

JULIE DROUIN

**Les lois g -et- h : définition, propriétés,
généralisations et applications**

Essai présenté
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de maîtrise en statistique
pour l'obtention du grade de Maître ès sciences (M.Sc.)

FACULTÉ DES SCIENCES ET DE GÉNIE
UNIVERSITÉ LAVAL
QUÉBEC

Août 2008

Résumé

Bien que la loi normale joue un rôle central en statistique, il arrive souvent qu'un ensemble de données soit incompatible avec le modèle gaussien, nécessitant de ce fait l'introduction de nouvelles lois capables de s'adapter à une grande variété de formes distributionnelles. Le présent essai se veut une étude des lois g -et- h , une famille de lois relativement peu connue mais offrant un potentiel considérable pour la simulation et la modélisation de données asymétriques et à queues lourdes. Nous traitons en un premier temps de la définition et des propriétés élémentaires des lois g -et- h dans un contexte univarié, puis nous discutons d'une généralisation multivariée de ces lois. Un survol des principales applications des lois g -et- h dans les domaines de la statistique, de la finance, de l'informatique et de la météorologie est également présenté.

Avant-propos

Je tiens ici à remercier toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de cet essai. Merci tout d'abord à mon directeur de recherche, M. Christian Genest, professeur au Département de mathématiques et de statistique de l'Université Laval. Ses conseils judicieux, ses encouragements et la confiance qu'il m'a témoignée tout au long de cet essai m'ont été d'une aide précieuse.

Merci également à l'équipe de professeurs et de professionnels du Département de mathématiques et de statistique, qui ont fait de mon séjour à l'Université Laval une expérience des plus enrichissantes. Je remercie tout particulièrement le professeur Lajmi Lakhel Chaieb d'avoir bien voulu relire et commenter cet essai.

Sur un plan plus personnel, je désire remercier les membres de ma famille pour leur soutien et leur amour. Merci particulièrement à mes parents, Lyne et Jean-Luc, qui ont toujours encouragé mes études et grâce à qui j'ai grandi dans un milieu épanouissant. Je remercie également mes amis pour leur présence et leur soutien. Un merci tout spécial à Edith, non seulement pour son aide avec le logiciel Matlab, mais surtout pour son amitié, ses encouragements et son éternel optimisme. Mes derniers remerciements s'adressent à mon amoureux, Dave, qui a toujours eu confiance en moi, m'a poussée à me dépasser et m'a aidée à garder le sourire dans les moments difficiles. Je le remercie simplement pour sa présence et son amour.

Ce travail a été financé par une bourse d'études supérieures du *Conseil de recherches en sciences naturelles et en génie du Canada* (CRSNG), ainsi que par des subventions de recherche accordées à M. Genest par le CRSNG, le *Fonds québécois de la recherche sur la nature et les technologies* et l'*Institut de finance mathématique de Montréal*. Que ces organismes trouvent ici l'expression de ma reconnaissance.

Table des matières

Résumé	ii
Avant-Propos	iii
Table des matières	v
Liste des tableaux	vi
Table des figures	vii
1 Introduction	1
2 Définition et propriétés des lois g-et-h univariées	3
2.1 Asymétrie et lois g	3
2.1.1 Lois g	4
2.1.2 Choix d'une valeur appropriée pour g	7
2.1.3 Lois log-normales	11
2.2 Allongement et lois h	11
2.2.1 Lois h	12
2.2.2 Choix d'une valeur appropriée pour h	13
2.3 Lois g -et- h	17
2.3.1 Choix de valeurs appropriées pour g et h	18
2.3.2 Méthodes d'estimation basées sur la vraisemblance	21
2.4 Lois g -et- h avec paramètres g et h variables	23
2.5 Moments	28
2.5.1 Lois g	29
2.5.2 Lois h	29
2.5.3 Lois g -et- h	30
2.5.4 Coefficients d'asymétrie et d'aplatissement	31
3 Généralisation multivariée de la loi g-et-h	33
3.1 Loi g -et- h multivariée	33
3.2 Propriétés de la loi g -et- h multivariée	38

3.3	Ajustement de la loi g-et-h multivariée à des données	39
4	Survol d'applications	41
4.1	Applications en statistique	41
4.2	Applications en finance	44
4.3	Application en informatique	45
4.4	Applications en météorologie	46
5	Conclusion	48
	Bibliographie	50
A	Valeurs lettrées	54
A.1	Exemple de calcul	55
B	Démonstrations des résultats (2.28) et (2.31)	56
B.1	Démonstration du résultat (2.28)	56
B.2	Démonstration du résultat (2.31)	57

Liste des tableaux

2.1	Valeurs de $Y_g(z_p)$ pour diverses valeurs de p et de g	5
2.2	Calculs pour déterminer g_p pour la loi du χ^2 à six degrés de liberté . . .	9
2.3	Résumé des résultats obtenus, quantiles approximés et résidus	10
2.4	Valeurs de $Y_h(z_p)$ pour diverses valeurs de p et de h	13
2.5	Calcul de $\ln(\psi_p)$ pour la loi de Cauchy centrée réduite	16
2.6	Ajustement pour un g constant dans l'exemple de la χ_6^2	19
2.7	Exactitude de l'approximation $\chi_6^2 \approx 5.348 + 7.968(e^{0.406Z} - 1)e^{-0.0164Z^2}$	21
2.8	Valeurs lettrées pour l'exemple des revenus de ménage	25
2.9	Calculs pour décrire l'asymétrie dans l'exemple des revenus de ménage	25
2.10	Ajustement des demi-étendues supérieures de l'exemple des revenus de ménage	26
2.11	Valeurs lettrées observées et ajustées pour l'exemple des revenus de ménage	28

Table des figures

2.1	Densité de probabilité pour six valeurs de g	6
2.2	Densité de probabilité de la loi du χ^2 à six degrés de liberté	8
2.3	Diagramme quantile-quantile de la χ^2_6 et de la loi g ajustée	9
2.4	Densités de probabilité de la χ^2_6 et de la loi g ajustée	10
2.5	Densité de probabilité de la loi h pour quatre valeurs de h	13
2.6	Densités de probabilité de la Cauchy centrée réduite et de la normale centrée réduite	15
2.7	Graphique de $\ln(\psi_p)$ en fonction de $z_p^2/2$ pour la loi de Cauchy centrée réduite	16
2.8	Densités de probabilité de la Cauchy et de la loi h ajustée	17
2.9	Graphique de $\ln(\text{UHS}_p^*)$ en fonction de $z_p^2/2$ pour l'exemple de la χ^2_6	20
2.10	Graphique de g_p en fonction de z_p^2 pour l'exemple des revenus de ménage	26
2.11	Graphique de $\ln(\text{UHS}_p^*)$ en fonction de z_p^2 pour l'exemple des revenus de ménage	27
2.12	Graphique de $\beta_1(g)$ en fonction de g pour les lois g	32
2.13	Graphique de $\beta_2(h)$ en fonction de h pour les lois h	32
3.1	Courbes de niveau des quantiles de normes ℓ_1 et ℓ_2 de la loi normale bivariée $\mathcal{N}_2(\mathbf{0}, \mathbf{I}_2)$	36
3.2	Courbes de niveau des quantiles de normes ℓ_1 et ℓ_2 de la loi normale bivariée $\mathcal{N}_2(\mathbf{0}, \mathbf{I}_2)$ après une transformation g-et-h	37

Chapitre 1

Introduction

Les lois normales, tant univariées que multivariées, jouent depuis plusieurs années un rôle prépondérant en statistique. Il existe néanmoins de nombreux phénomènes qui n'obéissent pas à ce type de loi, nécessitant de ce fait l'introduction de lois non gaussiennes capables de s'adapter à diverses formes. Citons par exemple les rendements boursiers et les vitesses extrêmes des vents, fréquemment étudiés en finance et en météorologie, qui présentent généralement des structures d'asymétrie et d'allongement incompatibles avec le modèle normal.

Suggérées en 1977 par le statisticien américain J. W. Tukey, les lois g -et- h ont notamment été étudiées par [Hoaglin et Peters \(1979\)](#), [Martinez et Iglewicz \(1984\)](#) et [Hoaglin \(1985\)](#) dans le cas univarié, puis par [Field et Genton \(2006\)](#) dans un contexte multivarié. Obtenues par une transformation d'une variable aléatoire normale centrée réduite, ces lois permettent une grande variété de formes distributionnelles. Elles peuvent ainsi s'ajuster à une vaste gamme d'ensembles de données et approximer de nombreuses lois théoriques non gaussiennes. Le présent essai se veut une étude de ces lois polyvalentes et des plus prometteuses.

Le chapitre [2](#) définit et présente les principales propriétés des lois g -et- h univariées. Deux sous-familles importantes de ces lois sont d'abord introduites : les lois g , qui permettent de décrire des variables aléatoires continues asymétriques, et les lois h , qui peuvent être utilisées pour la modélisation de lois symétriques ayant des queues plus lourdes que celles de la loi normale. Ces deux lois sont ensuite réunies pour former les lois g -et- h , une famille de lois permettant de décrire des structures d'asymétrie et d'allongement complexes.

Les travaux de [Field et Genton \(2006\)](#), qui proposent une généralisation multivariée des lois g -et- h , sont présentés au chapitre 3. Nous y traitons notamment de la définition et des propriétés élémentaires des lois \mathbf{g} -et- \mathbf{h} multivariées, ainsi que de la notion de quantile multivarié, une notion indispensable à l'ajustement de ces lois.

Le chapitre 4 expose pour sa part plusieurs applications des lois g -et- h univariées. Notre objectif ici n'est toutefois pas de présenter une liste exhaustive de telles applications, mais plutôt d'illustrer l'intérêt des lois g -et- h . Nous verrons en particulier que ces dernières ont été fréquemment utilisées pour la simulation et la modélisation de données asymétriques et à queues lourdes, et ce dans des domaines variés tels que la finance et la météorologie.

Chapitre 2

Définition et propriétés des lois g -et- h univariées

Dans ce chapitre, nous présentons en détail les lois g -et- h dans le cas univarié. La section 2.1 traite d'asymétrie et des lois g , alors que la section 2.2 illustre comment décrire un allongement positif et symétrique au moyen des lois h . Ces deux lois sont ensuite réunies pour former les lois g -et- h , définies à la section 2.3. La section 2.4 montre comment ces techniques peuvent être généralisées pour décrire des structures d'asymétrie et d'allongement plus complexes. Enfin, une discussion sur une famille de lois ne pouvant être complète sans information sur ses moments, la section 2.5 aborde ce sujet.

2.1 Asymétrie et lois g

La méthode présentée ici pour décrire une variable aléatoire continue asymétrique X est basée sur l'expression de cette dernière comme une fonction monotone d'une variable normale centrée réduite Z . Étant intéressés par la forme de la loi de X , nous devons d'abord tenir compte de sa localisation et de son échelle. Pour ce faire, écrivons

$$X = A + BY,$$

où A et B sont des scalaires et Y est une variable aléatoire « standard » de même forme que X . Dans cette standardisation, la médiane de Y est fixée à 0, de sorte que A est la médiane de X .

Écrivons Y comme une fonction de Z , $Y = Y(Z)$. L'asymétrie de Y peut être considérée comme produite par une fonction de remodelage G qui affecte les valeurs positives de Z différemment des valeurs négatives. Pour ce faire, posons

$$Y(Z) = G(Z) \times Z. \quad (2.1)$$

Outre le cas particulier $G(z) \equiv 1$, nous considérons habituellement G tel que $G(-z) \neq G(z)$ pour tout $z \neq 0$. De plus, l'asymétrie étant généralement plus apparente lorsque nous nous éloignons de la médiane, l'effet de remodelage de G devrait être plus léger près de 0 ; il est possible d'avoir ceci en exigeant que $Y(z) \approx z$ près de 0. En outre, dans le but d'imiter le comportement de plusieurs ensembles de données et lois théoriques, Y devrait être non bornée dans au moins une direction.

2.1.1 Lois g

Si la fonction $Y(z)$ doit satisfaire les conditions $Y(0) = 0$ et $Y(z) \approx z$ pour z près de 0, son expansion en série doit être de la forme $Y(z) = z + \dots$. Une famille pratique de fonctions possédant ces propriétés est donnée par

$$Y_g(z) = \frac{e^{gz} - 1}{g},$$

où le scalaire $g \neq 0$ contrôle l'asymétrie. Il est possible de vérifier le terme dominant de $Y_g(z)$ en rappelant que

$$e^{gz} = 1 + gz + \frac{(gz)^2}{2!} + \frac{(gz)^3}{3!} + \dots$$

Ainsi,

$$Y_g(z) = \frac{e^{gz} - 1}{g} = z + g \frac{z^2}{2!} + g^2 \frac{z^3}{3!} + \dots \quad (2.2)$$

Cette forme de $Y_g(z)$ revient à utiliser

$$G_g(z) = \frac{e^{gz} - 1}{gz} \quad (2.3)$$

comme multiplicateur de Z dans le membre de droite de l'équation (2.1). Les membres de cette famille de lois asymétriques sont dits de « loi g ».

Définition : Soit Z une variable aléatoire normale centrée réduite et g un scalaire. La variable aléatoire $Y_g(Z)$ donnée par

$$Y_g(Z) = G_g(Z) \times Z = \frac{e^{gZ} - 1}{gZ} \times Z$$

ou, de manière équivalente, par

$$Y_g(Z) = \frac{e^{gZ} - 1}{g} \quad (2.4)$$

est dite de loi g . Le paramètre g contrôle l'importance et la direction de l'asymétrie.

Lorsque $g \rightarrow 0$, il suit de l'équation (2.2) que $Y_g(Z) = Z$; ainsi, par extension, le cas $g = 0$ correspond à la loi normale, et donc à une absence d'asymétrie.

Comme $Y_g(z)$ est une fonction strictement croissante de z , il est facile d'obtenir les quantiles de la loi de $Y_g(Z)$ à partir des quantiles de la loi normale centrée réduite : si z_p est le p^e quantile de la loi normale centrée réduite, c'est-à-dire que $P(Z \leq z_p) = p$, alors $Y_g(z_p)$ est le p^e quantile de $Y_g(Z)$. Cette relation peut être utilisée pour examiner l'asymétrie correspondant à diverses valeurs de g .

Asymétrie pour diverses valeurs de g

Pour illustrer comment la variation de la valeur de g affecte l'asymétrie, le tableau 2.1 donne la valeur de $Y_g(z_p)$ pour diverses valeurs de p et pour $g = 0.2, 0.4, 0.6, 0.8$ et 1.0 .

p	z_p	$Y_g(z_p)$				
		$g = 0.2$	$g = 0.4$	$g = 0.6$	$g = 0.8$	$g = 1.0$
1/128	-2.4176	-1.917	-1.549	-1.276	-1.069	-0.911
1/64	-2.1539	-1.750	-1.444	-1.209	-1.027	-0.884
1/32	-1.8627	-1.555	-1.313	-1.122	-0.968	-0.845
1/16	-1.5341	-1.321	-1.147	-1.003	-0.884	-0.784
1/8	-1.1503	-1.028	-0.922	-0.831	-0.752	-0.683
1/4	-0.6745	-0.631	-0.591	-0.555	-0.521	-0.491
1/2	0	0	0	0	0	0
3/4	0.6745	0.722	0.774	0.831	0.894	0.963
7/8	1.1503	1.293	1.461	1.657	1.887	2.159
15/16	1.5341	1.795	2.118	2.517	3.015	3.637
31/32	1.8627	2.257	2.767	3.429	4.297	5.441
63/64	2.1539	2.692	3.417	4.402	5.752	7.618
127/128	2.4176	3.109	4.075	5.443	7.397	10.219

TAB. 2.1 – Valeurs de $Y_g(z_p)$ pour diverses valeurs de p et de g .

Un aperçu graphique de ces lois est donné à la figure 2.1, qui illustre les densités de probabilité de la loi g pour $g = 0, 0.2, 0.4, 0.6, 0.8$ et 1.0 . Cette figure permet de constater que les lois g sont plus asymétriques pour les grandes valeurs de g .

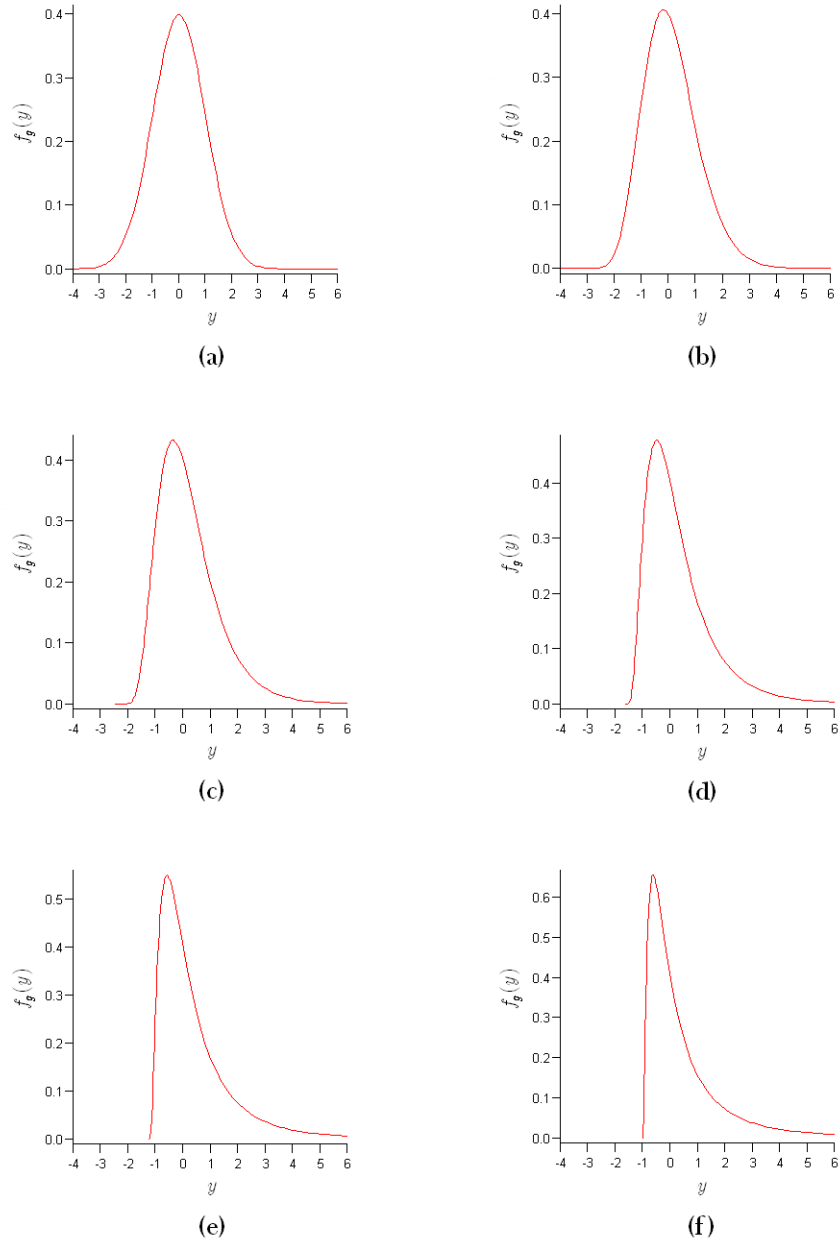


FIG. 2.1 – Graphique de la densité de probabilité $f_g(y)$ pour six valeurs de g . (a) $g = 0$, (b) $g = 0.2$, (c) $g = 0.4$, (d) $g = 0.6$, (e) $g = 0.8$, (f) $g = 1.0$.

2.1.2 Choix d'une valeur appropriée pour g

Lorsque nous travaillons avec des données, nous avons besoin de choisir une valeur appropriée pour g . La forme simple de Y_g rend ceci possible directement à partir des quantiles. L'idée de base est de trouver la valeur de g qui reproduit le plus fidèlement possible la médiane ainsi que les p^e et $(1-p)^e$ quantiles des données.

Supposons d'abord que la variable X est exactement de la forme $X = A + BY$, où $Y = (e^{gZ} - 1)/g$. Les p^e et $(1-p)^e$ quantiles de X vérifient alors les équations suivantes :

$$x_p = A + By_p = A + B \left(\frac{e^{gz_p} - 1}{g} \right)$$

et

$$x_{1-p} = A + By_{1-p} = A + B \left(\frac{e^{-gz_p} - 1}{g} \right)$$

pour $p \in (0, 0.5)$ et où nous avons utilisé le fait que $z_{1-p} = -z_p$. De plus, comme la médiane de Y , $y_{0.5}$, a été fixée à 0, nous avons aussi

$$x_{0.5} = A + By_{0.5} = A.$$

En utilisant ces résultats, nous obtenons

$$e^{-gz_p} = \frac{x_{1-p} - x_{0.5}}{x_{0.5} - x_p}. \quad (2.5)$$

Par conséquent, lorsque la supposition faite précédemment est valide, nous trouvons exactement

$$g = -\frac{1}{z_p} \ln \left(\frac{x_{1-p} - x_{0.5}}{x_{0.5} - x_p} \right), \quad (2.6)$$

où $z_p < 0$ et où g ne dépend pas de p .

Autrement dit, le membre de droite de l'équation (2.6) ne varie pas en fonction de p . Cette équation nous fournit donc une interprétation pour le paramètre g : celui-ci mesure l'asymétrie en termes du logarithme des distances des $(1-p)^e$ et p^e quantiles à la médiane. Dans ce qui suit, nous utiliserons la terminologie « demi-étendue » (« *half-spread* ») pour référer à la distance (positive) entre un quantile et la médiane. La p^e demi-étendue inférieure (« *lower half-spread* ») et la p^e demi-étendue supérieure (« *upper half-spread* ») sont données par $LHS_p = x_{0.5} - x_p$ et $UHS_p = x_{1-p} - x_{0.5}$ respectivement.

Dans le cas où l'hypothèse $X = A + BY$ n'est pas exactement vraie, le calcul (2.6) peut être effectué pour divers choix de p . Nous obtenons alors différentes valeurs de g , une pour chaque p . Si les valeurs de g varient peu en fonction de p , alors la loi de X est bien approximée par une loi de type g .

Mentionnons enfin que nous verrons à la section 2.4 comment tenir compte de l'asymétrie lorsque les valeurs de g varient de façon importante en fonction de p .

Exemple : Loi du khi-deux à six degrés de liberté

Supposons que l'on cherche à approximer une loi du χ^2 à six degrés de liberté, dont la densité de probabilité est représentée à la figure 2.2, par une loi g .

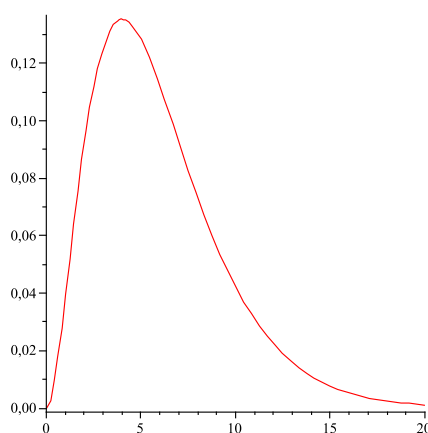


FIG. 2.2 – Densité de probabilité de la loi du χ^2 à six degrés de liberté.

Les deuxième et troisième colonnes du tableau 2.2 donnent les quantiles de la χ_6^2 pour certaines valeurs de p , et les colonnes de droite présentent les résultats des étapes successives du calcul. Par exemple, pour $p = 0.25$, les demi-étendues sont $5.348 - 3.455 = 1.893$ et $7.841 - 5.348 = 2.493$, leur ratio [comme à l'équation (2.5)] est $2.493/1.893 = 1.317$, dont le logarithme naturel est 0.2753. D'une table de la loi normale centrée réduite, $z_{0.25} = -0.6745$, de sorte que l'équation (2.6) conduit à $g_{0.25} = 0.2753/0.6745 = 0.4082$.

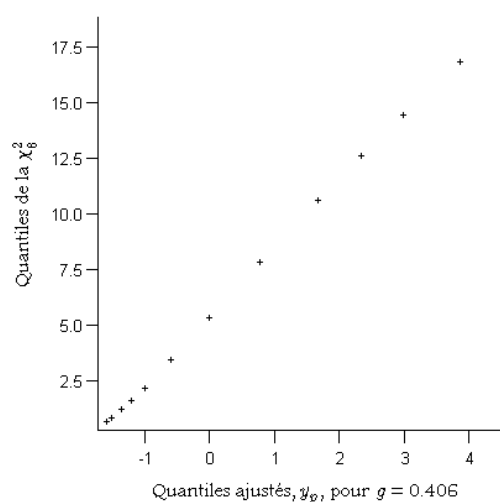
Bien que les valeurs de g_p (0.4082, 0.4070, 0.4063, 0.4055, 0.4043, 0.4032) décroissent régulièrement lorsque les valeurs de p deviennent plus extrêmes, seule la troisième décimale de g_p est affectée. Ainsi, une valeur constante de g est une bonne approximation pour décrire la structure d'asymétrie de la χ_6^2 et une valeur appropriée est 0.406, la médiane des six valeurs de g_p du tableau 2.2.

En optant pour une valeur de g , nous avons établi la forme correspondant à l'asymétrie de la χ_6^2 , mais nous n'avons fixé ni sa localisation, ni son échelle. Nous pouvons déterminer ces dernières et, par la même occasion, examiner grossièrement l'adéquation

p	Quantile		Demi-étendue		Logarithme du ratio	z_{1-p}	g_p
	Inférieur	Supérieur	Inférieure	Supérieure			
0.5	5.348						
0.25	3.455	7.841	1.893	2.493	0.2753	0.6745	0.4082
0.10	2.204	10.645	3.144	5.297	0.5216	1.2816	0.4070
0.05	1.635	12.592	3.713	7.244	0.6683	1.6449	0.4063
0.025	1.237	14.449	4.111	9.101	0.7947	1.9600	0.4055
0.01	0.872	16.812	4.476	11.464	0.9405	2.3263	0.4043
0.005	0.676	18.548	4.672	13.200	1.0386	2.5758	0.4032

TAB. 2.2 – Calculs pour déterminer g_p pour la loi du χ^2 à six degrés de liberté.

de la valeur de g choisie en utilisant un diagramme quantile-quantile; nous plaçons en ordonnée les quantiles disponibles de la χ_6^2 (5^e colonne du tableau 2.3) et en abscisse les quantiles correspondants de Y_g (3^e colonne du tableau 2.3), comme à la figure 2.3.

FIG. 2.3 – Diagramme quantile-quantile de la χ_6^2 et de la loi g ajustée ($g = 0.406$).

Le fait que la médiane soit 5.348 nous donne $A = 5.348$. L'ajustement d'un modèle de régression linéaire simple, avec une ordonnée à l'origine de 5.348, les quantiles de la χ_6^2 comme variable réponse et les quantiles ajustés y_p comme variable explicative, donne une pente de 2.985. Notre approximation de la χ_6^2 est donc

$$5.348 + 2.985 \left(\frac{e^{0.406Z} - 1}{0.406} \right). \quad (2.7)$$

Le tableau 2.3 présente un résumé des résultats obtenus et permet de comparer les quantiles de la χ_6^2 aux quantiles calculés au moyen de l'approximation (2.7). Cette

dernière peut également est illustrée à l'aide de la figure 2.4, qui représente les densités de probabilité de la χ_6^2 et de la loi g définie en (2.7). Cette figure permet de constater qu'à l'exception du centre de la loi, l'ajustement est très satisfaisant. (Le manque d'ajustement au niveau de la partie centrale s'explique par le fait que la valeur de g a été choisie à partir de quantiles provenant davantage des queues que du centre de la loi.)

p	z_p	y_p ($g = 0.406$)	$5.348 + 2.985y_p$	Quantile de la χ_6^2	Résidu
0.005	-2.5758	-1.597	0.581	0.676	0.095
0.01	-2.3263	-1.505	0.856	0.872	0.016
0.025	-1.9600	-1.352	1.312	1.237	-0.075
0.05	-1.6449	-1.200	1.766	1.635	-0.131
0.1	-1.2816	-0.999	2.366	2.204	-0.162
0.25	-0.6745	-0.590	3.587	3.455	-0.132
0.5	0	0	5.348	5.348	0
0.75	0.6745	0.776	7.664	7.841	0.177
0.9	1.2816	1.681	10.366	10.645	0.279
0.95	1.6449	2.340	12.333	12.592	0.259
0.975	1.9600	2.995	14.288	14.449	0.161
0.99	2.3263	3.871	16.903	16.812	-0.091
0.995	2.5758	4.546	18.918	18.548	-0.370

TAB. 2.3 – Résumé des résultats obtenus, quantiles approximés et résidus.

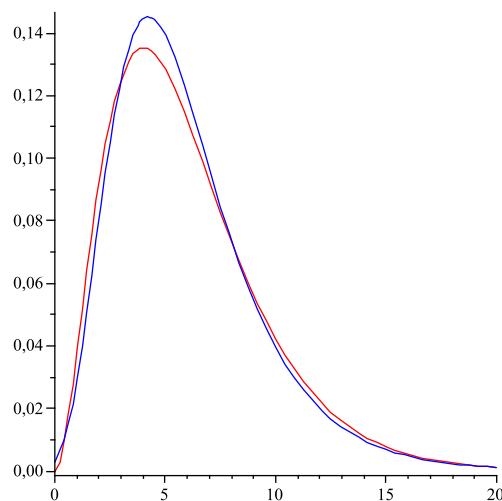


FIG. 2.4 – Densités de probabilité de la χ_6^2 (en rouge) et de la loi g définie en (2.7) (en bleu).

2.1.3 Lois log-normales

Un coup d'œil plus attentif à la définition

$$Y_g = \frac{e^{gZ} - 1}{g}$$

de l'équation (2.4) révèle que les lois g , avec g constant et positif, sont en fait des lois log-normales. Une façon pratique d'écrire une variable aléatoire X de la famille log-normale (Johnson *et coll.*, 1994, chapitre 14) est

$$X = \theta + e^{(Z-\gamma)/\delta}.$$

Si nous écrivons $X = A + BY = A + B(e^{gZ} - 1)/g$ (avec $B > 0$) et égalons les deux expressions pour X , nous obtenons les relations

$$\begin{aligned}\theta &= A - \frac{B}{g}, \\ \delta &= \frac{1}{g}, \\ \gamma &= -\frac{1}{g} \ln\left(\frac{B}{g}\right).\end{aligned}$$

Dans la dernière de ces relations, nous voyons pourquoi g doit être positif : l'argument du logarithme doit être positif. Ainsi, contrairement aux lois g , la famille des lois log-normales permet seulement une asymétrie positive. Mentionnons toutefois qu'une asymétrie négative peut facilement être obtenue à partir d'une loi log-normale en utilisant l'image miroir de cette dernière.

Sous réserve de la condition sur g , les relations entre (θ, δ, γ) et (A, B, g) montrent comment passer d'un ensemble de paramètres à l'autre. Ainsi, toute loi log-normale peut s'écrire comme une loi g , avec g positif, et vice versa. L'avantage des lois g réside cependant dans le fait qu'elles permettent de manipuler les asymétries négatives aussi facilement que les asymétries positives.

2.2 Allongement et lois h

Pour modéliser des lois symétriques ayant des queues plus lourdes que celles de la loi normale, nous pouvons remodeler une variable aléatoire normale centrée réduite en déployant ses ailes. L'idée de base est d'écrire

$$Y = H(Z) \times Z$$

et de choisir une fonction H qui permette d'étirer les queues, tout en préservant la symétrie. Ceci signifie que H doit être une fonction paire et positive, c'est-à-dire que nous devons avoir $H(-z) = H(z)$ et $0 < H(z) < \infty$ pour tout z fini. De plus, pour réaliser l'allongement, H doit être croissante pour $z \geq 0$.

2.2.1 Lois h

Une famille simple de fonctions ayant le comportement souhaité est donnée par

$$H_h(z) = e^{hz^2/2}, \quad (2.8)$$

de sorte que $Y = Ze^{hZ^2/2}$. Les membres de cette famille de lois sont dits de « loi h ».

Définition : Soit Z une variable aléatoire normale centrée réduite et h un scalaire. La variable aléatoire $Y_h(Z)$ donnée par

$$Y_h(Z) = Ze^{hZ^2/2} \quad (2.9)$$

est dite de loi h . Le paramètre h contrôle l'importance de l'allongement.

Lorsque $h = 0$, il suit de (2.9) que $Y_h(Z) = Z$; ainsi, $h = 0$ correspond à la loi normale, et donc à une absence d'allongement. En outre, les valeurs positives de h produisent un allongement positif qui augmente à mesure que h croît. [Une valeur négative de h n'est pas impossible, mais un traitement spécial peut être nécessaire car $Y_h(z)$ n'est plus monotone pour $z^2 > -1/h$.]

Dans le cas où $h \geq 0$, $Y_h(z)$ est une fonction strictement croissante de z et les quantiles de la loi de $Y_h(Z)$ peuvent facilement être obtenus à partir des quantiles de la loi normale centrée réduite. En effet, si z_p est le p^e quantile de la loi normale centrée réduite, alors $Y_h(z_p)$ est le p^e quantile de $Y_h(Z)$. En d'autres termes, si y_p dénote le p^e quantile de $Y_h(Z)$, alors

$$y_p = z_p e^{hz_p^2/2}. \quad (2.10)$$

Nous utilisons maintenant cette relation pour examiner l'allongement correspondant à diverses valeurs de h .

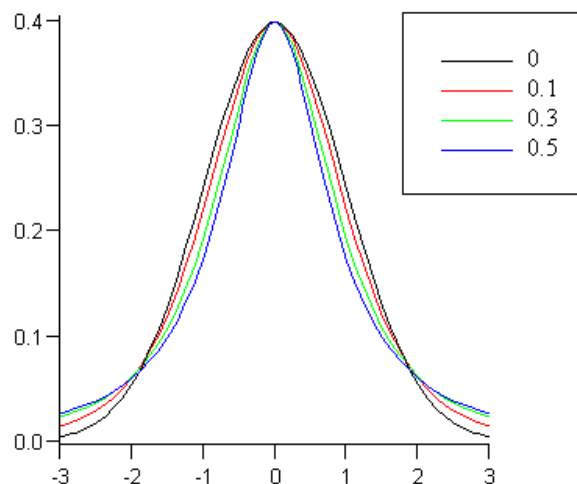
Allongement pour diverses valeurs de h

Pour illustrer comment la variation de la valeur de h affecte l'allongement, le tableau 2.4 donne la valeur de $Y_h(z_p)$, calculée à partir de l'équation (2.9), pour diverses valeurs de p et pour $h = 0.1, 0.2, 0.3, 0.4$ et 0.5 .

p	z_p	$Y_h(z_p)$				
		$h = 0.1$	$h = 0.2$	$h = 0.3$	$h = 0.4$	$h = 0.5$
3/4	0.6745	0.690	0.706	0.722	0.739	0.756
7/8	1.1503	1.229	1.313	1.403	1.499	1.601
15/16	1.5341	1.726	1.941	2.184	2.456	2.763
31/32	1.8627	2.216	2.635	3.135	3.728	4.435
63/64	2.1539	2.716	3.425	4.320	5.447	6.870
127/128	2.4176	3.238	4.337	5.809	7.781	10.423

TAB. 2.4 – Valeurs de $Y_h(z_p)$ pour diverses valeurs de p et de h .

Un aperçu graphique de certaines de ces lois est donné à la figure 2.5, qui illustre les densités de probabilité de la loi h pour $h = 0.1, 0.3$ et 0.5 , de même que pour $h = 0$ (qui correspond à la loi normale centrée réduite).

FIG. 2.5 – Graphique de la densité de probabilité de la loi h pour $h = 0, 0.1, 0.3$ et 0.5 .

En comparant les cinq dernières colonnes du tableau 2.4 à la deuxième colonne, nous pouvons voir l'importance de l'allongement. De plus, comme la figure 2.5 permet aussi de le constater, l'allongement augmente avec h .

2.2.2 Choix d'une valeur appropriée pour h

Supposons d'abord que la variable X est exactement de la forme $X = A + BY$, où A et B sont des paramètres de localisation et d'échelle et où $Y = Ze^{hZ^2/2}$.

Si A et B sont connus, nous pouvons travailler directement avec Y et trouver la valeur de h qui reproduit aussi fidèlement que possible le p^e quantile, y_p , de cette variable. En effet, lorsque la supposition précédente est valide, l'équation (2.10) nous permet de trouver exactement, pour $0 < p < 0.5$ ou $0.5 < p < 1$,

$$h = \frac{2 \ln(y_p/z_p)}{z_p^2}, \quad (2.11)$$

où h ne dépend pas de p .

Dans le cas où A et B sont connus, mais que l'hypothèse $X = A + BY$ n'est pas exactement vraie, le calcul (2.11) peut être effectué pour divers choix de p . Nous obtenons alors une valeur différente de h pour chaque p . Si ces valeurs varient peu en fonction de p , alors la loi de X est bien approximée par une loi de type h . (Nous verrons à la section 2.4 comment tenir compte de l'allongement si les valeurs de h varient de façon importante en fonction de p .)

Lorsque A et B ne sont pas connus, la symétrie nous permet d'éliminer A en utilisant des différences de quantiles symétriques. En effet, pour $0 < p < 0.5$, nous avons

$$x_p = A + Bz_p e^{hz_p^2/2} \quad (2.12)$$

et, en se rappelant que $z_p < 0$ et que $z_{1-p} = -z_p$,

$$x_{1-p} = A - Bz_p e^{hz_p^2/2}, \quad (2.13)$$

de sorte que

$$x_{1-p} - x_p = -2Bz_p e^{hz_p^2/2}. \quad (2.14)$$

En posant

$$(\text{pseudosigma})_p = \psi_p = \frac{x_{1-p} - x_p}{z_{1-p} - z_p} = \frac{x_{1-p} - x_p}{-2z_p}$$

et en réarrangeant (2.14), nous obtenons

$$\psi_p = \frac{x_{1-p} - x_p}{-2z_p} = B e^{hz_p^2/2}.$$

Ainsi, si l'hypothèse $X = A + BY$ est valide, un graphique de $\ln(\psi_p)$ en fonction de $z_p^2/2$, pour un ensemble de valeurs de p , sera linéaire avec ordonnée à l'origine $\ln B$ et pente h .

Dans le cas où cette hypothèse n'est pas exactement vraie mais qu'une loi de type h peut être ajustée de façon satisfaisante à la loi de X , un graphique de $\ln(\psi_p)$ en fonction de $z_p^2/2$, pour un ensemble de valeurs de p , sera approximativement linéaire avec ordonnée à l'origine $\ln B$ et pente h .

Exemple : Loi de Cauchy centrée réduite

Supposons que l'on cherche à utiliser une loi h pour approximer la loi de Cauchy centrée réduite, dont la densité de probabilité est donnée par

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty,$$

et est représentée à la figure 2.6, qui illustre également la densité de probabilité de la loi normale centrée réduite.

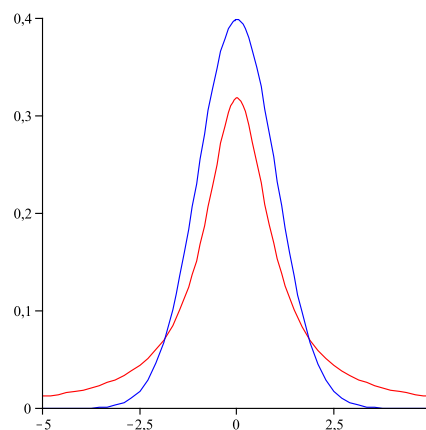


FIG. 2.6 – Densités de probabilité de la Cauchy centrée réduite (en rouge) et de la normale centrée réduite (en bleu).

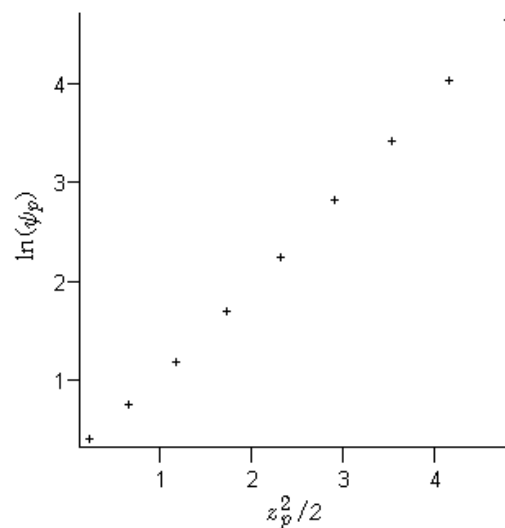
La loi de Cauchy centrée réduite est symétrique par rapport à 0, de sorte que $A = 0$ dans les équations (2.12) et (2.13). Nous avons cependant besoin d'une valeur appropriée de B .

Le tableau 2.5 donne les quantiles de la Cauchy pour certaines valeurs de p , ainsi que les valeurs de ψ_p , de $\ln(\psi_p)$ et de $z_p^2/2$. Par exemple, pour $p = 3/4$, nous avons $\psi_{3/4} = 1.0000/0.6745 = 1.483$, dont le logarithme naturel est 0.394. La figure 2.7 montre le graphique de $\ln(\psi_p)$ en fonction de $z_p^2/2$.

Un modèle de régression linéaire simple, avec les valeurs de $\ln(\psi_p)$ (5^e colonne du tableau 2.5) comme variable réponse et les valeurs de $z_p^2/2$ (6^e colonne du tableau 2.5) comme variable explicative, donne une pente de 0.939 et une ordonnée à l'origine de 0.109. Ainsi, pour les valeurs de p sélectionnées, la constante $h = 0.939$ résume de façon satisfaisante l'allongement de la loi de Cauchy. De plus, nous pouvons prendre $\ln B = 0.109$, c'est-à-dire $B = 1.115$, obtenant ainsi l'approximation

$$\text{Cauchy centrée réduite} \approx 1.115Z e^{0.939Z^2/2}. \quad (2.15)$$

p	Quantile de la Cauchy	z_p	ψ_p	$\ln(\psi_p)$	$z_p^2/2$
3/4	1.0000	0.6745	1.483	0.394	0.227
7/8	2.4142	1.1503	2.099	0.741	0.662
15/16	5.0273	1.5341	3.277	1.187	1.177
31/32	10.1532	1.8627	5.451	1.696	1.735
63/64	20.3555	2.1539	9.451	2.246	2.320
127/128	40.7355	2.4176	16.850	2.824	2.922
255/256	81.4832	2.6601	30.632	3.422	3.538
511/512	162.9726	2.8856	56.478	4.034	4.163
1023/1024	325.9483	3.0973	105.236	4.656	4.797

TAB. 2.5 – Calcul de $\ln(\psi_p)$ pour la loi de Cauchy centrée réduite.FIG. 2.7 – Graphique de $\ln(\psi_p)$ en fonction de $z_p^2/2$ pour la loi de Cauchy centrée réduite.

Cette approximation peut être illustrée au moyen de la figure 2.8, qui représente les densités de probabilité de la Cauchy centrée réduite et de la loi h définie en (2.15). Cette figure permet de constater qu'à l'exception du centre de la loi, l'ajustement est très satisfaisant. (Le manque d'ajustement au niveau de la partie centrale s'explique par le fait que la valeur de h a été choisie à partir de quantiles provenant davantage des queues que du centre de la loi.)

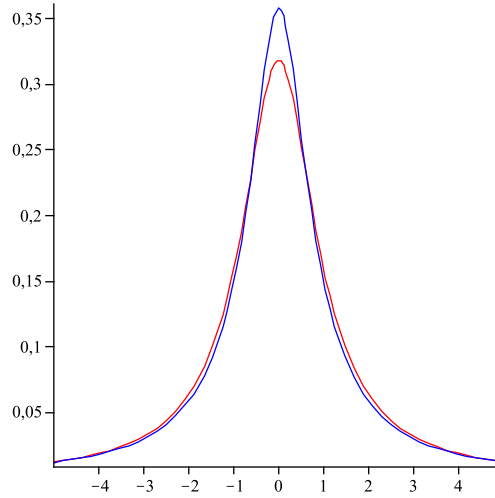


FIG. 2.8 – Densités de probabilité de la Cauchy centrée réduite (en rouge) et de la loi h définie en (2.15) (en bleu).

Mentionnons finalement qu'en général, nous pouvons vérifier l'adéquation de l'approximation obtenue en comparant les quantiles théoriques aux quantiles approximatés, comme nous l'avons fait au tableau 2.3 par exemple.

2.3 Lois g -et- h

Comme il arrive souvent qu'un ensemble de données soit à la fois asymétrique et allongé, nous devons être en mesure de traiter ces deux aspects simultanément. Dans les deux sections précédentes, nous avons pris en compte l'asymétrie et l'allongement en transformant une variable aléatoire normale centrée réduite Z , que nous avons multipliée par une fonction $G(Z)$ pour l'asymétrie et par une fonction $H(Z)$ pour l'allongement. Pour combiner ces deux aspects, nous utilisons de nouveau la multiplication, ce qui donne

$$Y = G(Z)H(Z)Z.$$

Pour le choix particulier des fonctions G et H des équations (2.3) et (2.8), nous obtenons

$$Y_{g,h}(Z) = Z \left(\frac{e^{gZ} - 1}{gZ} \right) e^{hZ^2/2}$$

ou, de manière équivalente,

$$Y_{g,h}(Z) = \left(\frac{e^{gZ} - 1}{g} \right) e^{hZ^2/2}. \quad (2.16)$$

Comme auparavant, nous tenons compte de la localisation et de l'échelle en posant

$$X = A + BY,$$

où A est la médiane de X et B est un paramètre d'échelle.

Les quantiles de $Y_{g,h}(Z)$ peuvent être facilement calculés dans le cas où $h \geq 0$. En effet, $Y_{g,h}$ est alors une fonction monotone croissante de Z et, par conséquent, une transformation bijective. Ainsi, le p^e quantile de la loi de $Y_{g,h}(Z)$ est simplement $Y_{g,h}(z_p)$, où z_p est le p^e quantile de la loi normale centrée réduite.

2.3.1 Choix de valeurs appropriées pour g et h

Un avantage de la forme multiplicative de la fonction (2.16) devient évident lorsque nous tentons de trouver une valeur appropriée pour g . Si, comme à l'équation (2.6) de la section 2.1, nous calculons g_p , la valeur de g qui correspond exactement à x_p et x_{1-p} , nous trouvons

$$\begin{aligned} \frac{x_{1-p} - x_{0.5}}{x_{0.5} - x_p} &= \frac{\left(A + B \frac{e^{-gz_p} - 1}{g} e^{hz_p^2/2}\right) - A}{A - \left(B \frac{e^{gz_p} - 1}{g} e^{hz_p^2/2} + A\right)} \\ &= e^{-gz_p}, \end{aligned}$$

où $0 < p < 0.5$. Ainsi, nous pouvons déterminer une valeur de g pour chaque $p \in (0, 1/2)$ sans avoir à connaître h . Tout comme à la section 2.1, nous avons

$$g_p = -\frac{1}{z_p} \ln \left(\frac{x_{1-p} - x_{0.5}}{x_{0.5} - x_p} \right)$$

et, si cela est possible, nous résumons l'asymétrie en utilisant une valeur constante de g .

Après avoir choisi une valeur pour g , nous pouvons tourner notre attention vers h . Considérons la demi-étendue supérieure, $\text{UHS}_p = x_{1-p} - x_{0.5}$, que nous pouvons écrire sous la forme

$$x_{1-p} - x_{0.5} = \frac{B}{g} (e^{-gz_p} - 1) e^{hz_p^2/2}.$$

En divisant cette dernière expression par $(e^{-gz_p} - 1)/g$, nous tenons compte de l'asymétrie (du moins lorsque g ne varie pas trop en fonction de p), ce qui nous permet alors de nous concentrer sur l'allongement. Ainsi, nous pouvons utiliser

$$\text{UHS}_p^* = \frac{g(x_{1-p} - x_{0.5})}{e^{-gz_p} - 1} = B e^{hz_p^2/2}$$

de la même façon que nous avons utilisé $x_{1-p} - x_{0.5}$ à la section 2.2 : un graphique de $\ln(\text{UHS}_p^*)$ en fonction de $z_p^2/2$ aura une ordonnée à l'origine $\ln B$ et une pente h .

Lorsque nous travaillons avec des données, et donc nécessairement avec des valeurs approximatives de g , il peut être utile d'utiliser les demi-étendues inférieure ($\text{LHS}_p = x_{0.5} - x_p$) et supérieure ($\text{UHS}_p = x_{1-p} - x_{0.5}$). Ceci peut être fait soit en prenant la moyenne de $\ln(\text{UHS}_p^*)$ et de $\ln(\text{LHS}_p^*)$, où $\text{LHS}_p^* = g(x_{0.5} - x_p)/(1 - e^{gz_p})$, soit en utilisant le fait que

$$\frac{g(x_{1-p} - x_p)}{e^{-gz_p} - e^{gz_p}} = Be^{hz_p^2/2},$$

où l'étendue totale, $x_{1-p} - x_p$, est employée.

Exemple : Loi du khi-deux à six degrés de liberté

Pour illustrer la procédure décrite ci-dessus, nous poursuivons l'exemple de la χ_6^2 débuté à la section 2.1.2 (voir le tableau 2.2). Le tableau 2.6 montre les calculs permettant d'obtenir les valeurs de $\ln(\text{UHS}_p^*)$, alors que la figure 2.9 donne le graphique de $\ln(\text{UHS}_p^*)$ en fonction de $z_p^2/2$.

p	UHS_p	$-z_p$	$(e^{-gz_p} - 1)/g$	$\ln(\text{UHS}_p^*)$	$z_p^2/2$
0.25	2.493	0.6745	0.7759	1.1672	0.227
0.10	5.297	1.2816	1.6812	1.1476	0.821
0.05	7.244	1.6449	2.3399	1.1301	1.353
0.025	9.101	1.9600	2.9954	1.1113	1.921
0.01	11.464	2.3263	3.8706	1.0858	2.706
0.005	13.200	2.5758	4.5458	1.0660	3.317

TAB. 2.6 – Ajustement pour un g constant ($g = 0.406$) dans l'exemple de la χ_6^2 .

L'ajustement d'un modèle de régression linéaire simple, avec les valeurs de $\ln(\text{UHS}_p^*)$ (5^e colonne du tableau 2.6) comme variable réponse et les valeurs de $z_p^2/2$ (6^e colonne du tableau 2.6) comme variable explicative, donne une ordonnée à l'origine de 1.174 et une pente de -0.0328 ; nous pouvons donc prendre $\ln B = 1.174$, c'est-à-dire $B = 3.235$, et $h = -0.0328$. Bien que nous ayons mis l'accent, tout au long de ce chapitre, sur les valeurs positives de h , nous avons mentionné à la section 2.2 qu'une valeur négative de h n'est pas impossible. Dans le cas de la χ_6^2 , l'approximation basée sur $g = 0.406$ et $h = -0.0328$ sera assez bonne mais, comme h est négatif, cette approximation ne peut pas être complètement précise. En effet, comme nous l'avons mentionné à la section 2.2, $Y_h(z)$ n'est pas monotone pour $z^2 > -1/h$. Dans notre exemple, cela se traduit par

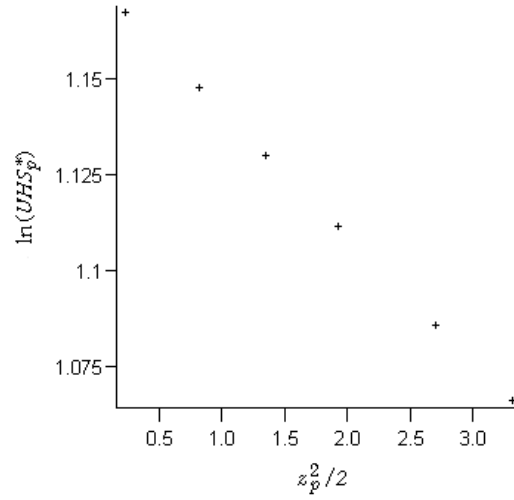


FIG. 2.9 – Graphique de $\ln(UHS_p^*)$ en fonction de $z_p^2/2$ pour l'exemple de la χ_6^2 .

la condition $z^2 > 30.49$ ou $|z| > 5.522$, une limitation qui n'est pas particulièrement sérieuse pour la plupart des utilisations, puisque $P\{|Z| > 5.522\} \approx 3.35 \times 10^{-8}$.

En rassemblant tous les éléments, nous obtenons

$$\chi_6^2 \approx 5.348 + 3.235 \left(\frac{e^{0.406Z} - 1}{0.406} \right) e^{-0.0328Z^2/2},$$

c'est-à-dire

$$\chi_6^2 \approx 5.348 + 7.968(e^{0.406Z} - 1)e^{-0.0164Z^2}. \quad (2.17)$$

Pour voir dans quelle mesure cette approximation est acceptable, considérons le tableau 2.7. À l'exception de $p = 0.005$ et de $p = 0.01$, les résidus sont tous inférieurs ou égaux à 0.01 en valeur absolue. Ainsi, l'approximation (2.17) est une façon assez précise de calculer les quantiles de la χ_6^2 à partir des quantiles de la loi normale centrée réduite. En outre, comme les valeurs de g_p ne sont pas tout à fait constantes et que nous avons choisi la valeur de h à partir de la demi-étendue supérieure de la loi en utilisant une valeur unique de g , il n'est pas très surprenant que l'approximation (2.17) soit légèrement moins précise à l'extrémité inférieure de la loi.

p	z_p	Quantile de la χ_6^2	Approximation	Résidu
0.005	-2.5758	0.676	0.713	-0.037
0.01	-2.3263	0.872	0.892	-0.020
0.025	-1.9600	1.237	1.242	-0.005
0.05	-1.6449	1.635	1.635	0
0.1	-1.2816	2.204	2.201	0.003
0.25	-0.6745	3.455	3.453	0.002
0.5	0	5.348	5.348	0
0.75	0.6745	7.841	7.839	0.002
0.9	1.2816	10.645	10.642	0.003
0.95	1.6449	12.592	12.589	0.003
0.975	1.9600	14.449	14.446	0.003
0.99	2.3263	16.812	16.806	0.006
0.995	2.5758	18.548	18.538	0.010

TAB. 2.7 – Exactitude de l'approximation $\chi_6^2 \approx 5.348 + 7.968(e^{0.406Z} - 1)e^{-0.0164Z^2}$.

2.3.2 Méthodes d'estimation des paramètres de la loi g -et- h basées sur la vraisemblance

La procédure décrite à la section 2.3.1 est une approche simple et pratique permettant de choisir des valeurs appropriées pour les paramètres A , B , g et h de la loi g -et- h . Il ne s'agit toutefois pas de l'unique façon de procéder. Dans cette section, nous présentons brièvement deux méthodes d'estimation basées sur la vraisemblance et discutées dans He (2005).

La densité de probabilité de la variable aléatoire

$$X = A + B \left(\frac{e^{gZ} - 1}{g} \right) e^{hZ^2/2} \quad (2.18)$$

de loi g -et- h est donnée par

$$\begin{aligned} f_X(x|A, B, g, h) &= f_Z(z|A, B, g, h) \left| \frac{dz}{dx} \right| \\ &= \frac{f_Z(z|A, B, g, h)}{\left| \frac{dx}{dz} \right|} \\ &= \frac{\frac{1}{(2\pi)^{1/2}} e^{-z^2/2}}{\left| B \left\{ e^{gz+hz^2/2} + hz \left(\frac{e^{gz}-1}{g} \right) e^{hz^2/2} \right\} \right|}. \end{aligned} \quad (2.19)$$

Cette densité n'est cependant pas une fonction explicite de x et doit par conséquent être évaluée numériquement. Le calcul consiste d'abord à résoudre l'équation (2.18) pour z , puis à substituer les solutions obtenues dans l'équation (2.19). Étant donné un échantillon indépendant et identiquement distribué x_1, \dots, x_n , la vraisemblance des données sous la loi g -et- h est

$$L_X(A, B, g, h|x) = \prod_{i=1}^n \frac{\frac{1}{(2\pi)^{1/2}} e^{-z_i^2/2}}{\left| B \left\{ e^{gz_i + hz_i^2/2} + hz_i \left(\frac{e^{gz_i} - 1}{g} \right) e^{hz_i^2/2} \right\} \right|}. \quad (2.20)$$

Lorsque $h > 0$, la vraisemblance (2.20) peut être maximisée en utilisant des procédures numériques. Par exemple, Rayner et MacGillivray (2002a) utilisent la méthode du simplexe de Nelder–Mead afin d'obtenir les estimateurs du maximum de vraisemblance pour deux généralisations de la loi g -et- h , les lois g -et- k et g -et- h généralisée, dont la définition est donnée au chapitre 5. Mentionnons toutefois qu'il peut être difficile d'obtenir les estimateurs du maximum de vraisemblance lorsque $h < 0$, x n'étant plus une fonction monotone de z dans ce cas.

He (2005) propose, pour sa part, une approche bayésienne pour l'estimation des paramètres A , B , g et h lorsque $h > 0$. Dans cette approche, des lois *a priori* doivent être précisées pour les paramètres de la loi g -et- h . Pour plus de simplicité, ces paramètres sont supposés *a priori* indépendants. La loi *a priori* choisie pour A et B est $p(A, B) \propto 1/B$, alors qu'une loi impropre est utilisée pour g et h , c'est-à-dire que $p(g, h) \propto 1$. Sous ces conditions, la loi *a priori* des paramètres de la loi g -et- h est $p(A, B, g, h) \propto 1/B$ et la loi *a posteriori* est donnée par

$$\begin{aligned} [A, B, g, h|x] &\propto L_X(A, B, g, h|x) p(A, B, g, h) \\ &\propto L_X(A, B, g, h|x)/B, \end{aligned} \quad (2.21)$$

où $[A, B, g, h|x]$ dénote la loi *a posteriori* des paramètres (A, B, g, h) étant donné x . À l'instar de la vraisemblance de l'équation (2.20), la densité *a posteriori* donnée par l'équation (2.21) n'est pas une fonction explicite de x et doit donc être évaluée numériquement. Pour ce faire, He (2005) propose l'utilisation de l'échantillonneur de Gibbs. Plus précisément, pour un échantillon indépendant et identiquement distribué de taille n , les étapes de l'échantillonneur de Gibbs à la t^e itération sont les suivantes :

1. Générer $A^{(t+1)}$ en simulant selon la loi

$$[A|B^{(t)}, g^{(t)}, h^{(t)}, x] \propto L_X(A, B^{(t)}, g^{(t)}, h^{(t)}|x).$$

2. Générer $B^{(t+1)}$ en simulant selon la loi

$$[B|A^{(t+1)}, g^{(t)}, h^{(t)}, x] \propto L_X(B, A^{(t+1)}, g^{(t)}, h^{(t)}|x)/B.$$

3. Générer $g^{(t+1)}$ en simulant selon la loi

$$[g|A^{(t+1)}, B^{(t+1)}, h^{(t)}, x] \propto L_X(g, A^{(t+1)}, B^{(t+1)}, h^{(t)}|x).$$

4. Générer $h^{(t+1)}$ en simulant selon la loi

$$[h|A^{(t+1)}, B^{(t+1)}, g^{(t+1)}, x] \propto L_X(h, A^{(t+1)}, B^{(t+1)}, g^{(t+1)}|x).$$

Chaque étape d'échantillonnage de la procédure précédente peut être exécutée au moyen de l'algorithme ARMS (« *adaptive rejection Metropolis sampling* »). Les moyenne/mode, variance et intervalles de crédibilité *a posteriori* des paramètres de la loi g -et- h peuvent être obtenus en utilisant les valeurs générées après convergence de la chaîne de Gibbs.

Remarquons que l'approche basée sur la vraisemblance, que ce soit par une procédure de maximisation numérique ou par l'algorithme de l'échantillonneur de Gibbs, peut nécessiter de nombreux calculs. De plus, ces deux procédures sont difficiles à mettre en œuvre lorsque $h < 0$. L'approche basée sur les quantiles, décrite à la section 2.3.1, ne pose quant à elle aucune difficulté lorsque $h < 0$, en plus d'être d'une très grande simplicité et de permettre, en général, un meilleur ajustement au niveau des queues de la loi. Comme nous le verrons à la section suivante, cette méthode peut également être généralisée pour décrire des structures d'asymétrie et d'allongement plus complexes.

2.4 Lois g -et- h avec paramètres g et h variables

Comme les sections précédentes l'ont montré, lorsque g_p et h_p sont à peu près constants en fonction de p , nous pouvons facilement travailler avec

$$Y = \frac{e^{gZ} - 1}{g} e^{hZ^2/2}.$$

Dans l'exemple de la χ_6^2 , cependant, la diminution régulière de g_p lorsque p décroît (voir le tableau 2.2) indique qu'une structure plus complexe pourrait être nécessaire. Il est possible d'obtenir cette plus grande généralité en permettant que g et h soient des fonctions de z , des polynômes en z^2 par exemple. Ainsi, nous pourrions envisager par exemple de prendre

$$g(z) = g_0 + g_2 z^2$$

et

$$h(z) = h_0 + h_2 z^2$$

et d'inclure, par la suite, des puissances plus élevées de z^2 si la situation le nécessite. Le fait d'exprimer g et h comme des fonctions de z^2 nous assure que H demeurera une fonction paire de z et que G continuera de ne pas être une fonction paire.

Pour déterminer g_0 et g_2 , nous commençons, comme nous l'avons fait aux sections 2.1 et 2.3, par calculer g_p . Un graphique de g_p en fonction de z_p^2 nous permettra ensuite de choisir des valeurs appropriées pour g_0 et g_2 . Les résidus, $g_p - g_0 - g_2 z_p^2$, pourront être utilisés pour évaluer la nécessité d'inclure des termes d'ordre supérieur.

Avant de pouvoir déterminer $h(z)$, nous devons prendre en compte l'asymétrie. L'idée est la même qu'à la section 2.3, où nous avons trouvé qu'un graphique de $\ln\{g(x_{1-p} - x_{0.5})/(e^{-gz_p} - 1)\}$ en fonction de $z_p^2/2$ devrait avoir une pente h et une ordonnée à l'origine $\ln B$. Cependant, nous devons maintenant permettre que $g(z)$ ne soit pas constant ; le diviseur devient ainsi $\{e^{-g(z_p)z_p} - 1\}/g(z_p)$ et le logarithme naturel de la demi-étendue supérieure ajustée,

$$\ln(\text{UHS}_p^*) = \ln \left\{ \frac{(x_{1-p} - x_{0.5})g(z_p)}{e^{-g(z_p)z_p} - 1} \right\},$$

devrait maintenant être égal à

$$\ln B + \frac{h_0}{2} z_p^2 + \frac{h_2}{2} z_p^4.$$

L'exemple suivant, tiré de Hoaglin (1985), illustrera les étapes permettant de déterminer $g(z)$ et $h(z)$.

Exemple : Revenus de ménage

Le Département d'urbanisme et du logement des États-Unis a réalisé une étude visant à mesurer les effets d'une aide financière directe (sous forme d'allocations de logement) aux familles afin de leur permettre de vivre dans un logement décent. Au début de l'enquête, les ménages participants ont déclaré leur revenu annuel. Le tableau 2.8 donne les valeurs lettrées¹ (« *letter values* ») correspondantes pour 994 ménages de Pittsburgh.

Le tableau 2.9 précise les demi-étendues inférieures et supérieures des revenus annuels rapportés par ces 994 ménages, ainsi que les calculs nécessaires pour décrire l'asymétrie. Comme les valeurs de g_p décroissent régulièrement et considérablement, une valeur constante de g ne fournirait pas une description très efficace de la structure d'asymétrie. Un graphique de g_p en fonction de z_p^2 (figure 2.10) révèle qu'une droite serait raisonnable, et l'ajustement d'un modèle de régression linéaire simple donne

$$g_p = 0.496 - 0.025z_p^2.$$

		Revenu du ménage (en dollars)	
$n = 994$		3480	
M	497.5		
F	249	2412	4944
E	125	1788	6443
D	63	1517	7284
C	32	1248	8350
B	16.5	963.5	8994
A	8.5	727.5	9754.5
Z	4.5	579	10210
Y	2.5	345	10675.5
	1	114	10874

TAB. 2.8 – Valeurs lettrées pour l'exemple des revenus de ménage.

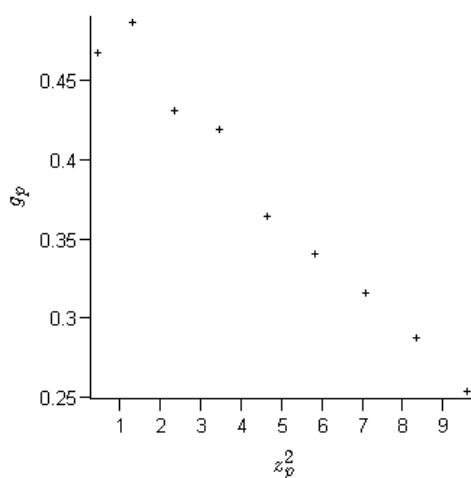
Étiquette	z_p	Demi-étendues		$g_p = -\frac{1}{z_p} \ln \left(\frac{\text{supérieure}}{\text{inférieure}} \right)$	z_p^2
		inférieure	supérieure		
F	-0.6745	1068	1464	0.468	0.455
E	-1.1503	1692	2963	0.487	1.323
D	-1.5341	1963	3804	0.431	2.353
C	-1.8627	2232	4870	0.419	3.470
B	-2.1539	2516.5	5514	0.364	4.639
A	-2.4176	2752.5	6274.5	0.341	5.845
Z	-2.6601	2901	6730	0.316	7.076
Y	-2.8856	3135	7195.5	0.288	8.327
X	-3.0973	3366	7394	0.254	9.593

TAB. 2.9 – Calculs pour décrire l'asymétrie dans l'exemple des revenus de ménage.

Pour voir si ces données sur les revenus de ménage comportent également un certain allongement, nous pouvons utiliser l'approche présentée au début de cette section afin de prendre en compte l'asymétrie. Le tableau 2.10 montre les étapes du calcul, et la figure 2.11 donne le graphique de $\ln(\text{UHS}_p^*)$ en fonction de z_p^2 .

Le graphique de $\ln(\text{UHS}_p^*)$ en fonction de z_p^2 (figure 2.11) peut suggérer une certaine courbure, mais celle-ci provient principalement du point pour les huitièmes (étiquette E), qui s'écartait déjà sensiblement de la droite de la figure 2.10. Ainsi, comme ce point ne s'adapte pas au modèle d'asymétrie utilisé, il continue de s'éloigner des autres points à la figure 2.11.

¹Une brève description de la notion de valeur lettrée ainsi qu'un exemple de calcul sont donnés à l'annexe A.


 FIG. 2.10 – Graphique de g_p en fonction de z_p^2 pour l'exemple des revenus de ménage.

Étiquette	UHS _p	$-z_p$	$g(z_p)$	$G^*(z_p) = \frac{e^{-g(z_p)z_p} - 1}{g(z_p)}$	$\ln\left(\frac{\text{UHS}_p}{G^*(z_p)}\right)$	z_p^2
F	1464	0.6745	0.485	0.798	7.515	0.455
E	2963	1.1503	0.463	1.519	7.576	1.323
D	3804	1.5341	0.437	2.185	7.462	2.353
C	4870	1.8627	0.409	2.793	7.464	3.470
B	5514	2.1539	0.380	3.334	7.411	4.639
A	6274.5	2.4176	0.350	3.802	7.409	5.845
Z	6730	2.6601	0.319	4.189	7.382	7.076
Y	7195.5	2.8856	0.288	4.499	7.377	8.327
X	7394	3.0973	0.256	4.726	7.355	9.593

 TAB. 2.10 – Ajustement des demi-étendues supérieures de l'exemple des revenus de ménage, $g(z) = 0.496 - 0.025z^2$.

L'ajustement d'une droite par la technique robuste décrite dans [Emerson et Hoaglin \(1985, section 7A\)](#) conduit à

$$7.522 - 0.0175z_p^2, \quad (2.22)$$

et les résidus correspondants ne présentent pas de courbure suffisante pour indiquer qu'un terme additionnel en z_p^4 serait nécessaire.

En interprétant les coefficients de (2.22) comme $\ln B = 7.522$ et $h_0/2 = -0.0175$, nous obtenons $B = 1848$ et $h_0 = -0.0350$. Ainsi, cet ensemble de données présente un léger allongement négatif. Le fait que la médiane soit égale à 3480 nous donne $A = 3480$

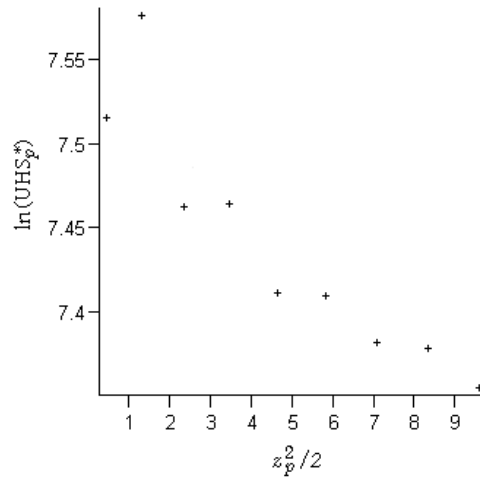


FIG. 2.11 – Graphique de $\ln(UHS_p^*)$ en fonction de z_p^2 pour l'exemple des revenus de ménage.

et complète la description en termes de la loi g -et- h ; le résultat est

$$3480 + 1848 \frac{e^{g(Z)Z} - 1}{g(Z)} e^{-0.0175Z^2}, \quad (2.23)$$

avec

$$g(Z) = 0.496 - 0.025Z^2. \quad (2.24)$$

Au lieu d'utiliser les demi-étendues supérieures pour déterminer $h(z)$ (comme nous l'avons fait au tableau 2.10), nous aurions pu utiliser les demi-étendues inférieures et appliquer un traitement similaire. Bien que nous ne rapportions pas ici les calculs et les graphiques, le résultat obtenu est essentiellement le même que précédemment.

Pour un examen global de la façon dont la loi obtenue s'ajuste aux données, nous pouvons comparer les valeurs lettrées observées à celles calculées à partir des équations (2.23) et (2.24). Le tableau 2.11 donne ces valeurs, ainsi que les résidus correspondants. À l'exception de trois valeurs lettrées dans la queue supérieure, l'ajustement est satisfaisant ; en fait, même ces résidus ne sont pas grands relativement aux valeurs observées.

Ainsi, en considérant g comme une simple fonction de z^2 et en travaillant avec les valeurs lettrées, nous avons réussi à réduire un échantillon de 994 revenus en une description de leur loi en termes de cinq constantes, explicitées aux équations (2.23) et (2.24). De cet ajustement, nous devrions être en mesure de récupérer de façon satisfaisante d'autres quantiles de la loi, particulièrement si ceux-ci n'exigent pas d'extrapolation au-delà des valeurs extrêmes observées.

Étiquette	Valeurs lettrées		Résidu
	Observée	Ajustée	
X	114	139	-25
Y	345	349	-4
Z	579	553	26
A	727.5	758	-31
B	963.5	974	-11
C	1248	1213	35
D	1517	1498	19
E	1788	1870	-82
F	2412	2425	-13
M	3480	3480	0
F	4944	4943	1
E	6443	6223	220
D	7284	7356	-72
C	8350	8338	12
B	8994	9162	-168
A	9754.5	9822	-68
Z	10210	10321	-111
Y	10675.5	10665	11
X	10874	10866	8

TAB. 2.11 – Valeurs lettrées observées et ajustées pour l'exemple des revenus de ménage.

2.5 Moments

Nous présentons dans cette section les moments de la loi g -et- h , ainsi que ses coefficients d'asymétrie et d'aplatissement. Avant de considérer le cas plus général où g et h sont tous deux non nuls, nous débutons avec les cas spéciaux plus simples, les lois g et les lois h . Mentionnons que tout au long de cette discussion, nous considérons uniquement des valeurs constantes de g et h . De plus, nous calculons les moments pour la variable aléatoire « standard » Y définie par (2.4), (2.9) et (2.16).

2.5.1 Lois g

Dans cette sous-famille, $Y = (e^{gZ} - 1)/g$. Les quatre premiers moments usuels sont

$$\begin{aligned} E(Y) &= \frac{e^{g^2/2} - 1}{g}, \\ \text{var}(Y) &= \frac{e^{g^2}(e^{g^2} - 1)}{g^2}, \end{aligned} \quad (2.25)$$

$$E[\{Y - E(Y)\}^3] = \frac{e^{(3/2)g^2}(e^{3g^2} - 3e^{g^2} + 2)}{g^3}, \quad (2.26)$$

$$E[\{Y - E(Y)\}^4] = \frac{e^{2g^2}(e^{6g^2} - 4e^{3g^2} + 6e^{g^2} - 3)}{g^4}. \quad (2.27)$$

Nous verrons à la section 2.5.3 un résultat général permettant d'obtenir les quatre moments précédents.

2.5.2 Lois h

Dans cette sous-famille, $Y = Ze^{hZ^2/2}$. [Martinez et Iglewicz \(1984\)](#) ont obtenu le moment d'ordre n par rapport à l'origine, valide pour $h < 1/n$:

$$E(Y^n) = \begin{cases} 0 & \text{pour } n \text{ impair,} \\ \frac{n!}{2^{n/2} (\frac{n}{2})!} \frac{1}{(1 - nh)^{\frac{n+1}{2}}} & \text{pour } n \text{ pair.} \end{cases} \quad (2.28)$$

(La démonstration de ce résultat est donnée à la section B.1 de l'annexe B.) En particulier, les deuxième et quatrième moments sont donnés par

$$E(Y^2) = \frac{1}{(1 - 2h)^{3/2}} \quad \text{pour } h < 1/2, \quad (2.29)$$

$$E(Y^4) = \frac{3}{(1 - 4h)^{5/2}} \quad \text{pour } h < 1/4. \quad (2.30)$$

Ainsi, la variance et le quatrième moment de Y ne sont finis que lorsque $h < 1/2$ et $h < 1/4$ respectivement.

2.5.3 Lois g -et- h

La loi g -et- h de paramètres g et h constants est la distribution de la variable

$$Y = \frac{e^{gZ} - 1}{g} e^{hZ^2/2}.$$

Martinez et Iglewicz (1984) ont montré que, lorsque $g \neq 0$ et $h < 1/n$, le moment d'ordre n est donné par

$$E(Y^n) = \frac{1}{g^n \sqrt{1 - nh}} \sum_{i=0}^n (-1)^i \binom{n}{i} e^{\{(n-i)g\}^2 / \{2(1-nh)\}}. \quad (2.31)$$

(La démonstration de ce résultat est donnée à la section B.2 de l'annexe B.) En particulier, l'espérance et la variance de Y , obtenues à partir de l'équation (2.31), sont données par

$$\begin{aligned} E(Y) &= \frac{1}{g\sqrt{1-h}} \left[e^{g^2/\{2(1-h)\}} - 1 \right] \quad \text{pour } h < 1, \\ \text{var}(Y) &= \frac{1}{g^2\sqrt{1-2h}} \left[e^{2g^2/(1-2h)} - 2e^{g^2/\{2(1-2h)\}} + 1 \right] \\ &\quad - \frac{1}{g^2(1-h)} \left[e^{g^2/\{2(1-h)\}} - 1 \right]^2 \quad \text{pour } h < 1/2. \end{aligned}$$

Compte tenu de la complexité des calculs, les moments centrés d'ordre plus élevé sont moins attrayants. Il est cependant possible de les obtenir numériquement pour des valeurs particulières de g et h en utilisant l'équation (2.31) et les identités qui relient les moments aux moments centrés.

Mentionnons finalement que, pour $g \neq 0$ et $h < 1/n$, le moment d'ordre n de

$$X = A + BY = A + B \left(\frac{e^{gZ} - 1}{g} \right) e^{hZ^2/2}$$

est donné par

$$E(X^n) = \sum_{i=0}^n \binom{n}{i} A^{n-i} B^i \frac{\sum_{r=0}^i (-1)^r \binom{i}{r} e^{\{(i-r)g\}^2 / \{2(1-ih)\}}}{g^i \sqrt{1-ih}}.$$

Ce résultat peut être obtenu en utilisant l'équation (2.31) et le fait que $X^n = (A + BY)^n = \sum_{i=0}^n \binom{n}{i} A^{n-i} B^i Y^i$, d'où

$$E(X^n) = \sum_{i=0}^n \binom{n}{i} A^{n-i} B^i E(Y^i).$$

2.5.4 Coefficients d'asymétrie et d'aplatissement

Nous nous intéressons ici aux troisième et quatrième moments centrés. Plus particulièrement, les coefficients d'asymétrie et d'aplatissement de la population sont donnés respectivement par

$$\sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \gamma_1$$

et

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \gamma_2,$$

où $\mu = E(X)$ et $\mu_k = E\{(X - \mu)^k\}$. Les mesures échantillonnables correspondantes sont définies de manière analogue en termes des moments échantillonnables. Rappelons que, pour une loi normale, $\sqrt{\beta_1} = 0$ et $\beta_2 = 3$.

Dans un souci d'exhaustivité, nous examinons ces mesures d'asymétrie et d'aplatissement dans les cas spéciaux les plus simples : les lois g et les lois h , avec g et h constants.

À partir des équations (2.25), (2.26) et (2.27), nous voyons que pour les lois g , nous avons

$$\sqrt{\beta_1(g)} = \frac{e^{3g^2} - 3e^{g^2} + 2}{(e^{g^2} - 1)^{3/2}},$$

$$\beta_2(g) = \frac{e^{6g^2} - 4e^{3g^2} + 6e^{g^2} - 3}{(e^{g^2} - 1)^2}.$$

Ces deux expressions se réduisent, après quelques calculs simples, à

$$\beta_1(g) = e^{3g^2} + 3e^{2g^2} - 4,$$

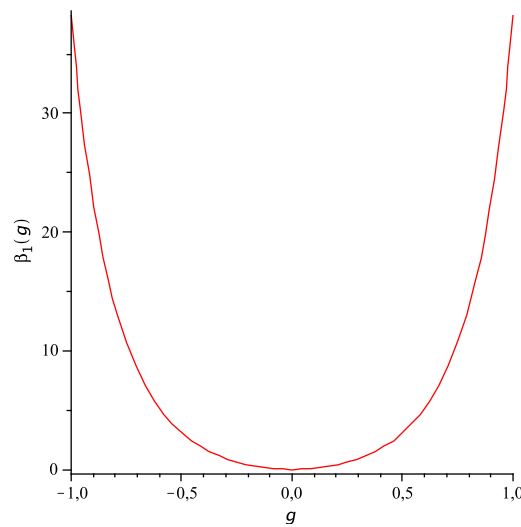
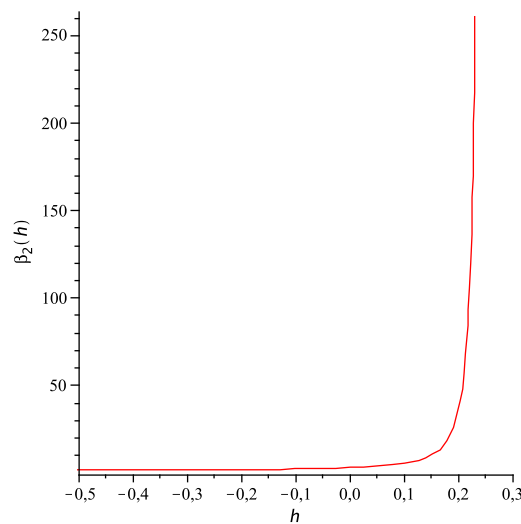
$$\beta_2(g) = e^{4g^2} + 2e^{3g^2} + 3e^{2g^2} - 3.$$

Une représentation graphique de $\beta_1(g)$ en fonction de g est donnée à la figure 2.12. Comme cette dernière le montre, la valeur de $\beta_1(g)$ croît rapidement lorsque le paramètre g s'éloigne de zéro, indiquant une asymétrie de plus en plus prononcée.

Les équations (2.29) et (2.30) permettent d'obtenir les coefficients d'asymétrie et d'aplatissement des lois h :

$$\beta_1(h) = 0 \quad \text{pour } h < 1/3,$$

$$\beta_2(h) = \frac{3(1 - 2h)^3}{(1 - 4h)^{5/2}} \quad \text{pour } h < 1/4.$$

FIG. 2.12 – Graphique de $\beta_1(g)$ en fonction de la valeur du paramètre g pour les lois g .FIG. 2.13 – Graphique de $\beta_2(h)$ en fonction de la valeur du paramètre h pour les lois h .

La figure 2.13 présente le graphique de $\beta_2(h)$ en fonction de h . Cette figure permet de constater que, pour les valeurs positives de h , le coefficient d'aplatissement $\beta_2(h)$ augmente rapidement à mesure que h croît.

Mentionnons finalement que, même avec g et h constants, la loi g -et- h générale conduit à une telle complexité algébrique que les formules pour ses coefficients d'asymétrie et d'aplatissement ne seraient pas très instructives.

Chapitre 3

Généralisation multivariée de la loi *g-et-h*

Une généralisation multivariée de la loi *g-et-h* est décrite dans ce chapitre, qui présente les travaux de [Field et Genton \(2006\)](#). Ces derniers utilisent les travaux récents de [Chaudhuri \(1996\)](#) et de [Chakraborty \(2001\)](#) pour définir la notion de quantile multivarié, une notion indispensable à l'ajustement d'une loi **g-et-h** multivariée à des données. Nous verrons en particulier que les quantiles définis en termes des normes ℓ_1 et ℓ_2 sont appropriés pour les transformations **g-et-h** multivariées.

Ce chapitre est structuré de la façon suivante. À la section [3.1](#), nous définissons la loi **g-et-h** multivariée basée sur des quantiles appropriés et nous illustrons, au moyen de courbes de niveau, diverses formes distributionnelles résultant de transformations **g-et-h**. À la section [3.2](#), nous discutons brièvement de deux caractéristiques de la loi **g-et-h** multivariée, à savoir sa densité et ses moments. Finalement, nous décrivons à la section [3.3](#) une procédure d'ajustement basée sur les quantiles.

3.1 Loi **g-et-h** multivariée

Un vecteur aléatoire $\mathbf{Y} \in \mathbb{R}^d$ est dit de loi **g-et-h** multivariée standard, où $\mathbf{g} = (g_1, \dots, g_d)^\top \in \mathbb{R}^d$ et $\mathbf{h} = (h_1, \dots, h_d)^\top \in \mathbb{R}_+^d$ contrôlent respectivement l'asymétrie et l'aplatissement, s'il peut être représenté par

$$\mathbf{Y} = (\tau_{g_1, h_1}(Z_1), \dots, \tau_{g_d, h_d}(Z_d))^\top = \boldsymbol{\tau}_{\mathbf{g}, \mathbf{h}}(\mathbf{Z}), \quad (3.1)$$

où $\mathbf{Z} = (Z_1, \dots, Z_d)^\top \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$ possède une loi normale multivariée centrée réduite et où la fonction univariée $\tau_{g,h}$ est définie par

$$\tau_{g,h}(Z) = \left(\frac{e^{gZ} - 1}{g} \right) e^{hZ^2/2}.$$

Notons que cette définition est cohérente avec le cas unidimensionnel ; en effet, (3.1) se réduit à (2.16) lorsque $d = 1$. Pour définir la loi \mathbf{g} -et- \mathbf{h} multivariée générale, soit Σ une matrice de variance-covariance quelconque et $\boldsymbol{\mu}$ un vecteur de localisation arbitraire. La loi \mathbf{g} -et- \mathbf{h} multivariée générale peut alors être représentée par

$$\mathbf{Y} = \Sigma^{1/2} \boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z}) + \boldsymbol{\mu}. \quad (3.2)$$

L'étape suivante consiste à considérer les quantiles de la loi \mathbf{g} -et- \mathbf{h} multivariée, ce qui nécessite d'abord l'introduction de la notion de quantile multivarié. Field et Genton (2006) utilisent pour ce faire des quantiles affine-équivalents définis en termes de minimisation de norme, une idée proposée par Chaudhuri (1996) et développée par Chakraborty (2001).

Pour le moment, supposons connu $\mathbf{q}_{\mathbf{Z}}(\mathbf{u})$, un quantile multivarié dans la direction \mathbf{u} basé sur la norme ℓ_p , avec $1 \leq p < \infty$, pour le vecteur aléatoire \mathbf{Z} de loi $\mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$. Les quantiles correspondants de la loi \mathbf{g} -et- \mathbf{h} multivariée générale, $\mathbf{Y} = \Sigma^{1/2} \boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z}) + \boldsymbol{\mu}$, sont donnés par

$$\mathbf{q}_{\mathbf{Y}}(\mathbf{u}) = \Sigma^{1/2} \boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{q}_{\mathbf{Z}}(\tilde{\mathbf{u}})) + \boldsymbol{\mu}, \quad (3.3)$$

où $\tilde{\mathbf{u}} = \Sigma^{-1/2} \mathbf{u} \|\mathbf{u}\|_r / \|\Sigma^{-1/2} \mathbf{u}\|_r$ et $1/p + 1/r = 1$, avec la convention que pour la norme ℓ_1 , c'est-à-dire pour $p = 1$, $r = \infty$ correspond à la norme infinie. Rappelons que si $\|\cdot\|_p$ dénote la norme ℓ_p pour $1 \leq p < \infty$ et si $\mathbf{x} = (x_1, \dots, x_d)^\top$, alors $\|\mathbf{x}\|_p = (|x_1|^p + \dots + |x_d|^p)^{1/p}$ et $\|\mathbf{x}\|_\infty = \max(|x_1|, \dots, |x_d|)$.

Avant de poursuivre, donnons quelques détails sur les quantiles multivariés de Chaudhuri (1996). Ce dernier a noté que, dans le cas unidimensionnel, pour tout $\alpha \in (0, 1)$ et $u = 2\alpha - 1$, le α^e quantile q pour un échantillon y_1, \dots, y_n peut être obtenu en minimisant la somme $\sum_{i=1}^n \{|y_i - q| + u(y_i - q)\}$. En d'autres termes, le α^e quantile q est donné par

$$\operatorname{argmin}_{q \in \mathbb{R}} \left(\sum_{i=1}^n |y_i - q| - nuq \right).$$

Chaudhuri a alors eu l'idée d'indicer les quantiles multivariés de \mathbb{R}^d par des éléments de la boule unité ouverte $B_p^{(d)} = \{\mathbf{u} \mid \mathbf{u} \in \mathbb{R}^d, \|\mathbf{u}\|_p < 1\}$ et de définir le quantile géométrique \mathbf{q} correspondant à $\mathbf{u} \in B_p^{(d)}$, pour un échantillon multivarié $\mathbf{y}_1, \dots, \mathbf{y}_n$ de \mathbb{R}^d ,

par

$$\operatorname{argmin}_{\mathbf{q} \in \mathbb{R}^d} \left(\sum_{i=1}^n \|\mathbf{y}_i - \mathbf{q}\|_p - n \sum_{j=1}^d u_j q_j \right), \quad (3.4)$$

où $\|\cdot\|_p$ dénote la norme ℓ_p pour $1 \leq p < \infty$. Seules les normes ℓ_1 et ℓ_2 seront considérées dans la suite.

Nous sommes maintenant en mesure de déterminer \mathbf{q}_Z , le quantile de norme ℓ_1 pour la loi normale multivariée $\mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$. Avec $p = 1$, $\|\mathbf{y}_i - \mathbf{q}\|_1 = \sum_{j=1}^d |y_{ij} - q_j|$, et (3.4) peut donc s'écrire

$$\operatorname{argmin}_{\mathbf{q} \in \mathbb{R}^d} \sum_{j=1}^d \left(\sum_{i=1}^n |y_{ij} - q_j| - n u_j q_j \right). \quad (3.5)$$

L'équation (3.5) nous permet de voir que la minimisation peut être effectuée séparément pour chaque dimension, de sorte que q_j est le $\{(u_j + 1)/2\}^e$ quantile pour la j^e composante des données. Ainsi, la j^e composante du quantile de norme ℓ_1 correspondant à \mathbf{u} pour la loi **g-et-h** multivariée standard est donnée par $\tau_{g_j, h_j}(z_{(u_j+1)/2})$, où z_u est le u^e quantile de la loi normale centrée réduite. Les quantiles de la loi **g-et-h** multivariée générale sont ensuite donnés par (3.3). Pour la norme ℓ_2 , le quantile pour \mathbf{Z} dans la direction \mathbf{u} est donné par $\gamma^{-1}(\|\mathbf{u}\|_2) \mathbf{u} / \|\mathbf{u}\|_2$, où

$$\gamma(t) = \sum_{j=0}^{\infty} \frac{\Gamma\left(\frac{d+1}{2} + j\right)}{\Gamma\left(\frac{d}{2} + j\right) j! 2^{j-1/2}} (2j - t^2) t^{2j-1} e^{-t^2/2}$$

et d est la dimension de \mathbf{Z} . Notons que les quantiles de norme ℓ_1 donnent les quantiles usuels dans le cas unidimensionnel, mais sont essentiellement calculés composante par composante pour les dimensions plus élevées. En revanche, les quantiles de norme ℓ_2 prennent en compte la dépendance entre les variables, mais donnent des quantiles moins communs dans le cas unidimensionnel. La figure 3.1 illustre les courbes de niveau (5 – 95%) des quantiles de norme ℓ_1 (à gauche) et de norme ℓ_2 (à droite) de la loi normale bivariée $\mathcal{N}_2(\mathbf{0}, \mathbf{I}_2)$. Ces courbes de niveau peuvent être modifiées de différentes façons en utilisant une transformation **g-et-h**, comme le montre la figure 3.2.

Afin d'ajuster une loi **g-et-h** multivariée, il est nécessaire d'être en mesure de calculer les quantiles d'un ensemble de données, qui doivent être affine-équivalents. Pour ce faire, Field et Genton (2006) utilisent une technique similaire à celle proposée par Chakraborty (2001). Dans ses travaux, ce dernier choisit un système de coordonnées induit par les données basé sur $d + 1$ des observations (avec la première observation, \mathbf{y}_{i_0} , servant d'origine). Soit $\mathbf{Y}(\alpha)$ la transformation, où $\alpha = \{i_0, \dots, i_d\} \subset \{1, \dots, n\}$ est le sous-ensemble des indices des $d + 1$ observations choisies et où la matrice $\mathbf{Y}(\alpha)$, de dimension $d \times d$, est constituée des colonnes $\mathbf{y}_{i_1} - \mathbf{y}_{i_0}, \dots, \mathbf{y}_{i_d} - \mathbf{y}_{i_0}$. Chakraborty transforme alors les données restantes pour obtenir $\mathbf{x}_i = (\mathbf{Y}(\alpha))^{-1} \mathbf{y}_i$, où $1 \leq i \leq n$ et

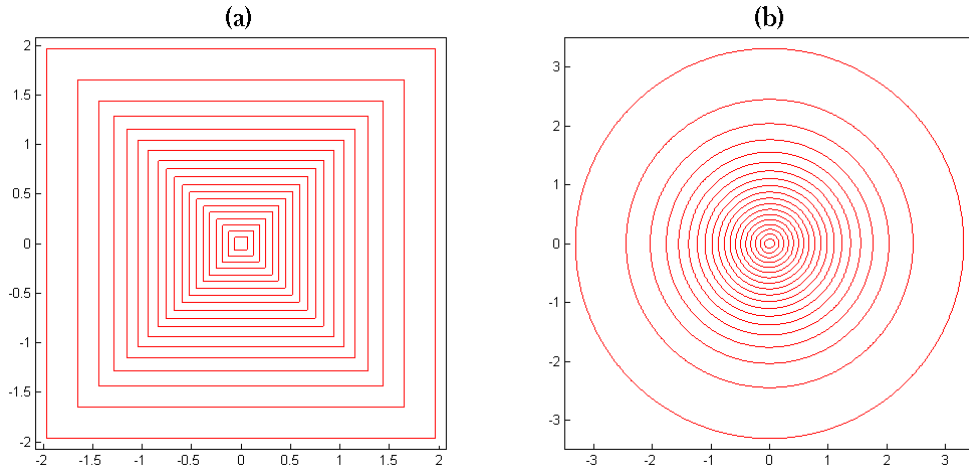


FIG. 3.1 – Courbes de niveau (5 – 95%) des quantiles de norme ℓ_1 (a) et de norme ℓ_2 (b) de la loi normale bivariée $\mathcal{N}_2(\mathbf{0}, \mathbf{I}_2)$.

$i \notin \alpha$, et il calcule les quantiles $\tilde{\mathbf{q}}(\tilde{\mathbf{u}})$ de norme ℓ_p pour \mathbf{x} dans la direction transformée $\tilde{\mathbf{u}} = \{\mathbf{Y}(\alpha)\}^{-1}\mathbf{u} \|\mathbf{u}\|_r / \|\{\mathbf{Y}(\alpha)\}^{-1}\mathbf{u}\|_r$, où $1/p + 1/r = 1$. Il obtient finalement les quantiles $\mathbf{Y}(\alpha)\tilde{\mathbf{q}}(\tilde{\mathbf{u}})$ dans la direction \mathbf{u} en effectuant la transformation inverse. Comme l'a montré Chakraborty, les quantiles ainsi obtenus sont affine-équivalents.

Dans leur article, [Field et Genton \(2006\)](#) modifient la technique de Chakraborty en remplaçant $\mathbf{Y}(\alpha)$ par $\hat{\Sigma}_{\text{MCD}}^{1/2}$, la racine carrée de l'estimateur MCD¹ (« *minimum covariance determinant* ») introduit par [Rousseeuw \(1985\)](#). En utilisant le fait que $\hat{\Sigma}_{\text{MCD}}^{1/2}$ est affine-équivalent et par un argument similaire à celui utilisé par Chakraborty (voir [Chakraborty, 2001](#), démonstration du théorème 2.1), il est possible de montrer que les quantiles obtenus sont affine-équivalents.

¹ Les estimateurs MCD de localisation et d'échelle, notés respectivement $\hat{\boldsymbol{\mu}}_{\text{MCD}}$ et $\hat{\Sigma}_{\text{MCD}}$, d'un échantillon multivarié $\mathbf{y}_1, \dots, \mathbf{y}_n$ de \mathbb{R}^d sont la moyenne et la matrice de variance-covariance calculées sur l'échantillon de m points parmi n ($1 \leq m \leq n$) qui minimise le déterminant de la matrice de variance-covariance correspondante. Il s'agit donc de trouver l'ensemble M^* tel que

$$M^* = \underset{M \subset \{\mathbf{y}_1, \dots, \mathbf{y}_n\}, \#M=m}{\operatorname{argmin}} \det(\hat{\Sigma}_M)$$

et d'en déduire les estimateurs

$$\hat{\boldsymbol{\mu}}_{\text{MCD}} = \frac{1}{m} \sum_{\mathbf{y}_i \in M^*} \mathbf{y}_i, \quad \hat{\Sigma}_{\text{MCD}} = \hat{\Sigma}_{M^*}.$$

Deux choix usuels pour m sont $m = \lfloor \frac{n+d+1}{2} \rfloor$ et $m = 0.75n$.

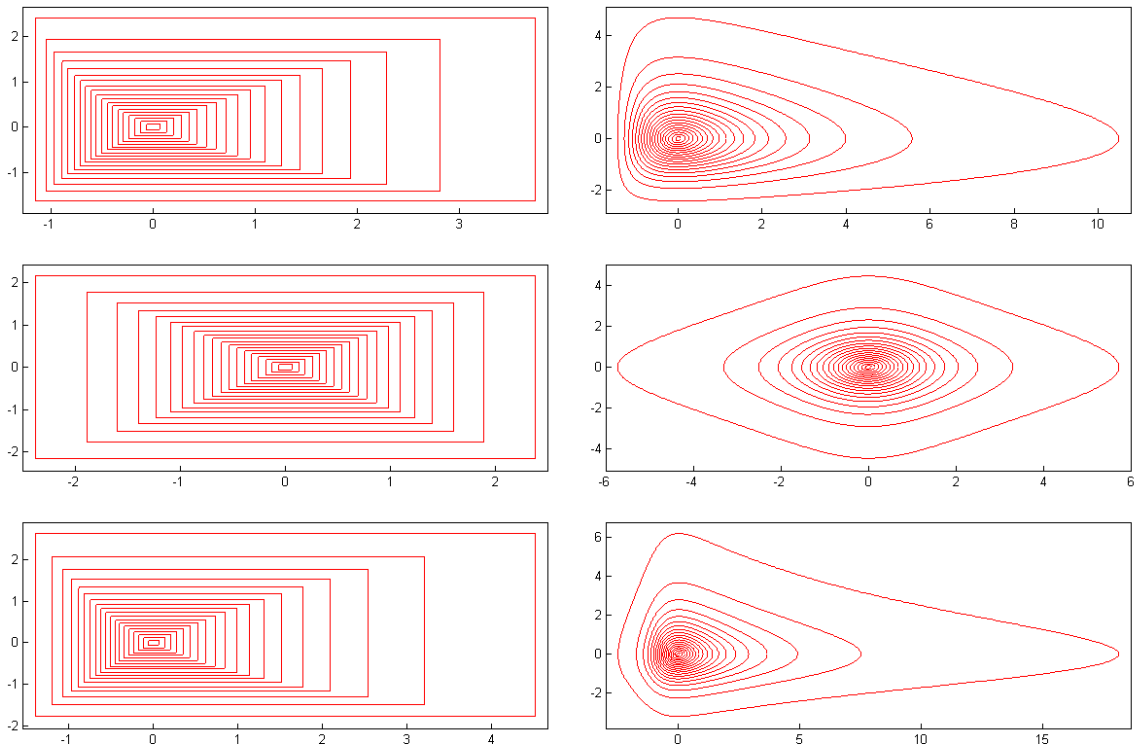


FIG. 3.2 – Courbes de niveau (5 – 95%) des quantiles de norme ℓ_1 (colonne de gauche) et de norme ℓ_2 (colonne de droite) de la loi normale bivariée $\mathcal{N}_2(\mathbf{0}, \mathbf{I}_2)$ après une transformation g -et- h : $g_1 = 0.6$, $g_2 = 0.2$, $h_1 = h_2 = 0$ (première ligne) ; $g_1 = g_2 = 0$, $h_1 = 0.1$, $h_2 = 0.05$ (deuxième ligne) ; $g_1 = 0.6$, $g_2 = 0.2$, $h_1 = 0.1$, $h_2 = 0.05$ (troisième ligne).

En résumé, les quantiles affine-équivalents peuvent être calculés de la façon suivante :

1. Calculer $\mathbf{x}_i = \hat{\Sigma}_{\text{MCD}}^{-1/2}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{\text{MCD}})$, où $\hat{\Sigma}_{\text{MCD}}$ et $\hat{\boldsymbol{\mu}}_{\text{MCD}}$ sont respectivement les matrice de variance-covariance et moyenne MCD estimées pour les données \mathbf{y} .
2. Calculer $\tilde{\mathbf{q}}(\tilde{\mathbf{u}})$, le quantile de norme ℓ_1 ou ℓ_2 dans la direction $\tilde{\mathbf{u}}$ pour les données \mathbf{x} , où $\tilde{\mathbf{u}} = \hat{\Sigma}_{\text{MCD}}^{-1/2} \mathbf{u} / \|\hat{\Sigma}_{\text{MCD}}^{-1/2} \mathbf{u}\|_r$ et où $1/p + 1/r = 1$, avec la convention que pour ℓ_1 , $r = \infty$ correspond à la norme infinie.
3. $\mathbf{q}(\mathbf{u}) = \hat{\Sigma}_{\text{MCD}}^{1/2} \tilde{\mathbf{q}}(\tilde{\mathbf{u}}) + \hat{\boldsymbol{\mu}}_{\text{MCD}}$.

3.2 Propriétés de la loi g -et- h multivariée

Dans cette section, nous discutons brièvement de la densité et des moments de la loi g -et- h multivariée. Mentionnons cependant que ces éléments sont d'une importance secondaire, l'objectif des lois g -et- h étant de modéliser les quantiles directement plutôt que de modéliser la densité.

Dans le cas univarié avec $g = 0$, la densité est donnée par

$$f_h(y) = \frac{\phi \circ \tau_{0,h}^{-1}(y)}{\tau'_{0,h} \circ \tau_{0,h}^{-1}(y)},$$

où ϕ est la densité normale centrée réduite. Bien qu'il n'existe pas de formule explicite pour la densité lorsque $g \neq 0$, la valeur de la fonction de répartition peut être évaluée numériquement en inversant la fonction quantile. Les moments dans le cas univarié, donnés par [Martinez et Iglewicz \(1984\)](#), sont étudiés plus en détail à la section 2.5.

Dans le cas multivarié, la définition (3.1) nous permet de constater que les composantes de $\mathbf{Y} = \boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z})$ sont indépendantes, de sorte que la densité conjointe est simplement le produit des densités marginales. Ainsi, il existe une forme explicite pour la densité conjointe uniquement dans le cas où tous les g_i sont nuls. Comme mentionné précédemment, la fonction de densité univariée peut être évaluée numériquement, ce qui nous permet donc d'obtenir la densité conjointe. Pour modéliser les cas de dépendance, nous pouvons simplement prémultiplier $\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z})$ par la matrice $\boldsymbol{\Sigma}^{1/2}$, comme à l'équation (3.2), pour obtenir la structure de corrélation désirée. Les moments marginaux de \mathbf{Y} découlent directement des moments univariés. L'espérance et la matrice de variance-covariance de $\mathbf{Y} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z}) + \boldsymbol{\mu}$ sont alors donnés par

$$\mathbb{E}(\mathbf{Y}) = \boldsymbol{\Sigma}^{1/2}\mathbb{E}\{\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z})\} + \boldsymbol{\mu} \quad (3.6)$$

et

$$\text{var}(\mathbf{Y}) = \boldsymbol{\Sigma}^{1/2}\text{var}\{\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z})\}(\boldsymbol{\Sigma}^{1/2})^\top. \quad (3.7)$$

Comme $\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}$ est une fonction non linéaire, la méthode Delta peut être employée pour approximer son espérance et sa matrice de variance-covariance, utilisées dans les expressions précédentes. Notons que, puisque $\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}$ agit sur \mathbf{Z} composante par composante, seules des approximations univariées sont nécessaires. Une approximation du premier ordre donne $\mathbb{E}\{\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z})\} \approx \mathbb{E}(\mathbf{Z}) = \mathbf{0}$ et $\text{var}\{\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z})\} \approx \text{var}(\mathbf{Z}) = \mathbf{I}_d$. Une approximation d'ordre 4 conduit, pour sa part, à $\mathbb{E}\{\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z})\} \approx \mathbf{d}$ et $\text{var}\{\boldsymbol{\tau}_{\mathbf{g},\mathbf{h}}(\mathbf{Z})\} \approx \mathbf{D}^2$, où \mathbf{D} est une matrice diagonale et où, pour $i = 1, \dots, d$,

$$d_i = \frac{1}{8}g_i^3 + \frac{3}{4}g_i h_i + \frac{1}{2}g_i \quad (3.8)$$

et

$$D_{ii}^2 = 1 + \frac{1}{6} g_i^6 + 2g_i^4 h_i + \frac{11}{12} g_i^4 + 6g_i^2 h_i^2 + \frac{11}{2} g_i^2 h_i + \frac{3}{2} g_i^2 + \frac{15}{4} h_i^2 + 3h_i. \quad (3.9)$$

3.3 Ajustement de la loi g -et- h multivariée à des données

Nous présentons dans cette section la procédure basée sur les quantiles empiriques que [Field et Genton \(2006\)](#) proposent pour l'ajustement d'une loi g -et- h multivariée.

Soit un échantillon $\mathbf{y}_1, \dots, \mathbf{y}_n$, où chaque observation est un élément de \mathbb{R}^d . Pour ajuster une loi g -et- h multivariée à cet échantillon, Field et Genton suggèrent, dans un premier temps, d'utiliser les estimateurs MCD pour estimer les paramètres $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$ de l'équation (3.2), ce qui correspond au cas où $\mathbf{g} = \mathbf{h} = \mathbf{0}$. Les paramètres \mathbf{g} et \mathbf{h} sont ensuite estimés en comparant les quantiles théoriques de la loi g -et- h multivariée aux quantiles empiriques calculés à partir des données. Les valeurs de $\boldsymbol{\mu}$ et de $\boldsymbol{\Sigma}$ peuvent alors être mises à jour en utilisant (3.6) et (3.7), de même que les approximations (3.8) et (3.9). Ceci conduit à l'algorithme suivant pour un ensemble \mathcal{U} de vecteurs de la boule unité ouverte :

Étape 0. Initialiser $\hat{\mathbf{g}} = \hat{\mathbf{h}} = \mathbf{0}$.

Étape 1. Calculer $\hat{\boldsymbol{\mu}}_{\text{MCD}}$ et $\hat{\boldsymbol{\Sigma}}_{\text{MCD}}$ sur $\mathbf{y}_1, \dots, \mathbf{y}_n$, et poser $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_{\text{MCD}}$, $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_{\text{MCD}}$.

Étape 2. Pour $\hat{\boldsymbol{\mu}}$ et $\hat{\boldsymbol{\Sigma}}$ fixés, calculer $\hat{\mathbf{g}}$ et $\hat{\mathbf{h}}$ par minimisation :

$$(\hat{\mathbf{g}}, \hat{\mathbf{h}}) = \underset{(\mathbf{g}, \mathbf{h})}{\operatorname{argmin}} \sum_{\mathbf{u} \in \mathcal{U}} \|\hat{\mathbf{q}}_{\mathbf{y}}(\mathbf{u}) - \hat{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{\tau}_{\mathbf{g}, \mathbf{h}}(\tilde{\mathbf{q}}_{\mathbf{z}}(\tilde{\mathbf{u}})) - \hat{\boldsymbol{\mu}}\|^2,$$

où le quantile empirique $\hat{\mathbf{q}}_{\mathbf{y}}(\mathbf{u})$ est calculé en utilisant la méthode décrite à la fin de la section 3.1.

Étape 3. Pour $\hat{\mathbf{g}}$ et $\hat{\mathbf{h}}$ fixés, mettre à jour $\hat{\boldsymbol{\mu}}$ et $\hat{\boldsymbol{\Sigma}}$ en utilisant les expressions

$$\hat{\boldsymbol{\Sigma}} \rightarrow \hat{\boldsymbol{\Sigma}}_{\text{MCD}}^{1/2} \mathbf{D}^{-2} (\hat{\boldsymbol{\Sigma}}_{\text{MCD}}^{1/2})^\top$$

et

$$\hat{\boldsymbol{\mu}} \rightarrow \hat{\boldsymbol{\mu}}_{\text{MCD}} - \hat{\boldsymbol{\Sigma}}^{1/2} \mathbf{d}.$$

Étape 4. Répéter les étapes 2 et 3 jusqu'à convergence.

Le choix de l'ensemble \mathcal{U} devrait être basé sur l'objectif de l'analyse. Dans la plupart des situations requérant l'utilisation des quantiles, l'intérêt se situe dans la queue de

la loi. Par conséquent, il est suggéré de mettre l'accent sur les quantiles de niveaux supérieurs à 90%.

Mentionnons finalement que Field et Genton présentent dans leur article deux applications de la loi **g-et-h** multivariée (voir [Field et Genton, 2006](#), sections 4.2 et 4.3). La première de ces applications vise à modéliser la loi de l'indice de masse corporelle et du poids maigre mesurés sur des athlètes australiens, une loi bivariée asymétrique, alors que la seconde illustre l'utilisation de la procédure proposée par les deux auteurs et décrite dans la présente section pour modéliser les vitesses maximales des vents dans le Pacifique Nord-Ouest.

Chapitre 4

Survol d'applications

Nous discutons dans ce chapitre de quelques applications des lois g -et- h univariées présentées au chapitre 2. Non seulement utilisées dans le domaine de la statistique (voir la section 4.1), ces dernières ont également été employées à maintes reprises pour la modélisation de données asymétriques et à queues lourdes, et ce dans des disciplines et secteurs d'activité variés. Ainsi, les sections 4.2 et 4.3 présentent des applications des lois g -et- h dans les domaines de la finance et de l'informatique, alors que la section 4.4 traite de l'utilisation de ces lois en météorologie.

4.1 Applications en statistique

Au cours des dernières années, plusieurs auteurs ont profité du fait que les lois g -et- h permettent une grande variété de formes distributionnelles pour étudier, par voie de simulation, la robustesse à la non-normalité de diverses procédures. Citons par exemple les travaux de [Guo et Luh \(2000\)](#) et de [Keselman *et coll.* \(2004\)](#), qui font usage de données simulées à partir de lois g -et- h pour évaluer la performance de nouvelles méthodes permettant de comparer deux échantillons indépendants en présence d'hétéroscédasticité et de données non gaussiennes. De façon similaire, des données simulées à partir de lois g -et- h permettent d'étudier les propriétés, en cas de non-normalité, d'une généralisation du modèle à effets aléatoires à un facteur (voir [Wilcox, 1994](#)) et de diverses méthodes visant à comparer K mesures prises sur des sujets provenant de deux groupes indépendants (voir [Wilcox et Keselman, 2001](#)).

Ce type d'approche peut également être trouvé dans les travaux de [Khan et Katti \(1991\)](#) et de [Posner et Baker \(2000\)](#). Dans le premier cas, les auteurs développent une procédure graphique permettant de détecter un écart à la normalité et d'obtenir un nouvel aperçu de la loi d'un ensemble de données. Dans ce contexte, des données générées à partir de lois g -et- h sont utilisées pour illustrer l'incidence de l'asymétrie et des queues lourdes sur les graphiques obtenus. Dans le second cas, [Posner et Baker \(2000\)](#) étudient et généralisent une approche de modélisation par équations structurelles pour l'analyse de durées de survie. Ils utilisent des données simulées à partir de trois lois non gaussiennes, dont une loi g -et- h , afin d'évaluer la robustesse de leur approche et d'étudier l'incidence des queues lourdes sur les résultats obtenus.

Remarquons que ce ne sont là que quelques exemples, des données simulées à partir de lois g -et- h ayant été utilisées pour évaluer l'effet de l'asymétrie et de l'allongement et étudier la robustesse à la non-normalité de procédures ou estimateurs dans plusieurs autres travaux, dont [Algina et coll. \(2006\)](#) et [Wilcox \(1998, 2006a,b,c,d\)](#).

Dans un autre ordre d'idées, [Raghunathan \(2000\)](#) utilise les lois g -et- h pour développer une approche bayésienne pour l'analyse de données d'enquête en présence de non-réponse non-ignorable. Plus précisément, les lois g -et- h sont utilisées pour modéliser la loi de la variable d'intérêt pour les non-répondants. L'auteur développe d'abord son approche pour un échantillonnage aléatoire simple, puis en présente une généralisation pour un échantillonnage stratifié en grappes. La méthodologie proposée est appliquée à des données provenant de l'enquête américaine « *Health and Retirement Survey* » pour estimer certaines caractéristiques de la loi de l'actif net total des adultes de 55 ans et plus.

Les lois g -et- h peuvent en outre être utilisées pour l'imputation de données manquantes dont la loi est continue, mais ne correspond pas à l'une des familles habituelles telles les lois normale, log-normale, Student ou gamma. Cette question est abordée par [He et Raghunathan \(2006\)](#), qui illustrent comment les lois g -et- h peuvent être utilisées pour l'analyse de données incomplètes par la technique de l'imputation multiple ([Rubin, 1987](#)). Cette dernière, qui est de plus en plus populaire pour le traitement des données manquantes, consiste à remplacer celles-ci par M ensembles possibles de valeurs. Chaque ensemble de données ainsi complété est par la suite analysé séparément, et les estimés ponctuels et erreurs-types sont combinés pour faire de l'inférence. [He et Raghunathan \(2006\)](#) suggèrent la procédure d'imputation multiple suivante :

1. Tirer un échantillon bootstrap $y_1^*, \dots, y_{n_1}^*$ des données observées y_1, \dots, y_{n_1} .
2. Utiliser la méthode basée sur les quantiles décrite à la section 2.3.1 pour estimer les paramètres A , B , g et h de la loi g -et- h à partir de l'échantillon bootstrap $y_1^*, \dots, y_{n_1}^*$.

3. Simuler des variables aléatoires normales centrées réduites indépendantes z_i ($i = n_1 + 1, \dots, n$) et estimer les valeurs manquantes $y_i = A + B \{(e^{gz_i} - 1)e^{hz_i^2/2}\}/g$.
4. Répéter M fois, de façon indépendante, les étapes 1 à 3 pour obtenir M ensembles de $n - n_1$ valeurs imputées. En pratique, un petit nombre d'imputations ($M = 5$ par exemple) suffit pour obtenir de bons résultats.

Les auteurs comparent les résultats de l'imputation multiple basée sur la loi g -et- h à ceux de l'analyse sur données complètes (qui offre les meilleurs résultats lorsque les données manquantes surviennent de manière complètement aléatoire), ainsi qu'aux résultats obtenus en utilisant des méthodes d'imputation multiple basées sur les lois normale et log-normale. Ces comparaisons, effectuées à l'aide de simulations et de données issues d'une enquête sur le comportement des adolescents à l'égard de la conduite avec facultés affaiblies, permettent de conclure que l'imputation par la loi g -et- h s'approche de l'analyse sur données complètes, tout en ayant un avantage indéniable sur les méthodes d'imputation plus classiques basées sur les lois normale et log-normale.

Plus de détails sur la procédure d'imputation multiple basée sur la loi g -et- h et utilisant l'estimation par les quantiles, soit la méthode décrite à la section 2.3.1, peuvent être trouvés dans les travaux de He (2005). Ce dernier propose de plus un algorithme d'imputation vraisemblantiste, c'est-à-dire utilisant une méthode d'estimation des paramètres basée sur la vraisemblance, pour les lois g -et- h avec $h > 0$. Il compare par voie de simulation la performance de ces deux procédures (vraisemblantiste et basée sur les quantiles), ce qui lui permet de conclure que ces approches offrent des rendements similaires.

He (2005) s'intéresse également au problème des données manquantes dans un contexte multivarié. En un premier temps, il propose de tenir compte de la non-normalité du terme d'erreur d'un modèle de régression en modélisant cette erreur au moyen d'une loi g -et- h , ce qui l'amène à développer des procédures d'imputation par la régression pour différentes structures de données manquantes. En plus d'illustrer son approche au moyen de données sur le comportement des adolescents à l'égard de la conduite avec facultés affaiblies, l'auteur en étudie la performance par voie de simulation.

En un second temps, He propose la généralisation multivariée suivante de la loi g -et- h :

$$Y_i = A_i + B_i \left(\frac{e^{g_i Z_i} - 1}{g_i} \right) e^{h_i Z_i^2/2}, \quad i = 1, \dots, d,$$

où $(Z_1, \dots, Z_d)^\top \sim \mathcal{N}_d(\mathbf{0}, \mathbf{R})$ possède une loi normale multivariée de matrice de corrélation \mathbf{R} . (Notons que cette définition diffère de la généralisation multivariée suggérée par Field et Genton, 2006, et décrite au chapitre 3.) Faisant usage de cette généralisation,

He propose une méthode d'imputation pour données multivariées non gaussiennes. Cette approche est ensuite appliquée à des données sur la conduite avec facultés affaiblies, et sa performance sous diverses lois multivariées est évaluée par voie de simulation.

4.2 Applications en finance

Les données présentant des structures d'asymétrie et d'allongement complexes sont fréquentes dans la littérature financière. Dans leur article, [Badrinath et Chatterjee \(1988\)](#) réalisent une étude exploratoire des propriétés de la loi de rendements boursiers en utilisant les lois g -et- h . Plus précisément, les auteurs montrent que, pour des périodes de temps suffisamment longues, la loi de l'indice boursier peut être décrite de façon satisfaisante par une loi g -et- h . Ils décrivent une procédure d'ajustement simple et robuste aux valeurs aberrantes, et illustrent leur approche au moyen de données quotidiennes et mensuelles provenant du CRSP (« *Center for Research in Security Prices* »).

Dans un article subséquent, [Badrinath et Chatterjee \(1991\)](#) utilisent les lois g -et- h pour explorer la nature de l'asymétrie et de l'allongement des lois des rendements boursiers quotidiens de plusieurs entreprises individuelles. Les auteurs sont ainsi appelés à étudier les structures d'allongement et d'asymétrie des rendements de diverses sociétés cotées à la Bourse de New-York ou à l'AMEX (« *American Exchange* ») et classées selon leur secteur d'activité, leur risque systématique et leur taille. Une fois de plus, l'ajustement de la loi g -et- h est acceptable.

Dans le même ordre d'idées, l'article de [Mills \(1995\)](#) présente une étude empirique des lois des rendements quotidiens de trois indices de la Bourse de Londres, les indices FT-SE 100, Mid 250 et 350, durant la période 1986–92. L'auteur conclut, à l'instar de [Badrinath et Chatterjee \(1988, 1991\)](#), que les lois g -et- h s'ajustent de façon très satisfaisante aux rendements des indices boursiers considérés.

Les lois g -et- h ont également été employées pour la modélisation de taux d'intérêt. Citons notamment les travaux de [Dutta et Babbel \(2002\)](#), dans lesquels les lois g -et- h et Bêta généralisée de seconde espèce (GB2) sont utilisées pour modéliser l'asymétrie et l'allongement de taux courte durée. Les auteurs observent ainsi que les lois des taux LIBOR (« *London Inter Bank Offer Rates* ») dollar américain 1 mois et 3 mois peuvent être décrites de façon précise au moyen de lois g -et- h . Ils concluent de plus que ces dernières sont probablement un choix plus intéressant que les lois GB2 en ce qui concerne la modélisation des taux LIBOR, tant au niveau de la qualité de l'ajustement que de la facilité avec laquelle l'estimation des paramètres peut être effectuée.

Les mêmes auteurs s'intéressent ensuite au marché des options. Plus précisément, [Dutta et Babbel \(2005\)](#) utilisent dans leur article la loi g -et- h pour obtenir une formule explicite d'évaluation des options européennes. Faisant usage de données sur les taux plafond, les auteurs analysent la performance du modèle obtenu et comparent ce dernier aux prix d'options basés sur les lois log-normale, Burr-3, Weibull et Bêta généralisée de seconde espèce. Ils concluent alors que la loi g -et- h permet une grande précision dans l'évaluation d'options, en plus d'offrir un rendement nettement supérieur aux autres lois.

Finalement, certains auteurs ont récemment abordé la question de la modélisation du risque opérationnel. Ainsi, [Dutta et Perry \(2007\)](#) effectuent une évaluation détaillée des méthodes couramment employées et de nouvelles techniques de mesure de ce risque basées sur la théorie des valeurs extrêmes, l'échantillonnage empirique ou l'ajustement d'une loi paramétrique (exponentielle, gamma, Pareto généralisée, log-logistique, log-normale tronquée, Weibull, Bêta généralisée de seconde espèce et g -et- h). Ces différentes approches sont ensuite appliquées à des données sur les pertes internes de sept institutions financières et comparées au moyen de diverses mesures de performance, ce qui permet aux auteurs de conclure que la loi g -et- h offre un meilleur rendement que les autres modèles évalués.

Pour terminer cette section, mentionnons que les travaux de [Dutta et Perry \(2007\)](#), qui présentent la loi g -et- h comme un candidat intéressant pour la modélisation du risque opérationnel, ont inspiré [Degen et coll. \(2007\)](#). En effet, ces derniers discutent dans leur article de quelques propriétés fondamentales des lois g -et- h , ainsi que de leur lien avec la théorie des valeurs extrêmes.

4.3 Application en informatique

[Liu et coll. \(2006\)](#) s'intéressent dans leurs travaux à la modélisation du trafic Internet. Plus précisément, les auteurs suggèrent une série chronologique autorégressive avec bruit distribué selon une loi g -et- h . Ils montrent que le modèle qu'ils proposent, qui peut être utilisé pour prédire les besoins en bande passante, permet de tenir compte simultanément de la périodicité, de l'autocorrélation et de la loi marginale non gaussienne du trafic Internet. Enfin, les auteurs illustrent leur approche au moyen de données réelles.

4.4 Applications en météorologie

Dupuis et Field (2004) abordent dans leur article la question de la modélisation des vitesses extrêmes des vents, dans le but de développer une procédure permettant d'identifier de manière fiable les observations aberrantes. Les auteurs présentent une méthode robuste pour ajuster une loi g -et- h à des données sur les vitesses des vents océaniques. De plus, ils proposent une procédure pour identifier les valeurs aberrantes et utilisent la statistique de Anderson–Darling pour développer une mesure de la qualité de l'ajustement d'un modèle. Cette mesure est ensuite utilisée pour comparer l'ajustement de la loi g -et- h à celui de deux lois plus classiques pour modéliser les vitesses extrêmes des vents, à savoir la loi des valeurs extrêmes généralisée (LVEG) et la loi de Pareto généralisée (LPG). Dupuis et Field concluent alors que la loi g -et- h est au moins aussi efficace pour modéliser les vitesses extrêmes des vents que les lois LVEG et LPG.

Dans un article subséquent, Field (2004) met de nouveau l'accent sur la modélisation des vitesses extrêmes des vents au moyen de la loi g -et- h . Il y présente une méthode pour modéliser les vitesses maximales des vents sur une période de temps fixée (par exemple, une semaine ou un mois). Cette approche fournit une alternative concurrentielle et flexible à la loi asymptotique habituelle, la loi des valeurs extrêmes généralisée. De plus, la procédure présentée est robuste et permet d'identifier les observations aberrantes. Dans la dernière section de son article, Field illustre sa méthode au moyen de données sur les vitesses des vents dans une région du centre de l'Atlantique.

Terminons cette section en décrivant brièvement la procédure robuste d'ajustement de la loi g -et- h utilisée par Dupuis et Field (2004), puis par Field (2004). L'idée de base de leur approche est de choisir g et h de manière à minimiser une certaine mesure de distance robuste entre les quantiles théoriques et observés. Pour ce faire, les auteurs fixent un ensemble de quantiles et utilisent la fonction ρ de Huber définie par

$$\rho(x) = \begin{cases} x^2/2 & \text{si } |x| < k, \\ k|x| - k^2/2 & \text{sinon,} \end{cases}$$

où k est un scalaire positif, pour comparer les quantiles théoriques (y_i) de la loi g -et- h aux quantiles observés (x_i) des données centrées ou centrées réduites. (Notons que les estimateurs des paramètres de localisation et d'échelle utilisés pour centrer et réduire les données doivent être robustes. Par exemple, la médiane et la médiane des écarts absolus à la médiane peuvent être utilisées.) De façon plus précise, les paramètres sont estimés en minimisant

$$\sum_i \rho\{(x_i - y_i)/y_i\} \tag{4.1}$$

par rapport à g et h , ainsi que par rapport au paramètre d'échelle B si l'estimation est réalisée à partir des données non réduites, la sommation (4.1) étant ici prise sur un ensemble fixé de quantiles.

Chapitre 5

Conclusion

Nous avons étudié dans cet essai les lois g -et- h , une famille de lois relativement peu connue mais offrant un grand potentiel pour la simulation et la modélisation de données non gaussiennes, tant univariées que multivariées. Ces lois permettent en outre d'approximer de nombreuses lois théoriques. En effet, [Martinez et Iglewicz \(1984\)](#) montrent comment plus de douze lois univariées, parmi lesquelles des lois logistique, Student, exponentielle, Bêta, Weibull et khi-deux, peuvent être approximées en choisissant de façon appropriée les paramètres g et h .

Nous avons également vu que les lois g -et- h ont été, jusqu'à présent, principalement utilisées en statistique, en finance et en météorologie. Nul doute cependant que ces lois très flexibles trouveront plusieurs autres domaines d'application dans les années à venir.

Soulignons pour terminer que certains auteurs se sont inspirés des remarques de [MacGillivray \(1992\)](#) sur la forme des lois g -et- h pour proposer deux adaptations de ces dernières, les lois g -et- h généralisée et g -et- k (voir [Rayner et MacGillivray, 2002a](#), et les références qui y sont mentionnées). Ces lois sont données respectivement par

$$Y_{g,h}^{\text{gen}}(Z) = A + BZ \left(1 + c \frac{1 - e^{-gZ}}{1 + e^{-gZ}} \right) e^{hZ^2/2}$$

et

$$Y_{g,k}(Z) = A + BZ \left(1 + c \frac{1 - e^{-gZ}}{1 + e^{-gZ}} \right) (1 + Z^2)^k,$$

où Z est une variable aléatoire normale centrée réduite, $A \in \mathbb{R}$ et $B > 0$ sont des paramètres de localisation et d'échelle et c est un scalaire (une valeur de $c = 0.8$ est généralement utilisée). De plus, à l'instar des lois g -et- h , le paramètre $g \in \mathbb{R}$ contrôle l'asymétrie, alors que $h > 0$ et $k > -1/2$ contrôlent l'allongement.

Pour plus d'informations sur les lois g -et- h généralisée et g -et- k ainsi que pour des exemples d'études de simulation réalisées à partir de lois g -et- k , le lecteur est invité à consulter les travaux de [Haynes *et coll.* \(1997\)](#) et [Khan et Rayner \(2003\)](#). Diverses méthodes d'estimation des paramètres de ces deux adaptations des lois g -et- h sont également proposées dans les travaux de [Rayner et MacGillivray \(2002a,b\)](#) et [Haynes et Mengersen \(2005\)](#), qui s'intéressent respectivement à des approches numérique vraisemblantiste, basée sur les quantiles et bayésienne.

Bibliographie

- ALGINA, J., KESELMAN, H. J. et PENFIELD, R. D. (2006). Confidence interval coverage for Cohen's effect size statistic. *Educational and Psychological Measurement*, 66(6): 945–960.
- BADRINATH, S. G. et CHATTERJEE, S. (1988). On measuring skewness and elongation in common stock return distributions : the case of the market index. *Journal of Business*, 61(4):451–472.
- BADRINATH, S. G. et CHATTERJEE, S. (1991). A data-analytic look at skewness and elongation in common-stock-return distributions. *Journal of Business & Economic Statistics*, 9(2):223–233.
- CHAKRABORTY, B. (2001). On affine equivariant multivariate quantiles. *Annals of the Institute of Statistical Mathematics*, 53(2):380–403.
- CHAUDHURI, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434):862–872.
- DEGEN, M., EMBRECHTS, P. et LAMBRIGGER, D. D. (2007). The quantitative modeling of operational risk : between g-and-h and EVT. *ASTIN Bulletin*, 37(2): 265–291. [En ligne], http://www.math.ethz.ch/~baltres/ftp/g-and-h_May07.pdf (Page consultée le .
- DUPUIS, D. J. et FIELD, C. A. (2004). Large wind speeds : modeling and outlier detection. *Journal of Agricultural, Biological, and Environmental Statistics*, 9(1):105–121.
- DUTTA, K. K. et BABEL, D. F. (2002). On measuring skewness and kurtosis in short rate distributions : the case of the US dollar London Inter Bank Offer Rates. Document de travail n° 02-25, Wharton Financial Institutions Center, The Wharton School, University of Pennsylvania, Philadelphia. [En ligne], <http://fic.wharton.upenn.edu/fic/papers/02/0225.pdf> (Page consultée le 4 juillet 2007).

- DUTTA, K. K. et BABEL, D. F. (2005). Extracting probabilistic information from the prices of interest rate options : tests of distributional assumptions. *Journal of Business*, 78(3):841–870.
- DUTTA, K. K. et PERRY, J. (2007). A tale of tails : an empirical analysis of loss distribution models for estimating operational risk capital. Document de travail n° 06-13, Federal Reserve Bank of Boston, Massachusetts. [En ligne], <http://www.bos.frb.org/economic/wp/wp2006/wp0613.htm> (Page consultée le 24 juillet 2008).
- EMERSON, J. D. et HOAGLIN, D. C. (1985). Resistant multiple regression, one variable at a time. Dans HOAGLIN, D. C., MOSTELLER, F. et TUKEY, J. W., éditeurs : *Exploring Data, Tables, Trends, and Shapes*, chapitre 7, pages 241–279. John Wiley & Sons, Inc.
- FIELD, C. A. (2004). Using the *gh* distribution to model extreme wind speeds. *Journal of Statistical Planning and Inference*, 122(1-2):15–22.
- FIELD, C. A. et GENTON, M. G. (2006). The multivariate **g**-and-**h** distribution. *Technometrics*, 48(1):104–111.
- GUO, J.-H. et LUH, W.-M. (2000). An invertible transformation two-sample trimmed *t*-statistic under heterogeneity and nonnormality. *Statistics & Probability Letters*, 49(1):1–7.
- HAYNES, M. A., MACGILLIVRAY, H. L. et MENGERSEN, K. L. (1997). Robustness of ranking and selection rules using generalised *g*-and-*k* distributions. *Journal of Statistical Planning and Inference*, 65(1):45–66.
- HAYNES, M. A. et MENGERSEN, K. L. (2005). Bayesian estimation of *g*-and-*k* distributions using MCMC. *Computational Statistics*, 20(1):7–30.
- HE, Y. (2005). *Multiple Imputation for Continuous Non-Normal Missing Data*. Thèse de doctorat, University of Michigan, Ann Arbor, Michigan.
- HE, Y. et RAGHUNATHAN, T. E. (2006). Tukey’s *gh* distribution for multiple imputation. *The American Statistician*, 60(3):251–256.
- HOAGLIN, D. C. (1985). Summarizing shape numerically : the *g*-and-*h* distributions. Dans HOAGLIN, D. C., MOSTELLER, F. et TUKEY, J. W., éditeurs : *Exploring Data Tables, Trends, and Shapes*, chapitre 11, pages 461–513. John Wiley & Sons, Inc.
- HOAGLIN, D. C. et PETERS, S. C. (1979). Software for exploring distribution shape. [En ligne], <http://dspace.mit.edu/bitstream/1721.1/909/1/P-0909-15603614.pdf> (Page consultée le 20 juin 2007).

- JOHNSON, N. L., KOTZ, S. et BALAKRISHNAN, N. (1994). *Continuous Univariate Distributions, Volume 1*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., 2^e édition.
- KESELMAN, H. J., OTHMAN, A. R., WILCOX, R. R. et FRADETTE, K. (2004). The new and improved two-sample t test. *Psychological Science*, 15(1):47–51.
- KHAN, A. et RAYNER, G. D. (2003). Robustness to non-normality of common tests for the many-sample location problem. *Journal of Applied Mathematics and Decision Sciences*, 7(4):187–206.
- KHAN, B. A. et KATTI, S. K. (1991). Graphical method to detect departure from normality with special application to Tukey's g and h distributions. *Biometrical Journal*, 33(7):829–840.
- LIU, Z., ALMHANA, J., CHOULAKIAN, V. et MCGORMAN, R. (2006). Periodic data traffic modeling and prediction-based bandwidth allocation. Proceedings of the 4th Annual Communication Networks and Services Research Conference, Moncton, New Brunswick.
- MACGILLIVRAY, H. L. (1992). Shape properties of the g -and- h and Johnson families. *Communications in Statistics – Theory and Methods*, 21(5):1233–1250.
- MARTINEZ, J. et IGLEWICZ, B. (1984). Some properties of the Tukey g and h family of distributions. *Communications in Statistics – Theory and Methods*, 13(3):353–369.
- MILLS, T. C. (1995). Modelling skewness and kurtosis in the London Stock Exchange FT-SE index return distributions. *The Statistician*, 44(3):323–332.
- POSNER, S. F. et BAKER, L. (2000). Evaluation and extensions of a structural equation modeling approach to the analysis of survival data. *Behavior Genetics*, 30(1):41–50.
- RAGHUNATHAN, T. E. (2000). A Bayesian approach for finite population inference with nonignorable nonresponse. Document de travail n° 094, Survey Methodology Program, Institute for Social Research, University of Michigan, Ann Arbor, Michigan. [En ligne], <http://www.isr.umich.edu/src/smp/Electronic%20Copies/94-gh.pdf> (Page consultée le 12 août 2008).
- RAYNER, G. D. et MACGILLIVRAY, H. L. (2002a). Numerical maximum likelihood estimation for the g -and- k and generalized g -and- h distributions. *Statistics and Computing*, 12(1):57–75.

- RAYNER, G. D. et MACGILLIVRAY, H. L. (2002b). Weighted quantile-based estimation for a class of transformation distributions. *Computational Statistics & Data Analysis*, 39(4):401–433.
- ROUSSEEUW, P. (1985). Multivariate estimation with high breakdown point. Dans GROSSMANN, W., PFLUG, G. C., VINCZE, I. et WERTZ, W., éditeurs : *Mathematical Statistics and Applications (Proceedings of the 4th Pannonian Symposium on Mathematical Statistics, Bad Tatzmannsdorf, Austria, 4-10 September, 1983)*, volume B, pages 283–297. D. Reidel Publishing Company.
- RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc.
- TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Series in Behavioral Science – Quantitative Methods. Addison-Wesley Publishing Company, Inc.
- WILCOX, R. R. (1994). A one-way random effects model for trimmed means. *Psychometrika*, 59(3):289–306.
- WILCOX, R. R. (1998). A note on the Theil–Sen regression estimator when the regressor is random and the error term is heteroscedastic. *Biometrical Journal*, 40(3):261–268.
- WILCOX, R. R. (2006a). Confidence intervals for prediction intervals. *Journal of Applied Statistics*, 33(3):317–326.
- WILCOX, R. R. (2006b). Pairwise comparisons of dependent groups based on medians. *Computational Statistics & Data Analysis*, 50(10):2933–2941.
- WILCOX, R. R. (2006c). Some results on comparing the quantiles of dependent groups. *Communications in Statistics – Simulation and Computation*, 35(4):893–900.
- WILCOX, R. R. (2006d). Testing the hypothesis of a homoscedastic error term in simple, nonparametric regression. *Educational and Psychological Measurement*, 66(1):85–92.
- WILCOX, R. R. et KESELMAN, H. J. (2001). Using trimmed means to compare k measures corresponding to two independent groups. *Multivariate Behavioral Research*, 36(3):421–444.

Annexe A

Valeurs lettrées

Introduites par [Tukey \(1977, chapitre 2\)](#), les valeurs lettrées (« *letter values* ») sont un ensemble particulier de quantiles théoriques ou échantillonnaires. Pour une loi théorique, elles commencent avec la médiane (aire sous la queue de 1/2) et les quartiles inférieur et supérieur (aire sous la queue de 1/4). Elles continuent ensuite en réduisant successivement l'aire sous la queue de moitié pour produire les huitièmes, les seizièmes, et ainsi de suite. Ainsi, les quantiles sélectionnés proviennent davantage des queues que du centre de la loi.

Dans un échantillon de taille n , les valeurs lettrées permettent d'extraire des observations, à certaines profondeurs simples, de l'échantillon ordonné

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}.$$

La *profondeur* d'une observation est sa position dans l'échantillon ordonné relativement à l'extrémité la plus près. Plus techniquement, la profondeur de $x_{(i)}$ est le minimum de i et de $n + 1 - i$. Ainsi, les observations $x_{(3)}$ et $x_{(n-2)}$ ont toutes deux une profondeur de 3.

Comme précédemment, les valeurs lettrées échantillonnaires commencent avec la médiane, à une profondeur de $(n + 1)/2$, et les observations extrêmes, à une profondeur de 1. Elles se poursuivent ensuite, de la médiane vers les extrêmes, selon une séquence de profondeurs définies par

$$\frac{\lfloor \text{profondeur précédente} \rfloor + 1}{2},$$

où $\lfloor x \rfloor$ dénote la partie entière de x , soit le plus grand entier inférieur ou égal à x . Si une des profondeurs calculées dans cette suite n'est pas un nombre entier, alors sa partie

fractionnaire doit être égale à $1/2$. Nous obtenons alors la valeur lettrée correspondante en prenant la moyenne de deux statistiques d'ordre successives.

Notons qu'à l'exception de la médiane, les valeurs lettrées viennent en paires : les valeurs lettrées de profondeur d sont $x_{(d)}$ et $x_{(n+1-d)}$. Mentionnons finalement que, pour référer aux valeurs lettrées, nous utilisons habituellement les étiquettes suivantes : M pour la médiane, F pour les quarts, E pour les huitièmes et ainsi de suite en ordre alphabétique inverse (D, C, B, A, Z, Y, X, ...).

A.1 Exemple de calcul

Considérons les observations ordonnées suivantes :

i :	1	2	3	4	5	6	7	8	9	10
$x_{(i)}$:	2.0	2.5	6.0	6.5	7.5	9.0	11.5	12.0	13.5	14.0

i :	11	12	13	14	15	16	17	18	19
$x_{(i)}$:	14.5	15.0	16.0	18.0	18.0	20.0	25.0	42.0	52.0

Les profondeurs et les valeurs lettrées sont alors données par

Étiquette	Profondeur	Valeurs lettrées
M	$10 = (19 + 1)/2$	14.0
F	$5.5 = (10 + 1)/2$	8.25 et 18.0
E	$3 = (5 + 1)/2$	6.0 et 25.0
D	$2 = (3 + 1)/2$	2.5 et 42.0
C	$1.5 = (2 + 1)/2$	2.25 et 47.0
	1	2.0 et 52.0

Cette information est souvent présentée sous la forme suivante :

$n = 19$	
M	10 14.0
F	5.5 8.25 18.0
E	3 6.0 25.0
D	2 2.5 42.0
C	1.5 2.25 47.0
	1 2.0 52.0

Annexe B

Démonstrations des résultats (2.28) et (2.31)

B.1 Démonstration du résultat (2.28)

Soit

$$Y = Ze^{hZ^2/2}.$$

Nous avons alors

$$Y^n = Z^n e^{nhZ^2/2}$$

et

$$\begin{aligned} E(Y^n) &= \int_{-\infty}^{\infty} z^n e^{nhz^2/2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^n \exp\left\{-\frac{1}{2}(1-nh)z^2\right\} dz. \end{aligned} \quad (\text{B.1})$$

En posant $w = (1-nh)^{1/2}z$, (B.1) devient

$$E(Y^n) = \frac{1}{\sqrt{2\pi}} \frac{1}{(1-nh)^{\frac{n+1}{2}}} \int_{-\infty}^{\infty} w^n e^{-w^2/2} dw.$$

Comme $w^n e^{-w^2/2}$ est une fonction impaire si n est impair, $E(Y^n) = 0$ pour n impair.

Si n est pair, nous avons

$$E(Y^n) = \frac{1}{\sqrt{2\pi}} \frac{2}{(1-nh)^{\frac{n+1}{2}}} \int_0^{\infty} w^n e^{-w^2/2} dw.$$

En posant $u = w^2/2$, nous obtenons

$$\begin{aligned}
\mathbb{E}(Y^n) &= \frac{1}{\sqrt{2\pi}} \frac{2}{(1-nh)^{\frac{n+1}{2}}} \int_0^\infty (2u)^{(n-1)/2} e^{-u} du \\
&= \frac{1}{\sqrt{2\pi}} \frac{2}{(1-nh)^{\frac{n+1}{2}}} 2^{(n-1)/2} \underbrace{\int_0^\infty u^{\frac{n+1}{2}-1} e^{-u} du}_{\Gamma(\frac{n+1}{2})} \\
&= \frac{1}{\sqrt{2\pi}} \frac{2}{(1-nh)^{\frac{n+1}{2}}} 2^{(n-1)/2} \Gamma\left(\frac{n+1}{2}\right) \\
&= \frac{1}{\sqrt{2\pi}} \frac{2}{(1-nh)^{\frac{n+1}{2}}} 2^{(n-1)/2} \sqrt{\pi} \frac{n!}{2^n \left(\frac{n}{2}\right)!} \\
&= \frac{n!}{2^{n/2} \left(\frac{n}{2}\right)!} \frac{1}{(1-nh)^{\frac{n+1}{2}}},
\end{aligned}$$

où nous avons utilisé le fait que

$$\Gamma\left(t + \frac{1}{2}\right) = \sqrt{\pi} \frac{(2t)!}{2^{2t} t!}$$

pour $t \in \mathbb{N}$. Ainsi, nous obtenons

$$\mathbb{E}(Y^n) = \begin{cases} 0 & \text{pour } n \text{ impair,} \\ \frac{n!}{2^{n/2} \left(\frac{n}{2}\right)!} \frac{1}{(1-nh)^{\frac{n+1}{2}}} & \text{pour } n \text{ pair,} \end{cases}$$

ce qui complète la démonstration du résultat (2.28).

B.2 Démonstration du résultat (2.31)

Soit

$$Y = \left(\frac{e^{gZ} - 1}{g}\right) e^{hZ^2/2}.$$

Nous avons alors, en utilisant la formule du binôme de Newton,

$$\begin{aligned}
Y^n &= \frac{e^{nhZ^2/2}}{g^n} (e^{gZ} - 1)^n \\
&= \frac{e^{nhZ^2/2}}{g^n} \sum_{i=0}^n (-1)^i \binom{n}{i} e^{(n-i)gZ}.
\end{aligned}$$

Ainsi,

$$\begin{aligned} E(Y^n) &= \frac{1}{g^n} \sum_{i=0}^n (-1)^i \binom{n}{i} \int_{-\infty}^{\infty} \exp \left\{ \frac{nhz^2}{2} + (n-i)gz \right\} \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{z^2}{2} \right) dz \\ &= \frac{1}{g^n} \sum_{i=0}^n (-1)^i \binom{n}{i} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2}(1-nh)z^2 + (n-i)gz \right\} dz. \quad (\text{B.2}) \end{aligned}$$

Pour simplifier la formule précédente, notons que si

$$T = \exp \left(\frac{1}{2}aZ^2 + bZ \right),$$

où a et b sont des constantes, alors

$$\begin{aligned} E(T) &= \int_{-\infty}^{\infty} \exp \left(\frac{az^2}{2} + bz \right) \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{z^2}{2} \right) dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2}(1-a)z^2 + bz \right\} dz. \end{aligned}$$

Après avoir complété le carré, l'expression pour $E(T)$ devient

$$\begin{aligned} E(T) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left\{ -\left(\frac{\sqrt{1-a}}{\sqrt{2}} z - \frac{1}{\sqrt{2}} \frac{b}{\sqrt{1-a}} \right)^2 + \frac{b^2}{2(1-a)} \right\} dz \\ &= \frac{1}{\sqrt{2\pi}} \exp \left\{ \frac{b^2}{2(1-a)} \right\} \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2} \left(\sqrt{1-a} z - \frac{b}{\sqrt{1-a}} \right)^2 \right\} dz \\ &= \exp \left\{ \frac{b^2}{2(1-a)} \right\} \frac{1}{\sqrt{1-a}} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1-a}} \exp \left[-\frac{1}{2} \left\{ \frac{z - \frac{b}{(1-a)}}{\frac{1}{\sqrt{1-a}}} \right\}^2 \right] dz}_1 \\ &= \exp \left\{ \frac{b^2}{2(1-a)} \right\} \frac{1}{\sqrt{1-a}} \quad (\text{B.3}) \end{aligned}$$

pour $a < 1$.

En insérant (B.3) dans (B.2) avec $a = nh$ et $b = (n-i)g$, nous obtenons

$$\begin{aligned} E(Y^n) &= \frac{1}{g^n} \sum_{i=0}^n (-1)^i \binom{n}{i} \exp \left[\frac{\{(n-i)g\}^2}{2(1-nh)} \right] \frac{1}{\sqrt{1-nh}} \\ &= \frac{1}{g^n \sqrt{1-nh}} \sum_{i=0}^n (-1)^i \binom{n}{i} e^{\{(n-i)g\}^2 / \{2(1-nh)\}} \end{aligned}$$

pour $nh < 1$, c'est-à-dire $h < 1/n$, ce qui complète la démonstration du résultat (2.31).