Известия НАН Армении, Математика, том 58, н. 3, 2023, стр. 78 – 83.

HELLINGER'S DISTANCE AND CORRELATION FOR A SUBCLASS OF STABLE DISTRIBUTIONS

M. T. MESROPYAN, V. G. BARDAKHCHYAN

Yerevan State University
E-mails: *mesropyanmesrop@list.ru; vardan.bardakchyan@ysu.am*

Abstract. We investigated correlation retrieval procedure from Hellinger's distance. We found monotone relation of Hellinger's distance and positive correlation in a sub-class of stable distributed random variables, with $\alpha > 1$ and $\mu = \beta = 0$. We implemented a technique suitable for the class of stable distributions, and showed that this positive relation holds even for the case of Levy distribution, i.e. $\alpha = 1/2$, $\beta = 1$ and $\mu = 0$.

MSC2020 numbers: 60E07; 62H20.
Keywords: Hellinger's distance; stable distributions; correlation coefficient.

1. Introduction and motivation

The problem of quantification of closeness to given distribution are analyzed using statistical (or probability) metrics and semi-metrics, such as Kolmogorov-Smirnov, KL-divergence etc ([1]). Closeness to normality is meaningful property for any set of random variables, as, thanks to central limit theorems, the wider the set, the closer to Gaussian random variable can be obtained from linear combinations of random variables considered (The converse procedure have applications in signal processing see for example [2]). In the paper ([3]) we analyzed method of constructing financial portfolio that is most Gaussian, based on the squared Hellinger's distance. The next step in that vein is to understand, to what extent adding new variable (new asset) may change the minimum distance found.

Hellinger's distance

Hellinger's distance ([4]) is a metric on the space of probability measures, defined

$$H(P,Q) = \left(\frac{1}{2}\int_{\Omega}(\sqrt{P}-\sqrt{Q})^2\right)^{1/2} \tag{1.1}$$

Given in random variables it quantifies the distance (dissimilarity) between them, and can be written as follows (here we give the continuous case)

$$H^2(X,Y) = \frac{1}{2}\int_{-\infty}^{\infty}(\sqrt{f_X(x)}-\sqrt{f_Y(x)})^2dx = 1-\int_{-\infty}^{\infty}\sqrt{f_X(x)f_Y(x)}dx \tag{1.2}$$

The notable property of Hellinger's distance is that it does not quantify the correlation between random variables as it does consider values taken itself, but rather the distributions.

Stated differently knowing $H^2(X,Y)$ one can say nothing about the correlation $\rho(X,Y)$.

Indeed if X and Y have the same distribution $H^2(X,Y)=0$, no matter are they independent or not.

However, if one consider, instead, $H^2(X,X+Y)$ much more is possible. Here we analyze the dependence structure of $H^2(X,Y)$ on $\rho(X,Y)$, and propose possible retrieval technique, for the subclass of stable distributions. Using triangle inequality and symmetry, we can thus state $H(X+Y,Z)\leq H(X,X+Y)+H(X,Z)$.

So we are interested in the $H(X,X+Y)$ as a bound for possible minimum. (The less is this increment, the more is confidence.)

Stable distributions

Class of stable distributions is a class with remarkable property of being closed (for each α-level) under summation of independent copies, i.e. if $X,Y\sim \mathrm{St}(\alpha)$ then $aX+bY\sim \mathrm{St}(\alpha)$ ([5]). The class is rather given with characteristic function (some of the members not having closed-form density function.)

Here we use the original (discontinuous in α) parametrization of the class.

$$\phi(t|\alpha,\beta,\mu,c)=\begin{cases} e^{it\mu-|ct|^{\alpha}\left(1-i\beta\tan\left(\frac{\pi\alpha}{2}\right)\mathrm{sign}(t)\right)}; & \alpha\in(0,1)\cup(1,2] \\ e^{it\mu-|ct|\left(1+\frac{2i\beta}{\pi}\ln|t|\mathrm{sign}(t)\right)}; & \alpha=1 \end{cases} \tag{1.3}$$

For some values of parameter, no expectation can be calculated, no Pearson correlation can be determined. Therefore, to understand the dependence of Hellinger's distance of correlation, we need to somehow determine correlation coefficient. Different regularization techniques can be considered to find the correlation. Here we analyze, rather, dependence on other parameter, which is directly connected with correlation. Thus, our motivation for the problem is twofold: first in connection with problem of financial portfolios properties, and second in definition of analogous measure to correlations using Hellinger's correlation. In the second section of the paper, we show the method of determination, and in third section we give main results for the subclass of stable distribution with special cases of Levy and normal distributions (some features of Hellinger's distance of Cauchy distribution are analyzed in [6]).

2. The determination method

We take the following approach of analyzing the correlations.

Let's consider Gaussian random variables X and Y with correlation coefficient ρ. As we have explicit form for Hellinger's distance between Gaussian random variables, and as $X+Y$ is also Gaussian, we can analyze the Hellinger's distance upon correlation directly.
More concretely

$$H^2(X, X+Y) = 1 - \sqrt{\frac{2\sigma_1\sqrt{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}}{2\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}} e^{\frac{1}{4}\frac{\mu_2^2}{2\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}} \tag{2.1}$$

Where $X \sim N(\mu_1, \sigma_1^2)$; $Y \sim N(\mu_2, \sigma_2^2)$.

The main result here, is that for positive correlations squared Hellinger's distance is either increasing or decreasing in ρ. And that there is only one possible minimum for negative correlations. In order to get similar result for more general α-stable distributions, we propose the following method. Given three random variables X_1,X_2,X_3 from the same family we construct

$$X = X_1 + X_2$$

$$Y = X_2 + X_3$$

Which are obviously non-independent (excluding the case when X_2 is non-random).

To analyze the dependence, we consider their squared Hellinger's distance

$$H^2(X, X+Y) = H^2(X_1 + X_2, X_1 + 2X_2 + X_3)$$

We want to analyze the dependence of Hellinger's distance of correlation coefficient. For that purpose, without changing the distributions of X and Y, we change their correlations.

$$X(k) := X_{1k} + kX_2$$

$$Y(k) := X_{3k} + kX_2$$

Taking X_{1k} and X_{3k} such that $X(k)$ and $Y(k)$ have the same distribution as X and Y respectively [1]. We hope Hellinger's distance will sense the change in k, in some way. More precisely, we expect monotonic dependence on k, for positive values.

Remark 2.1. . *Having distributions of X and X+Y does not guarantee explicit calculation for correlation coefficient. Any type of correlations can be modelled as above for stable distributions family, so hopefully one can assess "correlatedness" by means of Hellinger's distance.*

[1]Note that the same could be done by defining $X(\alpha) = (1-\alpha)X_{1\alpha} + \alpha X_2$ and $Y(\alpha) = (1-\alpha)X_{3\alpha} + \alpha X_2$. We used the above method for simplicity.

3. Main results

Main result can be formulated by the following theorem.

Theorem 3.1. *For stable distributions with $\alpha > 1$ and $\mu = \beta = 0$, Hellinger's distance $H^2(X(k), X(k) + Y(k))$ is monotonic in k.*

Proof. Instead of Hellinger's distance let's consider

$$\begin{aligned}&\int_{-\infty}^{\infty}\sqrt{f_{X(k)(x)}f_{X(k)+Y(k)}(x)}dx\\&\quad=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\sqrt{f_{X(k)(x)}\int_{-\infty}^{\infty}e^{-itx}e^{-(|c_{1k}|^\alpha+|2kc_2|^\alpha+|c_{3k}|^\alpha)|t|^\alpha}dtdx}\end{aligned}\tag{3.1}$$

Where we used the fact that $X(k) + Y(k) = X_{1k} + X_{3k} + 2kX_{2k}$ and that each of them are independent. Note also that the distribution of $X(k)$ does not change with k so we wrote explicitly only $f_{X(k)+Y(k)}(x)$. As we must have no change in distribution of $X(k)$ and $Y(k)$, the following relations must be satisfied

$$|c_{1k}|^\alpha + |c_2|^\alpha = |c_1|^\alpha + |c_2|^\alpha \tag{3.2}$$

$$|c_{3k}|^\alpha + |c_2|^\alpha = |c_3|^\alpha + |c_2|^\alpha \tag{3.3}$$

Thus we have

$$\begin{aligned}&\int_{-\infty}^{\infty}\sqrt{f_{X(k)(x)}f_{X(k)+Y(k)}(x)}dx\\&\quad=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\sqrt{f_{X(k)(x)}\int_{-\infty}^{\infty}e^{-itx}e^{-(|c_1|^\alpha+|c_3|^\alpha+(2^\alpha-2)k^\alpha|c_2|^\alpha)|t|^\alpha}dtdx}\end{aligned}$$

Next, as characteristic function depends only on $|t|$, it is even function, so we can take only cosine part of e^{itx}, as the sine part integrate to 0 (see [7]).

$$\begin{aligned}&\int_{-\infty}^{\infty}\sqrt{f_{X(k)(x)}f_{X(k)+Y(k)}(x)}dx\\&\quad=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}\sqrt{f_{X(k)(x)}\int_{0}^{\infty}cos(tx)e^{-(|c_1|^\alpha+|c_3|^\alpha+(2^\alpha-2)k^\alpha|c_2|^\alpha)|t|^\alpha}dtdx}\end{aligned}$$

We take derivative with respect to k, and analyze the sign of numerator

$$S := \int_0^\infty cos(tx)e^{-itx}e^{-(|c_1|^\alpha+|c_3|^\alpha+(2^\alpha-2)k^\alpha|c_2|^\alpha)|t|^\alpha}(2^\alpha-2)k^{\alpha-1}t^\alpha dt \tag{3.4}$$

Denoting $\theta(k) := |c_1|^\alpha + |c_3|^\alpha + (2^\alpha - 2)k^\alpha|c_2|^\alpha$, we rewrite

$$\begin{aligned}S&=\alpha(2^\alpha-2)k^{\alpha-1}\int_0^\infty cos(tx)e^{-\theta(k)t^\alpha}t^\alpha dt\\&=\alpha(2^\alpha-2)k^{\alpha-1}\sum_{j=1}^{\infty}\frac{(-1)^j}{(2j)!}x^{2j}\int_0^\infty t^{2j}e^{-\theta(k)t^\alpha}t^\alpha dt\end{aligned}\tag{3.5}$$

Changing the variable by $\theta(k)t^\alpha = z$, we get the following

$$\begin{aligned} S &= (2^\alpha - 2)k^{\alpha-1}\sum_{j=1}^{\infty}\frac{(-1)^j}{(2j)!}x^{2j}\left(\frac{1}{\theta(k)}\right)^{\frac{2j+1}{\alpha}}\int_0^\infty z^{\frac{2j+1}{\alpha}+1-1}e^{-z}dt \\ &= (2^\alpha - 2)k^{\alpha-1}\sum_{j=1}^{\infty}\frac{(-1)^j}{(2j)!}x^{2j}\left(\frac{1}{\theta(k)}\right)^{\frac{2j+1}{\alpha}}\Gamma(\frac{2j+1}{\alpha}+1) \end{aligned} \tag{3.6}$$

It only remains to state the above series are convergent. Indeed with $\alpha > 1$, $\frac{2j+1}{\alpha} < 2j$ starting from $j > \frac{1}{2(\alpha-1)}$. Note also that $\frac{\Gamma(\frac{2j+1}{\alpha}+1)}{(2j)!}$ will converge to 0 faster than $\frac{(\theta(k))^{\frac{2j+1}{\alpha}}}{x^2 j}$ no matter what x is taken. So the relation will be determined by the first several terms of sum. This completes the proof. □

Remark 3.1. *Note that if $\alpha < 1$ the series may diverge, thus we have no right to exchange integral and sum signs.*

Levy case

Considering the case of Levy distribution. i.e. the case where $\alpha = 1/2$, $\beta = 1$, we came to the similar result for Levy's starting from 0. Indeed

$$X \sim Levy(\mu, c) \iff f_X(x) = \frac{\sqrt{c}}{\sqrt{2\pi}}\frac{e^{-\frac{c}{2(x-\mu)}}}{(x-\mu)^{3/2}}$$

taking

$$\begin{gathered} X_i \sim Levy(0, c_i), i = \overline{1,3} \\ X = X_1 + X_2; Y = X_2 + X_3 \\ X \sim Levy(0,(\sqrt{c_1} + \sqrt{c_2})^2); Y\ Levy(0, (\sqrt{c_2} + \sqrt{c_3})^2) \end{gathered} \tag{3.7}$$

Defining

$$X(k) = X_{1k} + kX_2; Y(k) = kX_2 + X_{3k}$$

For $X(k)$ and $Y(k)$ to have the same distribution, the following relations must take place

$$\sqrt{c_{1,k}} = \sqrt{c_1} + \sqrt{c_2} - \sqrt{kc_2} > 0;\ \sqrt{c_{3,k}} = \sqrt{c_3} + \sqrt{c_2} - \sqrt{kc_2} > 0 \tag{3.8}$$

In that case

$$\begin{aligned} &H^2(X(k), X(k) + Y(k)) \\ &= 1 - \sqrt{2}\frac{\sqrt{(\sqrt{c_1} + \sqrt{c_2})(\sqrt{c_1} + 2\sqrt{c_2} + \sqrt{c_3} - 2\sqrt{kc_2} + \sqrt{2kc_2})}}{\sqrt{(\sqrt{c_1} + \sqrt{c_2})^2 + (\sqrt{c_1} + 2\sqrt{c_2} + \sqrt{c_3} - 2\sqrt{kc_2} + \sqrt{2kc_2})^2}} \end{aligned} \tag{3.9}$$

Which is obviously monotone increasing in k in the region where k can take values in accordance with (3.8) Note that in the case of Levy distribution the correlation coefficient can't be calculated directly, as neither expectation, nor variance are finitely determined. So to show that k is indeed closely related to correlation we

exploit regularization techniques with Esscher transform ([8]). Instead of X_i, we define X_{ih} using Esscher transform as follows

$$f_{X_{i,h}} = \frac{\sqrt{c_i}}{\sqrt{2\pi}} \frac{e^{-\frac{c_i}{2x}}}{x^{3/2}} \frac{e^{-hx}}{\int_0^{+\infty} \frac{\sqrt{c_i}}{\sqrt{2\pi}} \frac{e^{-\frac{c_i}{2x}}}{x^{3/2}} e^{-hx} dx} = \frac{\frac{e^{-\frac{c_i}{2x}}}{x^{3/2}} e^{-hx}}{\int_0^{+\infty} \frac{e^{-\frac{c_i}{2x}}}{x^{3/2}} e^{-hx} dx} \tag{3.10}$$

Next one computes correlation coefficient and let's h tend to 0 ([9]).

Using (3.10) and computing $X_h(k), Y_h(k)$, we get the following correlation

$$\begin{aligned} &\rho(X_h(k), Y_h(k)) = \\ &\frac{k^2 Var(X_{2h})}{\sqrt{k^2 Var(X_{1h})Var(X_{2h}) + Var(X_{1h})Var(X_{3h}) + k^2 Var(X_{2h})Var(X_{3h})}} \\ &= \frac{k^2\sqrt{c_2}}{\sqrt{k^2\sqrt{c_1c_2} + \sqrt{c_1c_3} + k^2\sqrt{c_2c_3}}} = \frac{\sqrt{c_2}}{\sqrt{\frac{1}{k^2}\sqrt{c_1c_2} + \frac{1}{k^4}\sqrt{c_1c_3} + \frac{1}{k^2}\sqrt{c_2c_3}}} \end{aligned} \tag{3.11}$$

Where we have computed variance with the following formula

$$Var(X_{i,h}) = \frac{\sqrt{c_i}}{2\sqrt{2}h^{3/2}} \tag{3.12}$$

Note that already h disappeared, even before taking the limit. Hence taking $h \to 0$, will wake no changes in (3.11).

Finally note that correlation is increasing function of k.

Список литературы

[1] S.T. Rachev, L. Klebanov, S. V. Stoyanov, F. Fabozzi, The Methods of Distances in the Theory of Probability and Statistics, Springer (2013).

[2] Aap. Hyvärinen, J. Karhunen and Er. Oja, Independent Component Analysis, 1st ed: Wiley (2001).

[3] M. Mesropyan, V. Mkrtchyan, "Assessing nromality of group of assets based on portfolio construction", Alternative, **3**, 14 – 21 (2021)

[4] L. Friedrich; M. Klaus-J., Statistical Decision Theory: Estimation, Testing, and Selection, Springer (2008).

[5] W. Linde, Probability in Banach Spaces: Stable and Infinitely Divisible Distributions, John Wiley & Sons (1986).

[6] F. Nielsen, K. Okamura, "On f-divergences between Cauchy distributions", arXiv:2101.12459, Submitted on 29 Jan 2021.

[7] L. Debnath, D. Bhatta, Integral Transforms and Their Applications, Chapman & Hall/CRC (2007).

[8] H. U. Gerber, "A Characterization of Certain Families of Distributions via Esscher Transforms and Independence", Journal of the American Statistical Association 1980, Springer, **75**, 372, 1015 – 1018 (2005).

[9] J.-Ph. Aguilar, C. Coste, H. Kleinert, J. Korbe, "Regularization and analytic option pricing under alpha-stable distribution of arbitrary asymmetry", Papers 1611.04320, arXiv.org, revised Nov (2016).

Поступила 12 декабря 2022

После доработки 10 марта 2023

Принята к публикации 15 марта 2023