Bornes de corrélation dans le cas de loi gaussienne : théorème de Lancaster
Amphi 7 - E. Gobet

Le résultat suivant donne une borne sur la corrélation extrémale de deux fonctions de variables gaussiennes corrélées. De manière surprenante, la transformation par des fonctions arbitraires réduit toujours la corrélation. Pour un exemple dans le cas de variables log-normales, voir diapo 20 du cours.
Ceci devrait mettre en garde le lecteur sur le calcul de corrélation et sa signification : la corrélation n'est qu'une mesure de dépendance (parmi d'autres), qui ne décrit pas toute la dépendance entre variables aléatoires.

Théorème. Soit (X, Y) un vecteur aléatoire de loi gaussienne, avec composante centrée réduite et corrélation $\rho \in [-1, 1]$. Alors pour toutes fonctions (f, g) telles que $\mathbb{E}(f^2(X)) < +\infty)$ et $\mathbb{E}(g^2(Y)) < +\infty)$, on a

$$|\mathbb{C}\text{or}(f(X), g(Y))| \leq \rho.$$

PREUVE. Rappelons l'expression de la Transformée de Laplace (TL) de la v.a. G de loi $\mathcal{N}(m, \sigma^2)$:

$$\mathbb{E}(e^{uG}) = e^{um+\frac{1}{2}u^2\sigma^2}. \tag{1}$$

La preuve passe par plusieurs résultats intermédiaires.

1. La TL appliquée à la v.a. X de loi $\mathcal{N}(0, 1)$ donne

$$\mathbb{E}(e^{uX-\frac{1}{2}u^2}) = 1. \tag{2}$$

2. Introduisons les polynômes d'Hermite définis par $P_k(x) = \partial_u^k[e^{ux-\frac{1}{2}u^2}]|_{u=0}$.
   Par exemple $P_0(x) = 1$, $P_1(x) = x$, $P_2(x) = x^2 - 1$, ...

3. En dérivant (2) par rapport à u en $u = 0$ et en justifiant la dérivation sous $\mathbb{E}$ par application répétée du théorème de Lebesgue, on obtient

$$0 = \partial_u^k \mathbb{E}(e^{uX-\frac{1}{2}u^2})|_{u=0} = \mathbb{E}(\partial_u^k[e^{uX-\frac{1}{2}u^2}]|_{u=0}) = \mathbb{E}(P_k(X)) \tag{3}$$

   pour $k \geq 1$.

4. On va montrer

$$\mathbb{E}(P_k(X)P_l(Y)) = \begin{cases} 0 & \text{si } k \neq l \\ \rho^k k! & \text{si } k = l \end{cases}, \tag{4}$$

   c'est-à-dire qu'en particulier $(P_k)_k$ est une famille orthogonale pour la loi gaussienne : $\langle P_k, P_l \rangle := \mathbb{E}(P_k(X)P_l(X)) = \delta_{k,l}$. Pour justifier (4), écrivons le développement en séries de

$$e^{ux-\frac{1}{2}u^2} = \sum_{k\geq 0} P_k(x)\frac{u^k}{k!}.$$

Alors, d'une part

$$\mathbb{E}(e^{uX-\frac{1}{2}u^2}e^{vY-\frac{1}{2}v^2}) = \mathbb{E}\left(\sum_{k\geq 0} P_k(X)\frac{u^k}{k!}\sum_{k\geq 0} P_k(Y)\frac{v^k}{k!}\right) = \sum_{k,l\geq 0}\frac{u^k}{k!}\frac{v^l}{l!}\mathbb{E}(P_k(X)P_l(Y)).$$

D'autre part, puisque $uX+vY$ est de loi $\mathcal{N}(0,u^2+v^2+2\rho uv)$, en appliquant (1) on a

$$\mathbb{E}(e^{uX-\frac{1}{2}u^2}e^{vY-\frac{1}{2}v^2}) = e^{\frac{1}{2}(u^2+v^2+2\rho uv)-\frac{1}{2}u^2-\frac{1}{2}v^2} = e^{\rho uv} = \sum_{k\geq 0}\frac{(\rho uv)^k}{k!}.$$

En identifiant les développements en u et v, on conclut à (4).

5. On peut passer à la preuve du théorème, en supposant d'abord $f(x) = \sum_{k=1}^n a_k P_k(x)$ et $g(y) = \sum_{k=1}^n b_k P_k(y)$, pour certains coefficients $(a_k)_k$ et $(b_k)_k$. Alors

$$\mathbb{E}(f(X)) = \mathbb{E}(g(Y)) = 0$$

par (3). Par ailleurs,

$$\begin{aligned}
|\mathbb{C}\text{ov}(f(X),g(Y))| = |\mathbb{E}(f(X)g(Y))| &= |\sum_{k=1}^n a_k b_k \rho^k k!| \quad \text{(en utilisant (4))}\\
&\leq |\rho|\sum_{k=1}^n |a_k||b_k|k!\\
&\leq |\rho|\sqrt{\sum_{k=1}^n a_k^2 k!}\sqrt{\sum_{k=1}^n b_k^2 k!} = |\rho|\sqrt{\mathbb{V}\text{ar}(f(X))}\sqrt{\mathbb{V}\text{ar}(g(Y))}
\end{aligned}$$

en appliquant l'inégalité de Cauchy-Schwarz à la dernière ligne. Le théorème est démontré dans ce cas.

6. Esquissons le cas général.
   - Toute v.a. de carré intégrable peut se décomposer sur la base des polynômes d'Hermite :
   $$f(X) = \sum_{k=0}^{+\infty} a_k P_k(X), \qquad g(Y) = \sum_{k=0}^{+\infty} b_k P_k(Y).$$
   - Puisque la corrélation ne concerne pas le centrage, on peut toujours recentrer $f(X)$ et $g(Y)$ et supposer donc $a_0 = b_0 = 0$.
   - Il suffit maintenant de reprendre les calculs précédents en remplaçant $\sum_{k=1}^{n}\cdots$ par $\sum_{k=1}^{\infty}\cdots$.

Référence : H.O. Lancaster, *Some properties of the bivariate normal distribution considered in the form of a contingency table*, Biometrika 44, 289-292, 1957.