ԲԱՆԲԵՐ ԵՐԵՎԱՆԻ ՀԱՄԱԼՍԱՐԱՆԻ. ՏՆՏԵՍԱԳԻՏՈՒԹՅՈՒՆ
2023. № 1. 64-71 *Տնտեսամաթեմատիկական մոդելավորում*
https://doi.org/10.46991/BYSU:G/2023.14.1.064

HELLINGER'S DISTANCE TO NORMAL DISTRIBUTION AS MARKET INVARIANT

MESROP MESROPYAN, VARDAN BARDAKHCHYAN

Main purpose of distance based portfolio constructions is for portfolio imitation. Here we used distance from normal distribution for other purpose. We attempted to find static market invariant within possible linear combinations of given random variables. We conjectured that "closeness" to normal distribution of possible portfolios in market may reliably represent market microstructure with possible correlations between assets. Taking the squared Hellinger's distance, we sought for each level of desired mean return the portfolio whose return distribution is closest to Gaussian, with variance taken from efficient frontier found by initially solving mean-variance problem. We found that minimal distance differs significantly from market to market. The sensitivity check showed small average sensitivity for 5% change of a 5% portion of data, small sensitivity for adding new variables in simulated market, and extreme sensitivity to bin numbers. Though distance to normality differed among markets, its sensitivity being small enough in average sometimes showed extreme changes.

Keywords: *Market invariant, Sensitivity analysis, Hellinger's distance, distance to normal distribution, financial portfolio theory, simulated market, Student's distribution*

1. Introduction and motivation

The hunt of market invariants has motivated authors to incorporate various interdisciplinary techniques.

All of them use slightly different definitions of market invariant, use different techniques to derive them, and for different purposes. Invariants in general are quantities which are not affected when some transformation of given class occur. However, in financial data they are perceived slightly differently. Some authors define market invariants to be random variables or random vectors the distribution of which does not change over time. Some authors treat market invariants as variables which are enough robust either remaining in small interval when time changes, or experiencing small or no change when data is transformed (e.g. scaled).

Also there are several purposes for determining market invariant. For example (Reddy and Sebastin 2008) used market invariants to determine market manipulation. Some authors used market invariants as a tool of market segmentation by their characteristics, i.e different values of market invariants corresponded to different segments of market.

Some authors define market invariants as universal constants.

Latest theory proposed by Kyle and Obizhaeva[1] take axiomatic approach to

[1] See (Kyle and Obizhaeva, Market microstructure invariance: Empirical hypotheses 2016)

define one such possible constant, which later proved good enough empirically[2].

They proposed $\frac{R}{N^{3/2}}$, with R total risk traded, and N number of daily bets (or bettors). This universal constant is applicable to a market in whole.

Other authors use more intuitive approach to finding invariants. Reddy and Sebastian[3] proposed as market invariant entropy. Their motivation was based on dynamical systems theory. Though they name two other possible measure used to identify (describe) non-linear system, namely Lyapunov exponent and correlation dimension. Pincak[4] analyzed foreign exchange markets (namely EUR/USD) series by means of string theory. Their inspiration was from later developments in physics. As invariant they proposed weighted sum of one lag correlations between returns in given time interval with weights following Gibbs distribution.

In these last two papers the market invariant is invariant of specific time series. On the other hand, Borghesi et.al.[5] considered correlations between log-returns of different groups of assets. After subtracting mean log-return, the correlation between new series was checked for invariance. (Four other form of correlations were analyzed and compared).

Our purpose of defining possible market invariant is pure for diagnostic use and differentiating between several markets. Here we treat market invariant as some measure specific to given market (given class of financial securities) that is enough robust to changes in data. Our analysis is based on portfolio construction, and as such techniques described here cannot be applied to one security time series.

Our analysis is based upon the consideration that log-returns are indeed invariants, most of authors take this as granted[6]. So we do not perform time series analysis or any other kind of modeling, and treat the log-returns of one time-series as i.i.d r.v.-s.

The main idea is to construct portfolio closest in distribution to Gaussian.

So we do make use of statistical distances. For each level of expected return, we find the minimum distance portfolio (expressed in portfolio weights). And among the collection of these "best" portfolios we determine one which has minimum of all distances. The distance found will be the measure (we believe to be a robust), differing from market to market.

In some sense the invariant estimated shows non-gaussianity of the market considered. Thus it is alike Negentropy[7] but with crucial differences that we hold constant only expectation and we take distance of linear combination of a strictly bounded set of random variables.[8]

---

[2] See (Bucci, et al. 2020, Kyle, Obizhaeva and Tuzun, Microstructure invariance in U.S. stock market trades 2020)

[3] See (Reddy and Sebastin 2008)

[4] See (Pincak 2013)

[5] See (Borghesi, Marsili and Micciché 2007)

[6] See for example (Attilio 2005)

[7] For details see in book (Hyvärinen, Karhunen and Oja 2001)

[8] We don't use Ex-Kurtosis measure as well due to the drawbacks described in (Hyvärinen, Karhunen and Oja 2001).

Here we use the Hellinger's distance[9] for simplicity of its form, though other statistical distances can also be used.

The use of normal distribution is inspired by central limit theorem (CLT), and thus by the hope that big markets will eventually yield these distances to be 0. The second reason to use normal distribution is based on pure practical considerations. The class of elliptic distributions is easier to analyze, as two parameters (mean and variance) is enough to describe portfolio and it is easier to make any form of predictions.

Market invariant based on portfolio construction means that we analyze properties of linear combinations (in our case convex combinations, with non-negative weights), instead of using any other functional form of given data. One reason for this is obvious: generally trading is linear combination of securities. If something is internally constant in linear combination of given securities returns, it will be useful to derive it explicitly.

The other reason is that constructing portfolio in the manner is pure static procedure (though based on historical data, when we analyze distributions), so if something remains close to constant, it wouldn't vary much when time passes. So internally formulation in form of static optimization is preferred without regard to specific future time horizons. While most authors take invariant to be internal characteristic of one or several time series over specific time intervals, we prefer to analyze one time characteristics.

We conduct sensitivity analysis on pure empirical basis, to check robustness of the distance. The techniques used can be treated as solving portfolio optimization problem, where squared Hellinger's distance is used as risk measure. As thus sensitivity analysis of minimal squared Hellinger's distance is in fact portfolio sensitivity analysis (like ones for mean-variance or VaR measures[10]). As Hellinger's distance is formulated by probability measures (distribution functions), the matrix perturbation techniques used by Best and Grauer[11] cannot be applied.

The paper is organized as follows. In section 2 the general problem of constructing portfolio with minimal Hellinger distance from normal distribution is stated, computational techniques are showed and result are briefly summarized. In section 3 result of sensitivity analysis are shown. The paper ends with discussion and conclusion.

2. Hellinger's distance based portfolio construction

Classical formulation of mean-variance portfolio is the following.

[9] For general approach to probability we refer to (Svetlozar, et al. 2011)

[10] Two main papers on this topic are (Gourieroux, Laurent and Scaillet 2000., Stoyanov, Rachev and Fabozzi 2013)

[11] See the seminal paper by authors (Best and Grauer 1991)

$$\begin{cases} E(X) = \bar{e} \\ Var(X) \to min \\ X = \sum X_i \\ w_i \geq 0 \\ \sum w_i = 1 \end{cases} \quad (1)$$

Solving the system above one get the portfolio with minimum variance for each level of expected return. X_i-s represent returns of each asset. We confine ourselves to the case of logarithmic return. The weights w_i-s represent the percentage of money in each asset. We assume only long positions i.e. $w_i \geq 0$.

Our goal of constructing portfolio with minimum Hellinger's distance (or squared one) can be formulated by following problem.

$$\begin{cases} E(X) = \bar{e} \\ H^2(X,N) \to min \end{cases} \quad (2)$$

Where $H(X,Y)$ – is Hellinger's distance between X, Y r.v.-s.

Here we use (X,Y) , $H(F_X,F_Y)$ and in absolutely continuous case $H(f_X,f_Y)$ interchangeably.

In other words, Hellinger's distance should be the least from normally distributed random variable. Parameters of desired normal distribution are got from solving (1) problem. So More generally we solve the following.

$$\begin{cases} E(X) = \bar{e} \\ H^2\left(X, N(e, Var(X^*(e))\right) \to min \\ w_i \geq 0 \\ \sum w_i = 1 \end{cases} \quad (3)$$

with $X^*(e)$ being solution of (1).

After solving (3) we take minimum of all Hellinger distances $(H^*)^2 = \min_e \left(H^2(e)\right)$.

For continuous case

$$H^2(f,g) = 1 - \int \sqrt{f(x)g(x)}dx \quad (4)$$

To solve the problem with discrete data we do binning process[12], which will yield by to continuous distribution represented by simple function.

The distance between these kind of distributions will take the following form:

$$\int_{a_j}^{a_{j+1}} \sqrt{f_Y(x)} = \frac{1}{\sigma^{\frac{1}{2}}(2\pi)^{\frac{1}{4}}} \int_{a_j}^{a_{j+1}} e^{\frac{(x-\mu)^2}{4\sigma^2}} dx \xrightarrow{\sigma_1 = \sigma\sqrt{2}} \sqrt{2}\sqrt{\sigma}(2\pi)^{\frac{1}{4}} \left(F_N(a_{j+1}|\mu,\sigma_1^2) - F_N(a_j|\mu,\sigma_1^2)\right) \quad (5)$$

Where X is the distribution got by binning process, thus take fixed values on subintervals of $[a,b)$. Let $[a_j,a_{j+1})$ be exactly the intervals where $f_X(x)$ take fixed values ($j = \overline{0,n-1}$), with $a_0 = a$ and $a_n = b$. Denoting those val-

[12] We exactly use binning, instead of kernel methods as of simplicity of calculations, and one more reason. The kernel methods have one more parameter which (as we checked) will add sensitivity to given solutions. The method is sensitive to what kernel is used.

ues by O_j for each interval, we will have the following form

$$H^2(X,Y) = 1 - \sum_{j=0}^{n-1} O_j \int_{a_j}^{a_{j+1}} \sqrt{f_Y(x)}dx \quad (6)$$

Generally O_j-s are any numbers, not bound to be different. We take the interval cut enough fine to have one value for each interval.

Whenever the counterpart's (Y's) distribution is normal (i.e. $f_Y(x)$ is density of normal distribution), we will the following:

$$\int_{a_j}^{a_{j+1}} \sqrt{f_Y(x)} = \frac{1}{\sigma^{\frac{1}{2}}(2\pi)^{\frac{1}{4}}} \int_{a_j}^{a_{j+1}} e^{\frac{(x-\mu)^2}{4\sigma^2}} dx \xrightarrow{\sigma_1=\sigma\sqrt{2}} \sqrt{2}\sqrt{\sigma}(2\pi)^{\frac{1}{4}} \left(F_N(a_{j+1}|\mu,\sigma_1^2) - F_N(a_j|\mu,\sigma_1^2)\right) \quad (7)$$

We noted that for different types of market the minimum squared Hellinger's distance is different.

More formally we found the following result for 3 different markets - Stock market of specific types of stock, exchange market and purely simulated market of correlated student's distributions with different degrees of freedom.

We used 4 randomly generated Student's distributions with 4, 3, 3 and 2 degrees of freedom. We got correlated variables with Cholesky decomposition of randomly taken correlation matrix. The data was taken to be 810 data, for each (approx. 3 years). We took 4 stocks of nearly the same branch (Cisco, Intel, Microsoft, NVIDIA), for the period of 3 years. And lastly we analyzed foreign exchange market data. We took data up to 7th month of 2020, of approximately 3 years (After that war started at Nagorno-Karabakh, which influenced our money market, and exchange rates).[13]. Squared Hellinger's distance for each market are shown in **Table I**.

Table 1

Squared Hellinger's distance calculated for 3 different markets

	Simulated data market	Stock Market	Exchange market
H^2	0.0127	0.027	0.0149

3. Sensitivity Analysis

Generally, under sensitivity analysis in optimization problems it is meant sensitivity checking depending of initial conditions or parameters. In this general form sensitivity of our problem seems infeasible as we deal with general distributions. So we checked sensitivity by randomly "editing" data.

We did 2 types of sensitivity check.

First we changed randomly taken portion of data, by random but restricted magnitude. We then computed percentage change, per 5 (smaller choices make to minor differences) of changed data. Here we report both absolute values of change and percentage.

Next we checked effect of changes in binning number. All the analysis was done by Python 3.10. We used Excel 2019 (for data preprocessing).

For the last part it is reasonable to expect that adding new variable will decrease the minimal Hellinger distance by CLT. However, when we speak about initial small number of variables it is not straightforward with adding one new.

[13] The data is in open access, so anybody can conduct the same analysis (we made use of Central bank of Armenia cba.am provided data for exchange market).

For binning procedure change of number of bins must be expected to bring to extreme changes in Hellinger.

For the first two procedures nothing exact can be expected.

We randomly changed a portion of initial data by random amount up to 1000 times and calculated average change in each case. The sensitivity in percentages and respective average absolute values are shown in **Table II**.

Table 2

Change in squared Hellinger due to change of 5% of data of maximum 5% of their value

	Simulated data	Stock	ForEx
Percentages	3.02%	1.01%	5.51%
Absolute values	0.00039125	0.000299	0.00092

We got the result for binning procedure and adding new variable (with student distribution with 3 degree of freedom, correlations chosen at random between (-1,1) but confining to correlation matrix remaining positive definite) for simulated data presented in **Table III**.

Table 3

Change in Hellinger due to change in bin number by 1 or adding new variable (results stated in percentages)

	Simulated data market
Binning	~10.1%
New variable	~2.04%

4. Discussion and conclusion

We have conducted a general sensitivity analysis for the minimum squared Hellinger's distance for 3 different financial markets.

There are several reasons to cast doubt about the results presented.

This sensitivity analysis lacks statistical background, i.e. we can't check statistically whether the results found reveal the true picture. However, we have conducted big enough number of simulation to check.

Second, the optimization procedure is not smooth one. Thus the initial guess of weights had crucial role. (We checked for enough fine grid, but we didn't combine finite difference techniques).

But generally the procedure revealed that only substantial changes can bring to changes minimum squared Hellinger's distance. Though we found generally less than 5% change in response to 5% change of initial data, one should note that we considered 5% change in only 5% of data. Anyway, small changes had extremely less effect. Obviously adding some dominant (in mean-variance sense) and normally distributed random variables to the market, it would drug the Hellinger down. So the measure proposed is showing how much efficient frontier is composed of close to normal elements.

Two things to note, is that obviously if all returns in market considered are normally distributed, one will get 0 Hellinger. However, though it can be supposed, that the bigger the market, the smaller Hellinger it would have, there are some source of doubt about it.

As we consider the distance to normal with mean and variance that of efficient frontier. It is generally known that Markowitz's mean-variance portfolio

tends to use only small number of instrument for each level mean return. If this small number can be combined to get closer to normal, the Hellinger will decrease substantially, otherwise the changes will be not that big.

Much further work should be done to find out whether the minimal Hellinger is indeed market invariants or not, both empirical and theoretical. One must keep in mind that this is not smooth (or even continuous) optimization problem, so convergent iterative methods will not be best choices.

For nonlinear combinations uniquely (or close to uniquely) representing market more steps are needed to be taken.

ՄԵՍՐՈՊ ՄԵՍՐՈՊՅԱՆ, ՎԱՐԴԱՆ ԲԱՐԴԱԽՉՅԱՆ – ***Հելինգերի հեռավորությունը նորմալ բաշխումից՝ որպես շուկայական ինվարիանտ*** – Հավանականության հեռավորությունների վրա կառուցվող արժեթղթերի պայուսակների հիմնական նպատակը այլոց պայուսակների իմիտացիան կամ կրկնօրինակումն է: Հոդվածում օգտագործվել են դրանք այլ նպատակով: Փորձ է արվել գտնելու ստատիկ շուկայական ինվարիանտ՝ տրված պատահական մեծությունների գծային կոմբինացիաներից: Առաջ է քաշվում այն վարկածը, որ հնարավոր պայուսակների նորմալ բաշխմանը «մոտ լինելը» թույլ կտա թվայնացնել շուկայական միկրոկառուցվածքը՝ արտացոլելով արժեթղթերի միջև կորելյացիաները: Նախապես լուծելով Մարկովիցի միջին վարիացիայի օպտիմիզացման խնդիրը՝ սպասվող եկամտաբերության ցանկացած մակարդակի համար փնտրվում է այն պայուսակը, որի եկամտաբերության բաշխումը ամենամոտն է Գաուսի այն բաշխումին, որի պարամետրերը վերցվում են միջին վարիացիա խնդրի արդյունավետ սահմանագծից: Այդ նպատակով օգտագործելով Հելինգերի հեռավորության քառակուսին՝ հեղինակները պարզել են, որ նվազագույն հեռավորությունը տարբեր շուկաներում էականորեն տարբեր է: Կատարվել է զգայունության վերլուծություն, որը ցույց է տվել, որ դիտարկվող հեռավորության չափը քիչ զգայուն է տվյալների 5%-ի 5% առավելագույն շեղման հանդեպ, քիչ զգայուն է շուկայի նոր պատահական մեծության ավելացման նկատմամբ և էականորեն զգայուն է զամբյուղների (հիստոգրամային) քանակի հանդեպ: Այդուհանդերձ, թեև նորմալ բաշխումից հեռավորությունը էականորեն տարբեր է ըստ շուկաների և ունի քիչ զգայունություն, այն երբեմն ցույց է տվել էական տարբերություններ՝ կախված նախնական պարամետրերի փոփոխություններից:

Բանալի բառեր - *շուկայական ինվարիանտ, զգայունության վերլուծություն, Հելինգերի հեռավորություն, նորմալ բաշխումից հեռավորություն, ֆինանսական պայուսակների տեսություն, սիմուլացված շուկաներ, Մթյուդենտի բաշխում*

МЕСРОП МЕСРОПЯН, ВАРДАН БАРДАХЧЯН – ***Расстояние Хеллингера от нормального распределения как рыночный инвариант.*** – Основная цель использования вероятностных метрик в теориях портфеля – имитация портфелей остальных участников рынка. Здесь мы используем расстояния от нормальных распределений в иных целях. Мы ищем рыночный инвариант среди всех линейных комбинаций данных случайных величин (прибыльности ценных бумаг). Мы предположили, что дальность от нормального распределения может считаться

кандидатом, которая характеризует рыночную микроструктуру, принимая во внимание все возможные корреляции между инструментами. Мы вычисляем квадрат расстояния Хеллингера данных портфелей от Гауссовского распределения для каждого уровня ожидаемой прибыльности, и вариацией, полученной от предварительного решения задачи Марковица, для данного уровня прибыльности. Мы установили, что такая дальность сильно отличается от рынка к рынку. Мы провели анализ чувствительности, который показал, низкую среднюю чувствительность к изменению порции данных (5% данных, с максимально 5%-и изменениями), низкую среднюю чувствительность к добавлению новой случайной переменной, и высокую чувствительность к бинированию (количеству разрядов в гистограмме). Хотя, в среднем, изучаемое расстояние показывает низкую чувствительность, иногда имелись резкие изменения от изменений начальных параметров.

Ключевые слова: *рыночный инвариант, анализ чувствительности, расстояние Хеллингера, расстояние от нормального распределения, теория финансовых портфелей, смоделированный рынок, распределение Стьюдента*