REVUE DE STATISTIQUE APPLIQUÉE

M. BARDOS
W. H. ZHU

Comparaison de l'analyse discriminante linéaire et des réseaux de neurones. Application à la détection de défaillance d'entreprises

Revue de statistique appliquée, tome 45, n° 4 (1997), p. 65-92
<http://www.numdam.org/item?id=RSA_1997__45_4_65_0>

Rev. Statistique Appliquée, 1997, **XLV** (4), 65-92

COMPARAISON DE L'ANALYSE DISCRIMINANTE LINÉAIRE ET DES RÉSEAUX DE NEURONES

Application à la détection de défaillance d'entreprises

M. Bardos (1) , W. H. Zhu (2)

(1) Banque de France, Observatoire des entreprises, 44 1356,
31, rue Croix des Petits Champs, 75001 Paris
frbdfdpq(a)ibmmail.com
(2) CISIA, 1 avenue Herbillon 94160 Saint Mandé,
CEREMADE, Université Paris IX Dauphine,
Place du Maréchal de Lattre De Tassigny, 75775 Paris Cedex 16
zhu(a)ceremade.dauphine.fr

RÉSUMÉ

Afin de construire un outil de détection précoce des défaillances d'entreprises, une analyse discriminante linéaire de Fisher, une régression logistique, et un réseau de neurones multicouche sont appliqués aux mêmes données individuelles économiques et financières. Les techniques et les résultats sont comparés*.

Mots-clés : *Analyse discriminante de Fisher, régression logistique, réseau de neurones multicouche, classification supervisée, sélection de variables, apprentissage, généralisation, validation, comparaison de performances, pourcentages de bons classements, probabilité a posteriori.*

ABSTRACT

In order to design a tool for the early detection of business failures, a Fisher linear discriminant analysis, a logistical regression and a multilayer neural network are applied to the same economic and financial data set. The techniques and results are compared*.

Keywords : *Fisher discriminant analysis, logistic regression, multilayer neural network, supervised classification, variables selection, learning, generalization, validation, performance comparison, correct allocation rate, probability a posteriori.*

* Les auteurs adressent leurs chaleureux remerciements à Patrick Gallinari pour son aide dans l'application du logiciel de LAFORIA sur nos données afin de mieux cerner la nature des difficultés que nous avons rencontrées dans l'amélioration des résultats sur les réseaux de neurones, et à Ludovic Lebart pour ses commentaires constructifs.
La présente rédaction n'engage que les auteurs.

Le but de cet article est de comparer plusieurs techniques d'analyse discriminante en les appliquant à un même ensemble de données. Le problème traité est la détection précoce des défaillances d'entreprises à partir de ratios économiques et financiers calculés sur données individuelles d'entreprises. Les techniques comparées sont :

– **l'analyse discriminante linéaire** largement utilisée en raison de sa robustesse [2] pour construire des fonctions scores en usage dans les banques, en vue d'accord de crédit ou de suivi de contentieux. On applique ici la méthode de Fisher et la régression logistique sur des variables quantitatives non codées ce qui conduit à une forme linéaire;

– **les réseaux de neurones**, dont l'une des applications est l'analyse discriminante, encore appelée dans ce contexte classification supervisée. Le modèle utilisé ici est le perceptron à une couche cachée.

On décrira d'abord la sélection initiale des variables (§ 1), puis on présentera les techniques et les résultats de l'analyse discriminante linéaire (§ 2) et des réseaux de neurones (§ 3), enfin on comparera les résultats (§ 4).

1. Les données

1.1. Les échantillons

1.1.1. Les entreprises

L'échantillon de base est constitué de PME de l'industrie, observées en 1990 et réparties en 2 groupes que l'on cherche à discriminer : **les entreprises défaillantes**, groupe noté D, **et les entreprises non défaillantes**, groupe noté N.

Les entreprises défaillantes ont déposé leur bilan en 1991 ou 1992 ou 1993 (c'est-à-dire 1, 2 ou 3 ans après la date d'observation), elles sont au nombre de 809.

Les entreprises non défaillantes sont des firmes qui n'ont aucun dépôt de bilan jusqu'à nouvel ordre. Les firmes non défaillantes étant beaucoup plus nombreuses que les firmes défaillantes, on a procédé à un tirage aléatoire de façon à constituer pour elles un ensemble représentatif et de taille comparable au groupe des firmes défaillantes. Les tailles respectives des 2 groupes dans l'échantillon de base, n_D et n_N, influencent les résultats de la régression logistique et des réseaux de neurones. On explicitera cette question au cours de la présentation et on utilisera deux échantillons différents de firmes non défaillantes, l'un comportant 1381 entreprises et l'autre 775 entreprises.

1.1.2. La validation

Quand on pratique la **validation croisée** on répartit l'échantillon de base (de taille n) en échantillon d'apprentissage et échantillon test, autant de fois qu'il est nécessaire. Ainsi, l'échantillon test peut être constitué d'un dixième de l'échantillon par tirage au sort; on réitère ce partage 10 fois et on calcule la moyenne et l'écart-type des pourcentages de bons classements sur les 10 échantillons tests ainsi construits;

ou bien on utilise la méthode «leaving one out» (celle-ci consiste à exclure une observation, construire un score sur les $n-1$ autres observations, puis tester ce score sur l'individu exclu; on réitère n fois, on obtient ainsi le taux de bons classements «leaving one out», taux calculé sur les n observations testées).

Pour les autres exercices comptables, de 1986 à 1992, on dispose d'échantillons indépendants de constitution analogue à l'échantillon de base. Ils sont utilisés comme **échantillons tests** pour la validation, et rassemblent 11 470 entreprises en tout dont 3 287 défaillantes et 8 183 non défaillantes.

1.2. Les variables

La détection précoce des défaillances d'entreprise repose ici sur des informations comptables qui ont le mérite d'être disponibles pour un très grand nombre d'entreprises. Grâce à elles, on construit des ratios économiques et financiers reposant, pour la plus grande part, sur la méthodologie d'analyse de la Centrale de bilans de la Banque de France, les autres concepts relevant de l'expérience acquise par les experts sur les études de cas.

Le large choix initial de 50 ratios intègre les principales préoccupations ayant trait à l'analyse des bilans du point de vue de la bonne santé de l'entreprise. Les thèmes sont la rentabilité, la structure productive, l'endettement financier, la structure du bilan, le crédit interentreprises, les dettes diverses, la solvabilité, la croissance d'activité, la croissance du financement.

1.3. La sélection préalable des variables

L'analyse discriminante multicritère nécessite la sélection préalable des variables sensibles au risque de défaillance, d'abord une à une, puis conjointement.

1.3.1 La sélection univariée

La sélection des ratios un à un repose sur la comparaison des distributions sur chacune des deux populations. En effet les ratios suivent rarement des lois connues, *a fortiori* des lois normales. La comparaison des moyennes par catégorie est donc insuffisante. Par ailleurs, pour certains ratios, l'approche de la défaillance n'affectera sérieusement qu'une partie des entreprises. L'étude du pouvoir discriminant des ratios commence donc par l'examen des histogrammes des fonctions de répartitions empiriques et l'examen des quantiles par des tests non paramétriques. On décèle ainsi les variables discriminantes et, parmi elles, celles qui permettent une discrimination linéaire ou à l'inverse non linéaire. Le rôle de l'échéance de la défaillance sur le pouvoir discriminant des ratios est également examiné. Connaissant mieux le type de discrimination permise par chaque ratio, on procède au test de Kolmogorov pour mesurer le pouvoir discriminant. On détermine ainsi une liste de ratios «clignotants» du risque.

1.3.2. La sélection conjointe en vue d'une analyse multicritère

La sélection conjointe dépend de la méthode d'analyse discriminante utilisée. Le choix d'une méthode linéaire interdira les ratios pour lesquels la discrimination est non linéaire. À partir d'un ensemble de variables discriminantes linéaires, on choisira un jeu de ratios non (ou peu) corrélés entre eux et qui représentent les différents critères d'appréciation de la bonne santé d'une entreprise. Plusieurs jeux alternatifs de variables pourront ainsi être définis. Pour chaque jeu possible, on pratique une sélection conjointe des meilleurs ratios en s'aidant de procédures pas à pas pour différents critères. On en citera trois : le λ de Wilks (§ 1.3.2.1), l'algorithme de Furnival et Wilson (§ 1.3.2.2), la statistique Wald pour les coefficients de la régression logistique (§ 1.3.2.3).

Dans le § 1.3.3. on énonce les principes qui permettent de sélectionner les ratios pour les méthodes non linéaires.

1.3.2.1. Le λ de Wilks

Sous l'hypothèse de multinormalité des variables dans chaque groupe et d'égalité des matrices de variance-covariance, on utilise une procédure de sélection pas à pas des variables fondée sur le λ de Wilks[1] .

La sélection commence sans variable dans le modèle. Au premier pas on choisit la variable qui a le plus grand pouvoir discriminant. Ensuite à chaque pas, le modèle est examiné. Si la variable du modèle qui contribue le moins à son pouvoir discriminant, mesuré par le λ de Wilks, tombe en dessous du seuil de signification préalablement choisi, alors la variable est enlevée. Par ailleurs, la variable, non encore dans le modèle, qui contribue le plus au pouvoir discriminant est entrée. Quand toutes les variables du modèle satisfont le critère et celles en dehors n'y satisfont pas, la sélection s'arrête.

Comme cette procédure rentre une variable à la fois à chaque pas, la sélection ne tient pas compte des interrelations avec les variables qui n'ont pas encore été sélectionnées si bien que des variables importantes peuvent être «oubliées» par la sélection. Le critère suivant tend à surmonter cet inconvénient.

1.3.2.2. L'algorithme de Furnival et Wilson

La discrimination linéaire à réaliser étant entre deux groupes, on peut sélectionner les meilleurs sous-ensembles de variables en réalisant la sélection des meilleurs ajustements par une régression linéaire multiple suivant trois critères possibles : R^2 maximum, R^2 ajusté maximum, $C(P)$ de Mallows minimum.

Parmi J variables, on peut choisir $2^J - 1$ sous-ensembles non vides. Pour chaque taille p de sous-ensemble de variables, le programme édite les meilleurs choix de sous-ensemble pour les trois critères ci-dessus. Appliqué tel que, ce programme serait très

[1] Le λ de Wilks est le déterminant de $T^{-1}W$, avec T matrice de variance-covariance totale et W matrice de variance-covariance intraclasse. Il suit une loi de Wilks de paramètre $(p, N-2, 1)$. Minimiser le λ de Wilks revient à maximiser la trace de $T^{-1}B$, ce qui, dans le cas de plus de 2 groupes, constitue une généralisation de la maximisation du D^2 de Mahalanobis (cf. Romeder [22]).

coûteux en temps calcul. On limite considérablement le nombre d'opérations en utilisant l'algorithme «leaps and bounds» de Furnival et Wilson (*cf.* Furnival, Wilson [11] et Richardot [21]).

1.3.2.3. La statistique de Wald pour les coefficients de la régression logistique

On dispose d'une mesure de la significativité des coefficients de la régression logistique : la statistique de Wald grâce à laquelle la probabilité que le paramètre estimé soit nul est calculée. On utilise ce critère dans la sélection pas à pas des variables explicatives.

Sous l'hypothèse que les résidus suivent une loi logistique, le choix des ratios peut s'effectuer de deux façons : la sélection ascendante choisit progressivement le ratio pour lequel la statistique de Wald est la plus élevée, la sélection descendante considère initialement tous les ratios et retire progressivement celui pour lequel la statistique de Wald est la plus faible.

La sélection pas à pas combine ces deux démarches. Il peut arriver qu'un ratio déjà choisi soit éliminé ultérieurement parce que sa statistique de Wald devient trop faible. La sélection s'arrête quand toutes les statistiques de Wald des ratios restants sont trop faibles.

1.3.3. Conclusion sur la sélection des variables

La sélection des variables est rendue délicate par le fait que les données ne vérifient jamais parfaitement les hypothèses des modèles décrits précédemment. On est donc obligé de se livrer à un certain tâtonnement, en comparant les résultats des analyses discriminantes pratiquées sur les diverses listes de variables sélectionnées.

Pour les méthodes non linéaires, on ajoutera aux ratios précédents les ratios au pouvoir discriminant fort mais non linéaire dont les fluctuations statistiques intertemporelles restent modérées et on attend qu'ils concourent à améliorer les taux de succès obtenus par les méthodes linéaires. C'est ainsi que pour les réseaux de neurones on part initialement d'un ensemble de ratios plus vaste que celui utilisé en analyse discriminante linéaire (*cf.* § 3). La question de la sélection des ratios les plus efficaces dans la discrimination sera alors traitée par des méthodes liées à la technique des réseaux de neurones.

2. L'analyse discriminante linéaire de Fisher et la régression logistique

2.1. L'analyse linéaire discriminante de Fisher

Dans le cas de deux groupes N et D, après avoir sélectionné les k ratios les plus discriminants pris conjointement, l'analyse discriminante linéaire de Fisher peut être présentée selon deux règles de décision.

2.1.1. Règle de décision fondée sur un critère métrique

Si l'entreprise est décrite par le vecteur a des k ratios $(a_1, a_2, \ldots, a_k)$ et si μ^N et μ^D sont les points moyens de chacun des 2 groupes, on affecte l'entreprise au groupe dont le point moyen est le plus proche. Par exemple on affecte l'entreprise «a» au groupe N si et seulement si $d(a, \mu^N) \leqslant d(a, \mu^D)$, où $d(a, \mu^N)$, (respectivement $d(a, \mu^D)$), désigne la distance de «a» à μ^N (respectivement μ^D).

Ceci se traduit par l'inégalité : $f(a) = (\mu^N - \mu^D)' T^{-1} \left(a - \frac{\mu^N + \mu^D}{2} \right) \geqslant 0$

T est la matrice de variance covariance totale; f est la fonction score; $f(a) = 0$ est l'équation de l'hyperplan qui sépare les groupes au mieux pour le critère de la distance; $\alpha = (\mu^N - \mu^D)' T^{-1}$ est le vecteur des k coefficients de la fonction, il ne dépend que des moyennes et de la matrice de variance-covariance[2], donc pas de la taille des groupes dans l'échantillon de base. Plus $f(a)$ est négative, plus la situation de l'entreprise est risquée.

La fonction de Fisher fournit une aide à l'interprétation de chaque cas. En effet, si on appelle point pivot $p = \frac{\mu^D + \mu^N}{2}$, avec $p = (p_1, p_2 \ldots p_k)$, on peut décomposer $f(a) = \alpha(a - p)$ sur les axes de coordonnées. Le score de l'entreprise «a» peut s'écrire :

$$f(a) = \alpha_1(a_1 - p_1) + \ldots + \alpha_i(a_i - p_i) + \ldots + \alpha_k(a_k - p_k)$$

$\alpha_i(a_i - p_i)$ est la contribution du $i^{\text{ème}}$ ratio au score $f(a)$.

Cette décomposition présente l'avantage d'indiquer quels sont les ratios les plus influents sur le score et constitue une aide précieuse à l'interprétation : les ratios favorables ont une contribution positive, au contraire, les ratios défavorables ont une contribution négative.

2.1.2. Règle de décision de Bayes de risque minimum

Dans le cas où les lois sont multinormales, homoscédastiques sur les 2 groupes, si on connaît la probabilité *a priori* de défaillance π_D, donc aussi $\pi_N = 1 - \pi_D$, et les coûts de mauvais classement $C_{D/N}$ et $C_{N/D}$[3], on affecte «a» de façon que le risque moyen soit minimum. La règle de décision devient donc :

$$a \in N \Longleftrightarrow f(a) \geqslant \ln \frac{\pi_D C_{N/D}}{\pi_N C_{D/N}}$$

Par rapport à la règle fondée sur la distance, cela revient à remplacer le seuil de décision 0 par le deuxième membre de l'inégalité.

---

[2] Le choix de la métrique W^{-1}, fondé sur la matrice de variance-covariance intraclasse, au lieu de T^{-1}, conduit au même vecteur propre donc aux mêmes affectations.

[3] On note $C_{N/D}$ le coût de classer non défaillante (N) une firme qui en réalité est défaillante (D), et $C_{D/N}$ le coût de classer défaillante (D) une entreprise qui ne le sera pas (N).

2.2. *La régression logistique*

La régression logistique est une méthode économétrique dans laquelle la variable endogène Y correspond au codage des entreprises : 0 si la firme est défaillante, 1 sinon; X est la matrice des variables exogènes (*cf.* Gourieroux [15]).

Pour l'entreprise i, on suppose que :

$$Y_i = \begin{cases} 0 & \text{si } \beta + \alpha X_i + u_i \leqslant 0 \quad \text{ou encore } u_i \leqslant -\beta - \alpha X_i \\ 1 & \text{si } \beta + \alpha X_i + u_i > 0 \quad \text{ou encore } u_i > -\beta - \alpha X_i \end{cases}$$

où β est une constante et α est le vecteur ligne des coefficients d'une combinaison linéaire à estimer.

Les u_i sont les perturbations supposées indépendantes, de moyenne nulle, de variance 1 ; elles sont supposées suivre une loi logistique de fonction de répartition F :

$$F(x) = \frac{1}{1 + e^{-x}}$$

Les paramètres α et β sont estimés par la méthode du maximum de vraisemblance.

La vraisemblance s'écrit :

$\prod_{i=1}^{n} p_i^{Y_i}(1 - p_i)^{1-Y_i}$ avec $n = n_D + n_N$, taille de l'échantillon,

et où

$$p_i = P(Y_i = 1/X_i) = P(u_i > -\beta - \alpha X_i) = 1 - F(-\beta - \alpha X_i) = \frac{1}{1 + e^{-\beta - \alpha X_i}}$$

$$1 - p_i = P(Y_i = 0/X_i) = P(u_i \leqslant -\beta - \alpha X_i) = \frac{1}{1 + e^{\beta + \alpha X_i}}$$

On a donc : $\text{logit } p_i = \ln \frac{p_i}{1 - p_i} = \beta + \alpha X_i$.

Contrairement à la méthode précédente, ici les tailles n_D et n_N des groupes D et N dans l'échantillon de base influencent la vraisemblance, donc l'estimation de la fonction discriminante et par conséquent de la probabilité *a posteriori* p_i (*cf.* § 2.3.2.).

2.3. *Les résultats des analyses linéaires discriminantes*

On compare les résultats des trois meilleures fonctions construites sur l'échantillon de l'exercice comptable 1990, composé de 809 défaillantes et 1381 non défaillantes, par les techniques décrites précédemment : L90 régression logistique, D90 fonction linéaire de Fisher (sélection des ratios par le λ de Wilks), S4 fonction linéaire de Fisher (sélection des ratios par l'algorithme de Furnival et Wilson).

2.3.1. Les pourcentages de bons classements

Les fonctions D90 ou S4 ont des résultats, généralement supérieurs à 70 % sauf en 1990 et 1986, et même souvent supérieurs à 73 % sur toute la période, assurant ainsi la stabilité des performances dans le temps.

La fonction L90 donne des pourcentages très déséquilibrés selon les groupes. Ceci tient au fait que la régression logistique est sensible à la taille des échantillons, ici $n_N > n_D$, et c'est le groupe des non défaillantes qui est le mieux classé sur l'échantillon de base comme sur les échantillons tests. Comme on le verra ci-dessous, un changement de seuil rétablit l'équilibre.

Pourcentages de bons classements pour le seuil 0

	Échantillon													
	1992		1991		1990		1989		1988		1987		1986	
nb entrep	318	1565	635	1523	809	1381	542	1173	388	979	306	845	289	717
Fonction	D	N	D	N	D	N	D	N	D	N	D	N	D	N
L 90	77,0	84,2	60,9	84,3	**52,6**	**85,5**	61,7	83,8	65,0	84,4	69,1	83,1	70,3	79,8
D 90	88,0	74,2	77,2	72,4	**70,1**	**74,0**	75,9	71,2	82,0	70,4	80,1	69,9	84,5	67,8
S 4	87,7	75,3	77,3	74,8	**68,3**	**75,3**	74,4	73,2	81,2	74,0	77,9	71,2	82,1	69,2

D : défaillantes; N : non défaillantes;

En gras : échantillon de base : résultat de la validation croisée par la méthode «leaving one out».

2.3.2. Les probabilités a posteriori *de défaillance*

2.3.2.1. Deux calculs possibles

La probabilité *a posteriori* de défaillance selon la valeur du score est un renseignement indispensable pour le décideur qui, ainsi, peut connaître la qualité de la prévision de faillite tout le long de l'échelle des scores.

Le calcul des probabilités *a posteriori* (c'est-à-dire connaissant la valeur du score) repose sur la formule de Bayes :

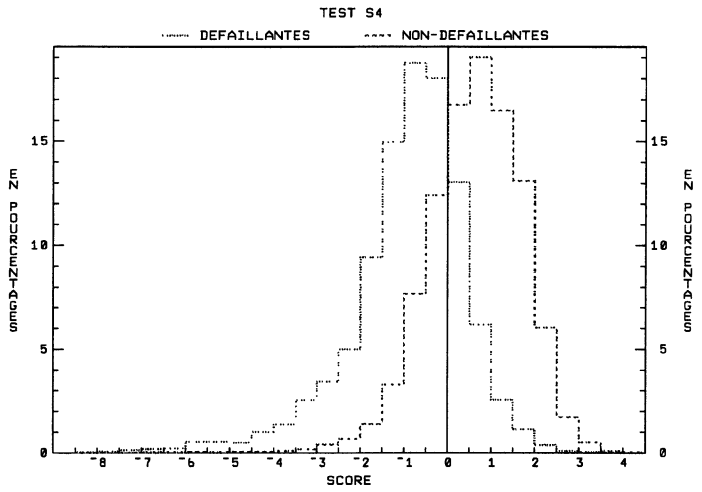
$$P[e \in D/s(e) \in r] = \frac{P[s(e) \in r/e \in D]P(e \in D)}{P[s(e) \in r]} = \frac{p_D \pi_D}{p_D \pi_D + p_N(1-\pi_D)}$$

où $s(e)$ est le score de l'entreprise e, et la région r peut être un intervalle $[a, b]$, et

$p_D = p_D(r) = P\big(s(e) \in r/e \in D\big)$ est la probabilité de r conditionnellement à l'appartenance à D.

$p_N = p_N(r) = P\big(s(e) \in r/e \in N\big)$ est la probabilité de r conditionnellement à l'appartenance à N.

L'apport de l'analyse discriminante est de fournir ces probabilités conditionnelles. Elles sont représentées sur le graphique ci-dessous sous forme d'histogrammes, réalisés avec l'ensemble de toutes les firmes disponibles (c'est-à-dire 11 470 entreprises dont 3 287 défaillantes et 8183 non défaillantes), pour la fonction S4.



On veut déterminer les régions de décision intéressantes, c'est-à-dire celles ayant une probabilité *a posteriori* soit très forte, soit très faible, tout en recueillant beaucoup d'entreprises. Cette probabilité *a posteriori* doit être examinée conjointement avec $P[s(e) \in r]$, probabilité pour une entreprise quelconque d'avoir son score dans la région r.

Pour la régression logistique* on peut utiliser

– soit la formule de Bayes s'exerçant sur les distributions des probabilités conditionnelles empiriques de la fonction L90,

– soit la modélisation de la probabilité *a posteriori* de défaillance pour l'entreprise i caractérisée par le vecteur X_i :

$$P[Y_i = 0/X_i] = 1 - p_i = \frac{1}{1 + e^{L90(X_i)}} = \frac{1}{1 + e^{\beta + \alpha X_i}}$$

2.3.2.2. Les résultats

Les tableaux ci-joints décrivent, par intervalle de score, la probabilité *a posteriori* de défaillance (première ligne du tableau), et la probabilité d'avoir un

---

* Pour l'analyse discriminante linéaire de Fisher, sous l'hypothèse de multinormalité et homoscédasticité, on peut également estimer la probabilité *a posteriori* par la formule : $P(D/X) = \frac{1}{1 + e^{\alpha X + \beta}}$, où $\alpha X + \beta$ est estimé comme indiqué au § 2.1.2. dans le cas où les coûts d'erreur de classements sont égaux.
Mais c'est surtout pour la régression logistique que la modélisation est couramment utilisée. C'est pourquoi dans les résultats on détaille la présentation de la régression logistique.

score dans l'intervalle (deuxième ligne du tableau). Dans les trois premiers tableaux on utilise, pour chacune des trois fonctions, la formule de Bayes sur les distributions empiriques construites sur l'ensemble des échantillons, tandis que dans le quatrième tableau on utilise la modélisation logistique en calculant la probabilité *a posteriori* moyenne sur chaque intervalle.

Au § 2.3.2.3. on réexamine cette modélisation en tenant compte du schéma d'échantillonnage.

Les tableaux 1 à 3 présentent des résultats similaires. Par contre, on constate un très grand décalage entre les tableaux 3 et 4 pour la probabilité *a posteriori*. Ceci peut provenir pour une part du fait que la fonction $L90$ est estimée sur l'échantillon de base, tandis que les probabilités conditionnelles utilisées dans le tableau 3 sont estimées sur l'ensemble des entreprises disponibles. Toutefois la stabilité des bons classements au cours de la période montre que l'influence de ce phénomène est sans doute très faible. Par contre, on va voir ci-dessous que l'influence du schéma d'échantillonnage est déterminante sur la régression logistique.

Probabilités a posteriori *et proportions d'entreprises évaluées en pourcentages sur l'ensemble des échantillons constitué de 11 470 entreprises*

Probabilités a priori[4] *:* $\pi_D = 0{,}105$ *et* $\pi_N = 0{,}895$
en utilisant les distributions conditionnelles et le théorème de Bayes

TABLEAU 1 : S4

Intervalle de score	− 3	−2,5	−2	−1,5	−1	−0,5	0	0,5	1	1,5
Probabilité de défaillance	65,1	48,3	44,5	34,8	22,3	14,5	8,4	3,7	1,8	1	0,7
Probabilité d'appartenance	1,8	1,1	2,2	4,5	8,8	13	16,4	17,7	15	11,9	7,6

TABLEAU 2 : D 90

Intervalle de score	− 3	−2,5	−2	−1,5	−1	−0,5	0	0,5	1	1,5
Probabilité de défaillance	67	56,6	43,7	35,6	22,8	13,7	7,6	3,4	1,6	0,9	0,6
Probabilité d'appartenance	1,5	0,9	2,1	4,7	9,6	14,6	18,2	17,2	14	9,4	7,8

[4] La probabilité *a priori* de défaillance pour les firmes de l'industrie est estimée par le taux de défaillance dans le secteur industrie fourni par l'INSEE multiplié par 3 car ici nous faisons de la prévision à 3 ans, l'observation des firmes défaillantes étant faite sur les trois années qui précèdent la défaillance.

TABLEAU 3 : L 90

Intervalle de score	− 2,5	−2	−1,5	−1	−0,5	−0	0,5	1	1,5	2	2,5
Probabilité de défaillance	67,8	57,4	44,2	30,7	20,5	13,4	7,1	3,4	1,7	1,2	0,8	0,6
Probabilité d'appartenance	1,6	0,6	1,3	2,4	5,4	9,8	13,5	16,6	15,1	12,6	8,9	12,2

TABLEAU 4 : L 90

En utilisant la modélisation logistique pour estimer la probabilité a posteriori *:* $P[Y_i = 0/X_i] = 1 - p_i = \frac{1}{1 + e^{\beta + aX_i}}$, *et sans tenir compte du schéma d'échantillonnage.*

Intervalle de score	− 2,5	−2	−1,5	−1	−0,5	−0	0,5	1	1,5	2	2,5
Probabilité moyenne de défaillance	96,8	90,5	85,2	77,7	67,9	56,2	43,8	32,1	22,3	14,8	9,5	4,7
Probabilité d'appartenance	1,6	0,6	1,3	2,4	5,4	9,8	13,5	16,6	15,1	12,6	8,9	12,2

2.3.2.3. Les résultats de la régression logistique en tenant compte du schéma d'échantillonnage.

On utilise ici un échantillonnage rétrospectif, c'est-à-dire qu'on dispose d'un ensemble de cas ($Y = 0$), les firmes défaillantes, que l'on compare au groupe témoin ($Y = 1$), les firmes non défaillantes. La vraisemblance est alors influencée par le taux de sondage (*cf.* Celeux, Nakache [8], Venditti [26]).

Soit T la variable indicatrice d'appartenance à l'échantillon. La probabilité *a posteriori* s'écrit alors :

$$P(Y = 0/X = x,\ T = 1) = \frac{P(Y = 0/X = x)P(T = 1/X = x,\ Y = 0)}{P(Y = 0/X = x)P(T = 1/X = x, Y = 0) + P(Y = 1/X = x)P(T = 1/X = x, Y = 1)}$$

Comme le sondage est indépendant de X, on a :

$$P(Y = 0/X = x,\ T = 1) = \frac{P(Y = 0/X = x)P(T = 1/Y = 0)}{P(Y = 0/X = x)P(T = 1/Y = 0) + P(Y = 1/X = x)P(T = 1/Y = 1)}$$

$P(T = 1/Y = 0) = \gamma_0 =$ taux de sondage dans le groupe D

$P(T = 1/Y = 1) = \gamma_1 =$ taux de sondage dans le groupe N

On note $\Pi(x) = P(Y = 1/X = x)$ la probabilité *a posteriori* d'être non défaillante pour une firme caractérisée[5] par x. On se gardera de confondre cette probabilité *a posteriori* avec les probabilités *a priori* π_N et π_D.

$$P(Y = 1/X = x, T = 1) = \frac{\gamma_1 \Pi(x)}{\gamma_0[1 - \Pi(x)] + \gamma_1 \Pi(x)} = \Pi^*(x)$$

En utilisant l'hypothèse qui fonde la régression logistique, à savoir la linéarité du logit : $\text{logit } \Pi(x) = \beta + \alpha X$

$$\text{logit } \Pi^*(x) = \ln \frac{\Pi^*(x)}{1 - \Pi^*(x)} = \ln \frac{\gamma_1 \Pi(x)}{\gamma_0[1 - \Pi(x)]} = \text{logit } \Pi(x) + \ln \frac{\gamma_1}{\gamma_0}$$

$$\text{logit } \Pi^*(x) = \beta + \ln \frac{\gamma_1}{\gamma_0} + \alpha X$$

$$\gamma_1 = P(T = 1/Y = 1) = \frac{P(T = 1 \text{ et } Y = 1)}{P(Y = 1)}$$
$$= \frac{P(Y = 1/T = 1)P(T = 1)}{P(Y = 1)} = \frac{n_N P(T = 1)}{n\pi_N}$$

de même $\gamma_0 = \dfrac{n_D P(T = 1)}{n\pi_D}$, d'où

$$\frac{\gamma_1}{\gamma_0} = \frac{n_N}{\pi_N} \times \frac{\pi_D}{n_D} = \frac{n_N}{n_D} \frac{\pi_D}{\pi_N}$$

$$\text{logit } \Pi^*(x) = \text{logit } \Pi(x) + \ln \frac{n_N}{n_D} + \ln \frac{\pi_D}{\pi_n}$$
$$= \alpha X + \beta + \ln \frac{n_N}{n_D} + \ln \frac{\pi_D}{\pi_N} = \alpha X + \beta'$$

avec :

$$\beta' = \beta + \ln \frac{n_N}{n_D} + \ln \frac{\pi_D}{\pi_N} = \beta + \ln \frac{1381}{809} + \ln \frac{0,105}{0,895} = \beta - 1,6081$$

[5] Cette notation simplifie l'écriture de l'introduction du § 2.2. où on faisait intervenir l'indice i de l'entreprise. On pourrait écrire :

$$p_i = \Pi(x_i) = P(Y_i = 1/X = x_i)$$

TABLEAU 5 : L' 90

En utilisant la modélisation logistique pour estimer la probabilité a posteriori *:* $P[Y_i = 0/X_i] = 1 - p_i = \frac{1}{1 + e^{\beta' + \alpha X_i}}$, *et en tenant compte du schéma d'échantillonnage.*

Intervalle de score	− 2,5	−2	−1,5	−1	−0,5	−0	0,5	1	1,5	2	2,5
Probabilité moyenne de défaillance	95,5	90,5	85,2	77,7	67,9	56,2	43,8	32,1	22,3	14,8	9,5	4,8
Probabilité d'appartenance	6,7	6,2	10,8	14,6	16,4	14,6	11,7	8,6	5,2	2,8	1,4	1,0

Les probabilités *a priori* sont comme au § 2.3.2.2., $\pi_D = 0,105$ et $\pi_N = 0,895$. On note L' 90 le nouveau score obtenu par régression logistique en tenant compte du schéma d'échantillonnage : $L'90 = \beta' + \alpha X$.

Les tableaux 4 et 5 permettent de vérifier le décalage vers les valeurs négatives de la distribution des scores L'90 comparativement aux scores L90 : les intervalles de la région négative ont une probabilité d'appartenance beaucoup plus importante que dans le tableau 4, et c'est l'inverse pour les régions positives. Les probabilités *a posteriori*[6] du tableau 5, sur les intervalles d'amplitude 0,1, restent par contre égales à celles du tableau 4. Mais elles diffèrent aux extrémités car ces classes extrêmes sont différemment dispersées qu'elles ne l'étaient au tableau 4.

C'est le tableau 5 qui correspond à une modélisation correcte de la régression logistique quand on veut utiliser la probabilité *a posteriori* du modèle. On peut constater à quel point il diffère du tableau 3 qui correspond lui à la réalité des données empiriques de l'échantillon global de 11470 entreprises. C'est le tableau 3 qui doit être utilisé pour estimer les probabilités *a posteriori* de défaillance dans le cas où on choisit la régression logistique. Cela revient à utiliser la régression logistique comme un algorithme de sélection des ratios et de détermination des coefficients. Mais on s'abstient d'utiliser les formules donnant la probabilité *a posteriori*, préférant l'estimation de celle-ci par le théorème de Bayes calculée sur un échantillon approchant les conditions asymptotiques.

2.3.3. Les courbes de performance

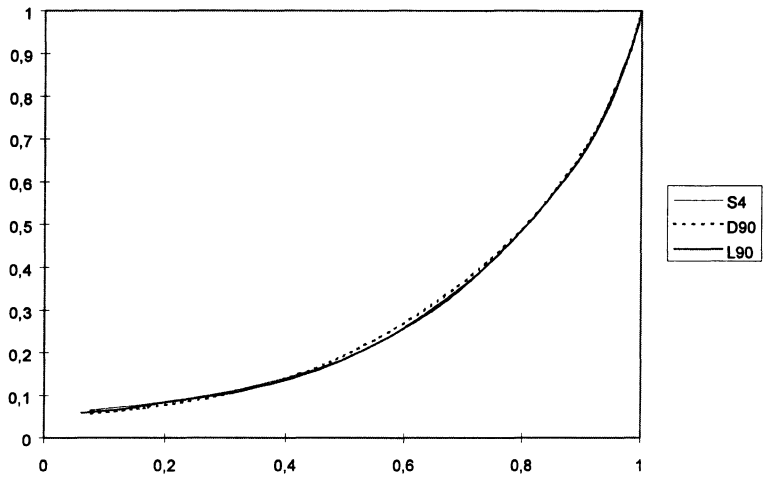
Les courbes de performance (*cf.* Gourieroux [16]) servent à comparer les performances des différents scores linéaires en rendant leurs échelles comparables. Si on considère la règle de décision : «Si $s(e) \geqslant s_o$, on prête à l'entreprise e», la part de marché est donnée par : $P[s(e) \geqslant s_o]$.

[6] La probabilité *a posteriori* est estimée sur chaque intervalle par la valeur prise par $1 - p_i$ au milieu du segment d'amplitude 0,1. Les classes extrêmes comportent plusieurs intervalles et le nombre de firmes dans ces intervalles diffèrent sensiblement dans les tableaux 4 et 5, d'où les valeurs différentes des probabilités *a posteriori* des classes extrêmes dans l'un et l'autre tableau.

Pour pouvoir comparer les performances de différents scores linéaires entre eux sans être gêné par leurs échelles généralement non comparables, on construit les courbes de performance de chacun des scores sur un même graphique.

Pour un score s, l'équation paramétrique, de paramètre s_0, de sa courbe de performance est :

$$x = P\big(s(e) \geqslant s_o\big)$$
$$y = \frac{P\big(e \in D / s(e) \geqslant s_o\big)}{P(e \in D)}$$



Avec cette définition des courbes de performance, le segment $y = 1$ correspondrait à un score ne discriminant pas mieux qu'un tirage aléatoire. Pour un deuxième score s', on peut comparer la position de sa courbe (C') par rapport à la courbe (C) du score s. Si la courbe (C') est en dessous de la courbe (C), le score s' discrimine mieux que le score s.

On utilise les probabilités estimées par la formule de Bayes, à partir des probabilités conditionnelles empiriques, pour tracer les courbes de performance des trois fonctions scores S4, D90 et L90. Celles-ci sont quasiment superposées, les performances des trois scores sont donc comparables. Ceci correspond certainement au fait que ces trois scores ont été sélectionnés comme les meilleurs (parmi les divers scores qui ont été construits mais ne sont pas présentés ici). On a pu constater qu'ils reposent sur un choix de ratios assez voisins.

2.4. Conclusion

Si l'hypothèse de multinormalité et d'homoscédasticité des lois conditionnelles est vérifiée, l'analyse discriminante linéaire de Fisher est optimale. Sinon, on attend

souvent de meilleures performances de la régression logistique du fait qu'elle s'applique à une famille plus large de fonctions de densité conditionnelles[7] .

Les hypothèses d'application plus larges de la régression logistique ont favorisé son usage. Pourtant sa qualité n'est pas meilleure que la méthode de Fisher comme on le voit sur les pourcentages de bons classements et les courbes de performance. On peut attribuer cela au fait que l'information utilisée par la méthode de Fisher est plus complète que celle utilisée par la régression logistique (*cf.* [26] Venditti).

La formule de la probabilité *a posteriori* fournie par la régression logistique nécessite des précautions liées au schéma d'échantillonnage. De plus elle ne fournit pas sur notre exemple une estimation satisfaisante de la probabilité *a posteriori.*

Les données d'entreprises ne vérifiant les hypothèses d'aucun des deux modèles (régression logistique et analyse discriminante linéaire de Fisher), c'est l'analyse discriminante linéaire de Fisher qui est préférable en raison de l'aide à l'interprétation qu'elle fournit.

3. Les réseaux de neurones

L'analyse discriminante linéaire (ADL), choisie pour sa robustesse intertemporelle sur les comptes d'entreprises, a donné sur les fichiers traités des taux de bons classements de l'ordre de 70 à 75% selon l'année étudiée. Les fonctions scores ainsi construites utilisent des variables «linéaires», c'est-à-dire qui constituent des critères monotones (croissant ou décroissant suivant la variable).

Dans le but d'améliorer ce résultat, on applique la technique du perceptron multicouche aux mêmes données. On peut alors utiliser non seulement les variables monotones sélectionnées pour l'ADL, mais aussi prendre en compte des non linéarités en introduisant d'autres variables.

Dans ce chapitre, on présente d'abord la technique des réseaux de neurones utilisée (§ 3.1.), puis on examine les divers paramètres à déterminer pour mettre en œuvre un réseau (§ 3.2.), enfin on examine les résultats (§ 3.3.).

---

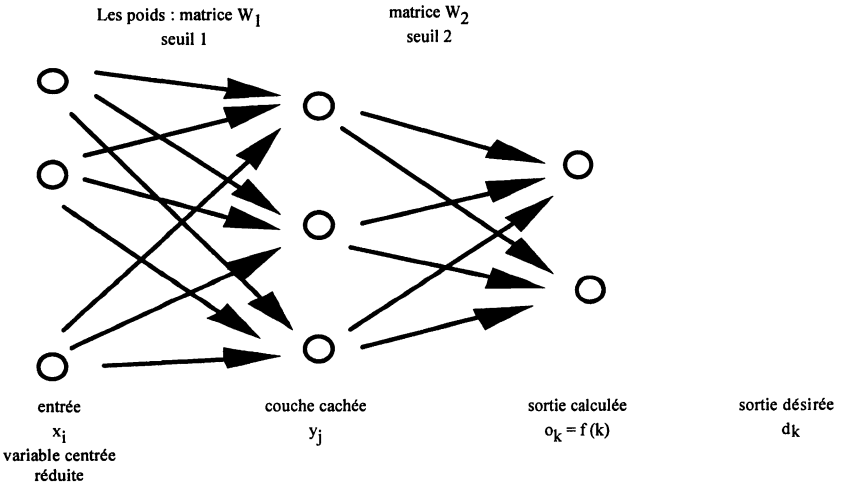
[7] La propriété du logit linéaire est équivalente à la linéarité du logarithme du rapport des fonctions de densité conditionnelles (*cf.* Caraux, Lechevallier [6]) :

$$\text{logit } \Pi(x) = \ln \frac{\Pi(x)}{1-\Pi(x)} = \beta + \alpha x \Longleftrightarrow \ln \frac{L_0(x)}{L_1(x)} = \beta^* + \alpha x$$

Cette propriété est vérifiée dans le cas particulier où les fonctions de densité conditionnelles sur chaque groupe, L_0 et L_1, sont multinormales et homoscédastiques.

3.1. Présentation des réseaux de neurones multicouches

3.1.1. Architecture du réseau de neurones



On utilise ici une architecture de réseaux multicouches : une couche d'entrée qui reçoit des variables explicatives x_i, une de sortie qui a autant de neurones o_k que de classes à discriminer (ici 2), et une couche cachée de neurones y_j.

Les neurones de niveau niv produisent une réponse sur chaque neurone de la couche de niveau niv + 1 par le calcul d'une somme pondérée des neurones de niveau niv auxquels il est connecté. Cette somme est ensuite transformée par une fonction sigmoïde g.

3.1.2. Fonction de transfert

Les variables initiales de la couche d'entrée x_i sont centrées réduites. Les neurones y_j de la couche cachée, et o_k de la couche de sortie, se calculent grâce aux formules ci-dessous, où on a supposé avoir 25 variables en entrée, 11 neurones sur la couche cachée, et 2 neurones sur la couche de sortie (comme c'est le cas dans un des exemples du § 3.3.).

$$y_j = g\left[\sum_{i=1}^{25} w_1(i,j)g(x_i) + \text{seuil}_1(j)\right]$$

avec $g(z) = \dfrac{1 - e^{-z/t}}{1 + e^{-z/t}}$ où $t = 0,5$.

Les neurones de sortie o_k, où $k = 1$ si on veut que l'entreprise soit identifiée comme défaillante, $k = 2$ dans le cas contraire, sont obtenus grâce à la fonction de

transfert[8] f à partir des x_i :

$$o_k = f(k,x) = \sum_{j=1}^{11} w_2(j,k) g \left[\sum_{i=1}^{25} w_1(i,j) g(x_i) + \text{seuil}_1(j) \right] + \text{seuil}_2(k)$$

L'apprentissage d'un tel réseau est supervisé. Il utilise un algorithme de rétropropagation du gradient de l'erreur (*cf.* Rummelhart, Hinton, Williams [23]) pour estimer les poids $w_1(i,j)$ $w_2(j,k)$ et les seuils $\text{seuil}_1(j)$ et $\text{seuil}_2(k)$. La fonction d'erreur est quadratique.

3.1.3. Algorithme de calcul des poids w : rétropropagation du gradient de l'erreur

Pour chaque entreprise présentée, le réseau cherche à minimiser, sur la couche de sortie, l'erreur quadratique E commise entre la réponse effective o_k et la réponse désirée d_k des 2 neurones, où d_k est la variable indicatrice du groupe k :

$$E = \sum_{k=1}^{2} (o_k - d_k)^2$$

Initialement les poids sont choisis au hasard. Puis le gradient de E est rétropropagé en modifiant la valeur des poids de façon à minimiser l'erreur E. Plus précisément les poids à déterminer sont les $W_1(i,j)$ et les $W_2(j,k)$. En appelant W un de ces poids, et t le numéro de l'itération de l'algorithme[9], on peut écrire qu'à l'étape t, le poids W dépend de l'étape antérieure $t-1$ par la formule suivante où est le paramètre d'apprentissage du réseau :

$$W(t) = W(t-1) + \Delta W(t)$$

avec $\Delta W(t) = -\varepsilon \dfrac{\partial E}{\partial W}$.

L'algorithme examine successivement toutes les entreprises de l'échantillon, puis recommence plusieurs fois le passage pour obtenir la convergence des poids vers un optimum. Toutefois on n'est pas assuré d'atteindre l'optimum global.

Ce sont les poids W qui portent la connaissance du problème à traiter.

---

[8] La fonction de transfert la plus classique est :

$$f(k,x) = g \left\{ \left(\sum_{j=1}^{11} w_2(j,k) g \left[\sum_{i=1}^{25} w_1(i,j) x_i + \text{seuil}_1(j) \right] + \text{seuil}_2(k) \right) \right\}.$$

Elle permettrait d'avoir des estimations de seuils plus précises.
Des essais sur le logiciel de LAFORIA en utilisant cette formulation ont été faits, mais les résultats n'ont pas été améliorés. Nous en sommes donc restés à notre formulation de la fonction de transfert pour présenter nos résultats. Toutefois d'autres formulations suggérées par Patrick Gallinari, en particulier l'entropie, feront l'objet d'un travail ultérieur.

[9] Le numéro t de l'itération correspond au calcul sur l'exemple t selon la terminologie des réseaux de neurones, c'est-à-dire au calcul sur l'individu t de l'échantillon selon la terminologie de l'analyse discriminante classique en analyse des données. Ici les individus sont ordonnés et passent successivement.

3.1.4. Probabilité a posteriori *de défaillance*

Comme Gish l'a montré (*cf.* Gish [13]), le réseau étant construit en utilisant l'erreur quadratique, il approxime la probabilité *a posteriori*.

Afin de normaliser les sorties du réseau on considère les nouveaux outputs :

$$p_1(x) = \frac{e^{f(1,x)}}{e^{f(1,x)} + e^{f(2,x)}} \quad \text{et} \quad p_2(x) = 1 - p_1(x)$$

Sous la condition de convergence du réseau vers l'optimum global, $p_1(x)$ peut être considéré comme la probabilité *a posteriori* de défaillance, dans le cas où la probabilité *a priori* de défaillance, π_D, est égale à la proportion de firmes défaillantes dans l'échantillon. La règle d'affectation est alors :

$$e \text{ est affectée à } o_1 \iff p_1(x) \geqslant 0,5$$

Cependant dans notre application la proportion d'entreprises défaillantes dans l'échantillon (36,9% ou 51,1% selon l'échantillon de non défaillantes retenu) est très différente de D. De plus, on n'est pas sûr que le réseau ait effectivement convergé vers l'optimum global. On sera donc amené à une évaluation différente de la probabilité *a posteriori*.

En effet, en utilisant les distributions conditionnelles de p_1 sur chaque catégorie de firme, on peut évaluer la probabilité *a posteriori* de défaillance selon la valeur de p_1 par le théorème de Bayes (*cf.* § 4.2).

3.2. Mise en œuvre du réseau

Pour optimiser la mise en œuvre du réseau on contrôle le choix de plusieurs éléments : le paramètre d'apprentissage du réseau ε, le choix d'un jeu de variables initiales, le nombre de neurones sur la couche cachée. Pour tous ces choix on utilise la validation croisée.

3.2.1. Les techniques de validation

Pour éviter le «sur-apprentissage» on scinde l'**échantillon de base** en échantillon d'apprentissage et échantillon test et on pratique la **validation croisée** (§ 1.1.2.). Une fois déterminé un réseau intéressant, on l'applique sur des **échantillons tests indépendants** de l'échantillon de base.

3.2.2. Le paramètre d'apprentissage ε

Le coefficient ε de l'algorithme de rétropropagation du gradient de l'erreur quadratique influence la qualité et la vitesse de convergence de l'algorithme qui optimise les poids W et les seuils s. On choisit ε de façon à concilier ces 2 objectifs. Pour cela on envisage différents choix de valeurs du paramètre ε, comprises entre

0,5 et 0,0001. En utilisant ici le maximum de variables (50 ratios) on construit les réseaux correspondant à ces différentes valeurs de ε. On retient la valeur qui maximise le taux moyen de bons classements et minimise l'erreur quadratique moyenne sur les échantillons tests : dans le cas du tableau 6 on choisit 0,001 qui allie ces deux objectifs.

TABLEAU 6
Pourcentages de biens classés et erreur en fonction de ε sur les échantillons tests

Coefficient ε	Pourcentage de biens classés	Erreur
0.5	73.10	0.548841
0.1	71.55	0.444563
0.05	69.72	0.439866
0.01	74.23	0.367107
0.005	72.25	0.380004
0.001	**74.93**	**0.333717**
0.0005	71.41	0.365224
0.0001	73.38	0.344127

3.2.3. Le choix des variables initiales

On dispose de 50 ratios. Parmi eux l'analyse discriminante linéaire avait déterminé plusieurs jeux intéressants. Par exemple ceux de la fonction S4 sont au nombre de 8.

On veut enrichir cet ensemble de 8 ratios pour capter des non linéarités. On se propose plusieurs choix initiaux :

- 8 variables de la fonction score S 4.
- 16 variables sélectionnées grâce à une ACP à partir des 50 ratios dont ils résument bien l'information et qui comprennent les 8 ratios précédents.

Ensuite on a encore élargi ce choix par 9 variables complémentaires :

- 25 variables (16 + 9).

Enfin, on a cherché à restreindre ce choix par une technique spécifique aux réseaux de neurones, l'analyse des sensibilités (*cf.* § 3.2.5.). On a ainsi obtenu deux autres jeux de variables en entrée :

- 18 variables,
- 14 variables.

3.2.4. Nombre de neurones de la couche cachée

Le nombre optimal de neurones dans la couche cachée est déterminé en utilisant les critères de choix du modèle (maximisation du pourcentage de biens classés et

minimisation de l'erreur). On présente au tableau 7 cette recherche sur les 16 ratios qui conduit dans ce cas à prendre 3 neurones.

3.2.5. Test de sensibilité des variables

Le réseau T est initialement construit avec p variables (ici $p = 25$). Son taux de bons classements est τ. Son erreur quadratique est E.

On neutralise la variable i en posant : $w_1(i,j) = 0$ pour $j = 1, ...J$, et on construit le réseau de neurones T_i, dont le taux de bons classements est τ_i et l'erreur quadratique est E_i.

On compare τ_i et τ, E_i et E. Les variables pour lesquelles τ_i est comparable à τ, et, E_i est très inférieur à E, sont enlevées.

À partir du réseau à 25 variables (réseau 2), on a ainsi construit : le réseau 3 (à 18 variables) et le réseau 4 (à 14 variables) ayant chacun 11 neurones sur la couche cachée.

3.3. Les résultats

3.3.1. Les réseaux à 8 ratios et à 16 ratios

Dans chaque cas on estime le réseau et ses performances par validation croisée, en détaillant les pourcentages de bons classements sur chaque catégorie d'entreprises, défaillantes et non défaillantes, puis globalement (cf. tableaux 8 et 9).

De ces deux réseaux, celui à 8 ratios est le plus performant : le réseau à perceptron multicouche (RPM) permet un progrès de 3,86 points par rapport à l'analyse discriminante linéaire (ADL) dans le pourcentage global de bons classements sur les échantillons tests (*cf.* tableau 8). On le désigne sous le nom de réseau 1. On constate qu'il classe mieux les non défaillantes. En effet, celles-ci sont plus nombreuses dans l'échantillon (809 défaillantes, 1381 non défaillantes). C'est pourquoi par la suite on utilise un autre échantillon de non défaillantes qui comporte 775 entreprises.

3.3.2. Les autres réseaux

Un réseau est construit à partir des 25 ratios, appelé réseau 2. Puis par analyse des sensibilités des variables on sélectionne deux sous ensembles de 18 et 14 ratios qui donneront respectivement les réseaux 3 et 4. On regroupe dans le tableau 10 les résultats des pourcentages de bons classements sur l'échantillon de base et sur les échantillons tests par catégorie d'entreprises. Pour mémoire et pour faciliter la comparaison on rappelle les pourcentages de bons classements de la fonction score la plus performante S4.

4. Comparaison des résultats

4.1. Pourcentages de bons classements

Pour rendre comparables les résultats on calcule un taux global de bons classements en pondérant par l'effectif de chaque groupe. Les résultats obtenus dans

TABLEAU 7

Pourcentages de biens classés et erreur sur les échantillons tests en fonction du nombre de neurones de la couche cachée

Nombre de neurones	Pourcentage de biens classés	Erreur
2	72,39	0,471251
3	**72,39**	**0,423197**
4	72,82	0,455237
5	70,85	0,514710
6	72,25	0,529572
7	72,25	0,550276
8	70,7	0,579803
9	69,72	0,599482
10	73,52	0,524924

TABLEAU 8

Réseau à 8 ratios
$\varepsilon = 0,001$, *couche cachée à 3 neurones*

	Échantillons tests	
	ADL	RPM
Défaillantes	67,87 %	61,80 %
Non défaillantes	74,27 %	82,84 %
Ensemble	71,07 %	74,93 %

ADL : analyse discriminante linéaire
RPM : réseau perceptron multicouche

TABLEAU 9

Réseau à 16 ratios
$\varepsilon = 0,001$, *couche cachée à 3 neurones*

	Échantillons tests	
	ADL	RPM
Défaillantes	69,40 %	60,54 %
Non défaillantes	72,90 %	78,32 %
Ensemble	71,60 %	71,61 %

ADL : analyse discriminante linéaire
RPM : réseau perceptron multicouche

TABLEAU 10
Réseaux de neurones : les résultats par groupe
Pourcentages de bons classements

Réseau	Échantillon													
	1992		1991		1990		1989		1988		1987		1986	
	D	N	D	N	D	N	D	N	D	N	D	N	D	N
nb entreprises	318	1565	635	1523	**809**	**1381**	542	1173	388	979	306	845	289	717
Réseau 1	82,7	77,4	71,3	79,4	**61,8**	**82,8**	67,7	79,4	75,3	79,8	75,2	76,2	81,7	73,8
S4 pour mémoire	87,7	75,3	77,3	74,8	**68,3**	**75,3**	74,4	73,2	81,2	74,0	77,9	71,2	82,1	69,2
nb entreprises	318	1565	635	1523	**809**	**775**	542	1173	388	979	306	845	289	717
Réseau 2	77,0	65,6	70,7	67,9	**72,9**	**77,6**	68,8	63,9	72,2	63,1	77,8	66,0	71,6	62,3
Réseau 3	77,2	67,1	74,5	66,3	**74,7**	**74,0**	69,4	67,3	75,8	66,9	77,1	66,0	82,4	62,6
Réseau 4	73,3	61,9	72,9	66,2	**69,1**	**77,0**	72,1	65,8	70,1	66,1	77,1	64,0	85,5	63,8

Réseau 1 (8 ratios) - Réseau 2 (25 ratios) - Réseau 3 (18 ratios) - Réseau 4 (14 ratios)
D : défaillantes N : non défaillantes
En gras : Échantillon de base, résultats de la validation croisée

le tableau 11 montrent que le meilleur outil est le réseau de neurones n°1, ensuite vient la fonction score S4, puis les réseaux 2,3,4 dont les résultats sont de plusieurs points inférieurs sur les échantillons tests.

TABLEAU 11
Réseaux de neurones : les résultats globaux
Pourcentages de bons classements

Réseau	Échantillon						
	1992	1991	1990	1989	1988	1987	1986
Réseau 1	78,8	77,0	**75,8**	75,7	78,5	75,9	76,1
S4 pour mémoire	77,4	75,5	**72,6**	73,6	76,0	73,0	72,9
Réseau 2	67,5	68,7	**75,2**	65,4	65,7	69,1	65,0
Réseau 3	68,8	68,7	**74,4**	68,0	69,4	69,0	68,2
Réseau 4	63,8	68,2	**73,0**	67,8	67,2	67,5	70,0

Réseau 1 (8 ratios) - Réseau 2 (25 ratios) - Réseau 3 (18 ratios) - Réseau 4 (14 ratios)
D : défaillantes N : non défaillantes
En gras : Échantillon de base, résultats de la validation croisée

Les constats de ce travail concernent d'abord la sélection des variables. Deux types de méthodes ont été appliquées : une méthode adaptée à l'analyse discriminante linéaire classique, une méthode spécifique des réseaux de neurones.

Ces deux méthodes ne sélectionnent pas les mêmes variables. Et ce sont les 8 ratios sélectionnés pour le score S4 qui nous donnent le meilleur réseau.

Le réseau 1 est plus performant que le score S4 sur l'échantillon de base comme sur les échantillons tests. La sélection par l'algorithme de Furnival et Wilson a permis pour les réseaux de neurones d'améliorer le pourcentage de bons classements. On peut dire ici que moins de variables bien choisies améliorent la simplicité et la performance des réseaux.

Le réseau 2, établi sur 25 variables discriminantes une à une et de caractère linéaire ou non linéaire, n'apporte pas d'amélioration au contraire.

On attendait de la sélection par les tests de sensibilité qu'elle permette une meilleure performance des nouveaux réseaux ainsi construits, mais il n'en est rien. En effet, les réseaux 3 et 4 (comme le réseau 2 d'ailleurs) présentent un surapprentissage, au sens où les pourcentages de bons classements baissent de 6 à 10 points quand on passe de l'échantillon de base (où les résultats sont établis par validation croisée) aux échantillons tests. Ceci n'est le cas ni pour le réseau 1, ni pour le score, pour lesquels les pourcentages de bons classements sur échantillons tests excèdent 72 %, tandis que pour les réseaux 2, 3, 4 ils sont inférieurs à 70 %.

4.2. Probabilités a posteriori

Dans ce § on s'en tient au meilleur des réseaux, le réseau 1. Pour la comparaison avec les résultats de l'analyse discriminante, on ne retiendra que la fonction S4, comme la plus performante.

4.2.1. Probabilité a posteriori pour le réseau de neurones n° 1

Reprenant l'expression normalisée de la sortie du réseau, $p_1(x)$, on en observe les distributions conditionnelles dans le tableau ci-dessous, sur le fichier le plus vaste de 11 470 entreprises.

TABLEAU 12
Distribution de p_1 selon la catégorie d'entreprises

groupe	nb	min	C1	D1	Q1	Q2	Q3	D9	C99	max
défaillantes	3287	0,03	0,08	0,26	0,46	0,63	0,73	0,81	0,88	0,92
non défaillantes	8183	0,03	0,03	0,06	0,09	0,21	0,46	0,64	0,79	0,90

nb : nombre d'entreprises min : minimum max : maximum
C_i : $i^{\text{ème}}$ centile D_i : $i^{\text{ème}}$ décile Q_i : $i^{\text{ème}}$ quartile

Comme il est attendu d'après la règle d'affectation (*cf.* § 3.1.4 et tableau de résultats avec les pourcentages de bons classements par catégorie § 3.3.2.), la proportion d'entreprises défaillantes pour lesquelles $p_1 > 0,5$ est supérieure à 50 %, tandis que la proportion de firmes non défaillantes pour lesquelles $p_1 > 0,5$ est inférieure à 50 %, et même inférieure à 25 %.

En utilisant le théorème de Bayes, on évalue la probabilité *a posteriori* de défaillance selon la valeur de p_1 (par exemple selon l'appartenance de p_1 à une

région r de l'intervalle $[0; 1]$).

$$P(e \in D/p_1 \in r) = \frac{P(p_1 \in r \text{ et } e \in D)}{P(p_1 \in r)} = \frac{P(p_1 \in r/D)\pi_D}{P(p_1 \in r/D)\pi_D + P(p_1 \in r/N))\pi_D}$$

En prenant pour r des intervalles d'amplitude 0,1, on obtient les répartitions suivantes de chacune des catégories de firmes ce qui permet de calculer les probabilités *a posteriori* de défaillance sur chaque intervalle, en prenant $\pi_D = 0,105$ et $\pi_N = 0,895$, comme au § 2.3.2. On présente ceci dans le tableau 13 en ordonnant par probabilité de défaillance décroissante pour respecter l'analogie de disposition avec les tableaux du § 2.3.2.

TABLEAU 13
Répartition des firmes par intervalle r de p_1 et probabilité a posteriori de défaillance

intervalle r	1-0,9	0,9-0,8	0,8-0,7	0,7-0,6	0,6-0,5	0,5-0,4	0,4-0,3	0,3-0,2	0,2-0,1	0 ,1-0	total
défaillantes	0,24	10,95	23,30	22,57	14,12	8,85	7,76	5,63	4,72	1,86	100
non défaillantes	0,01	0,77	4,64	7,83	8,15	7,96	9,67	12,64	22,11	26,23	100
probabilité a posteriori de défaillance	73,8	62,5	37,1	25,3	16,9	11,5	8,6	5,0	2,4	0,8	

On remarque que l'amplitude des valeurs prises par la probabilité *a posteriori* est analogue à ce qui avait été obtenu par S4.

4.2.2. Comparaison des performances du réseau de neurones n° 1 et de la fonction score S4

On établit le tableau croisé des probabilités *a posteriori* de défaillance par l'un et l'autre outil (tableau 14) et on étudie comment les firmes se répartissent dans ce tableau croisé.

On constate une proximité certaine des diagnostics par l'un et l'autre classement au sens où une entreprise qui a une forte probabilité de défaillance selon le score S4, a généralement aussi une forte probabilité de défaillance d'après le réseau 1. Bien sûr l'adéquation n'est pas très précise et le diagnostic est plus ou moins sévère selon l'outil examiné. Plus précisément, si on considère comme cas de contradiction forte de classement le cas d'une firme ayant une probabilité de défaillance forte par l'un des outils, et faible pour l'autre outil, on relève une proportion limitée de tels cas qui sont représentés par les cases grisées du tableau. Cela constitue 824 cas, soit 7,2 % du fichier. On constate qu'ils correspondent le plus souvent à une plus grande sévérité du réseau 1.

TABLEAU 14
Croisement des classements selon les probabilités a posteriori
(présentées en %)
obtenus par S4 et par le réseau 1

		Probabilité a posteriori de défaillance estimée par le réseau 1										
		73,8	62,5	37,1	25,3	6,9	11,5	8,6	5	2,4	0,8	Total
S4	P(D/S4)											
		4	62	121	77	**3**	0	0	0	0	0	267
−3	65,1	1,50	23,22	45,32	28,84	**1,12**	0,00	0,00	0,00	0,00	0,00	
		1	26	68	47	**3**	0	0	0	0	0	145
−2,5	48,3	0,69	17,93	46,90	32,41	**2,07**	0,00	0,00	0,00	0,00	0,00	
		2	24	92	95	5	**1**	0	0	0	0	219
−2	44,5	0,91	10,96	42,01	43,38	2,28	**0,46**	0,00	0,00	0,00	0,00	
		0	53	121	196	33	16	**4**	0	0	0	423
−1,5	34,8	0,00	12,53	28,61	46,34	7,80	3,78	**0,95**	0,00	0,00	0,00	
		0	41	122	333	169	70	23	**3**	0	0	761
−1	22,3	0,00	5,39	16,03	43,76	22,21	9,20	3,02	**0,39**	0,00	0,00	
		0	48	134	277	363	251	134	37	0	0	1244
−0,5	14,5	0,00	3,86	10,77	22,27	29,18	20,18	10,77	2,97	0,00	0,00	
		0	**6**	**153**	**262**	328	460	305	80	14	0	1608
0	8,4	0,00	**0,37**	**9,51**	**16,29**	20,40	28,61	18,97	4,98	0,87	0,00	
		0	0	**1**	**56**	**280**	471	522	333	133	4	1800
0,5	3,7	0,00	0,00	**0,06**	**3,11**	**15,56**	26,17	29,00	18,50	7,39	0,22	
		0	0	0	0	**13**	**36**	141	507	951	111	1759
1,0	1,8	0,00	0,00	0,00	0,00	**0,74**	**2,05**	8,02	28,82	54,06	6,31	
		0	0	0	0	0	**3**	14	48	604	764	1433
1,5	1	0,00	0,00	0,00	0,00	0,00	**0,21**	0,98	3,35	42,15	53,31	
		0	0	0	0	0	**1**	0	9	129	1672	1811
	0,7	0,00	0,00	0,00	0,00	0,00	**0,06**	0,00	0,50	7,12	92,32	
	Total	7	260	812	1343	1197	1309	1143	1017	1831	2551	11470

Dans chaque case : fréquence et pourcentage en ligne

En gras : cas de classement très différents

CONCLUSION

Finalement c'est la sélection de variables monotones qui a été la plus efficace. Au stade où nous en sommes, il semble donc que, sur données comptables d'entreprises, les non linéarités robustes sont beaucoup plus difficiles à capter que les linéarités. Certes nous n'avons pas encore utilisé toutes les techniques disponibles (*cf.* [9], [10]) pour perfectionner les réseaux comme :

- la réduction de la complexité du réseau par élimination des connexions dont les poids sont non significatifs,
- la régularisation par bruitage des données,
- l'utilisation de méthodes bayésiennes.

Dans la comparaison des performances des outils, l'étude de la sensibilité du diagnostic à une faible variation des comptes d'une entreprise est très nécessaire en cas d'utilisation opérationnelle. Plus précisément, une faible modification des comptes n'entraîne qu'un faible changement d'un score linéaire, donc de la probabilité de défaillance qui lui est associée. On a donc progressivité de l'appréciation sur l'entreprise. D'après l'étude des probabilités *a posteriori* du § 4.2. les réseaux de neurones ont à peu près les mêmes propriétés.

Bibliographie

[1] E. I. ALTMAN, G. MARCO, F. VARETTO (1994) : Corporate Distress Diagnosis : Comparisons using Linear Discriminant Analysis and Neural Networks (the italian experience), *Journal of Banking and Finance 18*, p. 505-529, North Holland;

[2] M. BARDOS (1989) : Trois méthodes d'analyse discriminante, *Cahiers économiques et monétaires n°33*, Banque de France, pp.151-189;

[3] M. BARDOS (1995) : *Les défaillances d'entreprises dans l'industrie : ratios significatifs, processus de défaillances, détection précoce*, Collection Entreprise B 95/03, Banque de France;

[4] J. BAETGE, C. KRAUSE (1994) : The Classification of Companies by Means of Neural Networks, *Recherches en comptabilité internationale*;

[5] Ch.M. BISHOP (1995) : *Neural Networks for Pattern Recognition*, Clarendon press Oxford;

[6] G. CARAUX, Y. LECHEVALLIER (1996) : Règles de décision bayésienne et méthodes statistiques de classement, *Revue d'Intelligence Artificielle*, volume 10, n°2 -3, 1996;

[7] J.F. CASTA, B. PRAT (1994) : Approche connexionniste de la classification des entreprises, Association française de comptabilité, congrès de Paris IX Dauphine, *Recherche en comptabilité internationale*;

[8] G. CELEUX, J.P. NAKACHE (1994) : *Analyse discriminante sur variables qualitatives*, Polytechnica;

[9] T. CIBAS, F. FOGELMAN SOULIÉ, P. GALLINARI, S. RAUDYS (1996) : Variable Selection with Neural Networks, *Neurocomputing*;

[10] M. COTRELL, B. GIRARD, Y. GIRARD, M. MANGEAS, C. MULLER (1995) : Neural Modeling for Time Series : a Statistical Stepwise Method for Weight Elimination, *IEEE Transactions on Neural Networks*, vol. 6, n°6, novembre 1995;

[11] G.M. FURNIVAL, R.W. WILSON (1974) : Regressions by leaps and bounds, *Technometrics*, 16, p 499-511;

[12] P. GALLINARI, O. GASCUEL (1996) : Statistique, apprentissage et généralisation; applications aux réseaux de neurones, *Revue d'Intelligence Artificielle*, volume 10, n°2-3, 1996;

[13] H. GISH (1990) : A probabilistic approach to the understanding and training of neural network classifiers, *Proceedings of IEEE Conference on Acoustics speech and signal processing*, p 1361-1364;

[14] R. GNANADESIKAN and panel of authors (1989) : Discriminant Analysis and Clustering, *Statistical Science*, vol. 4, N°1, p. 34-69;

[15] C. GOURIEROUX (1984) : *Économétrie des variables qualitatives*, Économica;

[16] C. GOURIEROUX (1992) : Courbes de performance, de sélection et de discrimination, *Annales d'économie et de statistique* n°28, INSEE;

[17] G.J. Mc LACHLAN (1992) : *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New-York;

[18] L. LEBART, A. MORINEAU, M. PIRON (1995) : *Statistique Exploratoire Multidimensionnelle*, Dunod;

[19] Y. LE CUN, J.S. DENKER, S.A. SOLLA (1990) : Optimal Brain Damage, *Advances in Neural Information Processing Systems*, vol. 2, p. 598-605;

[20] H. OOGHE, P. JOOS, D. DE VOS (1993) : *Towards an Improved Method of Evaluation of Financial Distress Models and Presentation of their Results*, Universteit GENT, Vakgroep bedrijfsfinancierning;

[21] Ph. RICHARDOT (1985) : Différents outils pour discrimination linéaire entre deux groupes, *thèse du 3e cycle, UER de mathématiques de la décision, Université de Paris-Dauphine*;

[22] J.M. ROMEDER (1973) : *Méthodes et programmes d'analyse discriminante*, Dunod;

[23] D. RUMMELHART, G.E. HINTON et R.J. WILLIAMS (1986) : Learning representations by error back propagation, in *Paralled distributed processing-s'exploration in the micro-structure of cognition*, MIT press;

[24] R. TAFFLER (1983) : The Assessment of Company Solvency and Performance Using a Statistical Model, *Accounting and Business Research*, pp. 295-308;

[25] J. ULMO (1973) : Différents aspects de l'analyse discriminante, *Revue de Statistique Appliquée*, vol. XXI, n°2, pp. 17-55;

[26] V. VENDITTI (1996) : Influence du schéma d'échantillonnage en regression logistique dans *Relecture de la procédure de regression logistique par le principe du maximum d'entropie* Journées de l'ASU à Québec;

[27] W. ZHU (1995) : *Méthodes statistiques et approche neuronale, stratégie et validation dans le cas de la discrimination*, thèse de doctorat, Université Paris Dauphine.