceci revient à faire l'hypothèse que les coefficients sont homogènes (idéalement tous égaux), c'est-à-dire à supposée que les rapports de cotes sont proportionnels. C'est ce que calcule implicitement la procédure LOGISTIC de SAS appliquée à une variable réponse Y ordinale en estimant un seul paramètre b mais $k-1$ termes constants correspondant à des translations de la fonctions logit. La procédure LOGISTIC fournit le résultat du test du score sur l'hypothèse H_0 de l'homogénéité des coefficients β_j.

Le coefficient b mesure donc l'association du facteur X avec la gravité de la maladie et peut s'interpréter comme suit : pour tout seuil de gravité choisi sur Y, la cote des risques d'avoir une gravité supérieure à ce seuil est e^b fois plus grande chez les exposés ($X = 1$) que chez les non exposés ($X = 0$).

Attention dans SAS, la procédure LOGISTIC adopte une paramétrisation $(-1, 1)$ analogue à celle de la procédure CATMOD mais différente de celle de GENMOD ou SAS/Insight $(0, 1)$. Ceci explique les différences observées dans l'estimation des paramètre d'une procédure à l'autre mais les modèles sont identiques.

4 Choix de modèle

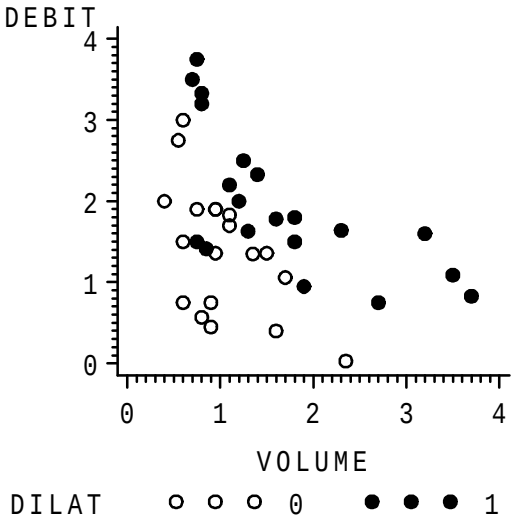

FIG. 3.1 – Dilatation : Nuage des modalités de Y dans les coordonnées des variables explicatives.

4.1 Recherche pas à pas

Principalement deux critères (test du rapport de vraisemblance et test de Wald, cf. bibliographie), sont utilisés de façon analogue au test de Fisher du modèle linéaire gaussien. Ils permettent de comparer un modèle avec un sous-modèle et d'évaluer l'intérêt de la présence des termes complémentaires. On suit ainsi une stratégie descendante à partir du modèle complet. L'idée est de supprimer, un terme à la fois, la composante d'interaction ou l'effet principal qui apparaît comme le moins significatif au sens du rapport de vraisemblance ou du test de Wald. Les tests présentent une structure hiérarchisée. SAS facilite cette recherche en produisant une décomposition (Type III) de ces indices permettant de comparer chacun des sous-modèles excluant un des termes avec le modèle les incluant tous.

Attention, du fait de l'utilisation d'une transformation non linéaire (logit), même si des facteurs sont orthogonaux, aucune propriété d'orthogonalité ne peut être prise en compte pour l'étude des hypothèses. Ceci impose l'élimination des termes un par un et la ré-estimation du modèle. D'autre part, un terme principal ne peut être supprimé que s'il n'intervient plus dans des termes d'interaction.

4.2 Critère

L'approche précédente favorise la qualité d'ajustement du modèle. Dans un but prédictif, certains logiciels, comme Splus/R ou Enterpirse Miner, proposent d'autres critères de choix (AIC, BIC). Une estimation de l'erreur de prévision par validation croisée est aussi opportune dans une démarche de choix de modèle.

5 Illustration élémentaire

5.1 Les données

On étudie l'influence du débit et du volume d'air inspiré sur l'occurence (codée 1) de la dilatation des vaisseaux sanguins superficiels des membres inférieurs. Un graphique élémentaire représentant les modalités de Y dans les coordonnées de $X^1 \times X^2$ est toujours instructif. Il montre une séparation raisonnable et de bon augure des deux nuages de points. Dans le cas de nombreuses variables explicatives quantitatives, une analyse en composantes principales s'impose. Les formes des nuages représentés, ainsi que l'allure des distributions (étudiées préalablement), incitent dans ce cas à considérer par la suite les logarithmes des variables. Une variable `un` ne contenant que des "1" dénombrant le nombre d'essais est nécessaire dans la syntaxe de `genmod`. Les données sont en effet non groupées.

```
proc logistic data=sasuser.debvol;
model dilat=l_debit l_volume;
```

```
run;
proc genmod data=sasuser.debvol;
model dilat/un=l_debit l_volume/d=bin;
run;
```

```
The LOGISTIC Procedure
                            Intercept
                Intercept       and
 Criterion        Only      Covariates    Chi-Square for Covariates
 AIC              56.040       35.216         .
 SC               57.703       40.206         .
 -2 LOG L         54.040       29.216(1)    24.824 with 2 DF (p=0.0001)
 Score              .            .          16.635 with 2 DF (p=0.0002)

            Parameter(2) Standard    Wald(3)     Pr >    Standardized  Odds
Variable DF  Estimate    Error   Chi-Square  Chi-Square  Estimate     Ratio
INTERCPT 1     2.8782    1.3214     4.7443     0.0294           .      .
L_DEBIT  1    -4.5649    1.8384     6.1653     0.0130    -2.085068    0.010
L_VOLUME 1    -5.1796    1.8653     7.7105     0.0055    -1.535372    0.006
```

Cette procédure fournit des critères de choix de modèle dont la déviance (1), le vecteur **b** des paramètres (2) et les statistiques des tests (3) comparant le modèle excluant un terme par rapport au modèle complet tel qu'il est décrit dans la commande.

```
        Criteria For Assessing Goodness Of Fit
     Criterion             DF         Value      Value/DF
     Deviance              36       29.2156        0.8115 (1)
     Scaled Deviance       36       29.2156        0.8115 (2)
     Pearson Chi-Square    36       34.2516        0.9514 (3)
     Scaled Pearson X2     36       34.2516        0.9514
     Log Likelihood         .      -14.6078             .

              Analysis Of Parameter Estimates
  Parameter    DF    Estimate (4)  Std Err   ChiSquare (5) Pr>Chi
  INTERCEPT     1     -2.8782      1.3214      4.7443      0.0294
  L_DEBIT       1      4.5649      1.8384      6.1653      0.0130
  L_VOLUME      1      5.1796      1.8653      7.7105      0.0055
  SCALE (6)     0      1.0000      0.0000           .       .
```

(1)	Déviance du modèle par rapport au modèle saturé.
(2)	Déviance pondérée si le paramètre d'échelle est différent de 1 en cas de sur-dispersion.
(3)	Statistique de Pearson, voisine de la déviance, comparant le modèle au modèle saturé .
(4)	Paramètres du modèle.
(5)	Statistique des tests comparant le modèle excluant un terme par rapport au modèle complet.
(6)	Estimation du paramètre d'échelle si la quasi-vraisemblance est utilisée.

5.2 Régression logistique ordinale

On étudie les résultats d'une étude préalable à la législation sur le port de la ceinture de sécurité dans la province de l'Alberta à Edmonton au Canada (Jobson, 1991). Un échantillon de 86 769 rapports d'accidents de voitures ont été compulsés afin d'extraire une table croisant :

i. Etat du conducteur : Normal ou Alcoolisé
ii. Sexe du conducteur
iii. Port de la ceinture : Oui Non
iv. Gravité des blessures : 0 : rien à 3 : fatales

Les modalités de la variable à expliquer concernant la gravité de l'accident sont ordonnées.

```
/* régression ordinale */
proc logistic data=sasuser.ceinture;
class sexe alcool ceinture;
```

```
model gravite=sexe alcool ceinture ;
weight effectif;
run;
```

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	Gr0	1	1.8699	0.0236	6264.9373	<.0001
Intercept	Gr1	1	2.8080	0.0269	10914.3437	<.0001
Intercept	Gr2	1	5.1222	0.0576	7917.0908	<.0001
sexe	Sfem	1	-0.3118	0.0121	664.3353	<.0001
alcool	A_bu	1	-0.5017	0.0190	697.0173	<.0001
ceinture	Cnon	1	-0.1110	0.0174	40.6681	<.0001

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
sexe	Sfem vs Shom	0.536	0.511	0.562
alcool	A_bu vs Ajeu	0.367	0.340	0.395
ceinture	Cnon vs Coui	0.801	0.748	0.858

6 Autres exemples

Les exemples sont décrits dans cette section à titre illustratif avec SAS ou R, une comparaison systématique des performances de chaque méthode est développée dans le dernier chapitre conclusif.

6.1 Cancer du sein

Les données (`Wisconsin BreastCancer Database`) sont disponibles dans la librairie `mlbench` du logiciel R. Elles servent très souvent de base de référence à des comparaisons de techniques d'apprentissage. Les variables considérées sont :

Cl.thickness Clump Thickness

Cell.size Uniformity of Cell Size

Cell.shape Uniformity of Cell Shape

Marg.adhesion Marginal Adhesion

Epith.c.size Single Epithelial Cell Size

Bare.nuclei Bare Nuclei

Bl.cromatin Bland Chromatin

Normal.nucleoli Normal Nucleoli

Mitoses Mitoses

Class "benign" et "malignant".

La dernière variable est celle à prédire, les variables explicatives sont ordinales ou nominales à 10 classes. Il reste 683 observations après la suppression de 16 présentant des valeurs manquantes.

Ce jeu de données est assez particulier car plutôt facile à ajuster. Une estimation utilisant toutes les variables conduit à des messages critiques indiquant un défaut de convergence et des probabilités exactement ajustées. En fait le modèle s'ajuste exactement aux données en utilisant toutes les variables aussi l'erreur de prévision nécessite une estimation plus soignée. Une séparation entre un échantillon d'apprentissage et un échantillon test ou une validation croisée permet une telle estimation (voir le chapitre 5).

On trouve alors qu'un modèle plus parcimonieux et obtenu par une démarche descendante, de sorte que les paramètres soient significatifs au sens d'un test du Chi2, conduit à des erreurs de prévision plus faibles sur un échantillon test indépendant qu'un modèle ajustant exactement les données. La qualité de l'ajustement du modèle se résume sous la forme d'une matrice de *confusion* évaluant les taux de bien et mal classés sur l'échantillon d'apprentissage tandis que l'erreur de prévision est estimée à partir de l'échantillon test.

```
# erreur d'ajustement
fitq.lm=glm(Class~Cl.thickness+Cell.size+Cell.shape ,data=datapq,family=binomial)
table(fitq.lm$fitted.values>0.5,datapq[,"Class"])

        benign malignant
  FALSE    345         6
  TRUE      13       182

# erreur de prévision
predq.lm=predict(fitq.lm,newdata=datestq) # prevision
table(predq.lm>0.5,datestq[,"Class"])

          benign malignant
  FALSE     84         5
  TRUE       2        46
```

Le taux d'erreur apparent estimé sur l'échantillon d'apprentissage est de 3,5% (0% avec le modèle complet) tandis que le taux d'erreur estimé sans biais sur l'échantillon test est de 5,1% (5,8 avec le modèle complet). Ces estimations demanderont à être affinées afin de comparer les méthodes entre elles.

6.2 Pic d'ozone

Plutôt que de prévoir la concentration de l'ozone puis un dépassement éventuel d'un seuil, il pourrait être plus efficace de prévoir directement ce dépassement en modélisant la variable binaire associée. Attention toutefois, ces dépassements étant relativement peu nombreux (17%), il serait nécessaire d'en accentuer l'importance par l'introduction d'une fonction coût ou une pondération spécifique. Ceci est un problème général losqu'il s'agit de prévoir des phénomènes très rares : un modèle trivial ne les prévoyant jamais ne commettrait finalement qu'une erreur relative faible. Ceci revient à demander au spécialiste de quantifier le risque de prévoir un dépassement du seuil à tord par rapport à celui de ne pas prévoir ce dépassement à tord. Le premier à des conséquences économiques et sur le confort des usagers par des limitations de trafic tandis que le 2ème a des conséquences sur l'environnement et la santé de certaines populations. Ce n'est plus un problème "statistique".

La recherche descendante d'un meilleur modèle au sens du critère d'Akaïke conduit au résultat ci-dessous.

```
         Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     831     744.34
O3_pr     1   132.89       830     611.46 9.576e-31
vmodule   1     2.42       829     609.04      0.12
s_rmh2o   1    33.71       828     575.33 6.386e-09
station   4    16.59       824     558.74 2.324e-03
TEMPE     1   129.39       823     429.35 5.580e-30
```

On peut s'interroger sur l'intérêt de la présence de la variable `vmodule` dont la présence rend plus faible la prévision de l'erreur au sens d'Akaïke mais dont le coefficient n'est pas significatif au sens du test du Chi2 ; ce critère étant lié à une qualité d'ajustement. L'erreur estimée sur l'échantillon test ne permet pas de départager ces modèles car les matrices de transition similaires conduisent à la même estimation du taux d'erreur de 11,5% tandis que la même erreur est de 13,2% pour le modèle MOCAGE. Un modèle de régression logistique faisant intervenir les interactions d'ordre 2 et optimisé par algorithme descendant aboutit à une erreur de 10,6% tandis que le modèle quantitatif de régression quadratique du chapitre précédent conduit à une erreur de 10,1% avec le même protocole et les mêmes échantillons d'apprentissage et de test.

```
Matrices de confusion de l'échantillon test pour différents modèles :
          0   1                    0   1                0   1                 0   1
  FALSE 163  19           FALSE 162  18        FALSE 163  17        FALSE 160   13
  TRUE    5  21           TRUE    6  22        TRUE    5  23        TRUE    8   27
logistique sans vmodule   avec vmodule      avec interactions      quantitatif
```

Notons que les erreurs ne sont pas "symétriques" et sont affectées du même biais : tous ces modèles "oublient" systématiquement plus de dépassements de seuils qu'ils n'en prévoient à tord. Une analyse

plus poussée de l'estimation de l'erreur de prédiction est évidemment nécessaire et ce sera, sur le plan méthodologique, l'objet du prochain chapitre. À ce niveau de l'étude, ce qui est le plus utile au météorologue, c'est l'analyse des coefficients les plus significativements présents dans la régression quadratique, c'est-à-dire avec les interactions. Ils fournissent des indications précieuses sur les faiblesses ou insuffisances de leur modèle physique.

6.3 Carte visa

Ces données sont présentées en détail dans Baccini et Besse (2000). Il s'agit de modéliser une variable binaire représentant la possession ou non de la carte visa premier en fonction du comportement bancaire d'un client. Comme dans l'exemple précédent, la possession de ce type de produit est rare ; aussi un échantillon spécifique, non représentatif, a été construit en surreprésentant la possession de ce type de produit.

Plusieurs stratégies peuvent être mises en œuvre sur ces données selon les transformations et codages réalisés sur les variables qualitatives. Elles sont explorées lors des différents TPs. La stratégie adoptée ici consiste à rechercher un "meilleur" modèle à l'aide de la procédure `SAS/STAT logistic` en association avec l'un des trois algorithmes de sélection (forward, backward ou stepwise).

La sélection de variables ainsi retenue est ensuite utilisée avec la procédure `genmod` aux sorties plus explicites qui est également mise en œuvre dans le module SAS Enterprise Miner. Le taux apparent d'erreur est évalué à partir du même échantillon d'apprentissage et donc de manière nécessairement biaisée par optimisme. Il mesure la qualité d'ajustement du modèle illustré par la matrice de confusion de l'échantillon ci-dessous associé à un taux d'erreur de 11,5%.

YVAR1(CARVPR) Frequency| Percent	PREDY 0	PREDY 1	Total
0	659	53	712
	61.65	4.96	66.60
1	70	287	357
	6.55	26.85	33.40
Total	729	340	1069
	68.19	31.81	100.00

La même démarche avec le logiciel R (voir les TP) conduit à un mmodèle qui, appliqué à l'échantillon test, fournit la matrice de confusion suivante avec un taux d'erreur de 17% supérieur à celui sur l'échantillon d'apprentissage qui est de 16%.

```
pred.vistest FALSE TRUE
       FALSE   125   22
       TRUE     12   41
```

On remarque que les échantillons tirés avec SAS ne conduisent pas du tout aux mêmes estimations d'erreurs qu'avec les échantillons tirés avec R. Ce n'est pas une question de logiciel, juste le hasard des tirages. Ceci implique qu'il faudra estimer plus finement le taux d'erreur de prévision afin de comparer les méthodes. Ceux-ci sont en effet entâchés d'une grande variance.

Chapitre 4

Modèle log-linéaire

1 Introduction

Comme dans le chapitre précédent, les modèles décrits dans ce chapitre s'intéressent plus particulièrement à la description ou l'explication d'observations constitués d'effectifs ; nombre de succès d'une variable de Bernouilli lors d'une séquence d'essais dans la cas précédent de la régression logistique, nombre d'individus qui prennent une combinaison donnée de modalités de variables qualitatives ou niveaux de facteurs, dans le cas présent. Ce modèle fait également partie de la famille du *modèle linéaire général* en étant associé à une loi de Poisson. Il est également appelé aussi *modèle log-linéaire* (voir Agresti (1990) pour un exposé détaillé) et s'applique principalement à la modélisation d'une table de contingence complète. Comme pour la régression logistique, les aspects au modèle linéaire général (estimation, tests, diagnostic) ont des stratégies de mise en œuvreest similaire au cas gaussien ; ils ne sont pas repris.

2 Modèle log-linéaire

2.1 Types de données

Les données se présentent généralement sous la forme d'une table de contingence obtenue par le croisement de plusieurs variables qualitatives et dont chaque cellule contient un effectif ou une fréquence à modéliser. Nous nous limiterons à l'étude d'une table élémentaire en laissant de côté des structures plus complexes, par exemple lorsque des zéros structurels, des indépendances conditionnelles, des propriétés de symétrie ou quasi-symétrie, une table creuse, sont à prendre en compte. D'autre part, sous sa forme la plus générale, le modèle peut intégrer également des variables quantitatives.

Ce type de situation se retrouve en analyse des correspondances simple ou multiple mais ici, l'objectif est d'expliquer ou de modéliser les effectifs en fonction des modalités prises par les variables qualitatives. L'objectif final pouvant être *explicatif* : tester une structure de dépendance particulière, ou *prédictif* avec choix d'un modèle parcimonieux.

2.2 Distributions

On considère la table de contingence complète constituée à partir de l'observation des variables qualitatives $X^1, X^2, \ldots, X^p$ sur un échantillon de n individus. Les effectifs $\{y_{jk\ldots l}; j = 1, J; k = 1, K; \ldots; l = 1, L\}$ de chaque cellule sont rangés dans un vecteur $\mathbf{y}$ à $I(I = J \times K \times \cdots \times L)$ composantes. Différentes hypothèses sur les distributions sont considérées en fonction du contexte expérimental.

Poisson

Le modèle le plus simple consiste à supposer que les variables observées Y_i suivent des lois de Poisson indépendantes de paramètre $\mu_i = E(Y_i)$. La distribution conjointe admet alors pour densité :

$$f(\mathbf{y}, \mu) = \prod_{i=1}^{I} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}.$$

La somme $N(N = y_+ = \sum_i y_i)$ des I variables aléatoires de Poisson indépendantes est également une variable de Poisson de paramètre $\mu_+ = \sum_i \mu_i$.

Multinomiale

En pratique, le nombre total n d'observations est souvent fixé a priori par l'expérimentateur et ceci induit une contrainte sur la somme des y_i. La distribution conjointe des variables Y_i est alors conditionnée par n et la densité devient :

$$f(\mathbf{y}, \mu) = \prod_{i=1}^{I} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \Big/ \frac{\mu_+^n e^{-\mu_+}}{n!}.$$

Comme $\mu_+^n = \sum_i \mu_+^{y_i}$ et $e^{-\mu_+} = \prod_i e^{-\mu_i}$, en posant $\pi_i = \frac{\mu_i}{\mu_+}$, on obtient :

$$f(\mathbf{y}, \mu) = n! \prod_{i=1}^{I} \frac{\pi_i^{y_i}}{y_i!} \quad \text{avec} \quad \sum_{i=1}^{I} \pi_i = 1 \text{ et } 0 \leq \pi_i \leq 1; i = 1, I.$$

On vérifie donc que $f(\mathbf{y}, \mu)$ est la fonction de densité d'une loi multinomiale dans laquelle les paramètres π_i modélisent les probabilités d'occurrence associées à chaque cellule. Dans ce cas, $E(Y_i) = n\pi_i$.

Produit de multinomiales

Dans d'autres circonstances, des effectifs marginaux lignes, colonnes ou sous-tables, peuvent être également fixés par l'expérimentateur comme dans le cas d'un sondage stratifié. Cela correspond au cas où une ou plusieurs variables sont contrôlées et ont donc un rôle explicatif ; leurs modalités sont connues *a priori*. Les lois de chacun des sous-éléments de la table, conditionnées par l'effectif marginal correspondant sont multinomiales. La loi conjointe de l'ensemble est alors un produit de multinomiales.

Conséquence

Trois modèles de distribution : Poisson, multinomial, produit de multinomiales, sont envisageables pour modéliser Y_i en fonction des conditions expérimentales. D'un point de vue théorique, on montre que ces modèles conduisent aux mêmes estimations des paramètres par maximum de vraisemblance. La différence introduite par le conditionnement intervient par une contrainte qui impose la présence de certains paramètres dans le modèle, ceux reconstruisant les marges fixées.

2.3 Modèles à 2 variables

Soit une table de contingence $(J \times K)$ issue du croisement de deux variables qualitatives X^1 à J modalités et X^2 à K modalités et dont l'effectif total n est fixé. La loi conjointe des effectifs Y_{jk} de chaque cellule est une loi multinomiale de paramètre π_{jk} et d'espérance :

$$E(Y_{jk}) = n\pi_{jk}.$$

Par définition, les variables X^1 et X^2 sont *indépendantes* si et seulement si :

$$\pi_{jk} = \pi_{+k}\pi_{j+}$$

où π_{j+} (resp. π_{+k}) désigne la loi marginale de X^1 (resp. X^2) :

$$\pi_{j+} = \sum_{k=1}^{K} \pi_{jk} \quad \text{et} \quad \pi_{+k} = \sum_{j=1}^{J} \pi_{jk}.$$

Si l'indépendance n'est pas vérifiée, on peut décomposer :

$$E(Y_{jk}) = n\pi_{jk} = n\pi_{j+}\pi_{+k}\frac{\pi_{jk}}{\pi_{j+}\pi_{+k}}.$$

Notons $\eta_{jk} = \ln(E(Y_{jk}))$. L'intervention de la fonction logarithme permet de linéariser la décomposition précédente autour du "modèle d'indépendance" :

$$\eta_{jk} = \ln n + \ln \pi_{j+} + \ln \pi_{+k} + \ln\left(\frac{\pi_{jk}}{\pi_{j+}\pi_{+k}}\right).$$

Ce modèle est dit *saturé* car, présentant autant de paramètres que de données, il explique exactement celles-ci. L'indépendance est vérifiée si le dernier terme de cette expression, exprimant une dépendance ou interaction comme dans le modèle d'analyse de variance, est nul pour tout couple (j, k).

Les logiciels mettent en place d'autres paramétrisations en faisant apparaître des effets différentiels, soit par rapport à une moyenne, soit par rapport à la dernière modalité.

Dans le premier cas, en posant :

$$\begin{aligned}
\beta_0 &= \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\eta_{jk} = \eta_{..},\\
\beta_j^1 &= \frac{1}{K}\sum_{k=1}^{K}\eta_{jk} - \eta_{..} = \eta_{j.} - \eta_{..},\\
\beta_k^2 &= \frac{1}{J}\sum_{j=1}^{J}\eta_{jk} - \eta_{..} = \eta_{.k} - \eta_{..},\\
\beta_{jk}^{12} &= \eta_{jk} - \eta_{j.} - \eta_{.k} + \eta_{..},
\end{aligned}$$

avec les relations :

$$\forall j, \forall k, \sum_{j=1}^{J}\beta_j^1 = \sum_{k=1}^{K}\beta_k^2 = \sum_{j=1}^{J}\beta_{jk}^{12} = \sum_{k=1}^{K}\beta_{jk}^{12} = 0,$$

le modèle saturé s'écrit :

$$\ln(E(Y_{jk})) = \eta_{jk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_{jk}^{12}.$$

Il se met sous la forme matricielle

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

où $\mathbf{X}$ est la matrice expérimentale (design matrix) contenant les indicatrices. L'indépendance est obtenue lorsque tous les termes d'interaction β_{jk}^{12} sont nuls.

La deuxième paramétrisation considère la décomposition :

$$\pi_{jk} = \pi_{JK}\frac{\pi_{Jk}}{\pi_{JK}}\frac{\pi_{jK}}{\pi_{JK}}\frac{\pi_{jk}\pi_{JK}}{\pi_{Jk}\pi_{jK}}.$$

En posant :

$$\begin{aligned}
\beta_0 &= \ln n + \ln \pi_{JK},\\
\beta_j^1 &= \ln \pi_{jK} - \ln \pi_{JK},\\
\beta_k^2 &= \ln \pi_{Jk} - \ln \pi_{JK},\\
\beta_{jk}^{12} &= \ln \pi_{jk} - \ln \pi_{jK} - \ln \pi_{Jk} + \ln \pi_{JK},
\end{aligned}$$

avec les mêmes relations entre les paramètres. Le modèle se met encore sous la forme :

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

et se ramène à l'indépendance si tous les paramètres β_{jk}^{12} sont nuls.

Si l'hypothèse d'indépendance est vérifiée, on peut encore analyser les effets principaux :

$$\text{si}, \forall j, \beta_j^1 = 0 \quad \text{alors}, \quad \pi_{jk} = \pi_{Jk} = \frac{1}{J}\pi_{+k}.$$

Il y a équiprobabilité des modalités de X^1. Même chose avec X^2 si les termes β_k^2 sont tous nuls.

Les paramètres du modèle log-linéaire sont estimés en maximisant la log-vraisemblance dont l'explicitation est reportée au chapitre suivant comme cas particulier de modèle linéaire généralisé. Pour les modèles simples, les estimations sont déduites des effectifs marginaux mais comme, dès que le modèle est plus compliqué, des méthodes itératives sont nécessaires, elles sont systématiquement mises en œuvre.

2.4 Modèle à trois variables

On considère une table de contingence $(J \times K \times L)$ obtenue par croisement de trois variables qualitatives X^1, X^2, X^3. La définition des paramètres est conduite de manière analogue au cas de deux variables en faisant apparaître des effets principaux et des interactions. Le modèle saturé se met sous la forme :

$$\ln(E(Y_{jkl})) = \eta_{jkl} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 + \beta_{jk}^{12} + \beta_{jl}^{13} + \beta_{kl}^{23} + \beta_{jkl}^{123}$$

et peut aussi est présenté sous forme matricielle.

Nous allons expliciter les sous-modèles obtenus par nullité de certains paramètres et qui correspondent à des structures particulières d'indépendance. Une façon classique de nommer les modèles consiste à ne citer que les interactions retenues les plus complexes. Les autres, ainsi que les effets principaux, sont contenues de par la structure hiérarchique du modèle. Ainsi, le modèle saturé est désigné par $(X^1X^2X^3)$ correspondant à la syntaxe `X1 | X2 | X3` de SAS.

Cas poissonnien ou multinomial

Seul le nombre total d'observations n est fixé dans le cas multinomial, ceci impose simplement la présence de β_0 dans le modèle.

i. Modèle partiel d'association ou de tout interaction d'ordre 2 : (X^1X^2, X^2X^3, X^1X^3)
Les termes β_{jkl}^{123} sont tous nuls, seules les interactions d'ordre 2 sont présentes. C'est le modèle implicitement considéré par l'analyse multiple des correspondances. Il s'écrit :

$$\eta_{jk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 + \beta_{jk}^{12} + \beta_{jl}^{13} + \beta_{kl}^{23}.$$

ii. Indépendance conditionnelle : (X^1X^2, X^1X^3)
Si, en plus, l'un des termes d'interaction est nul, par exemple $\beta_{kl} = 0$ pour tout couple (k, l), on dit que X^2 et X^3 sont indépendantes conditionnellement à X^1 et le modèle devient :

$$\eta_{jk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 + \beta_{jk}^{12} + \beta_{jl}^{13}.$$

iii. Variable indépendante : (X^1, X^2X^3)
Si deux termes d'interaction sont nuls : $\beta_{jl}\beta_{jk} = 0$ pour tout triplet (j, k, l), alors X^1 est indépendante de X^2 et X^3.

$$\eta_{jk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 + \beta_{kl}^{23}.$$

iv. Indépendance : (X^1, X^2, X^3)
Tous les termes d'interaction sont nuls :

$$\eta_{jk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3$$

et les variables sont mutuellement indépendantes.

Produit de multinomiales

- Si une variable est explicative, par exemple X^3, ses marges sont fixées, le modèle doit nécessairement conserver les paramètres
$$\eta_{jk} = \beta_0 + \beta_l^3 + \cdots$$
- Si deux variables sont explicatives, par exemple X^2 et X^3, le modèle doit conserver les termes :
$$\eta_{jk} = \beta_0 + \beta_k^2 + \beta_l^3 + \beta_{kl}^{23} + \cdots$$

La généralisation à plus de trois variables ne pose pas de problème théorique. Les difficultés viennent de l'explosion combinatoire du nombre de termes d'interaction et de la complexité des structures d'indépendance. D'autre part, si le nombre de variables est grand, on est souvent confronté à des tables de contingence creuses (beaucoup de cellules vides) qui rendent défaillant le modèle log-linéaire. Une étude exploratoire (correspondances multiples par exemple) préalable est nécessaire afin de réduire le nombre des variables considérées et celui de leurs modalités.

3 Choix de modèle

3.1 Recherche pas à pas

Principalement deux critères (test du rapport de vraisemblance et test de Wald), décrits en annexe pour un cadre plus général, sont considérés. Ces critères sont utilisés comme le test de Fisher du modèle linéaire gaussien. Ils permettent de comparer un modèle avec un sous-modèle et d'évaluer l'intérêt de la présence des termes complémentaires. On suit ainsi une stratégie descendante à partir du modèle complet ou saturé dans le cas du modèle log-linéaire. L'idée est de supprimer, un terme à la fois, la composante d'interaction ou l'effet principal qui apparaît comme le moins significatif au sens du rapport de vraisemblance ou du test de Wald. Les tests présentent une structure hiérarchisée. SAS facilite cette recherche en produisant une décomposition (Type III) de ces indices permettant de comparer chacun des sous-modèles excluant un des termes avec le modèle les incluant tous.

Attention, du fait de l'utilisation d'une transformation non linéaire (log), même si des facteurs sont orthogonaux, aucune propriété d'orthogonalité ne peut être prise en compte pour l'étude des hypothèses. Ceci impose l'élimination des termes un par un et la ré-estimation du modèle. D'autre part, un terme principal ne peut être supprimé que s'il n'intervient plus dans des termes d'interaction. Enfin, selon les conditions expérimentales qui peuvent fixer les marges d'une table de contingence, la présence de certains paramètres est imposée dans un modèle log-linéaire.

4 Exemples

4.1 Modèle poissonien

On étudie les résultats d'une étude préalable à la législation sur le port de la ceinture de sécurité dans la province de l'Alberta à Edmonton au Canada (Jobson, 1991). Un échantillon de 86 769 rapports d'accidents de voitures ont été compulsés afin d'extraire une table croisant :

i. Etat du conducteur : Normal ou Alcoolisé
ii. Port de la ceinture : Oui Non
iii. Gravité des blessures : 0 : rien à 3 : fatales

La procédure `genmod` est utilisée :

```
proc genmod data=sasuser.ceinture;
class co ce b ;
model effectif=co|ce|b @2  /type3 obstats dist=poisson;
run;
```

Une extraction des résultats donnent :

```
        Criteria For Assessing Goodness Of Fit
Criterion             DF         Value      Value/DF
Deviance               3        5.0136        1.6712

         LR Statistics For Type 3 Analysis
         Source       DF   ChiSquare  Pr>Chi

         CO            1   3431.0877  0.0001
         CE            1   3041.5499  0.0001
         CO*CE         1    377.0042  0.0001
         B             3  28282.8778  0.0001
         CO*B          3    474.7162  0.0001
         CE*B          3     42.3170  0.0001

           Analysis Of Parameter Estimates
Parameter          DF    Estimate     Std Err   ChiSquare  Pr>Chi
INTERCEPT           1      3.6341      0.1550    550.0570  0.0001
CO         A        1     -2.2152      0.1438    237.3628  0.0001
CE         N        1      1.8345      0.1655    122.8289  0.0001
```

```
    CO*CE      A  N     1      0.9343      0.0545    293.9236  0.0001
    B          0        1      5.7991      0.1552   1396.7752  0.0001
    B          1        1      2.7848      0.1598    303.6298  0.0001
    B          2        1      2.1884      0.1637    178.7983  0.0001
    CO*B       A  0     1     -1.4622      0.1354    116.5900  0.0001
    CO*B       A  1     1     -0.6872      0.1423     23.3154  0.0001
    CO*B       A  2     1     -0.5535      0.1452     14.5293  0.0001
    CE*B       N  0     1     -0.2333      0.1658      1.9807  0.1593
    CE*B       N  1     1     -0.0902      0.1708      0.2786  0.5976
    CE*B       N  2     1      0.0741      0.1748      0.1799  0.6715

                        Observation Statistics
EFFECTIF        Pred       Xbeta        Std     HessWgt      Lower      Upper
   12500       12497      9.4332   0.008930       12497      12280      12718
     604    613.3370      6.4189     0.0395    613.3370   567.6707   662.6770
     344    337.8089      5.8225     0.0530    337.8089   304.5010   374.7601
      38     37.8677      3.6341     0.1550     37.8677    27.9495    51.3053
   61971       61974     11.0345   0.004016       61974      61488      62464
   ...
```

Les résultats montrent que le modèle de toute interaction d'ordre 2 est acceptable (déviance) et il semble que tous les termes soient nécessaires, toutes les interactions doivent être présentes au sens du test de Wald.

Chapitre 5

Qualité de prévision

1 Introduction

La performance du modèle issu d'une méthode d'apprentissage s'évalue par sa *capacité de prévision* dite encore de *capacité de généralisation* dans la communauté informatique. La mesure de cette performance est très importante puisque, d'une part, elle permet d'opérer une *sélection de modèle* dans une famille associée à la méthode d'apprentissage utilisée et, d'autre part, elle guide le *choix de la méthode* en comparant chacun des modèles optimisés à l'étape précédente. Enfin, elle fournit, tous choix faits, une mesure de la qualité ou encore de la *confiance* que l'on peut accorder à la prévision en vue même, dans un cadre légal, d'une *certification*.

En dehors d'une situation expérimentale planifiée classique en Statistique, c'est-à-dire sans le secours de *modèles probabilistes*, c'est le cas, par principe, du *data mining*, trois types de stratégies sont proposés :

i. un partage de l'échantillon (apprentissage, validation, test) afin de distinguer estimation du modèle et estimations de l'erreur de prévision,

ii. une pénalisation de l'erreur d'ajustement faisant intervenir la complexité du modèle,

iii. un usage intensif du calcul (computational statistics) par la mise en œuvre de simulations.

Le choix dépend de plusieurs facteurs dont la taille de l'échantillon initial, la complexité du modèle envisagé, la variance de l'erreur, la complexité des algorithmes c'est-à-dire le volume de calcul admissible.

Pour répondre aux objectifs de la 2ème stratégie adaptée à un échantillon d'effectif trop restreint pour être éclater en trois partie, différents critères sont utilisées pour définir une qualité de modèle à fin prédictive.

- Le plus ancien est naturellement une estimation d'une *erreur de prévision* : risque quadratique ou taux de mal classés, comme mesure d'une distance moyenne entre le "vrai" ou le "meilleur" modèle et celui considéré. Ce risque quadratique se décomposant grossièrement en un carré de biais et une variance, l'enjeu est de trouver un bon compromis entre ces deux composantes en considérant un modèle parcimonieux.
- D'autres critères sont basés sur la dissemblance de Kullback entre mesure de probabilités. Ce critère mesure la qualité d'un modèle en considérant la dissemblance de Kullback entre la loi de la variable expliquée Y et celle de sa prévision $\widehat{Y}$ fournie par un modèle.
- La dernière approche enfin, issue de la théorie de l'apprentissage de Vapnik (1999), conduit à proposer une majoration de l'erreur de prévision ou risque ne faisant pas intervenir la loie conjointe inconnue ou des considérations asymtotiques mais une mesure de la compléxité du modèle appelée *dimension de Vapnik-Chernovenkis*.

Les travaux de Vapnik en théorie de l'apprentissage ont conduit à focaliser l'attention sur la présence ou l'absence de propriétés théoriques basiques d'une technique d'apprentissage ou de modélisation :

consistence qui garantit la capacité de généralisation. Un processus d'apprentissage est dit *consistant* si l'erreur sur l'ensemble d'apprentissage et l'erreur sur un jeu de données test convergent en probabilité vers la même limite lorsque la taille de l'échantillon d'apprentissage augmente.

vitesse de convergence. Une évaluation, quand elle est possible, de la vitesse de convergence de l'erreur

lorsque la taille augmente, est une indication sur la façon dont la généralisation s'améliore et informe sur la nature des paramètres, comme le nombre de variables explicatives, dont elle dépend.

contrôle Est-il possible, à partir d'un échantillon d'apprentissage de taille fini donc sans considérations asymtotiques, de contrôler la capacité de généralisation et donc de majorer le terme d'erreur de prévision ou risque ?

Une estimation de la qualité de la prévision est donc un élément central de la mise en place de la stratégie du *data mining*, telle qu'elle est décrite dans l'introduction (cf. chapitre 1 section 4) mais aussi dans beaucoup de disciplines concernées par la modélisation statistique. Le point important à souligner est que le "meilleur" modèle en un sens prédictif n'est pas nécessairement celui qui ajuste le mieux les données (cas de surajustement) ni même le "vrai" modèle si la variance des estimations est importante.

2 Erreur de prévision

2.1 Définition

Soit Y la variable à prédire, X la variable p-dimensionnelle ou l'ensemble des variables explicatives, F la loi conjointe de Y et X, $\mathbf{z} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ un échantillon et

$$Y = \phi(X) + \varepsilon$$

le modèle à estimer avec $E(\varepsilon) = 0, \text{Var}(\varepsilon) = \sigma^2$ et ε indépendant de X ; X, comme chacun des $\mathbf{x}_i$, est de dimension p.

L'erreur de prévision est définie par

$$\mathcal{E}_P(\mathbf{z}, F) = E_F[Q(Y, \widehat{\phi}(X))]$$

où Q est une *fonction perte*.

Si Y est quantitative, cette fonction perte est le plus généralement quadratique : $Q(y, \widehat{y}) = (y - \widehat{y})^2$, mais utilise parfois la valeur absolue : $Q(y, \widehat{y}) = |y - \widehat{y}|$. Cette dernière à l'avantage d'être plus *robuste*, car moins sensible aux valeurs extrêmes, mais nécessite des algorithmes d'optimisation plus complexes et pas nécessairement à solution unique.

Si Y est qualitative Q est une indicatrice de mal classé : $Q(y, \widehat{y}) = \mathbf{1}_{\{y \neq \widehat{y}\}}$.

Dans le cas quantitatif, l'estimation du modèle par minimisation de $\mathcal{E}_P$ revient à une approximation de la fonction ϕ et la solution est l'*espérance conditionnelle* (connaissant l'échantillon) tandis que, dans la cas qualitatif, c'est la classe la plus probable désignée par le *mode conditionnel* qui est prédite.

2.2 Décomposition

L'erreur de prévision se décompose dans le cas quantitatif[1]. Considérons celle-ci en un point $\mathbf{x}_0$.

$$\begin{aligned} \mathcal{E}_P(\mathbf{x}_0) &= E_F[(Y - \widehat{\phi}(\mathbf{x_0}))^2 \mid X = \mathbf{x}_0] \\ &= \sigma^2 + [E_F\widehat{\phi}(\mathbf{x}_0) - \phi(x)]^2 + E_F[\widehat{\phi}(\mathbf{x}_0) - E_F\widehat{\phi}(\mathbf{x}_0)]^2 \\ &= \sigma^2 + \text{Biais}^2 + \text{Variance}. \end{aligned}$$

Très généralement, plus un modèle (la famille des fonctions ϕ admissibles) est complexe, plus il est flexible et peu s'ajuster aux données observées et donc plus le biais est réduit. En revanche, la partie variance augmente avec le nombre de paramètres à estimer et donc avec cette complexité. L'enjeu, pour minimiser le risque quadratique ainsi défini, est donc de rechercher un meilleur compromis entre biais et variance : accepter de biaiser l'estimation comme par exemple en régression *ridge* pour réduire plus favorablement la variance.

[1] Plusieurs décompositions concurentes ont été proposées dans le cas qualitatif mais leur explicitation est moins claire.

2.3 Estimation

Le premier type d'estimation à considérer exprime la qualité d'ajustement du modèle sur l'échantillon observé. C'est justement, dans le cas quantitatif, ce critère qui est minimisé dans la recherche de moindres carrés. Ce ne peut être qu'une estimation biaisée, car trop *optimiste*, de l'erreur de prévision ; elle est liée aux données qui ont servi à l'ajustement du modèle et est d'autant plus faible que le modèle est complexe. Cette estimation ne dépend que de la partie "biais" de l'erreur de prévision et ne prend pas en compte la partie "variance" de la décomposition.

Cette estimation est notée :

$$\widehat{\mathcal{E}_P} = \frac{1}{n}\sum_{i=1}^{n} Q(y_i, \widehat{\phi}(\mathbf{x}_i)).$$

C'est simplement le *taux de mal classés* dans le cas qualitatif. Des critères de risque plus sophistiqués sont envisagés dans un contexte bayésien si des probabilités *a priori* sont connues sur les classes ou encore des coûts de mauvais classement (cf. chapitre 6).

La façon la plus simple d'estimer sans biais l'erreur de prévision consiste à calculer $\widehat{\mathcal{E}_P}$ sur un échantillon indépendant n'ayant pas participé à l'estimation du modèle. Ceci nécessite donc d'éclater l'échantillon en trois parties respectivement appelées *apprentissage*, *validation* et *test* :

$$\mathbf{z} = \mathbf{z}_{\text{Appr}} \cup \mathbf{z}_{\text{Valid}} \cup \mathbf{z}_{\text{Test}}.$$

i. $\widehat{\mathcal{E}_P}(\mathbf{z}_{\text{Appr}})$ est minimisée pour estimer un modèle,

ii. $\widehat{\mathcal{E}_P}(\mathbf{z}_{\text{Valid}})$ sert à la comparaison des modèles au sein d'une même famille afin de sélectionner celui qui minimise cette erreur,

iii. $\widehat{\mathcal{E}_P}(\mathbf{z}_{\text{Test}})$ est utilisée pour comparer entre eux les meilleurs modèles de chacune des méthodes considérées.

Cette solution n'est acceptable que si la taille de l'échantillon initiale est importante sinon :

- la qualité d'ajustement est dégradée car n est plus petit,
- la variance de l'estimation de l'erreur peut être importante et ne peut être estimée.

Si la taille de l'échantillon est insuffisante, le point ii ci-dessus : la sélection de modèle est basée sur un autre type d'estimation de l'erreur de prévision faisant appel soit à une pénalisation soit à des simulations.

3 Estimation avec pénalisation

3.1 C_p de Mallows

Le C_p de Mallows fut, historiquement, le premier critère visant à une meilleure estimation de l'erreur de prévision que la seule considération de l'erreur d'ajustement (ou le R^2) dans le modèle linéaire. Il repose sur une mesure de la qualité sur la base d'un risque quadratique. L'erreur de prévision se décompose en :

$$\mathcal{E}_P = \widehat{\mathcal{E}_P}(\mathbf{z}_{\text{Appr}}) + \text{Optim}$$

qui est l'estimation par resubstitution ou taux d'erreur apparent plus le biais par abus d'optimisme. Il s'agit donc d'estimer cette optimisme pour apporter une correction et ainsi une meilleure estimation de l'erreur recherchée. cette correction peut prendre plusieurs formes. Elle est liée à l'estimation de la variance dans la décomposition en biais et variance de l'erreur ou c'est encore une pénalisation associée à la complexité du modèle.

Son expression est détaillée dans le cas de la régression linéaire chapitre 2. On montre (cf. Hastie et col. 2001), à des fins de comparaison qu'il peut aussi se mettre sous une forme équivalente :

$$C_p = \widehat{\mathcal{E}_P} + 2\frac{d}{n}s^2$$

où d est le nombre de paramètres du modèles (nombre de variables plus un)), n le nombre d'observations, s^2 une estimation de la variance de l'erreur par un modèle de faible biais. Ce dernier point est fondamental pour la qualité du critère, il revient à supposer que le modèle complet (avec toutes les variables) est le "vrai" modèle ou tout du moins un modèle peu biaisé afin de conduire à une bonne estimation de de σ^2.

3.2 AIC, AIC$_c$, BIC

Contrairement au C_p associé à un risque quadratique, le critère d'information d'Akaïke (AIC) découle d'une expression de la qualité du modèle basée sur la dissemblance de Kullback. Il se présente sous une forme similaire mais plus générale que le C_p de Mallows. Il s'applique en effet à tout modèle estimé par maximisation d'une log-vraisemblance $\mathcal{L}$ et suppose que la famille de densités considérées pour modéliser la loi de Y contient la "vraie" densité de Y.

Après quelques développements incluant de nombreuses approximations (estimation de paramètres par maximum de vraisemblance, propriétés asymtotiques, formule de Taylor), le critère d"Akaïke se met sous la forme :

$$\text{AIC} = -2\mathcal{L} + 2\frac{d}{n}.$$

Dans le cas gaussien en supposant la variance connue, moindres carrés et déviance coincident, AIC est équivalent au C_p. Ce critère possède une version plus raffinée (AIC$_c$) dans le cas gaussien et plus particulièrement adaptée aux petits échantillons et asymptotiquement équivalente lorsque n est grand.

$$\text{AIC} = -2\mathcal{L} + \frac{n+d}{n-d-2}.$$

Une argumentation de type bayésien conduit à un autre critère BIC (*Bayesian information criterion*) qui cherche, approximativement (asymptotiquement), le modèle associé à la plus grande probabilité *a posteriori*. Dans le cas d'un modèle issu de la maximisation d'une log-vraisemblance, il se met sous la forme :

$$\text{BIC} = -2\mathcal{L} + \log(n)\frac{d}{n}.$$

On montre, dans le cas gaussien et en supposant la variance connue que BIC est proportionnel à AIC avec le facteur 2 remplacé par $\log n$. Ainsi, dès que $n > e^2 \approx 7,4$, BIC tend à pénaliser plus lourdement les modèles complexes. Asymptotiquement, on montre que la probabilité pour BIC de choisir le bon modèle tend vers 1 lorsque n tend vers l'infini. Ce n'est pas le cas d'AIC ni du C_p qui tendent alors à choisir des modèles trop complexes. Néanmoins à taille fini, petite, BIC risque de se limiter à des modèles trop simples.

Quelque-soit le critère adopté, il est facile de choisir le modèle présentant le plus faible AIC, AIC$_c$ ou BIC parmi ceux considérés. Globalement, si l'estimation du modèle découle d'une maximisation de la vraisemblance, estimation et choix de modèle reviennent à minimiser un critère de vraisemblance pénalisée s'écrit sous la forme :

$$\text{Crit} = f(\text{Vraisemblance}) + \text{Pénalisation}(d)$$

où f est une fonction décroissante de la vraisemblance ($-\log$) et la pénalisation une fonction croissante de la complexité du modèle.

Les critères ci-dessus ont pour la plupart été définis dans le cadre du modèle classique de régression multiple pour lequel il existe de nombreuses références et certains été généralisés ou adaptés à d'autres méthodes en étendant la notion de nombre de degrés de libertés à des situations où le nombre de paramètres du modèle n'est pas explicite (lissage ou régularisation).

Ainsi, pour les modèles non-linéaires voire plus complexes (non-paramétriques en dimension infinie), le nombre d de paramètres doit être remplacé par une mesure de complexité $p(\alpha)$. Par exemple, les modèles linéaires se mettent sous une forme : $\widehat{\boldsymbol{y}} = \mathbf{H}\mathbf{y}$ en incluant les méthodes de régularisation (*ridge*) ou de lissage (spline) où la matrice $\mathbf{H}$ dépend uniquement des $\mathbf{x}_i$. Dans ce cas, le nombre effectif de paramètres est défini comme la trace de la matrice $\mathbf{H}$: $d(\mathbf{H}) = \text{tr}(\mathbf{H})$. C'est encore d, le rang de $\mathbf{X}$ c'est-à-dire le nombre vecteurs de base (le nombre de variables + 1) si $\mathbf{H}$ est une matrice de projection orthogonale. Dans d'autres situations (perceptron), ce nombre de paramètres est plus difficile à contrôler car il fait intervenir les valeurs propres d'une matrice hessienne.

3.3 Dimension de Vapnik-Chernovenkis

Cet indicateur mesure la *complexité* d'une famille de fonctions candidates à la définition un modèle de prévision. Cette complexité est basée sur le pouvoir *séparateur* de la famille de fonction.

Considérons un échantillon $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ de $\mathbb{R}^p$. Il existe 2^n différentes manières de séparer cet échantillon en deux sous-échantillons. Par définition, on dit qu'un ensemble F de fonctions *hache* ou mieux *pulvérise* (shatters) l'échantillon si les 2^n séparations peuvent être construites par différents représentants de la famille de fonction F. Ainsi, par exemple, pour $p = 2$, les fonctions linéaires (droites) peuvent pulvériser 3 points mais pas 4.

DÉFINITION 5.1. — *Un ensemble de fonctions définis de $\mathbb{R}^p$ dans $\mathbb{R}$ est dit de VC dimension (Vapnik-Chernovenkis) h si :*

- *tout jeu de h vecteurs de $\mathbb{R}^p$ peut être pulvérisé.*
- *Aucun ensemble de $h + 1$ vecteurs ne peut être pulvérisé par cet ensemble de fonctions.*

Exemples

- La VC dimension de l'ensemble des hyperplans dans $\mathbb{R}^p$ est $p + 1$.
- La VC dimension de l'ensemble des fonctions $f(\mathbf{x}, w) = \text{sign}(\sin(w, x))$ avec $0 < c < x < 1$ où w est un paramètre libre, est infinie.
- La VC dimension de l'ensemble des indicatrices linéaires

$$f(\mathbf{x}, w) = \text{sign}\left(\sum_{j=1}^{p}(w_j x_j) + 1\right) \quad \text{avec } \|\mathbf{x}\| = 1$$

et satisfaisant la condition :

$$\|\boldsymbol{w}\|^2 = \sum_{j=1}^{p} w_j^2 \leq C$$

dépend de la constante C et peut prendre toutes les valeurs de 0 à p.

Attention, les VC dimensions ne sont pas égales au nombre de paramètres libres et sont souvent difficiles à exprimer pour une famille de fonctions données.

Vapnik (1999) prouve des résultats fondamentaux pour la théorie de l'apprentissage :

- Un processus d'apprentissage est consistant si et seulement si la famille de modèles considérés a une VC dimension h finie.
- La majoration de la différence entre l'erreur d'apprentissage (ou par resubstitution ou erreur apparente) et l'erreur de prévision dépend du rapport entre la VC dimension h et la taille n de l'ensemble d'apprentissage.
- L'inégalité de Vapnik, qui s'écrit sous une forme analogue à un intervalle de confiance, permet de contrôler l'erreur de prévision ou risque. Avec une probabilité $1 - rho$:

$$\mathcal{E}_P < \widehat{\mathcal{E}_P} + \sqrt{\frac{h(\log(\frac{2n}{h}) + 1) - \log\frac{\rho}{4}}{n}}.$$

Il est important de souligner que cette inégalité ne fait pas intervenir le nombre de variables explicatives p mais le rapport n/h. Elle ne fait pas intervenir non plus la loi conjointe inconnue du couple (Y, X). Le deuxième terme est grand (mauvaise précision) lorsque le rapport n/h est faible dû à une trop grande VC dimension et donc une famille de modèles trop complexe.

En pratique, il est important de minimiser simultanément les deux termes de l'inéquation. La stratégie à adopter est le *principe de minimisation structurée du risque* (SRM) qui consiste à faire de la VC dimension h une *variable contrôlée*. Ayant défini uen séquence ou structure de modèles emboîtés au sens de la VC dimension :

$$S_1 \subset S_2 \subset \cdots \subset S_k \quad \text{si les VC dimensions associées vérifient :} \quad h_1 < h_2 < \cdots < h_k.$$

Il s'agit de trouver la valeur h rendant le risque minimum et donc fournissant le meilleur compromis entre les deux termes de l'inégalité de Vapnik.

La complexité de la famille des modèles peut être controlée par différents paramètres de la technique d'apprentissage considérée : le nombre de neurones d'une couche dans un perceptron, le degré d'un polynôme, la contrainte sur les paramètres comme en régression ridge, une largeur de fenêtre ou paramètre de lissage...

4 Le cas spécifique de la discrimination

Les erreurs de prévisions précédentes ainsi que les critères de choix de modèles sont plus particulièrement adaptés à une situation de régression et donc une variable Y quantitative. Dans une situation de discrimination le seul critère de taux d'erreur de classement introduit précédemment n'est pas toujours bien adapté surtout, par exemple, dans le cadre de classes déséquilibrées : un modèle trivial qui ne prédit jamais une classe peu représentée ne commet pas un taux d'erreur supérieur au pourcentage de cette classe. Cette situation est souvent délicate à gérer et nécessite une pondérations des observations ou encore l'introduction de coûts de mauvais classement disymétrique afin de forcer le modèle à prendre en compte une petite classe.

4.1 Discrimination à deux classes

Dans le cas du problème le plus élémentaire à deux classes, d'autres critères sont proposés afin d'évaluer plus précisément une qualité de discrimination. La plupart des méthodes vues (régression logistique), ou à venir dans les chapitre qui suivent, évaluent, pour chaque individu i, un *score* ou une probabilité $\widehat{\pi}_i$ que cette individu prenne la modalité $Y = 1$ (ou succès, ou possession d'un actif, ou présence d'une maladie...). Cette probabilité ou ce score compris entre 0 et 1 est comparé avec une valeur seuil s fixée *a priori* (en général $0,5$) :

$$\text{Si} \quad \widehat{\pi}_i > c, \ \widehat{y}_i = 1 \quad \text{sinon} \quad \widehat{y}_i = 0.$$

Pour un échantillon de taille n dont l'observation de Y est connue ainsi que les scores $\widehat{\pi}_i$ fournis par un modèle, il est alors facile de construire la matrice dite de *confusion* croisant les modalités de la variable prédite au seuil s avec celles de la variable observée dans une table de contingence :

Prévision	Observation		Total
	$Y = 1$	$Y = 0$	
$\widehat{y}_i = 1$	$n_{11}(s)$	$n_{10}(s)$	$n_{1+}(s)$
$\widehat{y}_i = 0$	$n_{01}(s)$	$n_{00}(s)$	$n_{0+}(s)$
Total	n_+1	n_+0	n

Dans une situation classique de diagnostic médical ou en marketing les quantités suivantes sont considérées :

- Vrais positifs les $n_{11}(s)$ observations biens classées ($\widehat{y}_i = 1$ et $Y = 1$),
- Vrais négatifs les $n_{00}(s)$ observations biens classées ($\widehat{y}_i = 0$ et $Y = 0$),
- Faux négatifs les $n_{01}(s)$ observations mal classées ($\widehat{y}_i = 0$ et $Y = 1$),
- Faux positifs les $n_{10}(s)$ observations mal classées ($\widehat{y}_i = 1$ et $Y = 0$),
- Le taux d'erreur : $t(s) = \frac{n_{01}(s)+n_{10}(s)}{n}$,
- Le taux de vrais positifs ou *sensibilité* $= \frac{n_{11}(s)}{n_+1}$ ou taux de positifs pour les individus qui le sont effectivement,
- Le taux de vrais négatifs ou *spécificité* $= \frac{n_{00}(s)}{n_+0}$ ou taux de négatifs pour les individus qui le sont effectivement,
- Le taux de faux positifs $= 1 -$ Spécificité $= 1 - \frac{n_{00}(s)}{n_+0} = \frac{n_{10}(s)}{n_+0}$.

En revanche, en météorologie, d'autres taux sont utilisés :

- Le taux de bonnes prévisions : $H = \frac{n_{11}(s)}{n_{1+}(s)}$,
- Le taux de fausses alertes : $H = \frac{n_{10}(s)}{n_{+0}}$,
- Le score de Pierce : PSS$= H - F$, compris entre -1 et 1, évalue la qualité d'un modèle de prévision. Si ce score est supérieur à 0, le taux de bonnes prévisions est supérieur à celui des fausses alertes et plus il est proche de 1, meilleur est le modèle.

Le score de Pierce a été conçu pour la prévision d'évènements climatiques rares afin de pénaliser les modèles ne prévoyant jamais ces évènements ($H = 0$) ou encore générant trop de fausses alertes ($F = 1$). Le modèle idéal prévoyant tous les évènements critiques ($H = 1$) sans fausse alerte ($F = 0$). Des coûts de mauvais classement peuvent être introduits pour pondérer ce score.

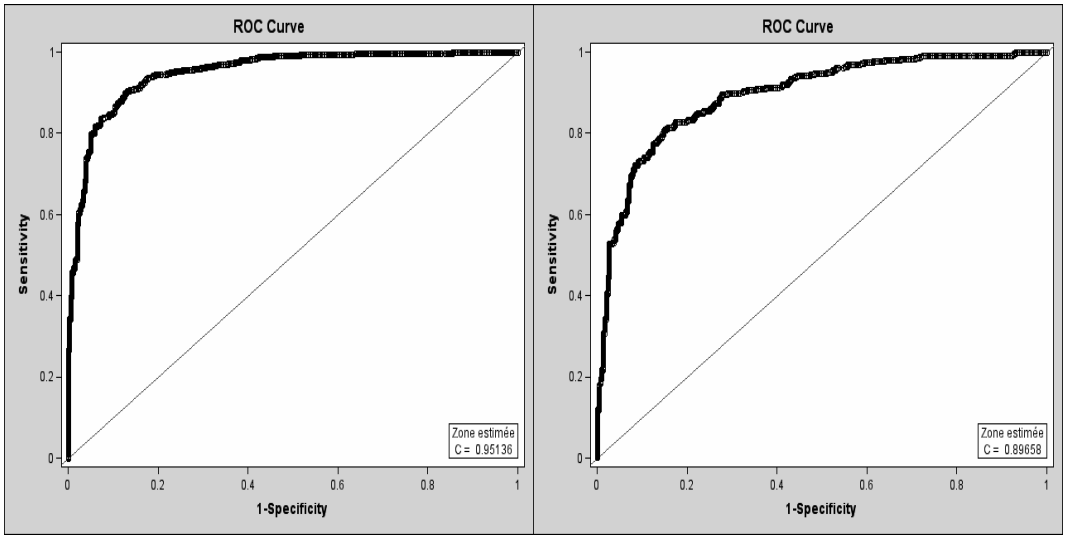

FIG. 5.1 – *Banque : Courbes ROC estimées sur l'échantillon d'apprentissage et sur l'échantillon test ainsi que les aires sous ces courbes (AUC).*

4.2 Courbe ROC et AUC

Les notions de *spécificité* et de *sensibilité* proviennent de la théorie du signal ; leurs valeurs dépendent directement de celle du seuil s. En augmentant s, la sensibilité diminue tandis que la spécificité augmente car la règle de décision devient plus exigeante ; un bon modèle associe grande sensibilité et grande spécificité pour la détection d'un "signal". Ce lien est représenté graphiquement par la courbe ROC (Receiver Operating Caracteristic) de la sensibilité ("probabilité" de détecter un vrai signal) en fonction de 1 moins la spécificité ("probabilité" de détecter un faux signal) pour chaque valeur s du seuil. Notons que la courbe ROC est une fonction monotone croissante :

$$1 - \frac{n_{00}(s)}{n_{+0}} < 1 - \frac{n_{00}(s')}{n_{+0}} \Rightarrow s < s' \Rightarrow \frac{n_{11}(s)}{n_{+1}} < \frac{n_{11}(s')}{n_{+1}}.$$

La figure 5.1 donne un exemple de courbes ROC pour associées au score d'appétence de la carte visa premier. Plus la courbe se rapproche du carré, meilleure est la discrimination, correspondant à la fois à une forte sensibilité et une grande spécificité. L'aire sous la courbe : AUC (*area under curve*) mesure la qualité de discrimination du modèle tandis qu'une analyse de la courbe aide au choix du seuil. Ainsi, dans l'exemple considéré, un seuil de $0,6$ ne pénalise pas énormément le nombre de positifs écartés tout en économisant des envois publicitaires par rapport à un seuil de $0,5$.

L'aire sous la courbe est calculée en considérant toutes les paires (i, i') formées d'un premier individu avec $y_i = 1$ et d'un second avec $y_{i'} = 0$. Une paire est dite concordante si $\widehat{\pi}_i > \widehat{\pi}_{i'}$; discordante sinon. Le nombre d'*ex æquo* est $n_{+0}n_{+1} - n_c - n_d$ où n_c est le nombre de paires concordantes et n_d le nombre de paires discordantes. Alors,

$$\text{AUC} = \frac{n_c + 0,5(n_{+0}n_{+1} - n_c - n_d)}{n_{+0}n_{+1}}.$$

On montre par ailleurs (voir par exemple Tenenhaus 2007) que le numérateur de cette expression est encore la Statistique de test de Mann-Whitney tandis que le coefficient de Gini, qui est le double de la surface entre la diagonale et la courbe vaut $2\text{AUC} - 1$.

Attention, pour comparer des modèles ou méthodes de complexités différentes, ces courbes doivent être estimées sur un échantillon test. Elles sont bien évidemment optimistes sur l'échantillon d'apprentissage. De plus, l'AUC ne donne pas un ordre total pour classer des modèles car les courbes ROC peuvent se croiser.

5 Estimation par simulation

La validation croisée est d'un principe simple, efficace et largement utilisée pour estimer une erreur moyennant un surplus de calcul. L'idée est d'itérer l'estimation de l'erreur sur plusieurs échantillons de *validation* puis d'en calculer la moyenne. C'est indispensable pour réduire la variance et ainsi améliorer la précision lorsque la taille de l'échantillon initial est trop réduite pour en extraire des échantillons de validation et test de taille suffisante.

Algorithm 2 Validation croisée

1: Découper aléatoirement l'échantillon en K parts (K-fold) de tailles approximativement égales selon une loi uniforme ;
2: **Pour** k=1 à K **Faire**
3: mettre de côté l'une des partie,
4: estimer le modèle sur les $K-1$ parties restantes,
5: calculer l'erreur sur chacune des observations qui n'ont pas participé à l'estimation
6: **Fin Pour**
7: moyenner toutes ces erreurs pour aboutir à l'estimation par validation croisée.

Plus précisément, soit $\tau : \{1,\dots,n\} \mapsto \{1,\dots,K\}$ la fonction d'indexation qui,pour chaque observation, donne l'attribution uniformément aléatoire de sa classe. L'estimation par *validation croisée* de l'erreur de prévision est :

$$\widehat{\mathcal{E}_{\text{CV}}} = \frac{1}{n}\sum_{i=1}^{n} Q(y_i, \widehat{\phi}^{(-\tau(i))}(x_i))$$

où $\widehat{\phi}^{(-k)}$ désigne l'estimation de ϕ sans prendre en compte la $k$ième partie de l'échantillon.

Le choix $K = 10$ est le plus courant, c'est souvent celui par défaut des logiciels (Splus). Historiquement, la validation croisée a été introduite par Allen avec $K = n$ (*delete-one cross validation*). Ce dernier choix n'est possible que pour n relativement petit à cause du volume des calculs nécessaires et l'estimation de l'erreur présente une variance souvent importante car chacun des modèles estimés est trop similaire au modèle estimé avec toutes les observations. En revanche, si K est petit (*i.e.* $K = 5$), la variance sera plus faible mais le biais devient un problème dépendant de la façon dont la qualité de l'estimation se dégrade avec la taille de l'échantillon.

Minimiser l'erreur estimée par validation croisée est une approche largement utilisée pour optimiser le choix d'un modèle au sein d'une famille paramétrée. $\widehat{\phi}$ est défini par $\widehat{\theta} = \arg\min_\theta \widehat{E_{\text{CV}}}(\theta)$.

5.1 Bootstrap

Cette section plus technique décrit des outils encore peu présents dans les logiciels commerciaux, elle peut être sautée en première lecture.

Introduction

L'idée, d'approcher par simulation (*Monte Carlo*) la distribution d'un estimateur lorsque l'on ne connaît pas la loi de l'échantillon ou, plus souvent, lorsque l'on ne peut pas supposer qu'elle est gaussienne, est l'objectif même du *bootstrap* (Efron, 1982).

Le principe fondamental de cette technique de rééchantillonnage est de substituer, à la distribution de probabilité inconnue F, dont est issu l'échantillon d'apprentissage, la distribution empirique F_n qui donne un poids $1/n$ à chaque réalisation. Ainsi on obtient un échantillon de taille n dit *échantillon bootstrap* selon la distribution empirique F_n par n tirages aléatoires avec remise parmi les n observations initiales.

Il est facile de construire un grand nombre d'échantillons bootstrap (*i.e.* $B = 100$) sur lesquels calculer l'estimateur concerné. La loi simulée de cet estimateur est une approximation asymptotiquement convergente sous des hypothèses raisonnables[2] de la loi de l'estimateur. Cette approximation fournit ainsi des

[2] Échantillon indépendant de même loi et estimateur indépendant de l'ordre des observations.

estimations du biais, de la variance, donc d'un risque quadratique, et même des intervalles de confiance (avec B beaucoup plus grand) de l'estimateur sans hypothèse (normalité) sur la vraie loi. Les grands principes de cette approche sont rappelés en annexe A.

Estimateur naïf

Soit $\mathbf{z}^*$ un échantillon bootstrap des données :

$$\mathbf{z}^* = \{(\mathbf{x}_1^*, y_1^*), \ldots, (\mathbf{x}_n^*, y_n^*)\}.$$

L'estimateur *plug-in* de l'erreur de prévision $\mathcal{E}_P(\mathbf{z}, F)$, pour lequel la distribution F est remplacée par la distribution empirique $\widehat{F}$ (cf. section A1.1) est défini par :

$$\mathcal{E}_P(\mathbf{z}^*, \widehat{F}) = \frac{1}{n}\sum_{i=1} nQ(y_i, \phi_{\mathbf{z}^*}(x_i))$$

où $\phi_{\mathbf{z}^*}$ désigne l'estimation de ϕ à partir de l'échantillon bootstrap. Il conduit à l'estimation bootstrap de l'erreur moyenne de prévision $E_F[\mathcal{E}_P(\mathbf{z}, F)]$ par

$$\mathcal{E}_{\text{Boot}} = E_{\widehat{F}}[\mathcal{E}_P(\mathbf{z}^*, \widehat{F})] = E_{\widehat{F}}\left[\frac{1}{n}\sum_{i=1} nQ(y_i, \phi_{\mathbf{z}^*}(x_i))\right].$$

Cette estimation est approchée par simulation :

$$\widehat{\mathcal{E}_{\text{Boot}}} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n}\sum_{i=1} nQ(y_i, \phi_{\mathbf{z}^{*b}}(x_i)).$$

L'estimation ainsi construite de l'erreur de prévision est généralement biaisée par optimisme car, au gré des simulations, les mêmes observations $(\mathbf{x}_i, y_i)$ apparaissent à la fois dans l'estimation du modèle et dans celle de l'erreur. D'autres approches visent à corriger ce biais.

Estimateur out-of-bag

La première s'inspire simplement de la validation croisée. Elle considère d'une part les observations tirées dans l'échantillon bootstrap et, d'autre part, celles qui sont laissées de côté pour l'estimation du modèle mais retenue pour l'estimation de l'erreur.

$$\widehat{\mathcal{E}_{\text{oob}}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{B_i}\sum_{b\in K_i} Q(y_i, \phi_{\mathbf{z}^{*b}}(x_i))$$

où K_i est l'ensemble des indices b des échantillons bootstrap ne contenant pas la $i$ème observation à l'issue des B simulations et $B_i = |K_i|$ le nombre de ces échantillons ; B doit être suffisamment grand pour que toute observation n'ait pas été tirée au moins une fois ou bien les termes avec $K_i = 0$ sont supprimés.

L'estimation $\widehat{\mathcal{E}_{\text{oob}}}$ résout le problème d'un biais optimiste auquel est confrontée $\widehat{\mathcal{E}_{\text{Boot}}}$ mais n'échappe pas au biais introduit pas la réduction tel qu'il est signalé pour l'estimation pas validation croisée $\widehat{\mathcal{E}_{\text{CV}}}$. C'est ce qui a conduit Efron et Tibshirani (1997) a proposer des correctifs.

Estimateur .632-bootstrap

La probabilité qu'une observation soit tirée dans un échantillon bootstrap est

$$P[\mathbf{x}_i \in \mathbf{x}^{*b}] = 1 - (1 - \frac{1}{n})^n \approx 1 - \frac{1}{e} \approx 0,632.$$

Très approximativement, la dégradation de l'estimation provoquée par le bootstrap et donc la surévaluation de l'erreur sont analogues à celle de la validation croisée avec $K = 2$. À la suite d'un raisonnement trop long pour être reproduit ici, Efron et Tibshirani (1997) proposent de compenser excès d'optimisme du taux apparent d'erreur et excès de pessimisme du bootstrap *out-of-bag* par une combinaison :

$$\widehat{\mathcal{E}_{.632}} = 0,368 \times \widehat{\mathcal{E}_P} + 0,632 \times \widehat{\mathcal{E}_{\text{oob}}}.$$

5.2 Remarques

- Toutes les estimations de l'erreur de prévision considérées (pénalisation, validation croisée, bootstrap) sont asymptotiquement équivalentes et il n'est pas possible de savoir laquelle concrètement sera, à n fini, la plus précise. Une large part d'arbitraire ou d'"expérience" préside donc le choix d'une estimation plutôt qu'une autre.
- Conceptuellement, le bootstrap est plus compliqué et pratiquement encore peu utilisé. Néanmoins, cet outil joue un rôle central dans les algorithmes récents de combinaison de modèles (cf. chapitre 9) en association avec une estimation *out-of-bag* de l'erreur. Il ne peut être négligé.
- L'estimateur .632-bootstrap pose des problèmes en situation de sur-ajustement aussi les mêmes auteurs ont proposé un rectifcatif complémentaire noté *.632+bootstrap*.
- Comme le signale Vapnik, la résolution d'un problème de modélisation : régression ou discrimination à fin prédictive doit, dans la mesure du possible, d'éviter de se ramener à un problème finalement beaucouop plus complexe comme celui de l'estimation d'une densité multidimensionnelle. C'est ainsi typiquement le cas en analyse discriminante non paramétrique.

Ce qu'il faut retenir en conclusion, c'est que l'estimation d'une erreur de prévision est une opération délicate aux conséquences importantes. Il est donc nécessaire

- d'utiliser le *même estimateur* pour comparer l'efficacité de deux méthodes,
- de se montrer très prudent, en dehors de tout système d'hypothèses probabilistes, sur le caractère absolu d'une estimation dans l'objectif d'une certification.

Dans ces deux dernières situations, le recours à un échantillon test de bonne taille est difficilement contournable alors qu'en situation de choix de modèle au sein d'une même famille, un estimateur (petit échantillon de validation, validation croisée) plus économique est adapté en supposant implicitement que le biais induit est identique d'un modèle à l'autre.

Chapitre 6

Analyse Discriminante Décisionnelle

1 Introduction

L'objet de ce chapitre est l'explication d'une variable qualitative Y à m modalités par p variables quantitatives $X^j, j = 1, \ldots, p$ observées sur unmême échantillon Ω de taille n. L'objectif de l'analyse discriminante décisionnelle déborde le simple cadre descriprif de l'analyse facorielle discriminante (AFD). Disposant d'un nouvel individu (ou de plusieurs, c'est la même chose) sur lequel on a observé les X^j mais pas Y, il s'agit maintenant de *décider* de la modalité $\mathcal{T}_\ell$ de Y (ou de la classe correspondante) de ce nouvel individu. On parle aussi de problème d'*affectation*. L'ADD s'applique donc également à la situation précédente de la régression logistique ($m = 2$) mais aussi lorsque le nombre de classes est plus grand que 2.

Pour cela, on va définir et étudier dans ce chapitre des *règles de décision* (ou d'affectation) et donner ensuite les moyens de les évaluer sur un seul individu ; $\mathbf{x} = (x^1, \ldots, x^p)$ désigne les observations des variables explicatives sur cet individu, $\{\mathbf{g}_\ell; \ell = 1, \ldots, m\}$ les barycentres des classes calculés sur l'échantillon et $\overline{\mathbf{x}}$ le barycentre global.

La matrice de covariance empirique se décompose en

$$\mathbf{S} = \mathbf{S}_e + \mathbf{S}_r.$$

où $\mathbf{S}_r$ est appelée variance intraclasse (within) ou résiduelle :

$$\mathbf{S}_r = \overline{\mathbf{X}}_r{}'\mathbf{D}\overline{\mathbf{X}}_r = \sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} w_i (\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)',$$

et $\mathbf{S}_e$ la variance interclasse (between) ou expliquée :

$$\mathbf{S}_e = \overline{\mathbf{G}}'\mathbf{D}\overline{\mathbf{G}} = \overline{X}_e'\mathbf{D}\overline{X}_e = \sum_{\ell=1}^{m} \overline{w_\ell}(\mathbf{g}_\ell - \overline{\mathbf{x}})(\mathbf{g}_\ell - \overline{\mathbf{x}})'.$$

2 Règle de décision issue de l'AFD

2.1 Cas général : m quelconque

DÉFINITION 6.1. — *On affectera l'individu x à la modalité de Y minimisant :*

$$d^2_{\mathbf{S}_r^{-1}}(\mathbf{x}, \mathbf{g}_\ell), \ell = 1, \ldots, m.$$

Cette distance se décompose en

$$d^2_{\mathbf{S}_r^{-1}}(\mathbf{x}, \mathbf{g}_\ell) = \|\mathbf{x} - \mathbf{g}_\ell\|^2_{\mathbf{S}_r^{-1}} = (\mathbf{x} - \mathbf{g}_\ell)'\mathbf{S}_r^{-1}(\mathbf{x} - \mathbf{g}_\ell)$$

et le problème revient donc à maximiser

$$\mathbf{g}_\ell' \mathbf{S}_r^{-1} \mathbf{x} - \frac{1}{2} \mathbf{g}_\ell' \mathbf{S}_r^{-1} \mathbf{g}_\ell.$$

Il s'agit bien d'une règle linéaire en $\mathbf{x}$ car elle peut s'écrire : $\mathbf{A}_\ell \mathbf{x} + \mathbf{b}_\ell$.

2.2 Cas particulier : $m = 2$

Dans ce cas, la dimension r de l'AFD vaut 1. Il n'y a qu'une seule valeur propre non nulle λ_1, un seul vecteur discriminant v^1 et un seul axe discriminant Δ_1. Les 2 barycentres $\mathbf{g}_1$ et $\mathbf{g}_2$ sont sur Δ_1, de sorte que v^1 est colinéaire à $\mathbf{g}_1 - \mathbf{g}_2$.

L'application de la règle de décision permet d'affecter $\mathbf{x}$ à $\mathcal{T}_1$ si :

$$\mathbf{g}_1' \mathbf{S}_r^{-1} \mathbf{x} - \frac{1}{2} \mathbf{g}_1' \mathbf{S}_r^{-1} \mathbf{g}_1 > \mathbf{g}_2' \mathbf{S}_r^{-1} \mathbf{x} - \frac{1}{2} \mathbf{g}_2' \mathbf{S}_r^{-1} \mathbf{g}_2$$

c'est-à-dire encore si

$$(\mathbf{g}_1 - \mathbf{g}_2)' \mathbf{S}_r^{-1} \mathbf{x} > (\mathbf{g}_1 - \mathbf{g}_2)' \mathbf{S}_r^{-1} \frac{\mathbf{g}_1 + \mathbf{g}_2}{2}.$$

Remarque

La règle de décision liée à l'AFD est simple mais elle est limitée et insuffisante notamment si les variances des classes ne sont pas identiques. De plus, elle ne tient pas compte de l'échantillonnage pour $\mathbf{x}$: tous les groupes n'ont pas nécessairement la même probabilité d'occurence.

3 Règle de décision bayésienne

3.1 Introduction

Dans cette optique, on considère que la variable Y, qui indique le groupe d'appartenance d'un individu, prend ses valeurs dans $\{\mathcal{T}_1, \ldots, \mathcal{T}_m\}$ et est munie d'une loi de probabilité $\pi_1, \ldots, \pi_m$. Les probabilités $\pi_\ell = P[\mathcal{T}_\ell]$ représentent les probabilités *a priori* des classes ou groupes ω_ℓ. On suppose que les vecteurs $\mathbf{x}$ des observations des variables explicatives suivent, connaissant leur classe, une loi de densité

$$f_\ell(\mathbf{x}) = P[\mathbf{x} \mid \mathcal{T}_\ell]$$

par rapport à une mesure de référence[1].

3.2 Définition

Une règle de décision est une application δ de Ω dans $\{\mathcal{T}_1, \ldots, \mathcal{T}_m\}$ qui, à tout individu, lui affecte une classe connaissant $\mathbf{x}$. Sa définition dépend du contexte de l'étude et prend en compte la

- connaissance ou non de coûts de mauvais classement,
- connaissance ou non des lois *a priori* sur les classes,
- nature aléatoire ou non de l'échantillon.

On désigne par $c_{\ell \mid k}$ le coût du classement dans $\mathcal{T}_\ell$ d'un individu de $\mathcal{T}_k$. Le *risque de Bayes* d'une règle de décision δ exprime alors le coût moyen :

$$R_\delta = \sum_{k=1}^{m} \pi_k \sum_{\ell=1}^{m} c_{\ell \mid k} \int_{\{\mathbf{x} \mid \delta(\mathbf{x}) = \mathcal{T}_\ell\}} f_k(\mathbf{x}) d\mathbf{x}$$

où $\int_{\{\mathbf{x} \mid \delta(\mathbf{x}) = \mathcal{T}_\ell\}} f_k(\mathbf{x}) d\mathbf{x}$ représente la probabilité d'affecté $\mathbf{x}$ à $\mathcal{T}_\ell$ alors qu'il est dans $\mathcal{T}_k$.

[1] La mesure de Lebesgues pour des variables réelles, celle de comptage pour des variables qualitatives

3.3 Coûts inconnus

L'estimation des coûts n'est pas du ressort de la Statistique et, s'ils ne sont pas connus, on suppose simplement qu'ils sont tous égaux. La minimisation du risque ou règle de Bayes revient alors à affecter tout $\mathbf{x}$ à la classe la plus probable c'est-à-dire à celle qui maximise la probabilité conditionnelle *a posteriori* : $P[\mathcal{T}_\ell \mid \mathbf{x}]$. Par le théorème de Bayes, on a :

$$P[\mathcal{T}_\ell \mid \mathbf{x}] = \frac{P[\mathcal{T}_\ell \text{ et } \mathbf{x}]}{P[\mathbf{x}]} = \frac{P[\mathcal{T}_\ell].P[\mathbf{x} \mid \mathcal{T}_\ell]}{P[\mathbf{x}]}$$

avec le principe des probabilités totales : $P[\mathbf{x}] = \sum_{\ell=1}^{m} P[\mathcal{T}_\ell].P[\mathbf{x} \mid \mathcal{T}_\ell]$.

Comme $P[\mathbf{x}]$ ne dépend pas de ℓ, la règle consistera à choisir $\mathcal{T}_\ell$ maximisant

$$P[\mathcal{T}_\ell].P[\mathbf{x} \mid \mathcal{T}_\ell] = \pi_\ell.P[\mathbf{x} \mid \mathcal{T}_\ell];$$

$P[\mathbf{x} \mid \mathcal{T}_\ell]$ est la probabilité d'observer $\mathbf{x}$ au sein de la classe $\mathcal{T}_\ell$. Pour une loi discrète, il s'agit d'une probabilité du type $P[\mathbf{x} = \mathbf{x}_k^l \mid \mathcal{T}_\ell]$ et d'une densité $f(\mathbf{x} \mid \mathcal{T}_\ell)$ pour une loi continue. Dans tous les cas nous utiliserons la notation $f_\ell(\mathbf{x})$.

La règle de décision s'écrit finalement sous la forme :

$$\delta(\mathbf{x}) = \arg \max_{\ell=1,\ldots,m} \pi_\ell f_\ell(\mathbf{x}).$$

3.4 Détermination des *a priori*

Les probabilités *a priori* π_ℓ peuvent effectivement être connues *a priori* : proportions de divers groupes dans une population, de diverses maladies… ; sinon elles sont estimées sur l'échantillon d'apprentissage :

$$\widehat{\pi}_\ell = w_\ell = \frac{n_\ell}{n} \quad \text{(si tous les individus ont le même poids)}$$

à condition qu'il soit bien un échantillon aléatoire susceptible de fournir des estimations correctes des fréquences. Dans le cas contraire il reste à considérer tous les π_ℓ égaux.

3.5 Cas particuliers

- Dans le cas où les probabilités *a priori* sont égales, c'est par exemple le cas du choix de probabilités non informatives, la règle de décision bayésienne revient alors à maximiser $f_\ell(\mathbf{x})$ qui est la vraisemblance, au sein de $\mathcal{T}_\ell$, de l'observation $\mathbf{x}$. La règle consiste alors à choisir la classe pour laquelle cette vraisemblance est maximum.
- Dans le cas où $m = 2$, on affecte $\mathbf{x}$ à $\mathcal{T}_1$ si :

  $$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1}$$

  faisant ainsi apparaître un rapport de vraisemblance. D'autre part, l'introduction de coûts de mauvais classement différents selon les classes amène à modifier la valeur limite π_2/π_1.

Finalement, il reste à estimer les densités conditionnelles $f_\ell(\mathbf{x})$. Les différentes méthodes d'estimation considérées conduisent aux méthodes classiques de discrimination bayésienne objets des sections suivantes.

4 Règle bayésienne avec modèle normal

On suppose dans cette section que, conditionnellement à $\mathcal{T}_\ell$, $\mathbf{x} = (x_1, \ldots, x_p)$ est l'observation d'un vecteur aléatoire gaussien $\mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)$; $\boldsymbol{\mu}_\ell$ est un vecteur de $\mathbb{R}^p$ et $\boldsymbol{\Sigma}_\ell$ une matrice $(p \times p)$ symétrique et définie-positive. La densité de la loi, au sein de la classe $\mathcal{T}_\ell$, s'écrit donc :

$$f_\ell(\mathbf{x}) = \frac{1}{\sqrt{2\pi}(\det(\boldsymbol{\Sigma}_\ell))^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}_\ell^{-1}(\mathbf{x} - \boldsymbol{\mu}_\ell)\right].$$

L'affectation de $\mathbf{x}$ à une classe se fait en maximisant $\pi_\ell.f_\ell(\mathbf{x})$ par rapport à l soit encore la quantité :

$$\ln(\pi_\ell) - \frac{1}{2}\ln(\det(\boldsymbol{\Sigma}_\ell)) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}_\ell^{-1}(\mathbf{x} - \boldsymbol{\mu}_\ell).$$

4.1 Hétéroscédasticité

Dans le cas général, il n'y a pas d'hypothèse supplémentaire sur la loi de $\mathbf{x}$ et donc les matrices $\mathbf{\Sigma}_\ell$ sont fonction de ℓ. Le critère d'affectation est alors *quadratique* en $\mathbf{x}$. Les probabilités π_ℓ sont supposées connues mais il est nécessaire d'estimer les moyennes $\boldsymbol{\mu}_\ell$ ainsi que les covariances $\mathbf{\Sigma}_\ell$ en maximisant, compte tenu de l'hypothèse de normalité, la vraisemblance. Ceci conduit à estimer la moyenne

$$\widehat{\boldsymbol{\mu}_\ell} = \mathbf{g}_\ell$$

par la moyenne empirique de $\mathbf{x}$ dans la classe l pour l'échantillon d'apprentissage et $\mathbf{\Sigma}_\ell$ par la matrice de covariance empirique $\mathbf{S}^*_{Rl}$:

$$\mathbf{S}^*_{Rl} = \frac{1}{n_\ell - 1} \sum_{i \in \Omega_\ell} (\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)'$$

pour ce même échantillon.

4.2 Homoscédasticité

On suppose dans ce cas que les lois de chaque classe partagent la même structure de covariance $\mathbf{\Sigma}_\ell = \mathbf{\Sigma}$. Supprimant les termes indépendants de l, le critère à maximiser devient

$$\ln(\pi_\ell) - \frac{1}{2}\boldsymbol{\mu}'_\ell \mathbf{\Sigma}_\ell^{-1} \boldsymbol{\mu}_\ell + \boldsymbol{\mu}'_\ell \mathbf{\Sigma}_\ell^{-1} \mathbf{x}$$

qui est cette fois *linéaire* en $\mathbf{x}$. Les moyennes $\boldsymbol{\mu}_\ell$ sont estimées comme précédemment tandis que $\mathbf{\Sigma}$ est estimée par la matrice de covariance intra empirique :

$$\mathbf{S}^*_R = \frac{1}{n - m} \sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} (\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)'.$$

Si, de plus, les probabilités π_ℓ sont égales, après estimation le critère s'écrit :

$$\overline{\mathbf{x}_\ell}' \mathbf{S}_R^{*-1} \mathbf{x} - \frac{1}{2} \overline{\mathbf{x}_\ell}' \mathbf{S}_R^{*-1} \overline{\mathbf{x}_\ell}.$$

On retrouve alors le critère de la section 2 issu de l'AFD.

4.3 Commentaire

Les hypothèses : normalité, éventuellement l'homoscédasticité, doivent être vérifiées par la connaissance *a priori* du phénomène ou par une étude préalable de l'échantillon d'apprentissage. L'hypothèse d'homoscédasticité, lorqu'elle est vérifiée, permet de réduire très sensiblement le nombre de paramètres à estimer et d'aboutir à des estimateurs plus fiables car de variance moins élevée. Dans le cas contraire, l'échantillon d'apprentissage doit être de taille importante.

5 Règle bayésienne avec estimation non paramétrique

5.1 Introduction

En Statistique, on parle d'estimation non paramétrique ou fonctionnelle lorsque le nombre de paramètres à estimer est infini. L'objet statistique à estimer est alors une fonction par exemple de régression $y = f(x)$ ou encore une densité de probabilité. Dans ce cas, au lieu de supposer qu'on a affaire à une densité de type connu (normale) dont on estime les paramètres, on cherche une estimation $\widehat{f}$ de la fonction de densité f. Pour tout x de $\mathbb{R}$, $f(x)$ est donc estimée par $\widehat{f}(x)$.

Cette approche très souple a l'avantage de ne pas nécessiter d'hypothèse particulière sur la loi (seulement la régularité de f pour de bonnes propriétés de convergence), en revanche elle n'est applicable qu'avec des échantillons de grande taille d'autant plus que le nombre de dimensions p est grand (*curse of dimensionality*).

Dans le cadre de l'analyse discriminante, ces méthodes permettent d'estimer directement les densités $f_\ell(\mathbf{x})$. On considère ici deux approches : la méthode du noyau et celle des k plus proches voisins.

5.2 Méthode du noyau

Estimation de densité

Soit $y_1, \ldots, y_n$ n observations équipondérées d'une v.a.r. continue Y de densité f inconnue. Soit $K(y)$ (le *noyau*) une densité de probabilité unidimensionnelle (sans rapport avec f) et h un réel strictement positif. On appelle estimation de f par la méthode du noyau la fonction

$$\widehat{f}(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right).$$

Il est immédiat de vérifier que

$$\forall y \in \mathbb{R}, \widehat{f}(y) \geq 0 \quad \text{et} \quad \int_{-\infty}^{+\infty} \widehat{f}(y)dy = 1;$$

h est appelé *largeur de fenêtre* ou paramètre de *lissage* ; plus h est grand, plus l'estimation $\widehat{f}$ de f est régulière. Le noyau K est choisi centré en 0, unimodal et symétrique. Les cas les plus usuels sont la densité gaussienne, celle uniforme sur $[-1, 1]$ ou triangulaire : $K(x) = [1 - |x|]\mathbf{1}_{[-1,1]}(x)$. La forme du noyau n'est pas très déterminante sur la qualité de l''estimation contrairement à la valeur de h.

Application à l'analyse discriminante

La méthode du noyau est utilisée pour calculer une estimation non paramétrique de chaque densité $f_\ell(\mathbf{x})$ qui sont alors des fonctions définies dans $\mathbb{R}^p$. Le noyau K^* dont donc être choisi multidimensionnel et

$$\widehat{f}_\ell(\mathbf{x}) = \frac{1}{n_\ell h^p} \sum_{i \in \Omega_\ell} K^*\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right).$$

Un noyau multidimensionnel peut être défini à partir de la densité usuelle de lois : multinormale $\mathcal{N}_p(0, \mathbf{\Sigma}_p)$ ou uniforme sur la sphère unité ou encore par produit de noyaux unidimensionnels :

$$K^*(\mathbf{x}) = \prod_{j=1}^{p} K(x^j).$$

5.3 k plus proches voisins

Cette méthode d'affectation d'un vecteur $\mathbf{x}$ consiste à enchaîner les étapes décrites dans l'algorithme ci-dessous. Pour $k = 1$, $\mathbf{x}$ est affecté à la classe du plus proche élément.

Algorithm 3 ***k*-nn**

Choix d'un entier k : $1 \geq k \geq n$.
Calculer les distances $d_{\mathbf{M}}(\mathbf{x}, \mathbf{x}_i)\,, \quad i = 1, \ldots, n$ où $\mathbf{M}$ est la métrique de Mahalanobis c'est-à-dire la matrice inverse de la matrice de variance (ou de variance intra).
Retenir les k observations $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(k)}$ pour lesquelles ces distances sont les plus petites.
Compter les nombres de fois $k_1, \ldots, k_m$ que ces k observations apparaissent dans chacune des classes.
Estimer les densités par

$$\widehat{f}_\ell(\mathbf{x}) = \frac{k_\ell}{kV_k(\mathbf{x})};$$

où $V_k(\mathbf{x})$ est le volume de l'ellipsoïde $\{\mathbf{z} | (\mathbf{z} - \mathbf{x})'\mathbf{M}(\mathbf{z} - \mathbf{x}) = d_{\mathbf{M}}(\mathbf{x}, \mathbf{x}_{(k)})\}$.

Comme toute technique, celles présentées ci-dessus nécessitent le réglage d'un paramètre (largeur de fenêtre, nombre de voisins considérés). Ce choix s'apparente à un choix de modèle et nécessite le même type d'approche à savoir l'optiomisation d'un critère (erreur de classement, validation croisée (cf. chapitre 5).

TAB. 6.1 – Cancer : estimations des taux d'erreurs de prévision obtenus par différents types d'analyse discriminante

Méthode	apprentissage	validations croisée	test
linéaire	1,8	3,8	3,6
kNN	2,5	2,7	2,9

TAB. 6.2 – Cancer : estimations des taux d'erreurs de prévision obtenus par différents types d'analyse discriminante

Méthode	apprentissage	validations croisée	test
linéaire	11,9	12,5	12,0
quadratique	12,7	14,8	12,5

6 Exemples

6.1 Cancer du sein

Par principe, l'analyse discriminante s'applique à des variables explicativers quantitatives. Ce n'est pas le cas des données qui sont au mieux ordinales. Il est clair que contruire une fonction de disrimination comme combinaison de ces variables n'a guère de sens. Néanmoins, en s'attachant uniquement à la qualité de prévision sans essayer de constuire une interprétation du plan ou de la surface de discrimination, il est d'usage d'utiliser l'analyse discriminante de façon "sauvage". Les résultats obtenus sont résumés dans le tableau 6.2. L'analyse dicriminante quadratique, avec matrice de variance estimée pour chaque classe n'a pas pu être calculée. Une des matrices n'est pas inversible.

6.2 Concentration d'ozone

Dans cet exemple aussi, deux variables sont qualitatives : le type de jour à 2 modalités ne pose pas de problème mais remplacer la station par un entier est plutôt abusif. D'ailleurs, les plus proches voisins ne l'acceptent, une transformation des données seraient nécessaire.

6.3 Carte visa

Comme pour les données sur le cancer, les données bancaires posent un problème car elles associent différents types de variables. Il est possible de le contourner, pour celles binaires, en considérant quantitative, l'indicatrice de la modalité (0 ou 1). Pour les autres, certaines procédures (DISQUAL pour discrimination sur variables qualitatives) proposent de passer par une analyse factorielle multiple des correspondances pour rendre tout quantitatif mais ceci n'est pas implémenté de façon standard dans les logiciels d'origine américaine.

Pour l'analyse discriminante, R ne propose pas de sélection automatique de variable mais inclut une estimation de l'erreur par validation croisée. Les résultats trouvés sont résumés dans le tableau 6.3. Seule une discimination linéaire semble fournir des résultats raisonnables, la recherche d'une discrimination quadratique n'apporte rien pour cs données. De son côté, SAS propose une sélection automatique (procédure stepdisc) mais les résultats obtenus ne sont pas sensiblement meilleurs après sélection.

TAB. 6.3 – Banque : estimations des taux d'erreurs de prévision obtenus par différents types d'analyse discriminante

Méthode	apprentissage	validations croisée	test
linéaire	16,5	18,3	18
quadratique	17,8	22,0	30
kNN	23,5	29,8	29

Chapitre 7

Arbres binaires

1 Introduction

Ce chapitre s'intéresse aux méthodes ayant pour objectif la construction d'*arbres binaires* de décision, modélisant une discrimination ou une régression. Complémentaires des méthodes statistiques plus classiques : analyse discriminante, régression linéaire, les solutions obtenues sont présentées sous une forme graphique simple à interpréter, même pour des néophytes, et constituent une aide efficace pour l'aide à la décision. Elles sont basées sur un découpage, par des hyperplans, de l'espace engendré par les variables explicatives. Nommées initialement partitionnement récursif ou segmentation, les développements importants de Breiman et col. (1984) les ont fait connaître sous l'acronyme de CART : Classification and Regression Tree ou encore de C4.5 (Quinlan, 1993) dans la communauté informatique. L'acronyme correspond à deux situations bien distinctes selon que la variable à expliquer, modéliser ou prévoir est qualitative (discrimination ou en anglais *classification*) ou quantitative (régression).

Ces méthodes ne sont efficaces que pour des tailles d'échantillons importantes et elles sont très calculatoires. Les deux raisons : modèle graphique de décision simple à interpréter, puissance de calcul nécessaire, suffisent à expliquer leur popularité récente. De plus, elles requièrent plutôt moins d'hypothèses que des méthodes statistiques classiques et semblent particulièrement adaptées au cas où les variables explicatives sont nombreuses. En effet, la procédure de sélection des variables est intégrée à l'algorithme construisant l'arbre, d'autre part, les interactions sont prises en compte. Néanmoins, cet algorithme suivant une stratégie pas à pas hiérarchisée, il peut, comme dans le cas du choix de modèle en régression, passer à coté d'un optimum global ; il se montre par ailleurs très sensible à des fluctuations d'échantillon rt nécessite une optimisation délicate de l'optimisation de la complexité par élagage. Ceci souligne encore l'importance de confronter plusieurs approches sur les mêmes données.

2 Construction d'un arbre binaire

2.1 Principe

Les données sont constituées de l'observation de p variables quantitatives ou qualitatives explicatives X^j et d'une variable à expliquer Y qualitative à m modalités $\{\mathcal{T}_\ell; \ell = 1 \ldots, m\}$ ou quantitative réelle, observées sur un échantillon de n individus.

La construction d'un arbre de discrimination binaire (cf. figure 2.1) consiste à déterminer une séquence de *nœuds*.

- Un nœud est défini par le choix conjoint d'une variable parmi les explicatives et d'une *division* qui induit une partition en deux classes. Implicitement, à chaque nœud correspond donc un sous-ensemble de l'échantillon auquel est appliquée une dichotomie.
- Une division est elle-même définie par une valeur seuil de la variable quantitative sélectionnée ou un partage en deux groupes des modalités si la variable est qualitative.
- À la racine ou nœud initial correspond l'ensemble de l'échantillon ; la procédure est ensuite itérée sur chacun des sous-ensembles.

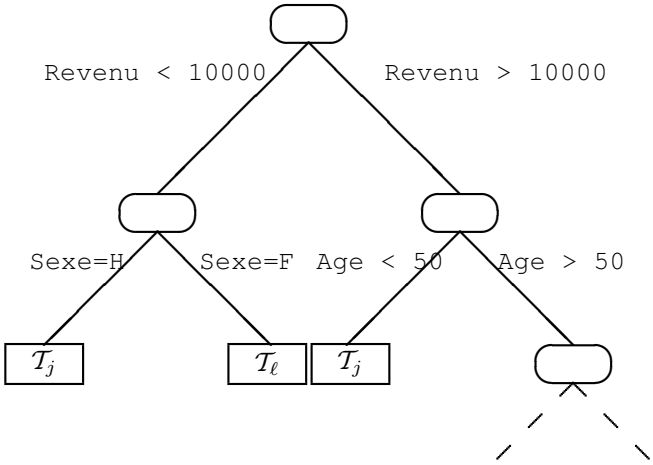

FIG. 7.1 – Exemple élémentaire d'arbre de classification.

L'algorithme considéré nécessite :

i. la définition d'un critère permettant de sélectionner la "meilleure" division parmi toutes celles *admissibles* pour les différentes variables ;

ii. une règle permettant de décider qu'un nœud est terminal : il devient ainsi une *feuille* ;

iii. l'affectation de chaque feuille à l'une des classes ou à une valeur de la variable à expliquer.

Le point (ii) est le plus délicat. Il correspond encore à la recherche d'un modèle parcimonieux. Un arbre trop détaillé, associé à une surparamétrisation, est instable et donc probablement plus défaillant pour la prévision d'autres observations. La contribution majeure de Breiman et col. (1984) est justement une stratégie de recherche d'arbre optimal. Elle consiste à

i. construire l'arbre maximal $A_{\max}$,

ii. ordonner les sous-arbres selon une séquence emboîtée suivant la décroissance d'un critère pénalisé de déviance ou de taux de mal-classés,

iii. puis à sélectionner le sous-arbre optimal ; c'est la procédure d'*élagage*.

Tous ces points sont détaillés ci-dessous.

2.2 Critère de division

Une division est dite *admissible* si aucun des deux nœuds descendants qui en découlent n'est vide. Si la variable explicative est qualitative ordinale avec m modalités, elle fournit $(m-1)$ divisions binaires admissibles. Si elle est seulement nominale le nombre de divisions passe à $2^{(m-1)}-1$. Une variable quantitative se ramène au cas ordinal.

Le critère de division repose sur la définition d'une fonction d'*hétérogénéité* ou de désordre explicitée dans la section suivante. L'objectif étant de partager les individus en deux groupes les plus homogènes au sens de la variable à expliquer. L'hétérogénéité d'un nœud se mesure par une fonction non négative qui doit être

i. nulle si, et seulement si, le nœud est homogène : tous les individus appartiennent à la même modalité ou prennent la même valeur de Y.

ii. Maximale lorsque les valeurs de Y sont équiprobables ou très dispersées.

La division du nœud k crée deux fils, gauche et droit. Pour simplifier, ils sont notés $(k+1)$ et $(k+2)$ mais une re-numérotation est nécessaire pour respecter la séquence de sous-arbres qui sera décrite dans la section suivante.

Parmi toutes les divisions admissibles du nœud k, l'algorithme retient celle qui rend la somme $D_{(k+1)} + D_{(k+2)}$ des désordres des nœuds fils minimales. Ceci revient encore à résoudre à chaque étape k de construc-

tion de l'arbre :

$$\max_{\{\textit{divisions de } X^j; j=1,p\}} D_k - (D_{(k+1)} + D_{(k+2)})$$

Graphiquement, la longueur de chaque branche peut être représentée proportionnellement à la réduction de l'hétérogénéité occasionnée par la division.

2.3 Règle d'arrêt

La croissance de l'arbre s'arrête à un nœud donné, qui devient donc terminal ou *feuille*, lorsqu'il est homogène c'est-à-dire lorsqu'il n'existe plus de partition admissible ou, pour éviter un découpage inutilement fin, si le nombre d'observations qu'il contient est inférieur à une valeur seuil à choisir en général entre 1 et 5.

2.4 Affectation

Dans le cas Y quantitative, à chaque feuille est associée une valeur : la moyenne des observations associées à cette feuille. Dans le cas qualitatif, chaque feuille ou nœud terminal est affecté à une classe $\mathcal{T}_\ell$ de Y en considérant le mode conditionnel :

- celle la mieux représentée dans le nœud et il est ensuite facile de compter le nombre d'objets mal classés ;
- la classe *a posteriori* la plus probable au sens bayesien si des probabilités *a priori* sont connues ;
- la classe la moins coûteuse si des coûts de mauvais classement sont donnés.

3 Critères d'homogénéité

Deux cas sont à considérer.

3.1 Y quantitative

On considère le cas plus général d'une division en J classes. Soit n individus et une partition en J classes de tailles $n_j; j = 1, \ldots, J$ avec $n = \sum_{j=1}^{J} n_j$. On numérote $i = 1, \ldots, n_j$ les individus de la $j$ème classe. Soit μ_{ij} (resp.y_{ij}) la valeur "théorique" (resp. l'observation) de Y sur l'individu (i, j) : le $i$ème de la $j$ème classe. L'hétérogénéité de la classe j est définie par :

$$D_j = \sum_{i=1}^{n_j} (\mu_{ij} - \mu_{.j})^2 \quad \text{avec} \quad \mu_{.j} = \sum_{i=1}^{n_j} \mu_{ij}.$$

L'hétérogénéité de la partition est définie par :

$$D = \sum_{j=1}^{J} D_j = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (\mu_{ij} - \mu_{.j})^2;$$

c'est l'inertie intra (homogène à la variance intraclasse) qui vaut $D = 0$ si et seulement si $\mu_{ij} = \mu_{.j}$ pour tout i et tout j.

La différence d'hétérogénéité entre l'ensemble non partagé et l'ensemble partagé selon la partition J est

$$\begin{aligned} \Delta &= \sum_{j=1}^{J} \sum_{i=1}^{n_j} (\mu_{ij} - \mu_{..})^2 - \sum_{j=1}^{J} \sum_{i=1}^{n_j} (\mu_{ij} - \mu_{.j})^2 \text{ où } \mu_{..} = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \mu_{ij}. \\ &= \sum_{j=1}^{J} n_j (\mu_{..} - \mu_{.j})^2; \end{aligned}$$

c'est encore homogène à la variance inter classe ou "désordre" des barycentres qui vaut $\Delta = n_1 n_2((\mu_{.1} - \mu_{.2})^2$ pour $J = 2$ dans le cas qui nous intéresse.

L'objectif, à chaque étape, est de maximiser Δ c'est-à-dire de trouver la variable induisant une partition en 2 classes associée à une inertie (variance) intraclasse minimale ou encore qui rend l'inertie (la variance) interclasse la plus grande.

Les quantités sont estimées :

$$D_j \quad \text{par} \quad \widehat{D_j} \quad = \quad \sum_{i=1}^{n_j}(y_{ij} - y_{.j})^2 \tag{7.1}$$

$$D \quad \text{par} \quad \widehat{D} \quad = \quad \sum_{j=1}^{J}\widehat{D_j} = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - y_{.j})^2. \tag{7.2}$$

Sous hypothèse gaussienne :

$$Y_{ij} = \mu_{.j} + u_{ij} \quad \text{avec} \quad + u_{ij} \sim \mathcal{N}(0, \sigma^2),$$

la log-vraisemblance

$$\log L = \text{Cste} - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \mu_{.j})^2$$

est rendue maximale pour

$$\mathcal{L}_\mu = \sup_\mu \log L = \text{Cste} - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - y_{.j})^2.$$

Pour le modèle saturé (une classe par individu) : $y_{ij} = \mu_{ij} + u_{ij}$, cet optimum devient :

$$\mathcal{L}_s = \sup_\mu \log L = \text{Cste} - \frac{n}{2}\log(\sigma^2).$$

La déviance (par rapport au modèle saturé) s'exprime alors comme :

$$\mathcal{D}_\mu = 2\sigma^2(\mathcal{L}_s - \mathcal{L}_\mu) = \widehat{D}.$$

Le raffinement de l'arbre est donc associé à une décroissance, la plus rapide possible, de la déviance. C'est l'optique retenue dans le logiciel Splus. On peut encore dire que la division retenue est celle qui rend le test de Fisher (analyse de variance), comparant les moyennes entre les deux classes, le plus significatif possible.

3.2 Y qualitative

Dans ce cas, la fonction d'hétérogénéité, ou de désordre d'un nœud, est définie à partir de la notion d'entropie, du critère de concentration de Gini ou encore d'une statistique de test du χ^2. En pratique, il s'avère que le choix du critère importe moins que celui du niveau d'élagage. Le premier critère (entropie) est souvent préféré (Splus) car il s'interprète encore comme un terme de déviance mais d'un modèle multinomial cette fois.

Entropie

On considère une variable à expliquer qualitative, Y à m modalités ou catégories $\mathcal{T}$ numérotées $\ell = 1, \ldots, m$. L'arbre induit une partition pour laquelle n_{+k} désigne l'effectif de la $k$ème classe ou $k$ème nœud. Soit

$$p_{\ell k} = P[\mathcal{T}_\ell \mid k] \quad \text{avec} \quad \sum_{\ell=1}^{m} p_{\ell k} = 1$$

la probabilité qu'un élément du $k$ème nœud appartienne à la ℓème classe.

Le *désordre* du $k$ème nœud, défini à partir de l'entropie, s'écrit avec la convention $0\log(0) = 0.$:

$$D_k = -2\sum_{\ell=1}^{m} n_{+k} p_{\ell k}\log(p_{\ell k})$$

tandis que l'hétérogénéité ou désordre de la partition est encore :

$$D = \sum_{k=1}^{K} D_k = -2 \sum_{k=1}^{K} \sum_{\ell=1}^{m} n_{+k} p_{\ell k} \log(p_{\ell k}).$$

Remarques :

- Cette quantité est positive ou nulle, elle est nulle si et seulement si les probabilités $p_{\ell k}$ ne prennent que des valeurs 0 sauf une égale à 1 correspondant à l'absence de mélange.
- Elle peut être remplacée par l'indice de Gini $1 - \sum_{\ell=1}^{m} p_{\ell k}^2$ qui conduit à une autre définition de l'hétérogénéité également utilisée mais qui ne s'interprète pas en terme de *déviance* d'un modèle comme dans le cas de l'entropie.

Désignons par $n_{\ell k}$ l'effectif observé de la ℓème classe dans le $k$ème nœud. Un nœud k de l'arbre représente un sous-ensemble de l'échantillon d'effectif $n_{+k} = \sum_{\ell=1}^{m} n_{\ell k}$.

Les quantités sont estimées :

$$D_k \quad \text{par} \widehat{D_k} \quad = \quad -2 \sum_{\ell=1}^{m} n_{+k} \frac{n_{\ell k}}{n_{+k}} \log \frac{n_{\ell k}}{n_{+k}} \tag{7.3}$$

$$D \quad \text{par} \quad \widehat{D} \quad = \quad \sum_{k=1}^{K} \widehat{D_k} = -2 \sum_{k=1}^{K} \sum_{\ell=1}^{m} n_{\ell k} \log \frac{n_{\ell k}}{n_{+k}}. \tag{7.4}$$

Considérons, pour chaque classe ou nœud k, un modèle multinomial à m catégories de paramètre :

$$p_k = (p_{1k}, \ldots, p_{mk}), \quad \text{avec} \quad \sum_{\ell=1}^{m} p_{\ell k} = 1.$$

Pour ce modèle, la logvraisemblance :

$$\log L = \text{Cste} + \sum_{k=1}^{K} \sum_{\ell=1}^{m} n_{\ell k} \log(p_{\ell k})$$

est rendue maximale pour

$$\mathcal{L}_\mu = \sup_{p_{\ell k}} \log L = \text{Cste} + \sum_{k=1}^{K} \sum_{\ell=1}^{m} n_{\ell k} \log \frac{n_{\ell k}}{n_{+k}}.$$

Pour le modèle saturé (une catégorie par objet), cet optimum prend la valeur de la constante et la déviance (par rapport au modèle saturé) s'exprime comme :

$$\mathcal{D} = -2 \sum_{k=1}^{K} \sum_{\ell=1}^{m} n_{\ell k} \log \frac{n_{\ell k}}{n_{+k}} = \widehat{D}.$$

Comme pour l'analyse discriminante décisionnelle, les probabilités conditionnelles sont définies par la règle de Bayes lorsque les probabilités *a priori* π_ℓ d'appartenance à la ℓème classe sont connues. Dans le cas contraire, les probabilités de chaque classe sont estimées sur l'échantillon et donc les probabilités conditionnelles s'estiment simplement par des rapports d'effectifs : $p_{\ell k}$ est estimée par $n_{\ell k}/n_{+k}$. Enfin, il est toujours possible d'introduire, lorsqu'ils sont connus, des coûts de mauvais classement et donc de se ramener à la minimisation d'un risque bayésien.

4 Élagage

Dans des situations complexes, la démarche proposée conduit à des arbres extrêmement raffinés et donc à des modèles de prévision très instables car fortement dépendants des échantillons qui ont permis

leur estimation. On se trouve donc dans une situation de sur-ajustement à éviter au profit de modèles plus parcimonieux donc plus robuste au moment de la prévision. Cet objectif est obtenu par une procédure d'*élagage* (*pruning*) de l'arbre.

Le principe de la démarche, introduite par Breiman et col. (1984), consiste à construire une suite emboîtée de sous-arbres de l'arbre maximum par élagage successif puis à choisir, parmi cette suite, l'arbre optimal au sens d'un critère. La solution ainsi obtenue par un algorithme pas à pas n'est pas nécessairement globalement optimale mais l'efficacité et la fiabilité sont préférées à l'optimalité.

4.1 Construction de la séquence d'arbres

Pour un arbre A donné, on note K le nombre de feuilles ou nœuds terminaux de A ; la valeur de K exprime la complexité de A. La mesure de qualité de discrimination d'un arbre A s'exprime par un critère

$$D(A) = \sum_{k=1}^{K} D_k(A)$$

où $D_k(A)$ est le nombre de mal classés ou la déviance ou le coût de mauvais classement de la $k$ème feuille de l'arbre A.

La construction de la séquence d'arbres emboîtés repose sur une pénalisation de la complexité de l'arbre :

$$C(A) = D(A) + \gamma K.$$

Pour $\gamma = 0$, $A_{\max} = A_K$ minimise $C(A)$. En faisant croître γ, l'une des divisions de A_K, celle pour laquelle l'amélioration de D est la plus faible (inférieure à γ), apparaît comme superflue et les deux feuilles obtenues sont regroupées (élaguées) dans le nœud père qui devient terminal ; A_K devient A_{K-1}.

Le procédé est itéré pour la construction de la séquence emboîtée :

$$A_{\max} = A_K \supset A_{K-1} \supset \cdots A_1$$

où A_1, le nœud racine, regroupe l'ensemble de l'échantillon.

Un graphe représente la décroissance ou éboulis de la déviance (ou du taux de mal classés) en fonction du nombre croissant de feuilles dans l'arbre ou, c'est équivalent, en fonction de la valeur décroissante du coefficient de pénalisation γ.

4.2 Recherche de l'arbre optimal

Les procédures d'élagage diffèrent par la façon d'estimer l'erreur de prédiction. Le graphe précédemment obtenu peut se lire comme un éboulis de valeur propre. Quand l'amélioration du critère est jugé trop petite ou négligeable, on élague l'arbre au nombre de feuilles obtenues. L'évaluation de la déviance ou du taux de mauvais classement estimée par resubstitution sur l'échantillon d'apprentissage est biaisée (trop optimiste). Une estimation sans biais est obtenue par l'utilisation d'un autre échantillon (validation) ou encore par validation croisée. La procédure de validation croisée présente dans ce cas une particularité car la séquence d'arbres obtenue est différente pour chaque estimation sur l'un des sous-échantillons. L'erreur moyenne n'est pas, dans ce cas, calculée pour chaque sous-arbre avec un nombre de feuilles donné mais pour chaque sous-arbre correspondant à une valeur fixée du coefficient de pénalisation. À la valeur de γ minimisant l'estimation de l'erreur de prévision, correspond ensuite l'arbre jugé optimal dans la séquence estimée sur tout l'échantillon d'apprentissage.

Le principe de sélection d'un arbre optimal est donc décrit dans l'algorithme ci-dessous.

5 Exemples

5.1 Cancer du sein

Un arbre de discrimination est estimé sur l'échantillon d'apprentissage, élagué par validation croisée et représenté dans la figure 7.2. La prévision de l'échantillon test par cet arbre conduit à la matrice de confusion :

Algorithm 4 Sélection d'arbre

Construction de l'arbre maximal $A_{\max}$.
Construction de la séquence $A_K \dots A_1$ d'arbres emboîtés.
Estimation sans biais (échantillon de validation ou validation croisée) des déviances $D(A_K), \dots, D(A_1)$.

Représentation de $D(A_k)$ en fonction de k ou de γ.
Choix de k rendant $D(A_k)$ minimum.

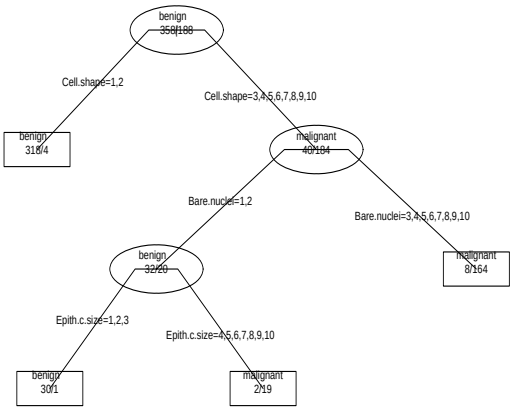

FIG. 7.2 – *Cancer : arbre de décision élagué par validation croisée (R).*

```
predq.tree  benign malignant
  benign        83         5
  malignant      3        46
```

avec un taux d'erreur estimé à 5,8%.

5.2 Concentration d'ozone

Arbre de régression

Un arbre de régression est estimé pour prévoir la concentration d'ozone. La librairie `rpart` du logiciel R prévoit une procédure d'élagage par validation croisée afin d'oprimiser le coefficient de pénalisation. L'arbre (figure 7.3) montre bien quelles sont les variables importantes intervenant dans la prévision. Mais, compte tenu de la hiérarchisation de celles-ci, due à la structure arborescente du modèle, cette liste n'est pas similaire à celle mise en évidence dans le modèle gaussien. On voit plus précisément ici la compexité des interactions entre la prédiction par MOCAGE et l'effet important de la température dans différentes situations. Les résidus de l'écchantillon test du modèle d'arbre de régression prennent une structure particulière (figure 7.4) car les observations communes à une feuille terminale sont affectées de la même valeur. Il y a donc une colonne par feuille. La précision de l'ajustement peut s'en trouver altérée ($R^2 = 0,68$) mais il apparaît que ce modèle est moins soumis au problème d'hétéroscédasticité très présent dans le modèle gaussien.

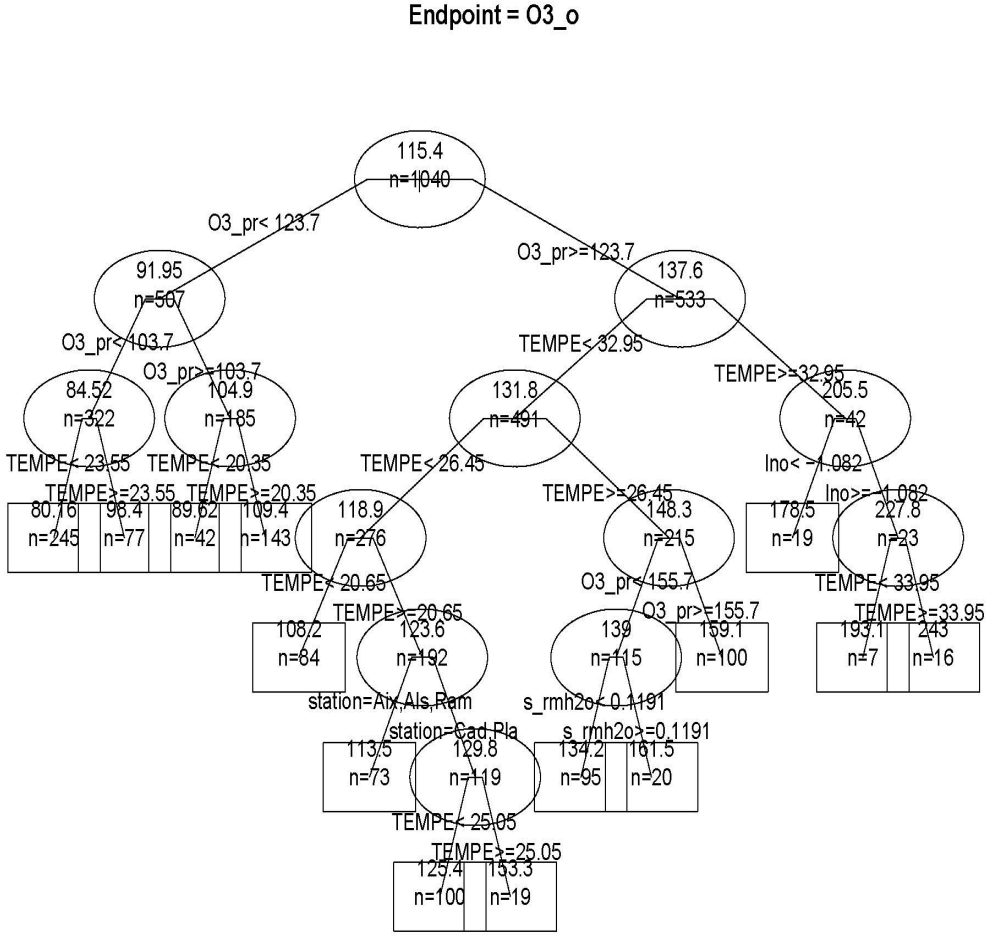

FIG. 7.3 – *Ozone : arbre de régression élagué par validation croisée (R).*

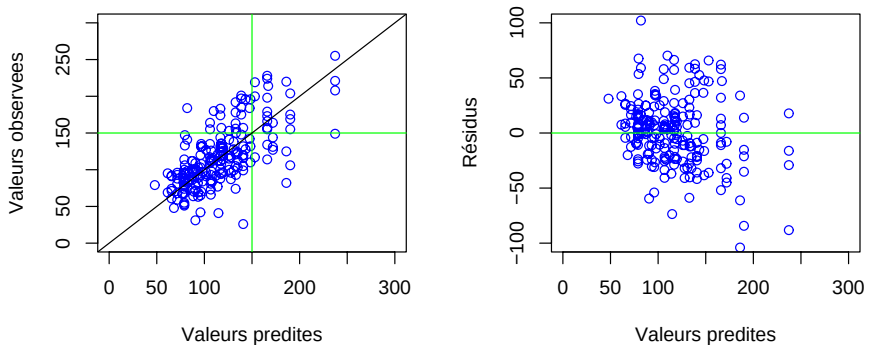

FIG. 7.4 – *Ozone : Valeurs observées et résidus de l'échantillon test en fonction des valeurs prédites.*

Arbre de discrimination

Un modèle est estimé afin de prévoir directement le dépassement d'un seuil. Il est de complexité similaire à l'arbre de régression mais ne fait pas jouer le même rôle aux variables. La température remplace la prévision MOCAGE de l'ozone comme variable la plus importante. Les prévisions de dépassement de seuil sur l'échantillon test sont sensiblement moins bonnes que celle de la régression, les taux sont de 14,4% avec l'arbre de régression et de 14,5% directement avec l'arbre de discrimination. Les matrices de confusion présentent les mêmes biais que les modèles de régression en omettant un nombre important de dépassements.

5.3 Carte Visa Premier

L'étude des données bancaires s'intéresse soit aux données quantitatives brutes soient à celles-ci après découpage en classes des variables quantitatives. Ce découpage rend des services en régression logistique car le modèle construit s'en trouve plus flexible : plus de paramètres mais moins de degrés de liberté, comme l'approximation par des indicatrices (des classes) de transformations non linéaires des variables. Il a été fait "à la main" en prenant les quantiles comme bornes de classe ; C'est un usage courant pour obtenir des classes d'effectifs égaux et répartit ainsi au mieux la précision de l'estimation des paramètres mais ce choix n'est pas optimal au regard de l'objectif de prévision. Dans le cas d'un modèle construit à partir d'un arbre binaire, il est finalement préférable de laisser faire celui-ci le découpage en classe c'est-à-dire de trouver les valeurs seuils de décision. C'est la raison pour laquelle, l'arbre est préférablement estimé sur els variables quantitatives et qualitatives initiales.

Le module SAS/STAT ne fournit pas d'estimation d'arbre de décision, il faut faire appel au module SAS Enterprise Miner. Celui-ci, par principe, propose le découpage de l'échantillon en trois parties apprentissage, validation et test. L'élagage de l'arbre estimé sur l'échantillon d'apprentissage est optimisé pour minimiser l'erreur estimée sur l'échantillon de validation. C'est le graphique de la figure **??**.

En revanche, la librairie `rpart` de R propose d'optimiser l'élagation par validation croisée. L'arbre ainsi obtenu est représenté dans la figure **??**

Cet arbre conduit à la matrice de confusion suivante sur l'échantillon test

```
vistest Cnon Coui
        Cnon  127    6
        Coui   10   57
```

avec un taux d'erreur estimé à 8%.

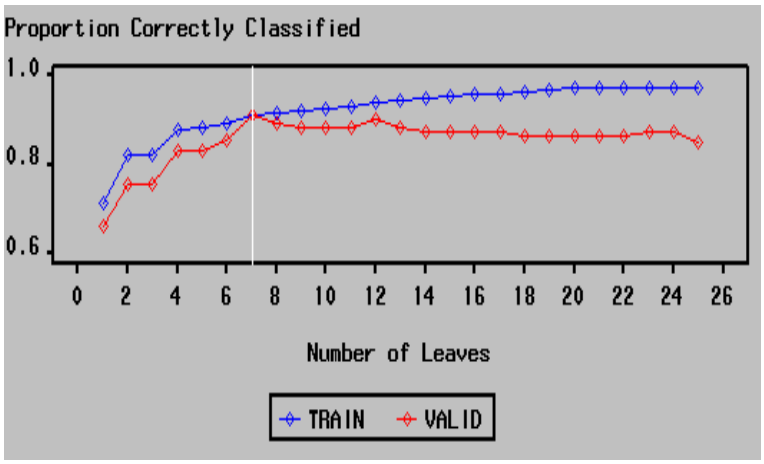

FIG. 7.5 – Banque : choix du nombre de feuilles par échantillon de validation (SEM, 2001).

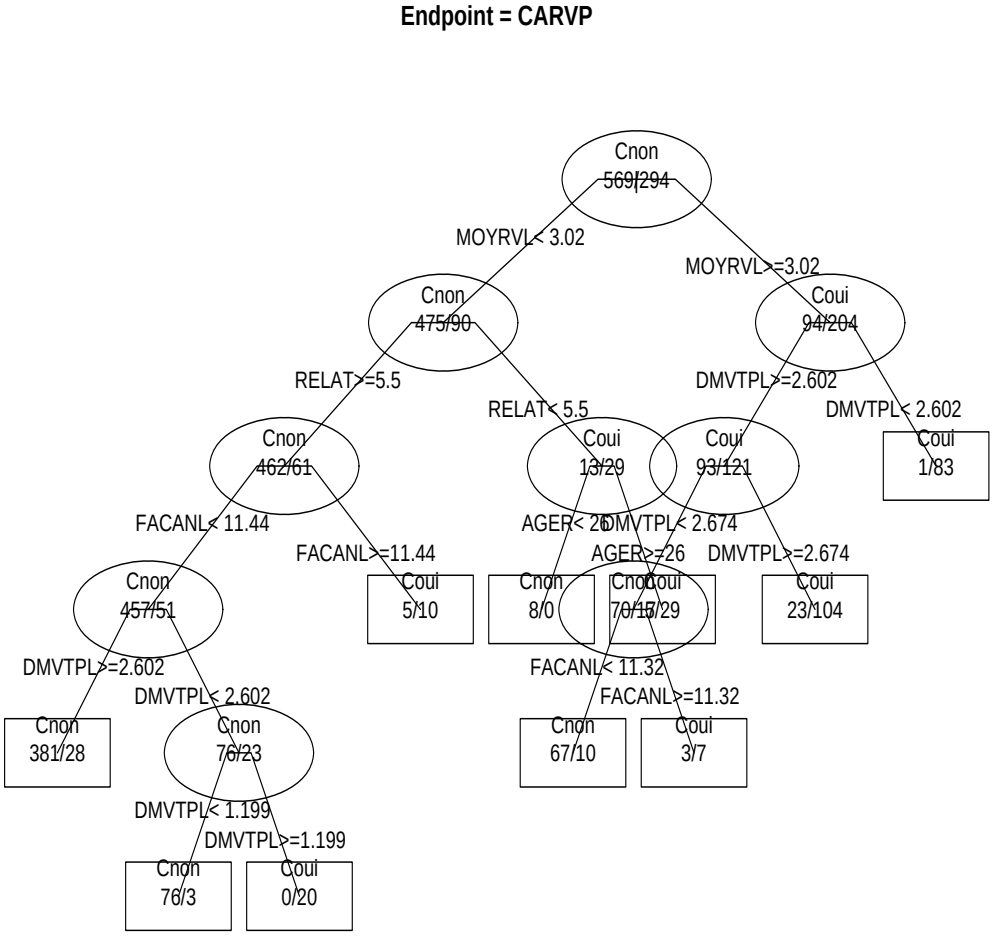

FIG. 7.6 – Banque : arbre de décision élagué par validation croisée dans R.

Chapitre 8

Méthodes connexionistes

1 Historique

Nous nous intéressons ici à une branche de l'Informatique fondamentale qui, sous l'appellation d'*Intelligence Artificielle*, a pour objectif de simuler des comportements du cerveau humain. Les premières tentatives de modélisation du cerveau sont anciennes et précèdent même l'ère informatique. C'est en 1943 que Mc Culloch (neurophysiologiste) et Pitts (logicien) ont proposé les premières notions de *neurone formel*. Ce concept fut ensuite mis en réseau avec une couche d'entrée et une sortie par Rosenblatt en 1959 pour simuler le fonctionnement rétinien et tacher de reconnaître des formes. C'est l'origine du *perceptron*. Cette approche dite *connexioniste* a atteint ses limites technologiques, compte tenu de la puissance de calcul de l'époque, mais aussi théoriques au début des années 70.

L'approche connexioniste à *connaissance répartie* a alors été supplantée par l'approche *symbolique* ou séquentielle qui promouvait les *systèmes experts* à *connaissance localisée*. L'objectif était alors d'automatiser le principe de l'expertise humaine en associant trois concepts :

- une *base de connaissance* dans laquelle étaient regroupées "toutes" les connaissances d'experts humains sous forme de propositions logiques élémentaires ou plus élaborées en utilisant des quantificateurs (logique du premier ordre).
- une *base de faits* contenant les observations du cas à traiter comme, par exemple, des résultats d'examens, d'analyses de sang, de salive pour des applications biomédicales de choix d'un antibiotique,
- un *moteur d'inférence* chargé d'appliquer les règles expertes sur la base de faits afin d'en déduire de nouveaux faits jusqu'à la réalisation d'un objectif comme l'élaboration du traitement d'un infection bactérienne.

Face aux difficultés rencontrées lors de la modélisation des connaissances d'un expert humain, au volume considérable des bases de connaissance qui en découlait et au caractère exponentiel de la complexité des algorithmes d'inférence mis en jeu, cette approche s'est éteinte avec les années 80. En effet, pour les systèmes les plus compliqués à base de calcul des prédicats du premier ordre, on a pu montrer qu'ils conduisaient à des problèmes NP complets et donc dont la solution pouvait être atteinte mais pas nécessairement en un temps fini !

L'essor technologique et surtout quelques avancées théoriques :

- algorithme d'estimation par rétropropagation de l'erreur par Hopkins en 1982,
- analogie de la phase d'apprentissage avec les modèles markoviens de systèmes de particules de la mécanique statistique (verres de spin) par Hopfield en 1982,

au début des années 80 ont permis de relancer l'approche connexioniste. Celle-ci a connu au début des années 90 un développement considérable si l'on considère le nombre de publications et de congrès qui lui ont été consacrés mais aussi les domaines d'applications très divers où elle apparaît. Sur de nombreux objectifs, justement ceux propres au data mining, les réseaux neuronaux ne rentrent pas nécessairement en concurrence avec des méthodes statistiques bientôt centenaires mais apportent un point de vue complémentaire qu'il est important de considérer (Thiria et col. 1997).

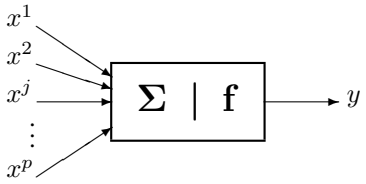

FIG. 8.1 – Représentation d'un neurone formel.

2 Réseaux de neurones

Un *réseau neuronal* est l'association, en un graphe plus ou moins complexe, d'objets élémentaires, les *neurones formels*. Les principaux réseaux se distinguent par l'organisation du graphe (en couches, complets...), c'est-à-dire leur architecture, son niveau de complexité (le nombre de neurones) et par le type des neurones (leurs fonctions de transition).

2.1 Neurone formel

De façon très réductrice, un neurone biologique est une cellule qui se caractérise par
- des synapses, les points de connexion avec les autres neurones, fibres nerveuses ou musculaires ;
- des dentrites, les "entrées" du neurones ;
- l'axone, la "sortie" du neurone vers d'autres neurones ou fibres musculaires ;
- le noyau qui active la sortie en fonction des stimuli en entrée.

Par analogie, le neurone formel est un modèle qui se caractérise par un état interne $s \in \mathcal{S}$, des signaux d'entrée $x_1, \ldots, x_p$ et une fonction de transition d'état

$$s = h(x_1, \ldots, x_p) = f\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right).$$

La fonction de transition opère une transformation d'une combinaison affine des signaux d'entrée, β_0 étant appelé le biais du neurone. Cette combinaison affine est déterminée par un *vecteur de poids* $[\beta_0, \ldots, \beta_p]$ associé à chaque neurone et dont les valeurs sont estimées dans la phase d'apprentissage. Ils constituent "la mémoire" ou "connaissance répartie" du réseau.

Les différents types de neurones se distinguent par la nature f de leur fonction de transition. Les principaux types sont :
- *linéaire* f est la fonction identité,
- *sigmoïde* $f(x) = 1/(1+e^x)$,
- *seuil* $f(x) = \mathbf{1}_{[0,+\infty[}(x)$,
- *stochastiques* $f(x) = 1$ avec la probabilité $1/(1+e^{-x/H})$, 0 sinon (H intervient comme une température dans un algorithme de recuit simulé),
- ...

Les modèles linéaires et sigmoïdaux sont bien adaptés aux algorithmes d'apprentissage comme celui de rétropropagation du gradient car leur fonction de transition est différentiable. Ce sont les plus utilisés. Le modèle à seuil est sans doute plus conforme à la "réalité" biologique mais pose des problèmes d'apprentissage. Enfin le modèle stochastique est utilisé pour des problèmes d'optimisation globale de fonctions perturbées ou encore pour les analogies avec les systèmes de particules. On ne le rencontre pas en *data mining*.

3 Perceptron multicouche

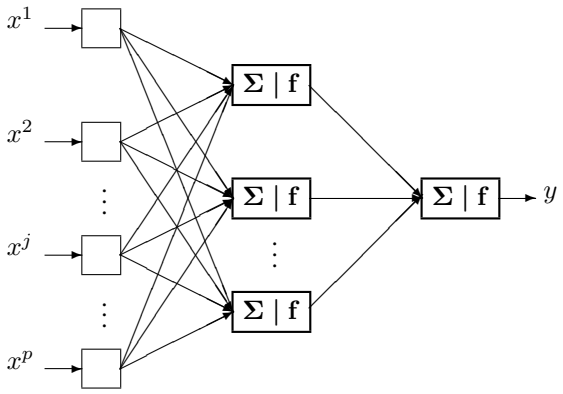

FIG. 8.2 – Exemple de perceptron multicouche élémentaire avec une couche cachée et une couche de sortie.

3.1 Architecture

Le perceptron multicouche (PMC) est un réseau composé de couches successives. Une *couche* est un ensemble de neurones n'ayant pas de connexion entre eux. Une couche d'entrée lit les signaux entrant, un neurone par entrée x_j, une couche en sortie fournit la réponse du système. Selon les auteurs, la couche d'entrée qui n'introduit aucune modification n'est pas comptablisée. Une ou plusieurs couches cachées participent au transfert. Un neurone d'une couche cachée est connecté en entrée à chacun des neurones de la couche précédente et en sortie à chaque neurone de la couche suivante.

Un perceptron multicouche réalise donc une transformation

$$y = \phi(x_1, \ldots, x_p; \boldsymbol{\beta})$$

où $\boldsymbol{\beta}$ est le vecteur contenant chacun des paramètres $\beta_{jk\ell}$ de la $j$ème entrée du $k$ème neurone de la ℓème couche ; la couche d'entrée ($\ell = 0$) n'est pas paramétrée, elle ne fait que distribuer les entrées sur tous les neurones de la couche suivante.

Par souci de cohérence, nous avons tâché de conserver les mêmes notations à travers les différents chapitres. Ainsi, les *entrées* d'un réseau sont encore notées $x^1, \ldots, x^p$ comme les variables explicatives d'un modèle tandis que les *poids* des entrées sont des paramètres $\boldsymbol{\beta}$ à estimer lors de la procédure d'*apprentissage* et que la *sortie* est la variable à expliquer ou cible du modèle.

3.2 Apprentissage

Supposons que l'on dispose d'une base d'apprentissage de taille n d'observations $(x_i^1, \ldots, x_i^p; y_i)$ des variables explicatives $X^1, \ldots, X^p$ et de la variable à prévoir Y. L'apprentissage est l'estimation $\widehat{\boldsymbol{\beta}}$ des paramètres du modèle solutions du problème des moindres carrés[1] :

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\mathbf{b}} Q(\mathbf{b}) \quad \text{avec} \quad Q(\mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \phi(x_i^1, \ldots, x_i^p; (b))]^2.$$

L'algorithme d'optimisation le plus utilisé est celui de rétropropagation du gradient basé sur l'idée suivante : en tout point $\mathbf{b}$, le vecteur gradient de Q pointe dans la direction de l'erreur croissante. Pour faire décroître Q il suffit donc de se déplacer en sens contraire. Il s'agit d'un algorithme itératif modifiant les poids de chaque neurone selon :

$$b_{jk\ell}(i) = b_{jk\ell}(i-1) + \Delta b_{jk\ell}(i)$$

[1] Équivalent à une maximisation de la vraisemblance dans le cas gaussien.

où la correction $\Delta b_{jk\ell}(i)$ est proportionnelle au gradient et à l'erreur attribuée à l'entrée concernée $\varepsilon_{jk\ell}(i)$ et incorpore un terme d'"inertie" $\alpha b_{jk\ell}(i-1)$ permettant d'amortir les oscillations du système :

$$\Delta b_{jk\ell}(i) = -\tau \varepsilon_{jk\ell}(i) \frac{\partial Q}{\partial b_{jk\ell}} + \alpha b_{jk\ell}(i-1).$$

Le coefficient de proportionnalité τ est appelé le *taux d'apprentissage*. Il peut être fixe à déterminer par l'utilisateur ou encore varier en cours d'exécution selon certaines règles paramétrées par l'utilisateur. Il paraît en effet intuitivement raisonnable que, grand au début pour aller plus vite, ce taux décroisse pour aboutir à un réglage plus fin au fur et à mesure que le système s'approche d'une solution. La formule de *rétropropagation de l'erreur* fournit, à partir des erreurs observées sur les sorties, l'expression de l'erreur attribuée à chaque entrée de la couche de sortie à la couche d'entrée.

La littérature sur le sujet propose quantités de recettes destinées à améliorer la vitesse de convergence de l'algorithme ou bien lui éviter de rester collé à une solution locale défavorable. Des propriétés (dynamique markovienne ergodique et convergence vers la mesure stationnaire) de cet algorithme impliquent une convergence presque sûre ; la probabilité d'atteindre une précision fixée *a priori* tend vers 1 lorsque la taille de l'échantillon d'apprentissage tend vers l'infini.

Une amélioration importante consiste à introduire une terme de pénalisation ou régularisation comme en *ridge* dans le critère à optimiser. Celui-ci devient alors :

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\mathbf{b}} Q(\mathbf{b}) + \delta \|b\|^2.$$

Le paramètre δ (*decay*) doit être fixé par l'utilisateur ; plus il est important et moins les paramètres ou poids peuvent prendre des valeurs "cahotiques" contribuant ainsi à limiter les risques de surapprentissage.

Algorithm 5 Rétropropagation du gradient

`Initialisation`
Les poids $b_{jk\ell}$ par tirage aléatoire selon une loi uniforme sur $[0,1]$.
Normaliser dans $[0,1]$ les données d'apprentissage.
Tant que $Q >$ errmax ou niter<itermax **Faire**
  Ranger la base d'apprentissage dans un nouvel ordre aléatoire.
  Pour chaque élément $i = 1, \ldots, n$ de la base **Faire**
    Calculer $\varepsilon(i) = y_i - \phi(x_i^1, \ldots, x_i^p; (b)(i-1))$ en propageant les entrées vers l'avant.
    L'erreur est "rétropropagée" dans les différentes couches afin d'affecter à chaque entrée une responsabilité dans l'erreur globale.
    Mise à jour de chaque poids $b_{jk\ell}(i) = b_{jk\ell}(i-i) + \Delta b_{jk\ell}(i)$
  Fin Pour
Fin Tant que

3.3 Utilisation

On pourra se reporter à l'abondante littérature sur le sujet (Haykin, 1994) pour obtenir des précisions sur les algorithme d'apprentissage et leurs nombreuses variantes. Il est important de rappeler la liste des choix qui sont laissés à l'utilisateur. En effet, même si les logiciels proposent des valeurs par défaut, il est fréquent que cet algorithme connaisse quelques soucis de convergence.

L'utilisateur doit donc déterminer

i. les variables d'entrée et la variable de sortie ; leur faire subir comme pour toutes méthodes statistiques, d'éventuelles transformations.

ii. L'architecture du réseau : le nombre de couches cachées (en général une ou deux) qui correspond à une aptitude à traiter des problèmes de non-linéarité, le nombre de neurones par couche cachée. Ces deux choix conditionnent directement le nombre de paramètres (de poids) à estimer. Ils participent à la recherche d'un bon compromis biais/variance c'est-à-dire à l'équilibre entre qualité d'apprentissage et qualité de prévision. À la louche, on considère en pratique qu'il faut un échantillon d'apprentissage au moins dix fois plus grand que le nombre de paramètres à estimer.

iii. Trois autres paramètres interviennent également sur ce compromis : le nombre maximum d'itérations, l'erreur maximum tolérée et un terme éventuel de régularisation (*decay*). En renforçant ces critères on améliore la qualité de l'apprentissage ce qui peut se faire au détriment de celle de la prévision.

iv. Le taux d'apprentissage ainsi qu'une éventuelle stratégie d'évolution de celui-ci.

En pratique, tous ces paramètres ne sont pas réglés simultanément par l'utilisateur. Celui-ci est confronté à des choix concernant principalement le contrôle du sur-apprentissage ; choix du paramètre : limiter le nombre de neurones ou la durée d'apprentissage ou encore augmenter le coefficient de pénalisation de la norme des paramètres ; choix du mode d'estimation de l'erreur : échantillon test, validation croisée ou bootstrap. Ces choix sont souvent pris par défaut dans la plupart des logiciels commerciaux. Il est important d'en connaître les implications.

Le nombre de couches reste restreint. On montre en effet que toute fonction que toute fonction continue d'un compact de $\mathbb{R}^P$ dans $\mathbb{R}^q$ peut être approchée avec une précision arbitraire par un réseau à une couche cachée en adaptant le nombre de neurones. Leccontrôle de la complexité du modèle ou plus généralement d'un sur-apprentissage peut se faire à l'aide de plusieurs paramètres : le nombre de neurones, une pénalisation de la norne du vecteur des poids ou paramètres comme en *ridge* (régularisation) ou encore par la durée de l'apprentissage. Ces paramètres sont optimisés en considérant un échantillon de validation et le plus simple consiste à arrêté l'apprentissage lorsque l'erreur sur l'échantillon de validation commence à se dégrader tandis que celle sur l'échantillon d'apprentissage ne peut que continuer à décroître.

Les champs d'application des PMC sont très nombreux : discrimination, prévision d'une série temporelle, reconnaissance de forme... Ils sont en général bien explicités dans les documentations des logiciels spécialisés.

Les critiques principales énoncées à l'encontre du PMC concernent les difficultés liés à l'apprentissage (temps de calcul, taille de l'échantillon, localité de l'optimum obtenu) ainsi que son statut de boîte noir. En effet, contrairement à un modèle de discrimination ou un arbre, il est *a priori* impossible de connaître l'influence effective d'une entrée (une variable) sur le système dès qu'une couche cachée intervient. Néanmoins, des techniques de recherche de sensibilité du système à chacune des entrées permettent de préciser les idées et, éventuellement de simplifier le système en supprimant certaines des entrées.

En revanche, ils possèdent d'indéniables qualités lorsque l'absence de linéarité et/ou le nombre de variables explicatives rendent les modèles statistiques traditionnelles inutilisables. Leur flexibilité alliée à une procédure d'apprentissage intégrant la pondération (le choix) des variables comme de leurs interactions peuvent les rendre très efficaces (Besse et col. 2001).

4 Exemples

Les réseaux de neurones étant des boîtes noires, les résultats fournis ne sont guère explicites et ne conduisent donc pas à des interprétations peu informatives du modèle. Seule une étude des erreurs de prévisions et, dans le cas d'une régression, une étude des résidus, permet de se faire une idée de la qualité du modèle.

4.1 Cancer du sein

La prévision de l'échantillon test par un réseau de neurones conduit à la matrice de confusion :

```
       benign malignant
FALSE      83         1
TRUE        3        50
```

et donc une erreur estimée de 3%.

4.2 Concentration d'ozone

La comparaison des résidus (figure 8.3 montre que le problème de non-linéarité qui apparaissait sur les modèles simples (MOCAGE, régression linéaire) est bien résolu et que ces résidus sont plutôt moins étendus, mais le phénomène d'hétéroscédasticité est toujours présent quelque soit le nombre de neurones

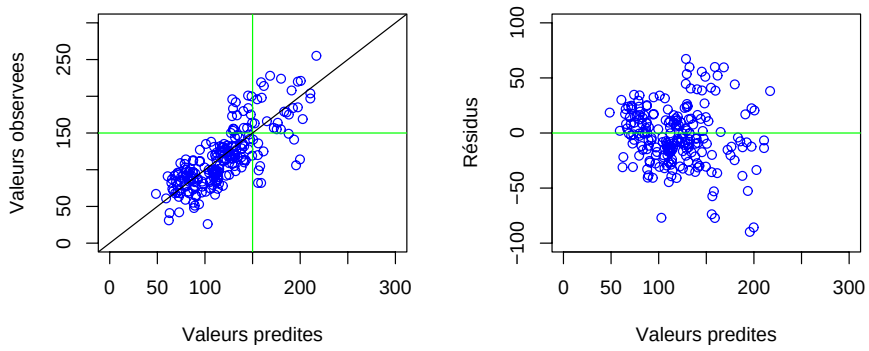

FIG. 8.3 – Ozone : Valeurs observées et résidus de l'échantillon test en fonction des valeurs prédites par un réseau de 10 neurones

utilisés. Il a été choisi relativement important (10) et conduit donc à un bon ajustement ($R^2 = 0,77$) mais devra être réduit pour optimiser la prévision.

Comme pour les arbres de décision, les réseaux de neurones ne proposent pas de modèles très efficaces sur cet exemple. Les taux d'erreur de prévision du dépassement du seuil sont de 14,4% à partir du modèle quantitatif et de 15,6% avec une prévision qualitative.

4.3 Carte visa

Une fonction de la librairie `e1071`, pratique mais très consomatrice de calculs, propose une automatisation de l'optimisation des paramètres (*decay*, nombre de neurones).

```
plot(tune.nnet(CARVP .,data=visapptq,size=2 :4,decay=0 :2))
```

Elle produit une carte de type contour permettant d'évaluer "à l'œil" les valeurs optimales. La prévision de l'échantillon test par ce réseau de neurones conduit à la matrice de confusion :

```
pred.vistest FALSE TRUE
       FALSE   110   16
       TRUE     27   47
```

et donc une erreur estimée de 21,5%.

Chapitre 9

Agrégation de modèles

1 Introduction

Ce chapitre décrit des algorithmes plus récemment apparus dans la littérature. Ils sont basés sur des stratégies adaptatives (*boosting*) ou aléatoires (*bagging*) permettant d'améliorer l'ajustement par une combinaison ou agrégation d'un grand nombre de modèles tout en évitant un sur-ajustement. Ces algorithmes se sont développés à la frontière entre apprentissage machine (*machine learning*) et Statistique. De nombreux articles comparatifs montrent leur efficacité sur des exemples de données simulées et surtout pour des problèmes réels complexes (voir par exemple Ghattas 2000) tandis que leurs propriétés théoriques sont un thème de recherche actif.

Deux types d'algorithmes sont décrits schématiquement dans ce chapitre. Ceux reposants sur une construction aléatoires d'une famille de modèle : *bagging* pour *bootstrap aggregating* (Breiman 1996), les forêts aléatoires (*random forests*) de Breiman (2001) qui propose une amélioration du *bagging* spécifique aux modèles définis par des arbres binaires (CART). Ceux basés sur le *boosting* (Freund et Shapiro,1996), reposent sur une construction *adaptative*, déterministe ou aléatoire, d'une famille de modèles.

Les principes du *bagging* ou du *boosting* s'appliquent à toute méthode de modélisation (régression, CART, réseaux de neurones) mais n'ont d'intérêt, et réduisent sensiblement l'erreur de prévision, que dans le cas de modèles *instables*, donc plutôt non linéaires. Ainsi, l'utilisation de ces algorithmes n'a guère de sens avec la régression multilinéaire ou l'analyse discriminante. Ils sont surtout mis en œuvre en association avec des arbres binaires comme modèles de base.

2 Famille de modèles aléatoires

2.1 *Bagging*

Principe et algorithme

Soit Y une variable à expliquer quantitative ou qualitative, $X^1, \ldots, X^p$ les variables explicatives et $\phi(\mathbf{x})$ un modèle fonction de $\mathbf{x} = \{x^1, \ldots, x^p\} \in \mathbb{R}^p$. On note n le nombre d'observations et

$$\mathbf{z} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$$

un échantillon de loi F.

L'espérance $\phi(.) = E_F(\widehat{\phi}_{\mathbf{z}})$ de l'estimateur définie sur l'échantillon $\mathbf{z}$, est un estimateur sans biais de variance nulle. Considérons B *échantillons indépendants* notés $\{\mathbf{z}_b\}_{b=1,B}$ et construisons une agrégation des modèles dans le cas où la variable à expliquer Y est :

- quantitative : $\widehat{\phi}_B(.) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\phi}_{\mathbf{z}_b}(.)$,
- qualitative : $\widehat{\phi}_B(.) = \arg\max_j \text{card}\left\{b \mid \widehat{\phi}_{\mathbf{z}_b}(.) = j\right\}$.

Dans le premier cas, il s'agit d'une simple moyenne des résultats obtenus pour les modèles associés à chaque échantillon, dans le deuxième, un *comité* de modèles est constitué pour *voter* et *élire* la réponse

la plus probable. Dans ce dernier cas, si le modèle retourne des probabilités associées à chaque modalité comme en régression logistique ou avec les arbres de décision, il est aussi simple de calculer des moyennes de ces probabilités.

Le principe est élémentaire, moyenner les prévisions de plusieurs modèles indépendants permet de réduire la variance et donc de réduire l'erreur de prévision.

Cependant, il n'est pas réaliste de considérer B échantillons indépendants. Cela nécessiterait généralement trop de données. Ces échantillons sont donc remplacés par B réplications d'échantillons *bootstrap* (cf. Annexe A) obtenus chacun par n tirages avec remise selon la mesure empirique $\widehat{F}$. Ceci conduit à l'algorithme ci-dessous.

Algorithm 6 Bagging

Soit $\mathbf{x}_0$ à prévoir et
$\mathbf{z} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ un échantillon
Pour $b = 1$ à B **Faire**
  Tirer un échantillon bootstrap $\mathbf{z}_b^*$.
  Estimer $\widehat{\phi}_{\mathbf{z}_b}(\mathbf{x}_0)$ sur l'échantillon bootstrap.
Fin Pour
Calculer l'estimation moyenne $\widehat{\phi}_B(\mathbf{x}_0) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\phi}_{\mathbf{z}_b}(\mathbf{x}_0)$ ou le résultat du vote.

Utilisation

Il est naturel et techniquement facile d'accompagner ce calcul par une estimation *bootstrap out-of-bag* (cf. chapitre 5 section 5.1) de l'erreur de prévision. Elle est une mesure de la qualité de généralisation du modèle et permet de prévenir une éventuelle tendance au sur-ajustement. C'est, pour éviter un biais, la moyenne des erreurs de prévision commises par chaque estimateur ; chacune des erreurs étant estimée sur les observations qui n'ont pas été sélectionnées par l'échantillon *bootstrap* correspondant.

En pratique, CART est souvent utilisée comme méthode de base pour construire une famille de modèles c'est-à-dire d'arbres binaires. Trois stratégies d'élagage sont alors possibles :

i. laisser construire et garder un arbre complet pour chacun des échantillons,

ii. construire un arbre d'au plus q feuilles,

iii. construire à chaque fois l'arbre complet puis l'élaguer par validation croisée.

La première stratégie semble en pratique un bon compromis entre volume des calculs et qualité de prévision. Chaque arbre est alors affecté d'un faible biais et d'une grande variance mais la moyenne des arbres réduit avantageusement celle-ci. En revanche, l'élagage par validation croisée pénalise lourdement les calculs sans gain substantiel de qualité.

Cet algorithme a l'avantage de la simplicité, il s'adapte et se programme facilement quelque soit la méthode de modélisation mise en œuvre. Il pose néanmoins quelques problèmes :

- temps de calcul important pour évaluer un nombre suffisant d'arbres jusqu'à ce que l'erreur de prévision*out-of-bag* ou sur un échantillon validation se stabilise et arrêt si elle tend à augmenter ;
- nécessiter de stocker tous les modèles de la combinaison afin de pouvoir utiliser cet outil de prévisionsur d'autres données,
- l'amélioration de la qualité de prévisionse fait au détriment de l'interprétabilité. Le modèle finalement obtenu devient une *boîte noire* comme dans le cas du perceptron.

2.2 Forêts aléatoires

Algorithme

Dans les cas spécifique des modèles CART (arbres binaires), Breiman (2001) propose une amélioration du *bagging* par l'ajout d'une randomisation. L'objectif est donc de rendre plus *indépendants* les arbres de l'agrégation en ajoutant du hasard dans le choix des variables qui interviennent dans les modèles. Cette approche semble plus particulièrement fructueuse dans des situations hautement multidimensionnelles, c'est-

à-dire lorsque le nombre de variables explicatives p est très important. C'est le cas lorsqu'il s'agit, par exemple, de discriminer des courbes, spectres, signaux, biopuces.

Algorithm 7 Forêts aléatoires

Soit $\mathbf{x}_0$ à prévoir et
$\mathbf{z} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ un échantillon
Pour $b = 1$ à B **Faire**
  Tirer un échantillon bootstrap $\mathbf{z}_b^*$
  Estimer un arbre sur cet échantillon avec randomisation des variables : la recherche de chaque nœud optimal est précédé d'un tirage aléatoire d'un sous-ensemble de q prédicteurs.
Fin Pour
Calculer l'estimation moyenne $\widehat{\phi}_B(\mathbf{x}_0) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\phi}_{\mathbf{z}_b}(\mathbf{x}_0)$ ou le résultat du vote.

Élagage

La stratégie d'élagage peut, dans le cas des forêts aléatoires, être plus élémentaire qu'avec le *bagging* en se limitant à des arbres de taille q relativement réduite voire même triviale avec $q = 2$ (*stump*). En effet, avec le seul *bagging*, des arbres limités à une seule fourche risquent d'être très semblables (fortement corrélés) car impliquant les mêmes quelques variables apparaissant comme les plus explicatives. La sélection aléatoire d'un nombre réduit de prédicteurs potentiels à chaque étape de construction d'un arbre, accroît significativement la variabilité en mettant en avant nécessairement d'autres variables. Chaque modèle de base est évidemment moins performant mais, l'union faisant la force, l'agrégation conduit finalement à de bons résultats. Le nombre de variables tirées aléatoirement n'est pas un paramètre sensible un choix par défaut de $q = \sqrt{p}$ est suggéré par Breiman (2001). Comme pour le *bagging*, l'évaluation itérative de l'erreur *out-of-bag* prévient d'un éventuel sur-ajustement si celle-ci vient à se dégrader.

Interprétation

Comme pour tout modèles construit par agrégation ou boîte noire, il n'y a pas d'interprétation directe. Néanmoins des informations pertinentes sont obtenues par le calcul et la représentation graphique d'indices proportionnels à l'importance de chaque variable dans le modèle agrégé et donc de sa participation à la régression ou à la discrimination. C'est évidemment d'autant plus utile que les variables sont très nombreuses. Plusieurs critères sont ainsi proposés pour évaluer l'importance de la $j$ème variable.

- Le premier (*Mean Decrease Accuracy* repose sur une permutation aléatoire des valeurs de cette variable. Il consiste à calculer la moyenne sur les observations *out-of-bag* de la décroissance de leur marge lorsque la variable est aléatoirement perturbée. La marge est ici la proportion de votes pour la vraie classe d'une observation moins le maximum des proportions des votes pour les autres classes. Il s'agit donc d'une mesure globale mais indirecte de l'influence d'une variable sur la qualité des prévisions. Plus la prévision est dégradée par la permutation des valeurs d'une variable, plus celle-ci est importante.
- Le deuxième (*Mean Decrease Gini*) est local, basé sur la décroissance d'entropie ou encore la décroissance de l'hétérogénéité définie à partir du critère de Gini. L'importance d'une variable est alors une somme pondérée des décroissances d'hétérogénéité induites lorqu'elle est utilisée pour définir la division associée à un nœud.
- Le troisième, qui n'a pas été retenu par Breiman, est plus rudimentaire, il s'intéresse simplement à la fréquence de chacune des variables apparaissant dans les arbres de la forêt.

Selon Breiman les deux premiers sont très proches, l'importance d'une variable dépend donc se sa fréquence d'apparition mais aussi des places qu'elle occupe dans chaque arbre. Ces critères sont pertinents pour une discrimination de deux classes ou, lorsqu'il y a plus de deux classes, si celles-ci sont relativement équilibrées. Dans le cas contraire, c'est-à-dire si une des classes est moins fréquente et plus difficile à discriminer, l'expérience montre que le troisième critère relativement simpliste présente un avantage : il donne une certaine importance aux variables qui sont nécessaires à la discrimination d'une classe difficile alors que celles-ci sont négligées par les deux autres critères.

3 Famille de modèles adaptatifs

3.1 Principes du *Boosting*

Le *boosting* diffère des approches précédentes par ses origines et ses principes. L'idée initiale, en apprentissage machine, était d'améliorer les compétences d'un *faible classifieur* c'est-à-dire celle d'un modèle de discrimination dont la probabilité de succès sur la prévision d'une variable qualitative est légèrement supérieure à celle d'un choix aléatoire. L'idée originale de Schapire (1990) a été affinée par Freund et Schapire (1996) qui ont décrit l'algorithme original *AdaBoost* (*Adaptative boosting*) pour la prévision d'une variable binaire. De nombreuses études ont ensuite été publiées pour adapter cet algorithme à d'autres situations : k classes, régression et rendre dompte de ses performances sur différents jeux de données (cf. Schapire, 2002) pour une bibliographie). Ces tests ont montré le réel intérêt pratique de ce type d'algorithme pour réduire sensiblement la variance (comme le *bagging*) mais aussi le biais de prévision comparativement à d'autres approches. Cet algorithme est même considéré comme la meilleure méthode "*off-the-shelf*" c'est-à-dire ne nécessitant pas un long prétraitement des données ni un réglage fin de paramètres lors de la procédure d'apprentissage.

Le *boosting* adopte le même principe général que le *bagging* : construction d'une famille de modèles qui sont ensuite agrégés par une moyenne pondéré des estimations ou un vote. Il diffère nettement sur la façon de construire la famille qui est dans ce cas récurrente : chaque modèle est une version *adaptative* du précédent en donnant plus de poids, lors de l'estimation suivante, aux observations mal ajustées ou mal prédites. Intuitivement, cet algorithme concentre donc ses efforts sur les observations les plus difficiles à ajuster tandis que l'agrégation de l'ensemble des modèles permet d'échapper au sur-ajustement.

Les algorithmes de *boosting* proposés diffèrent par différentes caractéristiques :

- la façon de pondérer c'est-à-dire de renforcer l'importance des observations mal estimées lors de l'itération précédente,
- leur objectif selon le type de la variable à prédire Y : binaire, qualitative à k classes, réelles ;
- la fonction perte, qui peut être choisie plus ou moins robuste aux valeurs atypiques, pour mesurer l'erreur d'ajustement ;
- la façon d'agréger, ou plutôt pondérer, les modèles de base successifs.

La littérature sur le sujet présente donc de très nombreuses versions de cet algorithme et il est encore difficile de dire lesquelles sont les plus efficaces et si une telle diversité est bien nécessaire. Il serait fastidieux de vouloir expliciter toutes les versions, ce chapitre en propose un choix arbitraire.

3.2 Algorithme de base

Décrivons la version originale du *boosting* pour un problème de discrimination élémentaire à deux classes en notant δ la fonction de discrimination à valeurs dans $\{-1, 1\}$. Dans cette version, le modèle de base retourne l'identité d'une classe, il est encore nommé Adaboost discret. Il est facile de l'adapter à des modèles retournant une valeur réelle comme une probabilité d'appartenance à une classe.

Les poids de chaque observations sont initialisés à $1/n$ pour l'estimation du premier modèle puis évoluent à chaque itération donc pour chaque nouvelle estimation. L'importance d'une observation w_i est inchangée si elle est bien classée, elle croît sinon proportionnellement au défaut d'ajustement du modèle. L'agrégation finale des prévisions : $\sum_{m=1}^{M} c_m \delta_m(\boldsymbol{x}_0)$ est une combinaison pondérée par les qualités d'ajustement de chaque modèle. Sa valeur absolue appelée *marge* est proportionnelle à la confiance que l'on peut attribuer à son signe qui fournit le résultat de la prévision.

Ce type d'algorithme est largement utilisé avec un arbre (CART) comme modèle de base. De nombreuses applications montrent que si le "classifieur faible" est un arbre trivial à deux feuilles (*stump*), *AdaBoost* fait mieux qu'un arbre sophistiqué pour un volume de calcul comparable : autant de feuilles dans l'arbre que d'itérations dans *AdaBoost*. Hastie et col. (2001) discutent la meilleure stratégie d'élagage applicable à chaque modèle de base. Ils le comparent avec le niveau d'interaction requis dans un modèle d'analyse de variance. Le cas $q = 2$ correspondant à la seule prise en compte des effets principaux. Empiriquement ils recommandent une valeur comprise entre 4 et 8.

Algorithm 8 AdaBoost (*adaptative boosting*)

Soit $\mathbf{x}_0$ à prévoir et
$\mathbf{z} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ un échantillon
`Initialiser` les poids $\boldsymbol{w} = \{w_i = 1/n\ ;\ i = 1, \ldots, n\}$.
Pour $m = 1$ à M **Faire**
Estimer δ_m sur l'échantillon pondéré par $\boldsymbol{w}$.
Calculer le taux d'erreur apparent :

$$\widehat{\mathcal{E}_p} = \frac{\sum_{i=1}^n w_i \mathbf{1}\{\delta_m(\boldsymbol{x}_i) \neq y_i\}}{\sum_{i=1}^n w_i}.$$

Calculer les logit : $c_m = \log((1 - \widehat{\mathcal{E}_p})/\widehat{\mathcal{E}_p})$.
Calculer les nouvelles pondérations : $w_i \leftarrow w_i . \exp\left[-c_m \mathbf{1}\{\delta_m(\boldsymbol{x}_i) \neq y_i\}\right] ; i = 1, \ldots, n$.
Fin Pour
Résultat du vote : $\widehat{\phi}_M(\boldsymbol{x}_0) = \text{signe}\left[\sum_{m=1}^M c_m \delta_m(\boldsymbol{x}_0)\right]$.

3.3 Version aléatoire

À la suite de Freund et Schapire (1996), Breiman (1998) développe aussi, sous le nom d'*Arcing* (adaptively resample and combine), une version aléatoire, et en pratique très proche, du *boosting*. Elle s'adapte à des classifieurs pour lesquels il est difficile voire impossible d'intégrer une pondération des observations dans l'estimation. Ainsi plutôt que de jouer sur les pondérations, à chaque itération, un nouvel échantillon est tiré avec remise, comme pour le bootstrap, mais selon des probabilités inversement proportionnelles à la qualité d'ajustement de l'itération précédente. La présence des observations difficiles à ajuster est ainsi renforcée pour que le modèle y consacre plus d'attention. L'algorithme *adaboost* précédent est facile à adapter en ce sens en regardant celui développé ci-dessous pour la régression et qui adopte ce point de vue.

3.4 Pour la régression

Différentes adaptations du *boosting* ont été proposées pour le cas de la régression, c'est-à-dire lorsque la variable à prédire est quantitative. Voici l'algorithme de Drucker (1997) dans la présentation de Gey et Poggi (2002) qui en étudient les performances empiriques en relation avec CART. Freund et Schapire (1996) ont proposé *Adaboost.R* avec le même objectif tandis que le point de vue de Friedman (2002) est décrit plus loin dans l'algorithme 10.

Précisions :

- Dans cet algorithme la fonction perte Q peut être exponentielle, quadratique ou, plus robuste, la valeur absolue. Le choix usuel de la fonction quadratique est retenu par Gey et Poggi (2002).
- Notons $L_m = \sup_{i=1,\ldots,n} l_m(i)$ le maximum de l'erreur observée par le modèle $\widehat{\phi}_m$ sur l'échantillon initial. La fonction g est définie par :

$$g(l_m(i)) = \beta_m^{1-l_m(i)/L_m} \tag{9.1}$$

$$\text{avec} \quad \beta_m = \frac{\widehat{\mathcal{E}_m}}{L_m - \widehat{\mathcal{E}_m}}. \tag{9.2}$$

- Selon les auteurs, une condition supplémentaire est ajoutée à l'algorithme. Il est arrêté ou réinitialisé à des poids uniformes si l'erreur se dégrade trop : si $\widehat{\mathcal{E}_m} < 0.5 L_m$.

L'algorithme génère M prédicteurs construits sur des échantillons bootstrap $\mathbf{z}_m^*$ dont le tirage dépend de probabilités $\mathbf{p}$ mises à jour à chaque itération. Cette mise à jour est fonction d'un paramètre β_m qui est un indicateur de la performance, sur l'échantillon $\mathbf{z}$, du $m$ième prédicteur estimé sur l'échantillon $\mathbf{z}_m^*$. La mise à jour des probabilités dépend donc à la fois de cet indicateur global β_m et de la qualité relative $l_m(i)/L_m$ de l'estimation du $i$ème individu. L'estimation finale est enfin obtenue à la suite d'une moyenne ou médiane des prévisions pondérées par la qualité respective de chacune de ces prévisions. Gey et Poggi (2002) conseille la médiane afin de s'affranchir de l'influence de prédicteurs très atypiques.

Algorithm 9 Boosting pour la régression

Soit $\mathbf{x}_0$ à prévoir et
$\mathbf{z} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ un échantillon
Initialiser $\boldsymbol{p}$ par la distribution uniforme $\mathbf{p} = \{p_i = 1/n \ ; \ i = 1, \ldots, n\}$.
Pour $m = 1$ à M **Faire**
Tirer avec remise dans $\mathbf{z}$ un échantillon $\mathbf{z}_m^*$ suivant $\mathbf{p}$.
Estimer $\widehat{\phi}_m$ sur l'échantillon $\mathbf{z}_m^*$.
Calculer à partir de l'échantillon initial $\mathbf{z}$:

$$\begin{aligned} l_m(i) &= Q\left(y_i, \widehat{\phi}_m(\mathbf{x}_i)\right) \quad i = 1, \ldots, n; \quad (Q : \text{fonction perte}) \\ \widehat{\mathcal{E}_m} &= \sum_{i=1}^{n} p_i l_m(i); \\ w_i &= g(l_m(i))p_i. \quad (g \text{ continue non décroissante}) \end{aligned}$$

Calculer les nouvelles probabilités : $p_i \leftarrow \frac{w_i}{\sum_{i=1}^{n} w_i}$.
Fin Pour
Calculer $\widehat{\phi}(\mathbf{x}_0)$ moyenne ou médiane des prévisions $\widehat{\phi}_m(\mathbf{x}_0)$ pondérées par des coefficients $\log(\frac{1}{\beta_m})$.

3.5 Modèle additif pas à pas

Le bon comportement du *boosting* par rapport à d'autres techniques de discrimination est difficile à expliquer ou justifier par des arguments théoriques. Un premier pas important en ce sens a été franchi par Breiman (1999) qui propose de considérer le *boosting* comme un algorithme global d'optimisation. Cette approche est reprise par Hastie et col. (2001) qui présentent le *boosting* dans le cas binaire sous la forme d'une approximation de la fonction ϕ par un modèle additif construit pas à pas :

$$\widehat{\phi}(\boldsymbol{x}) = \sum_{m=1}^{M} c_m \delta(\boldsymbol{x}; \gamma_m)$$

est cette combinaison où c_m est un paramètre, δ le classifieur (faible) de base fonction de $\boldsymbol{x}$ et dépendant d'un paramètre γ_m. Si Q est une fonction perte, il s'agit, à chaque étape, de résoudre :

$$(c_m, \gamma_m) = \arg\min_{(c,\gamma)} \sum_{i=1}^{n} Q(y_i, \widehat{\phi}_{m-1}(\boldsymbol{x}_i) + c\delta(\boldsymbol{x}_i; \gamma));$$

$\widehat{\phi}_m(\boldsymbol{x}) = \widehat{\phi}_{m-1}(\boldsymbol{x}) + c_m\delta(\boldsymbol{x}; \gamma_m)$ est alors une amélioration de l'ajustement précédent.

Dans le cas d'*adaboost* pour l'ajustement d'une fonction binaire, la fonction perte utilisée est $Q(y, \phi(\boldsymbol{x})) = \exp[-y\phi(\boldsymbol{x})]$. il s'agit donc de résoudre :

$$\begin{aligned} (c_m, \gamma_m) &= \arg\min_{(c,\gamma)} \sum_{i=1}^{n} \exp\left[-y_i(\widehat{\phi}_{m-1}(\boldsymbol{x}_i) + c\delta(\boldsymbol{x}_i; \gamma))\right]; \\ &= \arg\min_{(c,\gamma)} \sum_{i=1}^{n} w_i^m \exp\left[-cy_i\delta(\boldsymbol{x}_i; \gamma)\right] \\ \text{avec} \quad w_i^m &= \exp[-y_i\widehat{\phi}_{m-1}(\boldsymbol{x}_i)]; \end{aligned}$$

w_i^m ne dépendant ni de c ni de γ, il joue le rôle d'un poids fonction de la qualité de l'ajustement précédent. Quelques développements complémentaires montrent que la solution du problème de minimisation est ob-

tenue en deux étapes : recherche du classifieur optimal puis optimisation du paramètre c_m.

$$\begin{aligned}\gamma_m &= \arg\min_{\gamma} \sum_{i=1}^{n} \mathbf{1}\{y_i \neq \delta(\boldsymbol{x}_i;\gamma)\},\\ c_m &= \frac{1}{2}\log\frac{1-\widehat{\mathcal{E}}_p}{\mathcal{E}_p}\end{aligned}$$

avec $\widehat{\mathcal{E}}_p$ erreur apparente de prévision tandis que les w_i sont mis à jour avec :

$$w_i^{(m)} = w_i^{(m-1)} \exp[-c_m].$$

On montre ainsi qu'*adaboost* approche ϕ pas à pas par un modèle additif en utilisant une fonction perte exponentielle tandis que d'aubres types de *boosting* sont définis sur la base d'une autre fonction perte :

AdaBoost $Q(y,\phi(\boldsymbol{x})) = \exp[-y\phi(\boldsymbol{x})]$,

LogitBoost $Q(y,\phi(\boldsymbol{x})) = \log_2(1+\exp[-2y\phi(\boldsymbol{x})]$,

L^2**Boost** $Q(y,\phi(\boldsymbol{x})) = (y-\phi(\boldsymbol{x}))^2/2$.

D'autres fonctions pertes sont envisageables pour, en particulier, un algorithme plus robuste face à un échantillon d'apprentissage présentant des erreurs de classement dans le cas de la discrimination ou encore des valeurs atypiques (*outliers*) dans le cas de la régression. Hastie et col. (2001) comparent les intérêts respectifs de plusieurs fonctions pertes. Celles jugées robustes (entropie en discrimination, valeur absolue en régression) conduisent à des algorithmes plus compliqués à mettre en œuvre.

3.6 Régression et boosting

Dans le même esprit d'approximation adaptative, Friedman (2002) propose sous l'acronyme MART (*multiple additive regression trees*) un algorithme basé sur des arbres de régression pour traité le cas quantitatif en supposant la fonction perte seulement différentiable. Le principe de base est le même que pour *Adaboost*, construire une séquence de modèles de sorte que chaque étape, chaque modèle ajouté à la combinaison, apparaisse comme un pas vers une meilleure solution. Ce pas est franchi dans la direction du gradient, approché par un arbre de régression, de la fonction perte.

Algorithm 10 MART (Multiple additive regression trees)

Soit $\mathbf{x}_0$ à prévoir
Initialiser $\widehat{\phi}_0 = \arg\min_\gamma \sum_{i=1}^n Q(y_i,\gamma)$
Pour $m = 1$ à M **Faire**
  Calculer $r_i m = -\left[\frac{\delta Q(y_i,\phi(\boldsymbol{x}_i))}{\delta\phi(\boldsymbol{x}_i)}\right]_{\phi=\phi_{m-1}}$,
  Ajuster un arbre de régression aux $r_m i$ donnant les feuilles ou régions terminales $R_{jm}; j = 1,\ldots,J_m$.

  Pour $m = 1$ à M **Faire**
    Calculer $\gamma_{jm} = \arg\min_\gamma \sum_{\boldsymbol{x}_i\in R_{jm}} Q(y_i, \widehat{\phi}_{m-1}+\gamma)$.
  Fin Pour
  Mise à jour : $\widehat{\phi}_m(\boldsymbol{x}) = \widehat{\phi}_m(\boldsymbol{x}) \sum_{j=1}^{J_m} \gamma_{jm}\mathbf{1}\{\boldsymbol{x}\in R_{jm}\}$.
Fin Pour
Résultat : $\widehat{\phi}_M(\boldsymbol{x}_0)$.

L'algorithme est initialisé par un terme constant c'est-à-dire encore un arbre à une feuille. Les expressions du gradient reviennent simplement à calculer les résidus r_{mj} du modèle à l'étape précédente. Les termes correctifs γ_{jm} sont ensuite optimisés pour chacune des régions R_{jm} définies par l'arbre de régression ajustant les résidus. Un algorithme de discrimination est similaire calculant autant de probabilités que de classes à prévoir.

3.7 Compléments

De nombreuses adaptations ont été proposées à partir de l'algorithme initial. Elles font intervenir différentes fonctions pertes offrant des propriétés de robustesse ou adaptées à une variable cible Y quantitative ou qualitative à plusieurs classes : *Adaboost* M1, M2, MH ou encore MR. Schapire (2002) liste une bibliographie détaillée.

Sur-ajustement

Dans le dernier algorithme, le nombre d'itérations peut être contrôlé par un échantillon de validation. Comme pour d'autres méthodes (perceptron), il suffit d'arrêter la procédure lorsque l'erreur estimée sur cet échantillon arrive à se dégrader. Une autre possibilité consiste à ajouter un coefficient de rétrécissement (*shrinkage* comme en régression ridge). Compris entre 0 et 1, celui-ci pénalise l'ajout d'un nouveau modèle dans l'agrégation. Il joue le rôle du coefficient *decay* du perceptron) et, si sa valeur est petite $(< 0,1)$ cela conduit à accroître le nombre d'arbres mais entraîne des améliorations de la qualité de prévision. Le *boosting* est un algorithme qui peut effectivement converger exactement, donc vers une situation de sur-apprentissage. En pratique, cette convergence peut être rendue suffisamment lente pour être facilement contrôlée.

Interprétation

L'interprétabilité des arbres de décision sont une des raisons de leur succès. Leur lecture ne nécessite pas de compétences particulières en statistique. Cette propriété est évidemment perdue par l'agrégation d'arbres ou de tout autre modèle. Néanmoins, surtout si le nombre de variables est très grand, il est important d'avoir une indication de l'importance relative des variables entrant dans la modélisation.

Des critères d'importance des variables sont néanmoins faciles à calculer comme dans le cas des forêts aléatoires.

Instabilité

Tous les auteurs ont remarqué la grande instabilité des modèles construits à base d'arbres : une légère modification des données est susceptible d'engendrer de grandes modifications dans les paramètres (les seuils et feuilles) du modèle. C'est justement cette propriété qui rend cette technique très appropriée à une amélioration par agrégation. Breiman (1998), pour les arbres de classification, puis Gey et Poggi (2002), pour les arbres de régression, détaillent et quantifient en pratique l'influence de cette instabilité ainsi que celle de l'apport potentiel du *boosting* par rapport au *bagging*.

Propriétés

Les justifications théoriques des bons résultats du *boosting* et principalement la résistance au sur-ajustement sont encore l'objet de travaux intenses suivant différentes pistes. La difficulté vient de ce que l'application de ce type d'algorithme sur une méthode donnée, fait généralement mieux que l'asymptotique (en faisant croître la taille de l'échantillon) pour cette même méthode. Les approches usuelles de la statistique asymptotique sont mises en défaut et les bornes obtenues pour majorer les erreurs d'estimations ou de prévision sont trop grossières pour rendre compte de l'efficacité effective de la méthode. On trouve ainsi, empiriquement, que l'erreur de prévision ou de généralisation peut continuer à décroître longtemps après que l'erreur d'ajustement se soit annulée. Parmi les pistes explorées, une approche "stochastique" considère que, même déterministe, l'algorithme simule une dynamique markovienne (Blanchard, 2001). Une deuxième, rappelée ci-dessus, présente le *boosting* comme une procédure d'optimisation globale par une méthode de gradient (Friedman, 2001). D'autres enfin (par exemple Lugosi et Vayatis, 2001), plus probantes, utilisent des inégalités de Vapnik pour montrer que, sous des hypothèses raisonnables et vérifiées dans les cas usuels : convexité et régularité de la fonction perte (exponentielle), arbres binaires, la probabilité d'erreur du *boosting* converge avec la taille n de l'échantillon vers celle du classifieur bayésien c'est-à-dire celui, optimal, obtenu en supposant connue la loi conjointe de X et Y.

FIG. 9.1 – Cancer : Évolution des taux d'erreur (%) sur les échantillons d'apprentissage et de test en fonction du nombre d'arbres dans le modèle avec adaboost.

Logiciels

Le *bagging* est très facile à programmer dans R mais il existe une librairie (`ipred`) qui en propose des implémentations efficaces. L'algorithme de *boosting* (Freund et Schapire, 1996), ou plutôt la version de Friedman et col. (2000) a été développée et interfacée avec R dans la librairie `gbm` tandis que Friedman fait commercialiser ses outils par la société *Salford System*. Schapire diffuse lui le logiciel *Boost texter* sur sa page pour des utilisations non commerciales.

Les forêts aléatoires (Breiman, 2001), sont estimées par un programme écrit en fortran interfacé avec R et distribuées avec la librairie `randomForest` de R.

D'autres implémentations sont accessibles dans des boîtes à outils mathlab.

4 Exemples

4.1 Cancer du sein

La prévision de l'échantillon test par ces algorithmes conduit aux matrices de confusion :

```
             bagging(ipred)   adaboost(gbm)     random forest
            benign malignant  benign malignant  benign malignant
benign          83         3      84         1      83         0
malignant        3        48       2        50       3        51
```

et, respectivement, des erreurs estimées de 4,4 et 2,2% pour cet exemple et avec les échantillons (apprentissage et test) tirés.

Il est remarquable de noter l'évolution des erreurs d'ajustement et de test sur cet exemple (figure 9.1) en fonction du nombre d'arbres estimés par adaboost. L'erreur d'apprentissage arrive rapidement à 0 tandis que celle de test continue à décroître avant d'atteindre un seuil. Cet algorithme est donc relativement robuste au sur-apprentissage avant, éventuellement, de se dégrader pour des raisons, sans doute, de précision numérique. Ce comportement à été relevé dans beaucoup d'exemples dans la littérature.

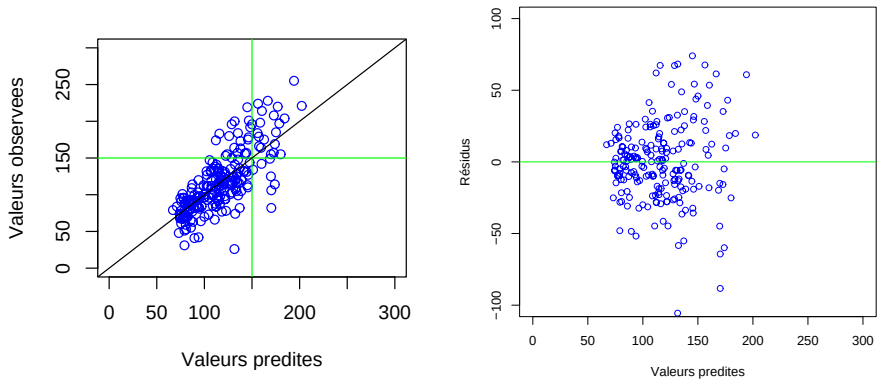

FIG. 9.2 – Ozone : Valeurs observées et résidus de l'échantillon test en fonction des valeurs prédites par une forêt aléatoire

4.2 Concentration d'ozone

Magré une bonne prévision quantitative, la prévision du dépassement de seuil reste difficile pour l'algorithme des forêts aléatoires. Par une régression ou une discrimination, le taux d'erreur obtenu est le même (12,5%) sur le même échantillon test et d'autres expérimentations sont nécessaires pour départager, ou non, les différentes méthodes. Il semble que, à travers plusieurs exemples, l'amélioration apportée à la prévision par des algorithmes d'agrégation de modèles soit nettement plus probante dans des situations difficiles c'est-à-dire avec beaucoup de variables explicatives et des problèmes de multicolinéarité.

Comme les réseaux de neurones, les algorithmes d'agrégation de modèles sont des boîtes noires. Néanmoins dans le cas des forêts, les critères d'importance donnent des indications sur le rôle de celles-ci. Les voici ordonnées par ordre croissant du critère basé sur celui de Gini pour la construction des arbres.

```
  jour station     lno    lno2 vmodule s_rmh2o   O3_pr   TEMPE
  2.54   13.58   21.78   23.33   24.77   31.19   43.87   67.66
```

Les variables prépondérantes sont celles apparues dans la construction d'un seul arbre.

4.3 Carte visa

Les arbres, qui acceptent à la fois des variables explicatives qualitatives et quantitatives en optimisant le découpage des variables quantitatives, se prêtent bien au traitement des données bancaires. on a vu qu'un seul arbre donnait des résultats semble-t-il très corrects. Naturellement les forêts constitués d'arbres se trouvent également performantes sur ces données en gagnant en stabilité et sans trop se poser de problème concernant l'optimisation de paramètres. Les TPs décrivent également les résultats proposés par les algorithmes de bagging et de boosting sur les arbres en faisant varier certains paramètres comme le *shrinkage* dans le cas du boosting.

Les graphiques de la figure 9.3 montrent bien l'insensibilité des forêts au sur-apprentissage. Les taux d'erreurs estimés, tant par bootstrap (out-of-bag), que sur un échantillon test, se stabilisent au bout de quelques centaines d'itérations. Il est même possible d'introduire dans le modèle toutes les variables quantitatives et qualitatives, avec certaines dupliquées, en laissant l'algorithme faire son choix. Cet algorithme conduit à un taux d'erreur de 10,5% sur l'échantillon test avec la matrice de confusion :

```
     Cnon Coui
Cnon  126   11
Coui   10   53
```

tandis que les coefficients d'importance :

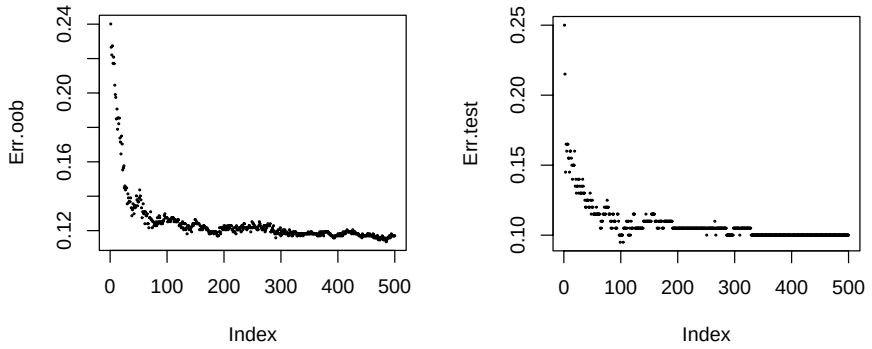

FIG. 9.3 – *Banque : Évolution du taux de mal classés estimés "out-of-bag" et sur l'échantillon test en fonction du nombre d'arbres intervenant dans la combinaison de modèles.*

```
QSMOY FACANL  RELAT DMVTPL QCREDL MOYRVL
20.97  26.77  29.98  36.81  40.31  50.01
```

mettent en évidence les variables les plus discriminantes. De son côté, le boosting (sans schrinkage) fournit des résultats tout à fait comparables avec un taux d'erreur de 11%.

4.4 Régime des souris

L'exemple reprend les données de Baccini et col. (2005) concernant les différences d'expression des gènes en croisant deux facteurs lors d'une expérience de régime alimentaire (5 régimes) chez des souris (2 génotypes). Ces données sont aussi introduites dans Baccini et Besse (2000). L'objectif des biologistes est de rechercher les gènes dont le comportement est le plus perturbé par les différentes situations de l'expérience : les génotypes ou les régimes. Il a été vu, par une simple analyse en composantes principales, que la distinction entre génotypes se visualise facilement ainsi que la caractérisation des gènes qui y participent. La discrimination des régimes est nettement plus difficile. Deux approches sont possibles pour répondre à cet objectif, la première consiste à exécuter une batterie de tests pour chercher les gènes significativement différentiellement exprimés en contrôlant "soigneusement" le niveau des tests à cause de leur multiplicité et donc de l'apparition factuelle de faux positifs. La deuxième (*wrapper method*) recherche le sous-ensemble de gènes conduisant à la meilleure discrimination à l'aide d'un classifieur donné. Compte tenu du nombre de gènes dans l'étude et de la difficulté à discriminer les régimes, les forêts aléatoires ont été privilégiées. L'avantage important de cette approche est sa robustesse aux problème de sur-apprentissage. L'indice d'importance est ensuite utilisé pour lister les gènes ou les représenter selon ce critère c'est-à-dire pour faire apparaître ceux qui, en moyenne sur l'ensemble des tirages *bootstrap*, contribuent le mieux à discriminer les modalités du facteur régime.

Dans le cas élémentaire de la discrimination des génotypes des souris, les gènes qui apparaissent les plus significatifs sont, par ordre décroissant : PMDCI, CAR1, THIOL, L.FABP, ALDH3, CYP3A11, PECI, GK, CYP4A10, ACBP, FAS, CPT2, BSEP, mHMGCoAS, ACOTH. La prévisiondes génotypes est presque sûre avec une estimation (out of bag) de l'erreur de prévisionde 2En revanche, la discrimination des régimes, beaucoup plus délicate, a été traitée conditionnellement au génotype. Le régime de rérérence est dans les deux cas le plus difficile à reconnaître. Le taux d'erreur obtenu est peu performant mais sans grande signification à cause du nombre de classes concernées. La figure 9.4 représente les gènes en fonction de leur importance pour la discrimination des régimes pour chacun des génotypes. C'est pour les souris PPARα que la discrimination des régimes est la plus difficile. Ce résultat s"interprètent sur le plan biologique comme une implication du récepteur PPARα dans les régulations géniques provoquées par les régimes alimentaires.

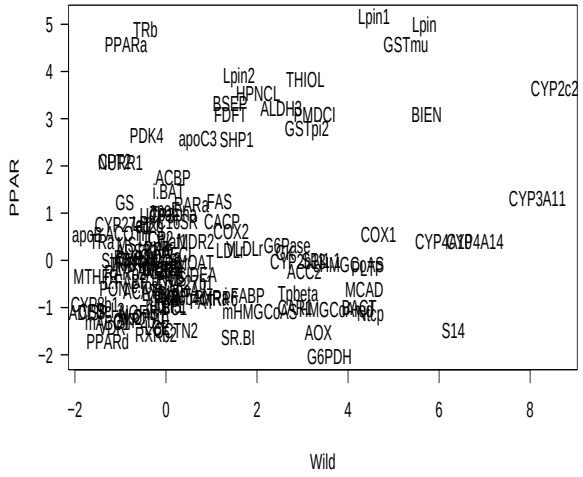

FIG. 9.4 – Souris : représentation des gènes en fonction de leur importance pour la discrimination des régimes à génotype fixé (WT sur l'axe horizontal et PPARα sur l'axe vertical).

Chapitre 10

Les *Support Vector Machines* (SVM)

1 Introduction

Les *Support Vector Machines* souvent traduit par l'appellation de Séparateur à Vaste Marge (SVM) sont une classe d'algorithmes d'apprentissage initialement définis pour la discrimination c'est-à-dire la prévision d'une variable qualitative initialement binaire. Ils ont été ensuite généralisés à la prévision d'une variable quantitative. Dans le cas de la discrimination d'une variable dichotomique, ils sont basés sur la recherche de l'*hyperplan de marge optimale* qui, lorsque c'est possible, classe ou sépare correctement les données tout en étant le plus éloigné possible de toutes les observations. Le principe est donc de trouver un classifieur, ou une fonction de discrimination, dont la capacité de généralisation (qualité de prévision) est la plus grande possible.

Cette approche découle directement des travaux de Vapnik en théorie de l'apprentissage à partir de 1995. Elle s'est focalisée sur les propriétés de généralisation (ou prévision) d'un modèle en contrôlant sa complexité. Voir à ce sujet le chapitre 5 section 3.3 concernant la dimension de Vapnik Chernovenkis qui est un indicateur du pouvoir séparateur d'une famille de fonctions associé à un modèle et qui en contrôle la qualité de prévision. Le principe fondateur des SVM est justement d'intégrer à l'estimation le contrôle de la complexité c'est-à-dire le nombre de paramètres qui est associé dans ce cas au nombre de vecteurs supports. L'autre idée directrice de Vapnik dans ce développement, est d'éviter de substituer à l'objectif initial : la discrimination, un ou des problèmes qui s'avèrent finalement plus complexes à résoudre comme par exemple l'estimation non-paramétrique de la densité d'une loi multidimensionnelle en analyse discriminante.

Le principe de base des SVM consiste de ramener le problème de la discrimination à celui, linéaire, de la recherche d'un hyperplan optimal. Deux idées ou astuces permettent d'atteindre cet objectif :

- La première consiste à définir l'hyperplan comme solution d'un problème d'optimisation sous contraintes dont la fonction objectif ne s'exprime qu'à l'aide de produits scalaires entre vecteurs et dans lequel le nombre de contraintes "actives" ou vecteurs supports contrôle la complexité du modèle.
- Le passage à la recherche de surfaces séparatrices non linéaires est obtenu par l'introduction d'une fonction noyau (*kernel*) dans le produit scalaire induisant implicitement une transformation non linéaire des données vers un espace intermédiaire (*feature space*) de plus grande dimension. D'où l'appellation couramment rencontrée de machine à noyau ou *kernel machine*. Sur le plan théorique, la fonction noyau définit un espace hilbertien, dit auto-reproduisant et isométrique par la transformation non linéaire de l'espace initial et dans lequel est résolu le problème linéaire.

Cet outil devient largement utilisé dans de nombreux types d'application et s'avère un concurrent sérieux des algorithmes les plus performants (agrégation de modèles). L'introduction de noyaux, spécifiquement adaptés à une problématique donnée, lui confère une grande flexibilité pour s'adapter à des situations très diverses (reconnaissance de formes, de séquences génomiques, de caractères, détection de spams, diagnostics...). À noter que, sur le plan algorithmique, ces algorithmes sont plus pénalisés par le nombre d'observations, c'est-à-dire le nombre de vecteurs supports potentiels, que par le nombre de variables. Néanmoins, des versions performantes des algorithmes permettent de prendre en compte des bases de données volumineuses dans des temps de calcul acceptables.

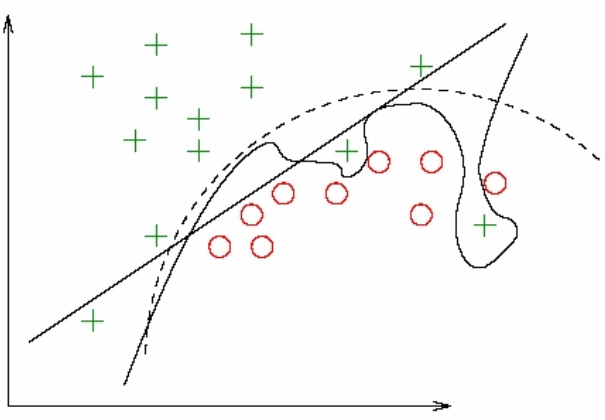

FIG. 10.1 – Sous-ajustement linéaire et sur-ajustement local (proches voisins) d'un modèle quadratique.

Le livre de référence sur ce sujet est celui de Schölkopf et Smola (2002). De nombreuses introduction et présentations des SVM sont accessibles sur des sites comme par exemple : `www.kernel-machines.org`. Guermeur et Paugam-Moisy (1999) en proposent une en français.

2 Principes

2.1 Problème

Comme dans toute situation d'apprentissage, on considère une variable Y à prédire mais qui, pour simplifier cette introduction élémentaire, est supposée dichotomique à valeurs dans $\{-1, 1\}$. Soit $\boldsymbol{X} = X^1, \ldots, X^p$ les variables explicatives ou prédictives et $\phi(\boldsymbol{x})$ un modèle pour Y, fonction de $\boldsymbol{x} = \{x^1, \ldots, x^p\} \in \mathbb{R}^p$. Plus généralement on peut simplement considérer la variable $\boldsymbol{X}$ à valeurs dans un ensemble $\mathcal{F}$.

On note

$$\boldsymbol{z} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$$

un échantillon statistique de taille n et de loi F inconnue. L'objectif est donc de construire une estimation $\widehat{\phi}$ de ϕ, fonction de $\mathcal{F}$ dans $\{-1, 1\}$, de sorte que la probabilité :

$$P(\phi(\boldsymbol{X}) \neq Y)$$

soit minimale.

Dans ce cas (Y dichotomique), le problème se pose comme la recherche d'une frontière de décision dans l'espace $\boldsymbol{\mathcal{F}}$ des valeurs de $\boldsymbol{X}$. De façon classique, un compromis doit être trouvé entre la *complexité* de cette frontière, qui peut s'exprimer aussi comme sa capacité à *pulvériser* un nuage de points par la VC dimension, donc la capacité d'*ajustement* du modèle, et les qualités de *généralisation* ou prévision de ce modèle. Ce principe est illustré par la figure 10.1.

2.2 Marge

La démarche consiste à rechercher, plutôt qu'une fonction $\widehat{\phi}$ à valeurs dans $\{-1, 1\}$, une fonction réelle f dont le signe fournira la prévision :

$$\widehat{\phi} = \text{signe}(f).$$

L'erreur s'exprime alors comme la quantité :

$$P(\phi(\boldsymbol{X}) \neq Y) = P(Y f(\boldsymbol{X}) \leq 0).$$

De plus, la valeur absolue de cette quantité $|Y f(\boldsymbol{X})|$ fournit une indication sur la confiance à accorder au résultat du classement.

On dit que $Y f(\boldsymbol{X})$ est la *marge* de f en $(\boldsymbol{X}, Y)$.

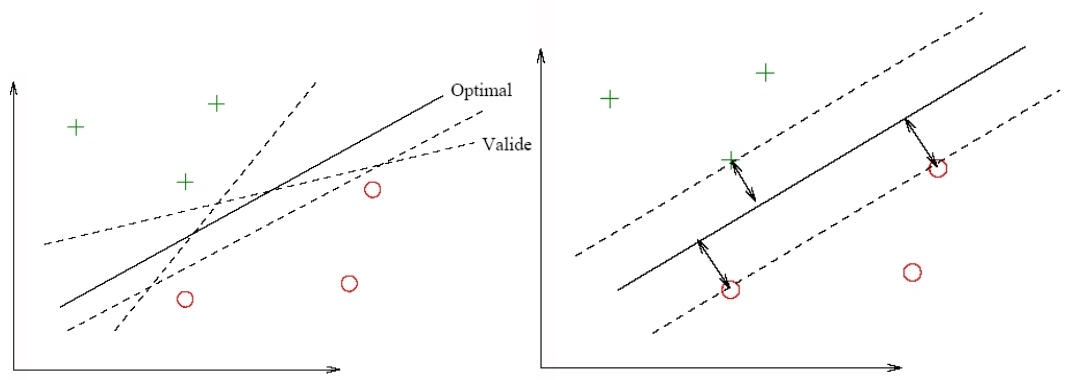

FIG. 10.2 – Recherche d'un hyperplan de séparation optimal au sens de la marge maximale.

2.3 Espace intermédiaire

Une première étape consiste à transformer les valeurs de $\boldsymbol{X}$, c'est-à-dire les objets de $\mathcal{F}$ par une fonction Φ à valeurs dans un espace $\mathcal{H}$ intermédiaire (*feature space*) muni d'un *produit scalaire*. Cette transformation est fondamentale dans le principe des SVM, elle prend en compte l'éventuelle non linéarité du problème posé et le ramène à la résolution d'une séparation linéaire. Ce point est détaillé dans une section ultérieure. Traitons tout d'abord le cas linéaire c'est-à-dire le cas où Φ est la fonction identité.

3 Séparateur linéaire

3.1 Hyperplan séparateur

La résolution d'un problème de séparation linéaire est illustré par la figure 10.2. Dans le cas où la séparation est possible, parmi tous les hyperplans solutions pour la séparation des observations, on choisit celui qui se trouve le plus "loin" possible de tous les exemples, on dit encore, de *marge* maximale.

Dans le cas linéaire, un hyperplan est défini à l'aide du produit scalaire de $\mathcal{H}$ par son équation :

$$\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0$$

où $\boldsymbol{w}$ est un vecteur orthogonal au plan tandis que le signe de la fonction

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$$

indique de quel côté se trouve le point $\boldsymbol{x}$ à prédire. Plus précisément, un point est bien classé si et seulement si :

$$yf(\boldsymbol{x}) > 0$$

mais, comme le couple $(\boldsymbol{w}, b)$ qui caractérise le plan est défini à un coefficient multiplicatif près, on s'impose :

$$yf(\boldsymbol{x}) \geq 1.$$

Un plan $(\boldsymbol{w}, b)$ est un séparateur si :

$$y_i f(\boldsymbol{x}_i) \geq 1 \quad \forall i \in \{1, \ldots, n\}.$$

La distance d'un point $\boldsymbol{x}$ au plan $(\boldsymbol{w}, b)$ est donnée par :

$$d(\mathbf{x}) = \frac{|\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b|}{\|\boldsymbol{w}\|} = \frac{|f(\boldsymbol{x})|}{\|\boldsymbol{w}\|}$$

et, dans ces conditions, la marge du plan a pour valeur $\frac{2}{\|\boldsymbol{w}\|^2}$. Chercher le plan séparateur de marge maximale revient à résoudre le problème ci-dessous d'optimisation sous contraintes (problème primal) :

$$\begin{cases} \min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \\ \text{avec } \forall i, y_i(<\boldsymbol{w}, \boldsymbol{x}_i> +b) \geq 1. \end{cases}$$

Le problème dual est obtenu en introduisant des multiplicateurs de Lagrange. La solution est fournie par un *point-selle* $(\boldsymbol{w}^*, b^*, \boldsymbol{\lambda}^*)$ du lagrangien :

$$L(\boldsymbol{w}, b, \boldsymbol{\lambda}) = 1/2\|w\|_2^2 - \sum_{i=1}^{n} \lambda_i \left[y_i \left(<\boldsymbol{w}, \boldsymbol{x}_i> +b\right) - 1\right].$$

Ce point-selle vérifie en particulier les conditions :

$$\lambda_i^* \left[y_i \left(<\boldsymbol{w}^*, \boldsymbol{x}_i> +b^*\right) - 1\right] = 0 \quad \forall i \in \{1, \ldots, n\}.$$

Les *vecteurs support* sont les vecteurs $\boldsymbol{x}_i$ pour lesquels la contrainte est active, c'est-à-dire les plus proches du plan, et vérifiant donc :

$$y_i \left(<\boldsymbol{w}^*, \boldsymbol{x}_i> +b^*\right) = 1.$$

Les conditions d'annulation des dérivées partielles du lagrangien permettent d'écrire les relations que vérifient le plan optimal, avec les λ_i^* non nuls seulement pour les points supports :

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \lambda_i^* y_i \boldsymbol{x}_i \quad \text{et} \quad \sum_{i=1}^{n} \lambda_i^* y_i = 0.$$

Ces contraintes d'égalité permettent d'exprimer la formule duale du lagrangien :

$$W(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j <\boldsymbol{x}_i, \boldsymbol{x}_j> .$$

Pour trouver le point-selle, il suffit alors de maximiser $W(\boldsymbol{\lambda})$ avec $\lambda_i \geq 0$ pour tout $i \in \{1, ...n\}$. La résolution de ce problème d'optimisation quadratique de taille n, le nombre d'observations, fournit l'équation de l'hyperplan optimal :

$$\sum_{i=1}^{n} \lambda_i^* y_i <\boldsymbol{x}, \boldsymbol{x}_i> +b^* = 0 \quad \text{avec} \quad b^0 = -\frac{1}{2}\left[<\boldsymbol{w}^*, sv_{class+1}> + <\boldsymbol{w}^*, sv_{class-1}>\right].$$

Pour une nouvelle observation $\boldsymbol{x}$ non apprise présentée au modèle, il suffit de regarder le signe de l'expression :

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \lambda_i^* y_i \left\langle \boldsymbol{x}, \boldsymbol{x}_i \right\rangle + b^*$$

pour savoir dans quel demi-espace cette forme se trouve, et donc quelle classe il faut lui attribuer.

3.2 Cas non séparable

Lorsque les observations ne sont pas séparables par un plan, il est nécessaire d'"assouplir" les contraintes par l'introduction de termes d'erreur ξ_i qui en contrôlent le dépassement :

$$y_i \left\langle \boldsymbol{w}, \boldsymbol{x}_i \right\rangle + b \geq +1 - \xi_i \quad \forall i \in \{1, \ldots, n\}.$$

Le modèle attribue ainsi une réponse fausse à un vecteur $\boldsymbol{x}_i$ si le ξ_i correspondant est supérieur à 1. La somme de tous les ξ_i représente donc une borne du nombre d'erreurs.

Le problème de minimisation est réécrit en introduisant une pénalisation par le dépassement de la contrainte :

$$\begin{cases} \min \frac{1}{2}\|\boldsymbol{w}\|^2 + \delta \sum_{i=1}^{n} \xi_i \\ \forall i, y_i \left\langle \boldsymbol{w}, \boldsymbol{x}_i \right\rangle + b \geq +1 - \xi_i \end{cases}$$

Remarques

- Le paramètre δ contrôlant la pénalisation est à régler. Plus il est grand et plus cela revient à attribuer une forte importance à l'ajustement. Il est le paramètre qui ajuste le compromis entre bon ajustement et bonne généralisation.
- Le problème dans le cas non séparable se met sous la même forme duale que dans la cas séparable à une différence près : les coefficients λ_i sont tous bornés par la constante δ de contrôle de la pénalisation.
- De nombreux algorithmes sont proposés pour résoudre ces problèmes d'optimisation quadratique. Certains, proposant une décomposition de l'ensemble d'apprentissage, sont plus particulièrement adaptés à prendre en compte un nombre important de contraintes lorsque n, le nombre d'observation, est grand.
- On montre par ailleurs que la recherche des hyperplans optimaux répond bien au problème de la "bonne" généralisation. On montre aussi que, si l'hyperplan optimal peut être construit à partir d'un petit nombre de vecteurs supports, par rapport à la taille de la base d'apprentissage, alors la capacité en généralisation du modèle sera grande, indépendamment de la taille de l'espace.
- Plus précisément, on montre que, si les $\boldsymbol{X}$ sont dans une boule de rayon R, l'ensemble des hyperplans de marge fixée δ a une VC-dimension bornée par

$$\frac{R^2}{\delta^2} \quad \text{avec } \|\boldsymbol{x}\| \leq R.$$

- L'erreur par validation croisée (*leave-one-out*) et bornée en moyenne par le nombre de vecteurs supports. Ces bornes d'erreur sont bien relativement prédictives mais néanmoins trop pessimistes pour être utiles en pratique.

4 Séparateur non linéaire

4.1 Noyau

Revenons à la présentation initiale du problème. Les observations faites dans l'ensemble $\mathcal{F}$ (en général $\mathbb{R}^p$) sont considérées comme étant transformées par une application non linéaire Φ de $\mathcal{F}$ dans $\mathcal{H}$ muni d'un produit scalaire et de plus grande dimension.

Le point important à remarquer, c'est que la formulation du problème de minimisation ainsi que celle de sa solution :

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \lambda_i^* y_i \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle + b^*$$

ne fait intervenir les éléments $\boldsymbol{x}$ et $\boldsymbol{x}'$ que par l'intermédiaire de *produits scalaires* : $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$. En conséquence, il n'est pas nécessaire d'expliciter la transformation Φ, ce qui serait souvent impossible, à condition de savoir exprimer les produits scalaires dans $\mathcal{H}$ à l'aide d'une fonction $k : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ symétrique appelée *noyau* de sorte que :

$$k(\boldsymbol{x}, \boldsymbol{x'}) = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x'}) \rangle .$$

Bien choisi, le noyau permet de matérialiser une notion de "proximité" adaptée au problème de discrimination et à sa structure de données.

Exemple

Prenons le cas trivial où $\boldsymbol{x} = (x_1, x_2)$ dans $\mathbb{R}^2$ et $\Phi(\boldsymbol{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$ est explicite. Dans ce cas, $\mathcal{H}$ est de dimension 3 et le produit scalaire s'écrit :

$$\begin{aligned} \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') \rangle &= x_1^2 x_1'^2 + 2x_1x_2x_1'x_2' + x_2^2 x_2'^2 \\ &= (x_1x_1' + x_2x_2')^2 \\ &= \langle \boldsymbol{x}, \boldsymbol{x}' \rangle^2 \\ &= k(\boldsymbol{x}, \boldsymbol{x'}). \end{aligned}$$

Le calcul du produit scalaire dans $\mathcal{H}$ ne nécessite pas l'évaluation explicite de Φ. D'autre part, le plongement dans $\mathcal{H} = \mathbb{R}^\ni$ peut rendre possible la séparation linéaire de certaines structures de données (cf. figure 10.3).

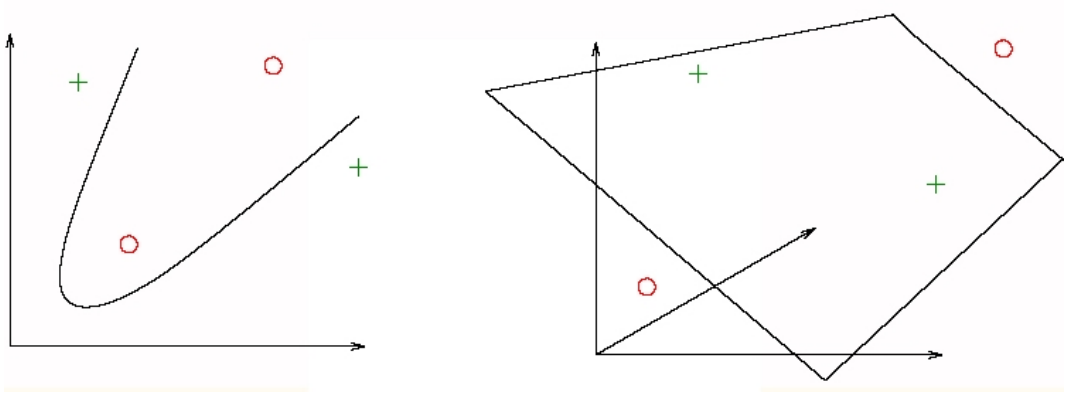

FIG. 10.3 – Rôle de l'espace intermédiaire dans la séparation des données.

4.2 Condition de Mercer

Une fonction $k(.,.)$ symétrique est un noyau si, pour tous les $\boldsymbol{x}_i$ possibles, la matrice de terme général $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ est une matrice définie positive c'est-à-dire quelle définit une matrice de produit scalaire.

Dans ce cas, on montre qu'il existe un espace $\mathcal{H}$ et une fonction Φ tels que :

$$k(\boldsymbol{x}, \boldsymbol{x'}) = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x'}) \rangle .$$

Malheureusement, cette condition théorique d'existence est difficile à vérifier et, de plus, elle ne donne aucune indication sur la construction de la fonction noyau ni sur la transformation Φ. La pratique consiste à combiner des noyaux simples pour en obtenir des plus complexes (multidimensionnels) associés à la situation rencontrée.

4.3 Exemples de noyaux

- Linéaire
$$k(\boldsymbol{x}, \boldsymbol{x'}) = \langle \boldsymbol{x}, \boldsymbol{x'} \rangle$$
- Polynômial
$$k(\boldsymbol{x}, \boldsymbol{x'}) = (c + \langle \boldsymbol{x}, \boldsymbol{x'} \rangle)^d$$
- Gaussien
$$k(\boldsymbol{x}, \boldsymbol{x'}) = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{x'}\|^2}{2\sigma^2}}$$

Beaucoup d'articles sont consacrés à la construction d'un noyau plus ou moins exotique et adapté à une problématique posée : reconnaissance de séquences, de caractères, l'analyse de textes... La grande flexibilité dans la définition des noyaux, permettant de définir une notion adaptée de similitude, confère beaucoup d'efficacité à cette approche à condition bien sur de construire et tester le bon noyau. D'où apparaît encore l'importance de correctement évaluer des erreurs de prévision par exemple par validation croisée.

Attention, les SVM à noyaux RBF gaussiens, pour lesquels, soit on est dans le cas séparable, soit la pénalité attribuée aux erreurs est autorisée à prendre n'importe quelle valeur, ont une VC-dimension infinie.

4.4 SVM pour la régression

Les SVM peuvent également être mis en oeuvre en situation de régression, c'est-à-dire pour l'approximation de fonctions quand Y est quantitative. Dans le cas non linéaire, le principe consiste à rechercher une estimation de la fonction par sa décomposition sur une base fonctionnelle. la forme générale des fonctions calculées par les SVM se met sous la forme :

$$\phi(\boldsymbol{x}, \boldsymbol{w}) = \sum_{i=1}^{\infty} w_i v_i(\boldsymbol{x}).$$

Le problème se pose toujours comme la minimisation d'une fonction coût, mais, plutôt que d'être basée sur un critère d'erreur quadratique (moindres carrés), celle-ci s'inspire des travaux de Huber sur la recherche de modèles robustes et utilise des écarts absolus.

On note $|.|_\epsilon$ la fonction qui est paire, continue, identiquement nulle sur l'intervalle $[0, \epsilon]$ et qui croit linéairement sur $[\epsilon, +\infty]$. La fonction coût est alors définie par :

$$E(\boldsymbol{w}, \gamma) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \phi(\boldsymbol{x}_i, \boldsymbol{w})|_\epsilon + \gamma \|\boldsymbol{w}\|^2$$

où γ est, comme en régression ridge, un paramètre de régularisation assurant le compromis entre généralisation et ajustement. De même que précédemment, on peut écrire les solutions du problèmes d'optimisation. Pour plus de détails, se reporter à Schölkopf et Smola (2002). Les points de la base d'apprentissage associés à un coefficient non nul sont là encore nommés vecteurs support.

Dans cette situation, les noyaux k utilisés sont ceux naturellement associés à la définition de bases de fonctions. Noyaux de splines ou encore noyau de Dériclet associé à un développement en série de Fourier sont des grands classiques. Ils expriment les produits scalaires des fonctions de la base.

5 Exemples

Même si les SVM s'appliquent à un problème de régression, nous n'illustrons que le cas plus classique de la discrimination.

5.1 Cancer du sein

La prévision de l'échantillon test par un Séparateur à Vaste marge conduit à la matrice de confusion :

```
ign malignant
  benign         83          1
  malignant       3         50
```

et donc une erreur estimée de 3%.

5.2 Concentration d'ozone

Un modèle élémentaire avec noyau par défaut (gaussien) et une pénalisation de 2 conduit à une erreur de prévision estimée à 12,0% sur l'échantillon test. La meilleure prévision de dépassement de seuil sur l'échantillon test initial est fournie par des SVM d'ε-régression. Le taux d'erreur est de 9,6% avec la matrice de confusion suivante :

```
          0   1
  FALSE 161  13
  TRUE    7  27
```

Ce résultat serait à confirmer avec des estimations sytématiques de l'erreur. Les graphiques de la figure 10.4 montre le bon comportement de ce prédicteur. Il souligne notamment l'effet "tunnel" de l'estimation qui accepte des erreurs autour de la diagonale pour se concentrer sur les observations plus éloignées donc plus difficiles à ajuster.

5.3 Carte Visa

Les données bancaires posent un problème car elles mixent variables quantitatives et qualitatives. Celles-ci nécessiteraient la construction de noyaux très spécifiques. Leur traitement par SVM n'est pas détaillé ici.

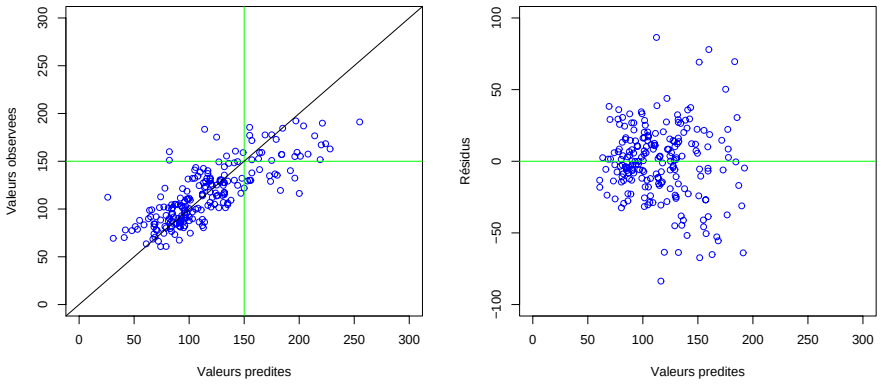

FIG. 10.4 – Ozone : Valeurs observées et résidus en fonction des valeurs prédites pour l'échantillon test.

Chapitre 11

Conclusion

Ce chapitre se propose de résumer les grandes lignes de ce cours dans une vue synthétique : méthodes et stratégies dans l'objectif d'une comparaison globale des méthodes sur les différents jeux de données (cancer, pollution, carte visa). Il évoque enfin les pièges fréquents de telles démarches et revient sur la place du statisticien.

1 Stratégies du data mining

Les chapitres précédents décrivent les outils de base du prospecteur de données tandis que les logiciels commerciaux en proposent une intégration plus ou moins complète, plus ou moins conviviale de mise en œuvre. En pratique, l'enchaînement de ces techniques permet la mise en place de *stratégies de fouille* bien définies. Celles-ci dépendent essentiellement des *types* de variables considérés et des *objectifs* poursuivis.

Types de variables

Explicatives L'ensemble des p variables explicatives ou prédictives est noté X, il est constitué de variables

- $X_{\mathbb{R}}$ toutes quantitatives[1],
- X_E toutes qualitatives,
- $X_{\mathbb{R}\cup E}$ un mélange de qualitatives et quantitatives.

À expliquer La variable à expliquer ou à prédire ou *cible* (target) peut être

- Y quantitative,
- Z qualitative à 2 modalités,
- T qualitative.

Objectifs

Trois objectifs principaux sont poursuivis dans les applications classiques de data mining :

i. **Exploration multidimensionnelle** ou réduction de dimension : production de graphes, d'un sous-ensemble de variables représentatives X_r, d'un ensemble de composantes C_q préalables à une autre technique.

ii. **Classification** (clustering) ou segmentation : production d'une variable qualitative T_r.

iii. **Modélisation (Y ou Z)/Discrimination (Z ou T)** production d'un modèle de prévision de Y (resp. Z, T).

D'autres méthodes plus spécifiques à certaines problématiques peuvent apparaître (analyse sensorielle, analyse conjointe, SARIMA... mais leur usage reste limité à des contextes bien particuliers.

Outils

Les méthodes utilisables se classent en fonction de leur objectif et des types de variables prédictives et cibles.

[1] Une variables explicative qualitative à 2 modalités (0,1) peut être considérée comme quantitative ; c'est l'indicatrice des modalités.

Exploration

`ACP` $X_{\mathbb{R}}$ et $\emptyset$
`AFCM` X_E et $\emptyset$
`AFD` $X_{\mathbb{R}}$ et T

Classification

`CAH` $X_{\mathbb{R}}$ et $\emptyset$
`NuéeDyn` $X_{\mathbb{R}}$ et $\emptyset$
`RNKoho` $X_{\mathbb{R}}$ et $\emptyset$

Modélisation

i. Modèle linéaire généralisé
   `RLM` $X_{\mathbb{R}}$ et Y
   `ANOVA` X_E et Y
   `ACOVA` $X_{\mathbb{R}\cup E}$ et Y
   `Rlogi` $X_{\mathbb{R}\cup E}$ et Z
   `Lglin` X_T et T
ii. Analyse discriminante
   `ADpar/nopar` $X_{\mathbb{R}}$ et T
iii. Classification and regression Tree
   `ArbReg` $X_{\mathbb{R}\cup E}$ et Y
   `ArbCla` $X_{\mathbb{R}\cup E}$ et T
iv. Réseaux neuronaux
   `percep` $X_{\mathbb{R}\cup E}$ et Y ou T
v. Agrégation de modèles
   Bagging $X_{\mathbb{R}\cup E}$ et Y ou T
   RandFor $X_{\mathbb{R}\cup E}$ et Y ou T
   Boosting $X_{\mathbb{R}\cup E}$ et Y ou T
vi. Support Vector Machine
   SVM-R $X_{\mathbb{R}\cup E}$ et Y
   SVM-C $X_{\mathbb{R}\cup E}$ et T

Stratégies

Les stratégies classiques de la fouille de données consistent à enchaîner les étapes suivantes :

i. **Extraction** de l'entrepôt des données éventuellement par sondage pour renforcer l'effort sur la qualité des données plutôt que sur la quantité.
ii. **Exploration**

- Tri à plat, étape élémentaire mais essentielle de vérification des données, de leur cohérence. Étude des distributions, transformation, recodage éventuel des variables quantitatives, regroupement de modalités des variables qualitatives, élimination de certaines variables (trop de données manquantes, quasi constantes, redondantes...). Gérer rigoureusement les codes des variables et de leurs modalités.
- Étude bivariée Recherche d'éventuelles relations non linéaires. Si les variables sont trop nombreuses, sélectionner les plus liées à la variable cible. Complétion des données manquantes.

iii. **Analyse**

Classification :
Pas de variable à expliquer
- En cas de variables $X_{\mathbb{R}\cup E}$ ou X_T, la classification est exécutée sur les C_q issues d'une AFCM des variables codées en classes.
- Caractérisation des classes par les variables initiales à l'aide des outils de discrimination.

Modélisation/Discrimination :
Une variable à expliquer Y, Z ou T
- Extraction d'un échantillon *test*,
- Estimation, optimisation (validation croisée) des modèles pour chacune des méthodes utilisables.
- Comparaison des performances des modèles optimaux de chaque méthode sur l'échantillon test.

iv. *Exploitation* du modèle et diffusion des résultats. Finalement, une fois que la bonne méthode associé au bon modèle ont été choisie, tout l'échantillon est regroupé pour faire une dernière estimation du modèle qui sera utilisé en exploitation.

2 Comparaison des résultats

2.1 Cancer du sein

Le programme d'estimation des modèles écrit en R a été automatisé afin de répéter 50 fois l'opération consistant à extraire aléatoirement 20% des observations pour constituer un échantillon test ; le reste constituant l'échantillon d'apprentissage. L'optimisation des paramètres est réalisée par validation croisée. Chaque

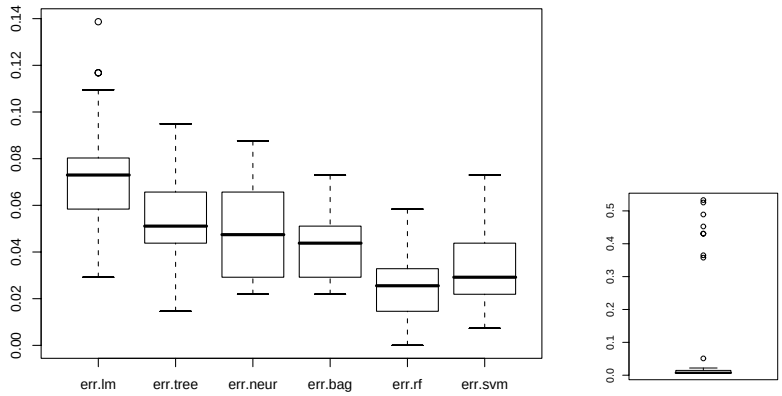

FIG. 11.1 – Cancer : Diagrammes boîtes des taux d'erreurs observés sur 50 échantillons tests et pour chaque méthode : régression logistique, arbre de décision, réseau de neurones, bagging, random forest, svm. Le boosting est mis de côté pour des problèmes d'échelle et de comportement erratique.

TAB. 11.1 – Banque : Moyennes des taux d'erreurs de classement calculés sur 30 échantillons test pour chaque modèle de prévision

Méthode	Adaboost	Arbre	Régression	Perceptron	Forêt
Moyenne	9.7	11.8	12.5	13.4	10.6
Écart-type	2.0	2.3	2.0	2.3	2.2

échantillon test fournit donc une estimation sans biais de l'erreur de prévision. La distribution de ces erreurs est alors représentée par des diagrammes en boîtes (cf ; fig. 11.1). Les résultats montrent le bon comportement des forêts aléatoires et les très bons résultats du boosting en général mais cet algorithme, sur cet exemple, peut réserver des surprises mal contrôlées et ici pas encore expliquées.

2.2 Concentration d'ozone

Toujours avec le même protocole, 50 échantillons tests on été successivement tirés afin d'estimer sans biais les erreurs de prévision. Les résultats sont présentés dans la figure 11.2. Les techniques d'agrégation (random forest) sont performantes mais pas de façon très significative. En fait, le problème ne présentant que peu de variables explicatives, une simple régression quadratique donne des résultats très satisfaisants et surtout facilement interprétables ; ils sont en effet chargés d'un sens "physique" pour le météorologue qui peut donc directement relever les faiblesses du modèle physique à la base de MOCAGE. Il semble bien que dans cet exemple, le nombre de variables explicatives n'est pas très important et le vrai modèle physique sous-jacent peu exotique. Dans ce cas, la régression quadratique est la plus appropriée. Remarque : la prévision des dépassements peut conduire à d'autres choix de méthode ou de stratégie en prévoyant directement le dépassement sans passer par la régression de la concentration. Ce point est laissé en attente car le nombre de dépassements observés (plus de 180) dans les stations est relativement rare donc difficiles à prévoir. Ceci nécessite plus de précautions : repondération des dépassements.

2.3 Carte visa

Trente échantillons tests ont successivement été tirés afin d'observer les distributions des taux de mauvais classement obtenus par différentes méthodes : arbre de décision, régression logistique, réseaux de neurones, *boosting* et forêt aléatoire.

Les algorithmes d'agrégation de modèles fournissent des résultats qui, en moyenne, se montrent sensi-

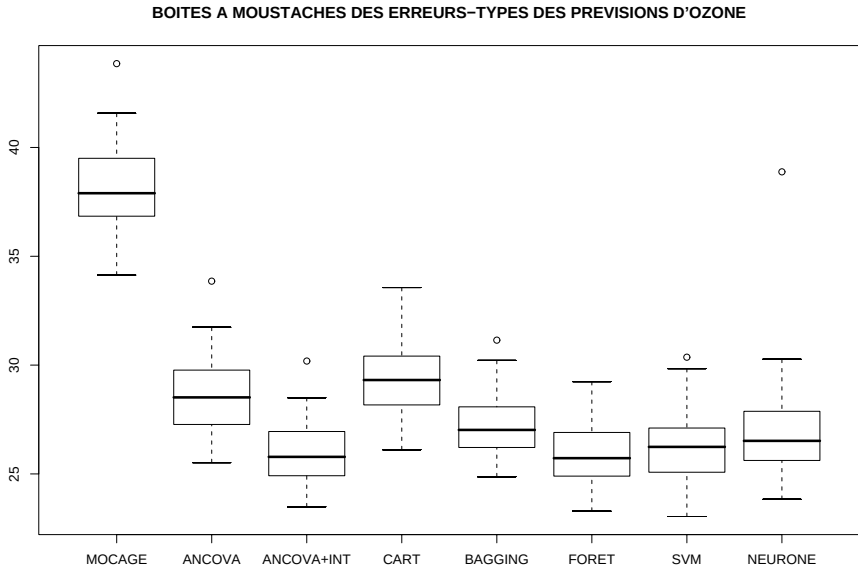

FIG. 11.2 – Ozone : Diagrammes boîtes des taux d'erreurs observés sur 50 échantillons tests et pour chaque méthode : mocage, régression linéaire, quadratique, arbre de décision, bagging, random forest, svm réseau de neurones.

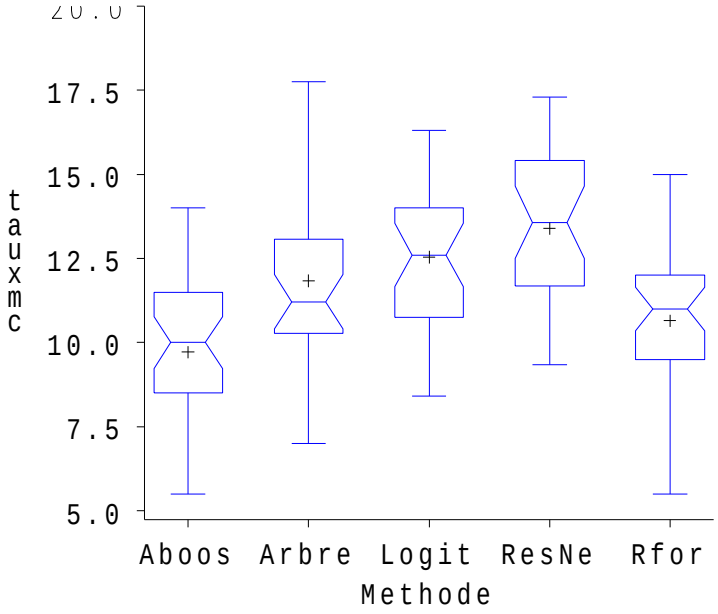

FIG. 11.3 – Banque : Diagrammes boîtes des taux d'erreurs observés sur 30 échantillons tests et pour chaque méthode.

blement plus performants (cf. figure 11.3 et tableau 11.1) sur un échantillon test. Les écarts-types, dépendant de la taille de l'échantillon test, y sont relativement stables. Les moyennes montrent, sur cet exemple, que le *boosting* prédit un peu mieux que les forêts aléatoires sans que des raisons bien spécifiques viennent l'expliquer. Bien sûr, ce qui est gagné en prédictibilité est perdu en interprétabilité par rapport à un modèle classique. Néanmoins le gain réalisé est souvent étonnant et des indices d'importance des variables restent disponibles.

3 Pièges

Les principaux pièges qui peuvent être rencontrés au cours d'une prospection peuvent être le résultat d'un *acharnement* en quête de sens (*data snooping*). Cela signifie qu'à force de creuser, contrairement à un prospecteur minier à la recherche de diamants bien réels, le prospecteur en données disposant d'un grand nombre de variables finit bien, en mode exploratoire, par trouver des relations semblant hautement significatives. Par exemple, au seuil classique, 5% des tests sont, à tord, significatifs et conduisent à des "faux positifs" ou des fausses corrélations. Il suffit donc d'en faire beaucoup, de croiser beaucoup de variables, pour nécessairement trouver du "sens" dans des données. Encore une fois, il est préférable d'éviter le fonctionnement "Shadock" (cf. figure 11.4) : *je n'ai qu'une chance sur un milliard de réussir ; je me dépêche donc de rater le plus d'essais possibles*.

En phase de modélisation, une sur-paramétrisation ou un sur-ajustement du modèle peut parfaitement expliquer des données sans pour autant que les résultats soient extrapolables ou généralisables à d'autres données que celles étudiées. Les résultats de prévision seront donc entachés d'une forte erreur relative liée à la variance des estimations des paramètres. C'est toujours le problème de trouver un bon compromis entre le biais d'un modèle plus ou moins faux et la variance des estimateurs. Nous insistons donc sur les indispensables phases de choix de modèles et comparaison des méthodes.

4 Rôle du statisticien

4.1 Des compétences multiples

Une bonne pratique du *Data Mining* nécessite de savoir articuler toutes les méthodes entrevues dans ce document. Rude tâche, qui ne peut être entreprise qu'à la condition d'avoir très bien spécifié les objectifs de l'étude. On peut noter que certaines méthodes poursuivent les mêmes objectifs prédictifs. Dans les bons cas, données bien structurées, elles fourniront des résultats très similaires, dans d'autres une méthode peut se révéler plus efficace compte tenu de la taille de l'échantillon ou géométriquement mieux adaptée à la topologie des groupes à discriminer ou encore en meilleure interaction avec les types des variables. Ainsi, il peut être important et efficace de découper en classes des variables prédictives quantitatives afin d'approcher de façon sommaire une version *non-linéaire* du modèle par une combinaison de variables indicatrices. Cet aspect est par exemple important en régression logistique ou avec un perceptron mais inutile avec des arbres de décisions qui intègrent ce découpage en classes dans la construction du modèle (seuils optimaux). D'autre part, les méthodes ne présentent pas toutes les mêmes facilités d'interprétation. Il n'y a pas de meilleur choix *a priori*, seul l'expérience et un protocole de *test* soigné permettent de se déterminer. C'est la raison pour laquelle des logiciels généralistes comme SAS (module Enterprise Miner) ne font pas de choix et offrent ces méthodes en parallèle pour mieux s'adapter aux données, aux habitudes de chaque utilisateur (client potentiel) et à la mode.

4.2 De l'utilité du statisticien

Le travail demandé déborde souvent du rôle d'un statisticien car la masse et la complexité des données peuvent nécessiter le développement d'interfaces et d'outils graphiques sophistiqués permettant un accès aisés aux données, comme à des résultats, par l'utilisateur finale à l'aide par exemple d'un simple navigateur sur l'intranet de l'entreprise. Néanmoins, au delà de ces aspects plus "informatiques", l'objectif principal reste une "quête de sens" en vue de faciliter les prises de décision tout en en préservant la fiabilité. Ainsi, la présence ou le contrôle d'une expertise statistique reste incontournable car la méconnaissance des limites et pièges des méthodes employées peut conduire à des aberrations discréditant la démarche et rendant caducs les investissements consentis. En effet, il faut bien admettre, et faire admettre, que, même si un petit quart

FIG. 11.4 – Shadoks : Tant qu'à pomper, autant que cela serve à quelque chose !

d'heure suffit pour se familiariser avec une interface graphique conviviale, la bonne compréhension des méthodes employées nécessite plusieurs heures de cours ou réflexion à Bac+5. Il devient tellement simple, avec les outils disponibles, de lancer des calculs, que certains n'hésitent pas à comparer prospecteur de données et chauffeur de voiture en arguant qu'il n'est pas nécessaire d'être un mécanicien accompli pour savoir conduire. Néanmoins, la conduite d'une modélisation, d'une segmentation, d'une discrimination, imposent à son auteur des choix plus ou moins implicites qui sont loin d'être neutres et qui dépassent largement en complexité celui du choix d'un carburant par le conducteur à la pompe.

Bibliographie

[1] A. AGRESTI : *Categorical data analysis*. Wiley, 1990.

[2] A. ANTONIADIS, J. BERRUYER et R. CARMONA : *Régression non linéaire et applications*. Economica, 1992.

[3] J.-M. AZAÏS et J.-M. BARDET : *Le modèle linéaire par l'exemple : régression, analyse de la variance et plans d'expéreinces illustrés avec R, SAS et Splus*. Dunod, 2005.

[4] A. BACCINI et P. BESSE : Data mining : 1. exploration statistique, 2000. www.ups-tlse.fr/Besse/enseignement.html.

[5] A. BACCINI, P. BESSE, S. DÉJEAN, P. MARTIN, C. ROBERT-GRANIÉ et M. SAN CRISTOBAL : Stratégies pour l'analyse statistique de données transcriptomiques. *Journal de la Société Française de Statistique*, 146:4–44, 2005.

[6] P.C. BESSE, C. LE GALL, N. RAIMBAULT et S. SARPY : Statistique et data mining. *Journal de la Société Française de Statistique*, 142:5–36, 2001.

[7] G. BLANCHARD : Generalization error bounds for aggregate classifiers. *In Proceedings of the MSRI international conference on nonparametric estimation and classification*, page , 2001.

[8] L. BREIMAN : Bagging predictors. *Machine Learning*, 26(2):123–140, 1996.

[9] L. BREIMAN : Arcing classifiers. *Annals of Statistics*, 26:801–849, 1998.

[10] L. BREIMAN : Prediction games and arcing algorithms. *Neural Computation*, 11:1493–1517, 1999.

[11] L. BREIMAN : Random forests. *Machine Learning*, 45:5–32, 2001.

[12] L. BREIMAN, J. FRIEDMAN, R. OLSHEN et C. STONE : *Classification and regression trees*. Wadsworth & Brooks, 1984.

[13] P.-A. CORNILLON et E. MATZNER-LØ BER : *Régression, Théorie et applications*. Springer, 2007.

[14] H. DRUCKER : Improving regressors using boosting techniques. *In* M. KAUFMANN, éditeur : *Proceedings of th 14th International Conference on Machine Learning*, pages 107–115, 1997.

[15] B. EFRON : *The Jackknife, the Bootstrap and other Resampling Methods*. SIAM, 1982.

[16] B. EFRON et R. TIBSHIRANI : Improvements on cross-validation : The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997.

[17] B. EFRON et R.J. TIBSHIRANI : *An introduction to the bootstrap*. Chapman and Hall, 1993.

[18] Y. FREUND et R.E. SCHAPIRE : Experiments with a new boosting algorithm. *In Machine Learning : proceedings of the Thirteenth International Conference*, pages 148–156. Morgan Kaufman, 1996. San Francisco.

[19] Y. FREUND et R.E. SCHAPIRE : Experiments with a new boosting algorithm. *Journal of Computer and System Sciences*, 55:119–139, 1997.

[20] J. H. FRIEDMAN : Greedy function approximation : a gradient boosting machine. *Annals of Statistics*, 29:1189–1232., 2001.

[21] J. H. FRIEDMAN : Stochastic gradient boosting. *Computational Statisrics and Data Analysis*, 38: , 2002.

[22] J. H. FRIEDMAN, H. HASTIE et R. TIBSHIRANI : Additive logistic regression : a statistical view of boosting. *The Annals of Statistics*, 28:337–407, 2000.

[23] S. GEY et J.-M. POGGI : Boosting and instabillity for regression trees. Rapport technique 36, Université de Paris Sud, Mathématiques, 2002.

[24] B. GHATTAS : Agrégation d'arbres de classification. *Revue de Statistique Appliquée*, 48(2):85–98, 2000.

[25] Y. GUERMEUR et H. PAUGAM-MOISY : Théorie de l'apprentissage de vapnik et svm, support vector machines. *In* M. SEBBAN et G. VENTURINI, éditeurs : *Apprentissage automatique*, pages 109–138. Hermes, 1999.

[26] T. HASTIE, R. TIBSHIRANI et J FRIEDMAN : *The elements of statistical learning : data mining, inference, and prediction*. Springer, 2001.

[27] T.J. HAYKIN : *Neural network, a comprehensive foundation*. Prentice-Hall, 1994.

[28] J.D. JOBSON : *Applied Multivariate Data Analysis*, volume I : Regression and experimental design. Springer-Verlag, 1991.

[29] G. LUGOSI et N. VAYATIS : On the bayes-risk consistency of boosting methods. *Preprint*, : , 2001.

[30] P. MCCULLAGH et J.A. NELDER : *Generalized Linear Models*. Chapman & Hall, 1983.

[31] J.R. QUINLAN : *C4.5 – Programs for machine learning*. Morgan Kaufmann, 1993.

[32] B.D. RIPLEY : *Pattern recognition and neural networks*. Cambridge University Press, 1996.

[33] G. SAPORTA : *Probabilités, Analyse des Données et Statistique*. Technip, deuxième édition, 2006.

[34] SAS : *SAS/STAT User's Guide*, volume 2. Sas Institute Inc., fourth édition, 1989. version 6.

[35] SAS : *SAS/INSIGHT User's Guide*. Sas Institute Inc., third édition, 1995. version 6.

[36] R. SCHAPIRE : The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

[37] R. SCHAPIRE : The boosting approach to machine learning. an overview. *In MSRI workshop on non linear estimation and classification*, page , 2002.

[38] B SCHÖLKOPF et A SMOLA : *Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

[39] SEM : *SAS/ Enterprise Miner User's Guide*. Sas Institute Inc., 2001. version 8.

[40] M. TENENHAUS : *Statistique : méthodes pour décrire, expliquer et prévoir*. Dunod, 2007.

[41] S. THIRIA, Y. LECHEVALLIER, O. GASCUEL et S. CANU : *Statistique et méthodes neuronales*. Dunod, 1997.

[42] S. TUFFÉRY : *Data Mining et Statistique décisionnelle : l'intelligence des données*. Technip, 2007.

[43] V.N. VAPNIK : *Statistical learning theory*. Wiley Inter science, 1999.

Annexes

Chapitre A

Introduction au bootstrap

1 Introduction

La motivation du *bootstrap*[1] (Efron, 1982 ; Efron et Tibshirani, 1993) est d'approcher par simulation (*Monte Carlo*) la distribution d'un estimateur lorsque l'on ne connaît pas la loi de l'échantillon ou, plus souvent lorsque l'on ne peut pas supposer qu'elle est gaussienne. L'objectif est de remplacer des hypothèss probabilistes pas toujours vérifiées ou même invérifiables par des simulations et donc beaucoup de calcul.

Le principe fondamental de cette technique de rééchantillonnage est de substituer à la distribution de probabilité inconnue F, dont est issu l'échantillon d'apprentissage, la distribution empirique $\widehat{F}$ qui donne un poids $1/n$ à chaque réalisation. Ainsi on obtient un échantillon de taille n dit *échantillon bootstrap* selon la distribution empirique $\widehat{F}$ par n tirages aléatoires avec remise parmi les n observations initiales.

Il est facile de construire un grand nombre d'échantillons bootstrap sur lesquels calculer l'estimateur concerné. La loi simulée de cet estimateur est une approximation asymptotiquement convergente sous des hypothèses raisonnables[2] de la loi de l'estimateur. Cette approximation fournit ainsi des estimations du biais, de la variance, donc d'un risque quadratique, et même des intervalles de confiance de l'estimateur sans hypothèse (normalité) sur la vraie loi.

1.1 Principe du *plug-in*

Soit $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ un échantillon de taille n issue d'une loi inconnue F sur $(\Omega, \mathcal{A})$. On appelle *loi empirique* $\widehat{F}$ la loi discrète des singletons $(x_1, \ldots, x_n)$ affectés des poids $1/n$:

$$\widehat{F} = \sum_{i=1}^{n} \delta_{x_i}.$$

Soit $A \in \mathcal{A}$, $P_F(A)$ est estimée par :

$$\widehat{(P)}_F(A) = P_{\widehat{F}}(A) = \sum_{i=1}^{n} \delta_{x_i}(A) = \frac{1}{n}\text{Card}x_i \in A.$$

De manière plus générale, soit θ un paramètre dont on suppose que c'est une fonction de la loi F. on écrit donc $\theta = t(F)$. Par exemple, $\mu = E(F)$ est un paramètre de F suivant ce modèle. Une *statistique* est une fonction (mesurable) de l'échantillon. Avec le même exemple :

$$\widehat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

et $\overline{x}$ est la statistique qui estime μ. On dit que c'est un estimateur "plug-in" et, plus généralement,

[1] Cette appellation est inspirée du baron de Münchhausen (Rudolph Erich Raspe) qui se sortit de sables mouvants par traction sur ses *tirants de bottes*. En France "bootstrap" est parfois traduit par *à la Cyrano* (acte III, scène 13) en référence à ce héros qui prévoyait d'atteindre la lune en se plaçant sur une plaque de fer et en itérant le jet d'un aimant.

[2] Échantillon indépendant de même loi et estimateur indépendant de l'ordre des observations.

DÉFINITION A.1. — *On appelle estimateur* plug-in *d'un paramètre θ de F, l'estimateur obtenu en remplaçant la loi F par la loi empirique :*

$$\widehat{\theta} = t(\widehat{F}).$$

comme dans le cas de l'estimation de μ : $\widehat{\mu} = E(\widehat{F}) = \overline{x}$.

1.2 Estimation de l'écart-type de la moyenne

Soit X une variable aléatoire réelle de loi F. On pose :

$$\mu_F = E_F(X), \quad \text{et} \quad \sigma_F^2 = \text{Var}_F(X) = E_F[(X - \mu_F)^2];$$

Ce qui s'écrit :

$$X \sim (\mu_F, \sigma_F^2).$$

Soit $(X_1, \ldots, X_n)$ n variables aléatoires i.i.d. suivant aussi la loi F. Posons $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Cette variable aléatoire a pour espérance μ_F et pour variance σ_F^2/n. On dit aussi que la statistique

$$\overline{X} \sim (\mu_F, \sigma_F^2/n).$$

Remarquons qu'en moyennant plusieurs valeurs ou observations, on réduit la variance inhérente à une observation. De plus, sous certaines conditions sur la loi F et comme résultat du théorème de la limite centrale, $\overline{X}$ converge en loi vers la loi normale.

L'estimateur plug-in de σ_F est défini par :

$$\begin{aligned} \widehat{\sigma}^2 &= \widehat{\sigma_F}^2 = \sigma_{\widehat{F}}^2 = \text{Var}_{\widehat{F}}(X) \\ &= E_{\widehat{F}}[(X - E_{\widehat{F}}(X))^2] = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2. \end{aligned}$$

L'estimateur plug-in de σ_F est (légèrement) différent de celui du maximum de vraisemblance. L'estimateur plug-in est en général biaisé mais il a l'avantage d'être simple et de pouvoir s'appliquer à tout paramètre θ même lorsque l'on ne peut pas calculer la vraisemblance du modèle.

2 Estimation bootstrap d'un écart-type

Soit $\widehat{\theta} = s(x)$ un estimateur quelconque (M.V. ou autre) de θ pour un échantillon x donné. On cherche à apprécier la précision de $\widehat{\theta}$ et donc à estimer son écart-type.

2.1 Échantillon bootstrap

Avec les mêmes notation, $\widehat{F}$ est la distribution empirique d'un échantillon $\boldsymbol{x} = \{x_1, \ldots, x_n\}$.

DÉFINITION A.2. — *On appelle* échantillon bootstrap *de $\boldsymbol{x}$ un échantillon de taille n noté*

$$\boldsymbol{x}^* = \{x_1^*, \ldots, x_n^*\}$$

suivant la loi $\widehat{F}$; $\boldsymbol{x}^*$ est un ré-échantillon de $\boldsymbol{x}$ avec remise.

2.2 Estimation d'un écart-type

DÉFINITION A.3. — *On appelle* estimation bootstrap *de l'écart-type $\widehat{\sigma_F}(\widehat{\theta})$ de $\widehat{\theta}$, son estimation plug-in : $\sigma_{\widehat{F}}(\widehat{\theta})$.*

Mais, à part dans le cas très élémentaire où, comme dans l'exemple ci-dessus, θ est une moyenne, il n'y a pas de formule explicite de cet estimateur. Une approximation de l'estimateur bootstrap (ou plug-in) de l'écart-type de $\widehat{\theta}$ est obtenue par une simulation (Monte-Carlo) décrite dans l'algorithme ci-dessous.

Pour un paramètre θ et un échantillon $\boldsymbol{x}$ donnés, on note $\widehat{\theta} = s(\boldsymbol{x})$ l'estimation obtenue sur cet échantillon. Une *réplication bootstrap* de $\widehat{\theta}$ est donnée par : $\widehat{\theta}^* = s(\boldsymbol{x}^*)$.

$\widehat{\sigma}_B$ est l'approximation bootstrap de l'estimation plug-in recherchée de l'écart-type de $\widehat{\theta}$.

Algorithm 11 Estimation bootstrap de l'écart-type

Soit $\boldsymbol{x}$ un échantillon et θ un paramètre.
Pour $b = 1$ à B **Faire**
Sélectionner 1 échantillon bootstrap $\boldsymbol{x}^{*b} = \{x_1^{*b}, \ldots, x_n^{*b}\}$. par tirage avec remise dans x.
Estimer sur cet échantillon : $\widehat{\theta}^*(b) = s(\boldsymbol{x}^{*b})$.
Fin Pour
Calculer l'écart-type de l'échantillon ainsi construit :

$$\widehat{\sigma}_B^2 = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\theta}^*(b) - \widehat{\theta}^*(.))^2$$

$$\text{avec} \quad \widehat{\theta}^*(.) = \frac{1}{B}\sum_{b=1}^{B}(\widehat{\theta}^*(b).$$

2.3 Estimation du biais

Avec les mêmes notations :

$$\theta = t(F) \quad \text{et} \quad \widehat{\theta} = s(\boldsymbol{x}),$$

le biais d'un estimateur s'exprime comme

$$\mathcal{B}_F(\widehat{\theta}) = E_F[s(\boldsymbol{x})] - t(F).$$

Un estimateur est sans biais si $E[\widehat{\theta}] = \theta$. Le biais est aussi une mesure de la précision d'un estimateur et on a vu que, généralement, les estimateurs plug-in étaient biaisés.

DÉFINITION A.4. — *On appelle estimateur bootstrap du biais, l'estimateur plug-in :*

$$\widehat{\mathcal{B}_F}(\widehat{\theta}) = \mathcal{B}_{\widehat{F}}(\widehat{\theta}) = E_{\widehat{F}}[s(\boldsymbol{x}^*)] - t(\widehat{F}).$$

Comme pour l'écart-type, il n'existe généralement pas d'expression analytique et il faut avoir recours à une approximation par simulation.

Algorithm 12 Estimation bootstrap du biais

Soit $\boldsymbol{x}$ un échantillon et θ un paramètre.
Pour $b = 1$ à B **Faire**
Sélectionner 1 échantillon bootstrap $\boldsymbol{x}^{*b} = \{x_1^{*b}, \ldots, x_n^{*b}\}$. par tirage avec remise dans $\boldsymbol{x}$.
Estimer sur cet échantillon la réplication bootstrap de $\widehat{\theta}$: $\widehat{\theta}^*(b) = s(\boldsymbol{x}^{*b})$.
Fin Pour
Approcher $E_{\widehat{F}}[s(\boldsymbol{x}^*)]$ par $\widehat{\theta}^*(.) = \frac{1}{B}\sum_{b=1}^{B}(\widehat{\theta}^*(b)$
L'approximation bootstrap du biais est : $\widehat{\mathcal{B}_B}(\widehat{\theta}) = \widehat{\theta}^*(.) - \widehat{\theta}$.

3 Compléments

En résumé, on peut dire que le bootstrap repose sur une hypothèse très élémentaire : $\widehat{\theta}^*$ se comporte par rapport à $\widehat{\theta}$ comme $\widehat{\theta}$ par rapport à θ. La connaissance de $\widehat{\theta}^*$ (distribution, variance, biais...) renseigne alors sur celle de $\widehat{\theta}$.

Beaucoup d'autres compléments sont à rechercher dans la littérature et en particulier dans Efron et Tibshirani (1993). Il est ainsi possible de définir des intervalles de confiance bootstrap en considérant la distribution et les quantiles de $\widehat{\theta}^*$ ou même encore des tests à partir des versions bootstrap de leur statistique.

Le bootstrap rapidement décrit ici est dit "non-paramétrique" car la loi empirique $\widehat{F}$ est une estimation non-paramétrique de F. Dans le cas où F serait connue à un paramètre près, il existe également une version dite *paramétrique* du bootstrap.

Pour des estimateurs plus compliqués (fonctionnels) comme dans le cas de la régression non-paramétrique par noyau ou spline, il est facile de construire graphiquement une enveloppe bootstrap de l'estimateur à partir de réplications de l'échantillon. Celle-ci fournit généralement une bonne appréciation de la qualité de l'estimateur obtenu. Attention, dans le cas de la régression il est en principe plus justifié de répliquer le tirage sur les *résidus* plutôt que sur les observations. Ce sont les résidus qui sont en effet supposés i.i.d. et qui vérifient donc les hypothèses nécessaires mais cette approche devient très sensible à l'hypothèse sur la validité du modèle. Il est finalement d'usage de considérer un échantillon bootstrap issu des données initiales (Efron et Tibshirani) :

$$\mathbf{z}^{*b} = \{(\mathbf{x}_1^{*b}, y_1^{*b}), \ldots, (\mathbf{x}_n^{*b}, y_n^{*b})\};$$

c'est ce qui a été choisi dans ce document.

Enfin, l'estimation bootstrap est justifiée par des propriétés asymptotiques (convergence en loi) lorsque le nombre de réplications (B) croit conjointement avec la taille de l'échantillon (n).

Table des matières