STATISTICS IN TRANSITION new series, December 2020

Vol. 21, No. 5, pp. 99–118, DOI 10.21307/stattrans-2020-057

Received – 26.03.2020; accepted – 14.05.2020

Modelling bid-ask spread conditional distributions using hierarchical correlation reconstruction

Jarosław Duda[1], Henryk Gurgul[2], Robert Syrek[3]

ABSTRACT

While we would like to predict exact values, the information available, being incomplete, is rarely sufficient - usually allowing only conditional probability distributions to be predicted. This article discusses hierarchical correlation reconstruction (HCR) methodology for such a prediction using the example of bid-ask spreads (usually unavailable), but here predicted from more accessible data like closing price, volume, high/low price and returns. Using HCR methodology, as in copula theory, we first normalized marginal distributions so that they were nearly uniform. Then we modelled joint densities as linear combinations of orthonormal polynomials, obtaining their decomposition into mixed moments. Then we modelled each moment of the predicted variable separately as a linear combination of mixed moments of known variables using least squares linear regression. By combining these predicted moments, we obtained the predicted density as a polynomial, for which we can e.g. calculate the expected value, but also the variance to determine the uncertainty of the prediction, or we can use the entire distribution for, e.g. more accurate further calculations or generating random values. 10-fold cross-validation log-likelihood tests were conducted for 22 DAX companies, leading to very accurate predictions, especially when individual models were used for each company, as significant differences were found between their behaviours. An additional advantage of using this methodology is that it is computationally inexpensive; estimating and evaluating a model with hundreds of parameters and thousands of data points by means of this methodology takes only a second on a computer.

Key words: machine learning, conditional distribution, bid-ask spread, liquidity.

JEL Classification: C49, C58, G15.

1. Introduction

Liquidity is one of the key measures of financial market quality. The notion liquidity denotes a desirable function that should reflect a well-organized financial market. By liquid market we understand a market for which there exists a prompt and secure channel between the supply and demand of assets accompanied by low transaction costs. Providing a rigorous scientific definition of market liquidity happens to be a challenging aim. Liquidity is the main index of the health of a given stock market and the condition of the associated investment industry, using funds from this stock market. It is clear that more active trading leads

[1]Jagiellonian University, Poland. E-mail: jaroslaw.duda@uj.edu.pl.
ORCID: https://orcid.org/0000-0001-9559-809X.

[2]AGH University of Science and Technology, Poland. E-mail: henryk.gurgul@gmail.com.
ORCID: https://orcid.org/0000-0002-6192-2995.

[3]Jagiellonian University, Poland. E-mail: robert.syrek@uj.edu.pl.
ORCID: https://orcid.org/0000-0002-8212-8248.

to lower trading costs, more intensive flows of information and more activity concerning the relevant stocks displayed by potential investors. It is worth mentioning the role of speculators who can significantly increase the liquidity of the market, but may not necessarily have a positive impact on it.

In some recent contributions the definitions of market liquidity are based on the bid-ask spread and an estimation of its components. However, the difference between bid and ask quotes for an asset provides a liquidity measure with respect to a dealer market. Not to a broker market. Nevertheless, it is possible to compute approximations that replicate the difference between bid and ask quotes even in broker markets. Therefore, intradaily measures of liquidity can describe the main feature of a market, such as the arrival of new information in the hands of market participants. There are several definitions of liquidity. In each study on liquidity the initial goal is to formulate a definition of liquidity and justify it. The notion liquidity is related on the one hand to the transaction time - i.e. the duration of transactions, and on the other to transaction costs, understood as the price paid by investors for the supply of liquidity.

The common definition widely used by both researchers and market participants states that an asset is liquid if it can be sold quickly at a minimal cost. This definition of liquidity for a particular asset can be generalized for the whole market. A similar definition can also be applied to the stock market as a whole. In this sense, a market is liquid if it is possible to buy and sell assets at a minimal cost without a significant delay from the placement of the order. When assessing the liquidity of the stock market, in relation to incurring the lowest transaction costs, it is also important to take into consideration other elements than the size of the spread, which affect the cost of concluding buy/sell transactions, such as commissions and exchange fees; or taxation on capital gains; market volatility. However, in this contribution we focus on the spread which reflects to some extent the listed factors.

In the literature different measures of asset liquidity are known. These measures of liquidity take into account various alternative elements of the measurement approach. Some measures focus on the trading volume while other indices are based on the execution-cost relation of liquidity. The measures related to volume information reflect the price impact of transactions. After combining them into scalar measures they denote the liquidity on the whole market. However, the indices based on execution costs enable the properties of an asset to be evaluated. This is possible by analyzing the cost paid to the market maker (dealer or specialist) for matching the supply and demand.

The value added of this study is twofold. First of all, in order to find the characteristics of the future bid-ask spread we use a new methodology that has not been used for a financial time series before. Secondly, on the basis of empirical data from the German stock index DAX we have confirmed the advantages of this approach.

The most important conclusions concerning liquidity are based on the bid-ask spread and its variations. We aim to use our hierarchical correlation reconstruction (HCR) methodology in spread bid-ask description and forecasting. A more detailed outline of the advantages of this new methodology is at the end of the next Section. The content of the paper is organized as follows. In the next Section, the literature overview is presented. The third Section includes data and methodology. In the fourth Section, the empirical results are presented. The last Section provides conclusions.

2. Literature review

The pioneer in estimation of bid-ask spread, most often used measure of liquidity, was Roll Roll (1984). The model derived by Roll has been very useful tool of bid-ask description since the mid-eighties. The followers tried to improve and extend this approach. In Roll's model the spread is approximated based on return autocovariance.

According to Butler et al. (2005), lower liquidity implies higher transaction costs if the share capital increases. Moreover, a higher return on equity, or cost of equity, is expected.

Lesmond et al. (1999) belong to the first researchers who tested the quality of measures of stock liquidity. The contributors compared them based on different stocks. Bid-ask spread was used as a benchmark measure. Armitage et al. (2014) in contribution based on empirical data for Ukraine (2005-2006) found that the proportion of nontrading days, the proportion of zero-return days, stock volatility, and measure of Amihud (2002) exhibit high correlations with this spread. In conclusion the contributors stated that these indicators are good enough to measure liquidity for Ukraine. The findings of Armitage et al. (2014) regarding turnover are in line with those of Lesmond et al. (1999) for other emerging markets. In addition, they found that the proportion of zero-return days is a better measure for emerging markets than for developed markets.

In their studies of the determinants of the cost of trading, Armitage et al. (2014), Stoll (2000), Naik and Yadav (2003) and Gajewski and Gresse (2007) proved that the effective bid-ask spreads mentioned above depend on stock liquidity. Stock liquidity was measured by the number of non-trading days per year and the average number of trades per day. It turned out that higher liquidity stocks had narrower bid-ask spreads, as assumed. In the opinion of these and other scholars these effective spreads are related to the risk of the stock. The last is measured by return volatility. The more risky stocks exhibit usually wider bid-ask spreads. However, the opposite relationship between cost and trade size was observed for dealership markets like the London Stock Exchange (LSE) and NASDAQ. Some results are not consistent, e.g. on the basis of the data for the LSE, Reiss and Werner (1996) demonstrated that larger trades (but not to large) receive better prices. However, for unusually large orders this empirical observation is not true. Hansch et al. (1999) reported that on the LSE the price rise in relation to this spread is smallest for small trades, larger for medium-sized trades and largest for large ones. Huang and Stoll (1996) calculated that the mean spread for small trades amounts to almost 20 cents but for large trades it is smaller approximately by 30-35 percent. They discovered an asymmetry in the cost of trading between buyer- and seller-initiated trades. In addition, the authors analysing the company spreads on NYSE and NASDAQ in their paper, found out that spreads on NASDAQ are higher than on NYSE.

Chan and Lakonishok (1993) claim that in a portfolio for sale the number of stocks is limited. They try to convince the readers that the decision to sell must not convey negative information. On the contrary, according to the authors purchases are usually implied by firm specific information which is available.

Stoll (2000) conjectured that the spread depends on some factors related to a stock's liquidity and risk. On the basis of data from the USA, he performed a panel regression of this spread using five determinants as explanatory variables, namely trading volume, the

number of trades per day, free float, return variance and stock price. The models fit the data well since all explanatory variables are significant and the determination coefficient is over 0.6.

Naik and Yadav (2003) conducted similar research and obtained interesting results for the London Stock Exchange. Unfortunately, their results are not in line with later findings reported in Gajewski and Gresse (2007), who used data from Euronext Paris and the London Stock Exchange. As a new explanatory variable they included the imbalance between purchase and sale orders. They established that trading volume, return variance, and order imbalance were significant and exhibited the expected signs. However, free float, stock price, and the number of trades per day turned out not to be significant.

In some research the bid-ask spread is used as a measure of stock market liquidity employed in market microstructure studies. In Christie and Schultz (1994); Huang and Stoll (1996); Bessembinder (2003) the bid-ask spread is used to conduct inter-market comparisons of trading costs. The efficiency of rules and regulations aimed at reducing the cost of trading can be proven by checking the rules and regulations and their impact on the bid-ask spread.

In a more recent study Chen et al. (2017) proposed a non-parametric method to estimate the spread on the basis of the Roll (1984)) model. A further development can be found in Abdi and Ranaldo (2017), who incorporate the Corwin and Schultz (2012) model into the Roll model to derive a new estimator.

In the next part of this paper we shall focus on scarce bid-ask spreads, predicted on the basis of data which is more accessible, such as closing price, volume, high/low price, returns. Very preliminary results of this paper are in unpublished working paper by Duda et al. (2019).

In our calculations, we use hierarchical correlation reconstruction (HCR) methodology: each moment of the predicted variable is independently modelled as a linear combination of mixed moments of the variables used, then they are finally combined into the predicted (conditional) probability distribution. A basic use of predicting the entire distribution is to predict a value, e.g. as its expected value, additionally also estimating the uncertainty from its variance. Another use may be to handle more sophisticated situations such as a binomial distribution with two (or more) separate maxima: when predicting the expected value might not be a good choice (it may have a much lower density), a better prediction might be, e.g. one of the maxima, or may be both: providing a prediction as an alternative of two (or more) possibilities.

We can also use the entire predicted density, e.g. for a more accurate additional calculation, estimating the quantiles, or generating random values. HCR methodology combines the advantages of classical statistics and machine learning. While the former allows for well controlled and interpretable but relatively small (rough) models/descriptions, machine learning allows for very accurate descriptions using huge models, but usually lacks uniqueness of solution, control and interpretability of coefficients, and often is computationally costly. HCR allows one to work on huge models obtained from (unique) least-squares optimization, using well interpretable coefficients: as mixed moments of variables, starting, e.g. with moments of single variables and the correlation coefficients. The results for 22 DAX companies seem to be promising, especially using individual models for each company. An

additional advantage of this methodology is that it is computationally inexpensive; such complex models for these data can be estimated and evaluated in a second.

3. Data set and basic concepts

This Section discusses the data set and reminds one of the standard concepts, to be used for describing the methodology used in the next Section.

3.1. Data set and variables

Daily data for DAX companies from the 1999-2013 period were used (source in Acknowledgment); they were selected as they have at least 2000 data points: Deutsche Telekom AG (DTE), Daimler AG (DAI), SAP SE (SAP), Siemens AG (SIE), Deutsche Post AG (DPW), Allianz SE (ALV), BMW AG St (BMW), Infineon Technologies AG (IFX), Volkswagen AG Vz (VOW3), Fresenius SE & Co. KGaA (FRE), Henkel AG & Co. KGaA Vz (HNK3), Continental (CON), Merck KGaA (MRK), Münchener Rück AG (MUV2), Deutsche Börse AG (DB1), Deutsche Lufthansa AG (LHA), Fresenius Medical Care AG & Co. KGaA St (FME), Deutsche Bank AG (DBK), Fresenius Medical Care AG & Co. KGaA St (HEI), RWE AG St (RWE), Beiersdorf Aktiengesellschaft (BEI), Thyssenkrupp AG (TKA).

The basic set of variables is P - closing price, V - volume, R - return, H,L - high/low price. However, it turned out that trying to exploit dependence on R and L alone did improve evaluation, hence finally the basic model considered: '123' uses only P as '1'-st variable, V as '2'-nd variable and normalized $(H-L)/P$ as '3'-rd variable. It might be worth noting that the paper presents average spreads on the German stock market in question. This type of data is also applied in the cited references.

3.2. Bid-ask spread and some of its standard predictors

Bid-ask spread is the difference between the lowest asking price (*ask*, offered by a seller) and the highest bid price (*bid*, offered by a buyer). While this value is important because it is a main measure of market quality (Mestel et al. (2018); Gurgul and Machno (2017)), this information is usually publicly unavailable. Therefore, there is an interest in being able to predict this value on the basis of other, more accessible data.

At this point, one can present an important account that the smaller the spread, the more efficiently the market operates, and its liquidity understood by the volume of trading in securities also increases indirectly (Roll (1984)).

We consider bid-ask spread as a standard measure of liquidity. More specifically, we work on relative quoted spread, which is normalized by dividing by midpoint $(ask+bid)/2$: $S = \frac{ask-bid}{(ask+bid)/2}$.

Simple examples of its predictors based on the 5 basic variables are *AMI* (Amihud (2002); Fong et al. (2017)), *HLR* (Będowska-Sójka and Echaust (2019); Gurgul and Syrek (2019)):

$$AMI = \ln\left(1+\frac{|R|}{P\cdot V}\right) \qquad HLR = 2\frac{H-L}{H+L} \tag{1}$$

They are intended for a simpler task than that discussed: to predict values, while here we want to predict entire conditional probability distributions. We can reduce the predicted probability distributions into predicted values, e.g. as the expected value, median, or positions of maxima (especially for multimodal distributions). Fig. 1 presents comparisons using such predictions reduced with the expected value.

However, in practice such a prediction is often further processed through several functions, generally $E(f(X)) \neq f(E(X))$ for nonlinear, hence it is more accurate to process the probability distribution (e.g. on a lattice) through the functions before, e.g. taking the expected value.

3.3. Normalization to nearly uniform marginal distributions

Like in copula theory, in HCR methodology it is convenient to initially normalize all variables to nearly uniform marginal distributions in $[0,1]$, hence below we shall only work on such normalized variables, which beside usually better prediction also allows for better presentation of evaluation: e.g. density without prediction is 1, log-likelihood is 0.

This standard normalization requires estimation of the cumulative distribution function (CDF), individually for each variable, and this CDF function to be applied to the original values. Finally, having a prediction we can go back to the original variable using CDF^{-1}, for example as in the original Duda and Szulc (2018) article, although for simplicity we omit this step here - working only on normalized variables. Also, AMI, HLR predictions underwent such normalization for the purpose of Fig. 1 visual performance comparison - which means that a perfect predictor would give a diagonal plot.

The empirical distribution function (EDF) was used for this normalization here: for each variable its n observed values are sorted, then i-th value in such an order obtains $(i-0.5)/n$ normalized value. Hence, values become their estimated quantiles this way, a difference of two normalized values describes the percentage of population between these two values.

Having predicted density for normalized variable, we can transform it to the original variable, e.g. by discretizing this density to probability distribution on a $\{(i-0.5)/n\}_{i=1,\ldots,n}$ lattice, and assigning probability of its i-th position to i-th ordered original value. For simplicity it is omitted in this article.

3.4. Evaluation: log-likelihood with 10-fold cross-validation

The most standard evaluation of probability distributions is log-likelihood as in ML estimation: the average (natural) logarithm of the (predicted) density in the actually observed value. Hence, we will use this evaluation here.

Working on variables normalized to $\rho \approx 1$ marginal distributions, without prediction they would have practically zero log-likelihood. This allows to imagine the gains from predictions as an averaged improvement over this $\rho \approx 1$, as in Fig. 2. For example, the best observed log-likelihood ≈ 1 corresponds to $\approx \exp(1) \approx 2.7$ density: 2.7 times as good as without the prediction, the same as if we could squeeze a $[0,1]$ range 2.7 times to a 0.37 wide range. Sorting the predicted densities into the actually observed values, we can obtain additional information regarding the distribution of prediction, as presented in this Figure.

Here, we predict the conditional density - denoted as $\rho(Y = y|X = x)$ for the density of Y predicted on the basis of the known value of X. Hence its evaluation can be seen as an estimation of $E_{XY}(\ln(\rho(Y|X))$, which is minus conditional entropy $-H(Y|X)$. While it is unknown here, random variables have some concrete value of conditional entropy - we can

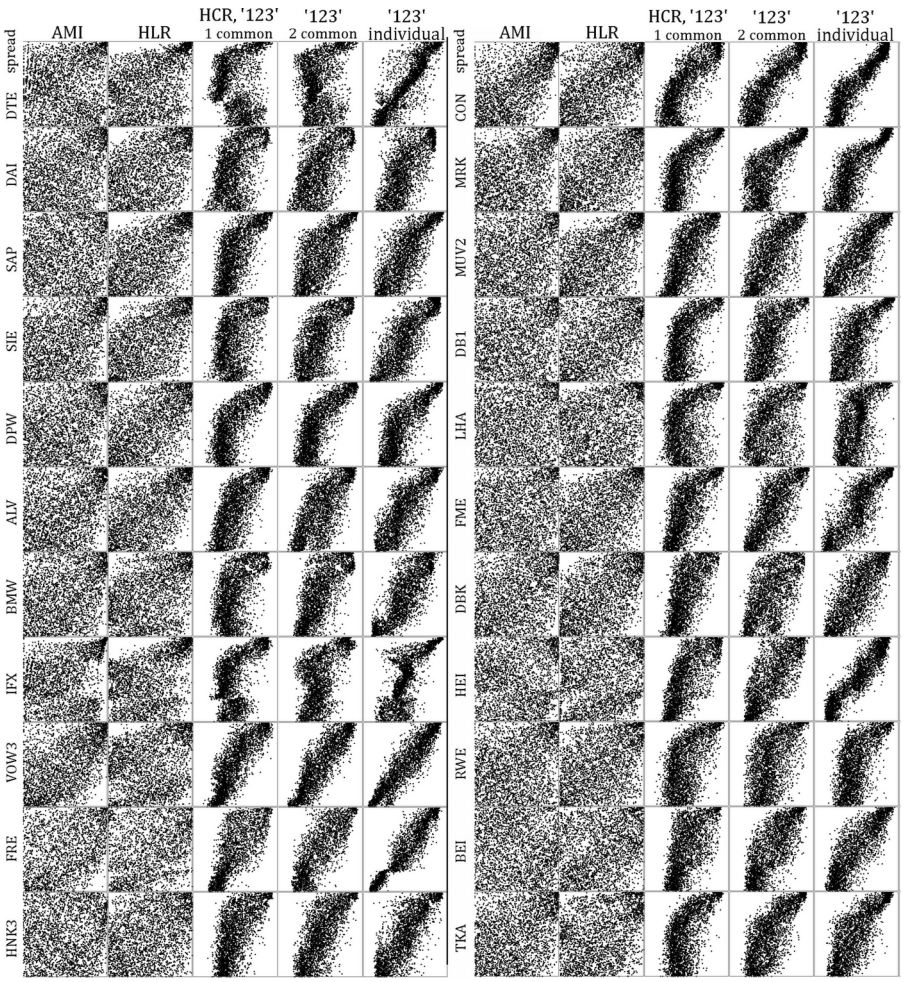

Figure 1: Comparison of spread predictors on data set for visual evaluation: a perfect predictor would give a diagonal scatter plot, a completely useless one would give a uniform distribution. All variables are normalized to nearly uniform marginal distributions, including outcomes of standard methods: *AMI*, *HLR*. The following 3 columns use the expected values of predicted densities from the discussed '123' model (using $P,V,(H-L)/P$ variables, $8 \cdot 53 = 424$ coefficients). The "1 common" column uses one model for all, "2 common" groups companies into two subsets and uses one of two models (as in Fig. 7, using models comL, comR from Fig. 6). The last column uses models individually optimized for each company.

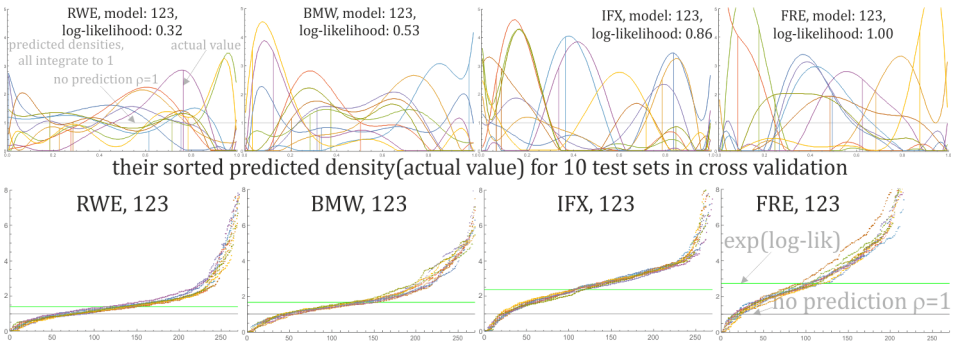

Figure 2: Top: examples of predicted conditional densities. Bottom: evaluation of such a prediction. While log-likelihood only provides averaged $\ln(\rho(y^i|x^i))$, sorted $\rho(y^i|x^i)$ values are presented here, allowing to additionally see, e.g. how frequently such prediction is below $\rho = 1$ threshold of using no prediction. Colours denote one of 10 rounds of 10-fold cross-validation, visualizing dependence of randomly splitting into the training and test set.

hopefully try to approach it with better and better models.

Here, we are focusing on large models that use hundreds of coefficients, estimated from thousands of data points. Hence we need to be careful not to overfit: represent only behaviour which indeed generalizes - is not just a statistical artefact of the training set. Machine learning also builds large models, usually evaluating them using cross-validation: a randomly split data set into a training and test set, the training set is used to build the model, then the test (or validation) set is used to evaluate this model.

However, this evaluation depends on the random splitting into the training and test set. Standard 10-fold cross-validation is used here to weaken this random effect: the data set is randomly split into 10 nearly equal size subsets, the evaluation is an average from 10 cross-validations: using successive subsets as the test set and the remaining ones as the training set. However, a scale ≈ 0.01 randomness of such an evaluation is still observed, hence for log-likelihoods only two digits after the comma are presented.

4. The HCR-based methodology used

This Section discusses the methodology used, which is an expansion of the one used in Duda and Szulc (2018). To predict conditional distribution $\rho(Y|X)$ we decompose X and Y variables into mixed moments and model separately each moment of Y using least-squares linear regression of moments of X, then combine them into the predicted conditional probability distribution of Y.

4.1. Decomposing joint distribution into mixed moments

After normalizing the marginal distributions of all variables to nearly uniform on $[0,1]$, for d variables their joint distribution on $[0,1]^d$ would also be nearly uniform if they were statistically independent. Distortion from uniform joint distribution corresponds to statistical

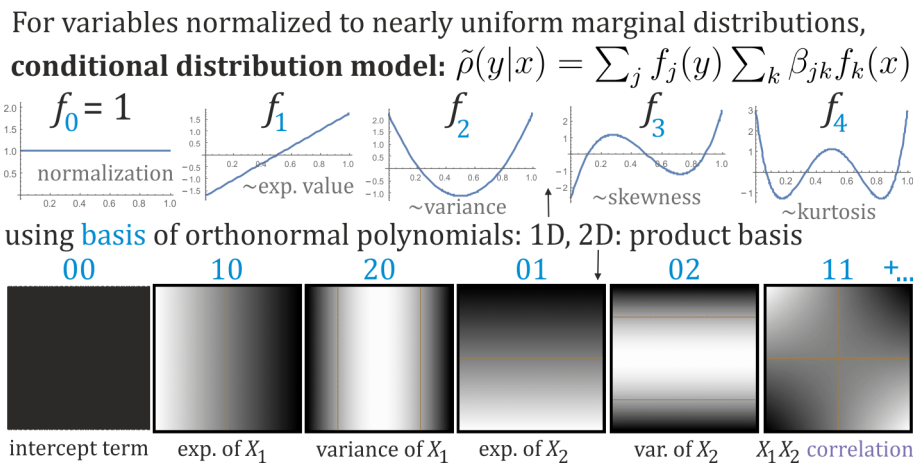

Figure 3: General concept, some first functions of the 1 and 2 dimensional basis of orthornormal polynomials used ($f_{j_1 j_2}(x) = f_{j_1}(x_1) f_{j_2}(x_2)$), and application example. For simplicity we assume working on variables normalized to nearly uniform marginal densities on $[0,1]$. We would like to model distortion from this uniform distribution for the predicted variable Y on the basis of the context X: as a linear combination, e.g. of orthornormal polynomials here, for which coefficients have similar interpretation as moments/cumulants: a_1 shifts right/left like the expected value, a_2 increases/decreases the probability of extreme values as variance, etc.

dependencies between these variables - we would like to model and exploit it.

In HCR we model it as just a linear combination using an orthornormal basis, e.g. of polynomials, which gives the coefficients a similar interpretation as moments and mixed moments: the dependencies between moments for multiple variables. In Fig. 3 the general concept of the HCR methodology is presented.

The first orthonormal ($\int_0^1 f_i(x) f_j(x) dx = \delta_{ij}$) polynomials (rescaled Legendre) for $[0,1]$ are $f_0 = 1$ and f_1, f_2, f_3, f_4 correspondingly (plotted in Fig. 3):

$$\sqrt{3}(2x-1), \sqrt{5}(6x^2-6x+1), \sqrt{7}(20x^3-30x^2+12x-1), 3(70x^4-140x^3+90x^2-20x+1)$$

We could alternatively use, e.g. $1, \sqrt{2}\cos(\pi x k)$ for $k \geq 1$ orthonormal basis. However, experimentally this usually leads to inferior evaluation.

Decomposing density $\rho(x) = \sum_j a_j f_j(x)$, we need $a_0 = 1$ normalization to integrate to 1. Due to orthogonality, $\int_0^1 f_j(x) dx = 0$ for $j > 0$, hence the following coefficients do not affect normalization. As we can see in their plots in Fig. 3, positive a_1 shifts density toward right - acting analogously as the expected value. Positive a_2 increases the probability of extreme values at the cost of central values - analogously as variance. Skewness-like higher order asymmetry is brought by a_3 and so on - we can intuitively interpret these coefficients as moments (cumulants). This is only an approximation, but useful for interpreting these models.

In multiple dimensions we can use the product basis:

$$f_j(x) = f_{j_1}(x_1) \cdot \ldots \cdot f_{j_d}(x_d) \qquad \text{for } j = (j_1, \ldots, j_d) \tag{2}$$

leading to a model of joint distribution:

$$\rho(x) = \sum_{j \in B} f_j(x) = \sum_{j \in B} a_j f_{j_1}(x_1) \cdot \ldots \cdot f_{j_d}(x_d) \tag{3}$$

where $B \subset \mathbb{N}^d$ is the basis of the mixed moments we are using for our modelling. It is required that it contains $(0, \ldots, 0)$ for normalization. Besides, there is freedom in choosing this basis, which allows one to hierarchically decompose the statistical dependencies of multiple variables into mixed moments: describing marginal distribution first, then pairwise dependencies, and so on for dependencies of growing numbers of variables.

Fig. 3 contains the first 5 functions of such a product basis for $d = 2$ variables: f_{00} corresponds to normalization and requires $a_{00} = 1$. The coefficients of f_{10}, f_{20} describe the expected value and the variance of the first variable, f_{01} and f_{02} analogously of the second. Then we can start including moment dependencies, starting with a_{11}, which determines the decrease/increase in the expected value of one variable with the growth in the expected value of the second variable - analogously to the correlation coefficient. We also have dependencies between higher moments, such as asymmetric a_{12}, which relates the expected value of the first variable and the variance of the second.

And analogously for more variables, e.g. a_{010010} describes the correlation between the 2nd and 5th out of 6 variables. Finally, we can hierarchically decompose the statistical dependencies between multiple variables into their mixed moments. However, to completely describe the general joint distribution, we would need $B = \mathbb{N}^d$ infinite number of mixed moments for complete expansion - for practical modelling we need to choose the finite basis B of moments to focus on.

4.2. Estimation using least squares linear regression

Having a data sample $\mathscr{X}$, we would like to estimate such mixed moments as coefficients for the linear combination of an orthonomal basis of functions, e.g. polynomials. Smoothing the sample using kernel density estimation, finding a linear combination which minimizes the square distance to such a smoothed sample, and performing limit to zero width of the kernel used, we obtain a convenient and inexpensive MSE estimation Duda (2018): independently for each coefficient j as just the average over the data set of value of the corresponding function:

$$a_j = \frac{1}{|\mathscr{X}|} \sum_{x \in \mathscr{X}} f_j(x) \tag{4}$$

We could use this model for predicting conditional distribution: substitute the known variables to the modelled joint distribution, after normalization obtaining the (conditional) density of the unknown variables.

However, for the bid-ask spread prediction problem, a slightly better evaluation was obtained using the generalizing alternative approach of Duda and Szulc (2018), which allows

one to additionally exploit subtle variable dependencies, hence we will focus on this.

Specifically, to model $\rho(Y=y|X=x)$, let us use separate bases of (mixed) moments: B_X for X, B_Y for Y, and model relations between them. While more sophisticated models could be considered for such relations including neural networks, for simplicity and interpretability we focus on linear models here, treating $f_j(x)$ as interpretable features:

$$\rho(y|x) = \sum_{j\in B_Y} f_j(y) a_j(x) \qquad \text{for} \qquad a_j(x) = \sum_{k\in B_X} \beta_{jk} f_k(x) \tag{5}$$

hence the model is defined by the $|B_Y| \times |B_X|$ matrix β.

It allows for good interpretability: β_{jk} coefficient is linear contribution of k-th mixed moment of X to j-th (mixed) moment of Y. We focus on one-dimensional Y, but the formalism allows one to analogously predict density for multidimensional Y.

To find the β we use least-squares optimization here - it is very inexpensive, can be used independently for each $j \in B_Y$ thanks to the use of an orthonormal basis, and intuitively it is a proper heuristic: least-squares optimization estimates the mean - exactly as we would like for coefficient estimation (4). However, this is not necessarily the optimal choice - it might also be worth exploring more sophisticated ways.

This least-squares optimization has to be performed separately for each $j \in B_Y$. Denoting $\beta_{j\cdot} = (\beta_{jk})_{k\in B_X}$ as a coefficient vector for j-th moment and $\mathscr{D} = \{(y^i, x^i)\}_{i=1..n}$ as (e.g. training) data set of (y,x) pairs:

$$\beta_{j\cdot} = \text{argmin}_v \sum_{(y,x)\in\mathscr{D}} \left(\sum_{k\in B_X} f_k(x) v_k - f_j(y) \right)^2 = \text{argmin}_v \left\| Mv - b^j \right\|^2$$

$$\text{for } M = [f_k(x^i)]_{i=1..n, k\in B_X}, \qquad b^j = (f_j(y^i))_{i=1..n}$$

matrix M and vector b^j for $j \in B_Y$. This least-squares optimization has a unique solution:

$$\beta_{j\cdot} = (M^T M)^{-1} M^T b^j \tag{6}$$

Separately calculated for each $j \in B_Y$, leading to the entire model as β matrix with $\beta_{j\cdot}$ rows.

4.3. Applying the model, enforcing nonnegativity

We can apply the found model β to (e.g. test) data points as in (5), obtaining the predicted conditional density for y on $[0,1]$ as a polynomial. However, sometimes it can drop below 0, so let us refer to it as $\tilde{\rho}$ and then enforce the non-negativity required for densities:

$$\tilde{\rho}(y|x) = \sum_{j\in B_Y} f_j(y) \sum_{k\in B_X} \beta_{jk} f_k(x) \tag{7}$$

This polynomial always integrates to 1. However, it can occasionally be below zero, which should be interpreted as corresponding to a low positive density. This interpretation to non-negative density ρ is referred to as calibration, and can be optimized on the basis of the data

set. For simplicity only the following was used:

$$\rho(y|x) = \max\left(\tilde{\rho}(y|x), 0.03\right)/N \tag{8}$$

where N normalization factor is chosen to integrate to 1: $N = \int_0^1 \max\left(\tilde{\rho}(y|x), 0.03\right) dy$. The 0.03 threshold was experimentally chosen as a compromise for the data set used, its tuning can slightly improve evaluation.

4.4. Basic basis selection

The optimal choice of the basis is a difficult open question. As the basic choice the combinatorial family was used:

$$\mathscr{B}((m_1,\ldots,m_d),s,r) := \left\{ j \in \mathbb{N}^d : \forall_i j_i \leq m_i,\ \sum_{i=1}^{d} j_i \leq s,\ \sum_{i=1}^{d} \operatorname{sgn}(j_i) \leq r \right\} \tag{9}$$

where m_i chooses how many first moments to use for i-th variable, s bounds the sum of used moments (and formally the degree of the corresponding polynomial), r bounds the number of nonzero j_i: to include the dependencies of up to r variables.

For example the '123' model infers 8 moments $B_Y = \mathscr{B}((8),8,1)$ from 3 variables using a compromise: $B_X = \mathscr{B}((4,4,4),5,3)$ of size $|B_X| = 53$ basis, directly written, e.g. in Fig. 6.

4.5. '123' model using basic variables

The initial plan for this article was to improve prediction from standard models: AMI, HLR, trying to predict the conditional distribution of spread from their values using the methodology under discussion. However, the results were disappointing, especially for AMI, as we can see in Fig. 1.

Therefore, we decided to use the original variables (P,V,L,H,R) instead, which turned out to lead to essentially better predictions. A search for parameters using $\mathscr{B}$ basic basis selection (9) was performed manually to maximize the averaged log-likelihood in 10-fold cross-validation. This search finally leads to $B_X = \mathscr{B}((4,4,4),5,3)$ basis for only 3 variables: $P, V, (H-L)/P$ to predict up to the 8-th moment of Y. Surprisingly, adding dependence on R and L alone worsened the evaluation - their dependence did not generalize from training to test sets, hence they are not used in the final model.

The top of Fig. 2 contains examples of conditional densities predicted. The predicted $\tilde{\rho}(y|x^i) = \sum_j f_j(y) \sum_k \beta_{jk} f_k(x^i)$ polynomial for i-th data point undergoes $\rho = \max(\tilde{\rho}, 0.03)/N$ to remove negative densities, and normalization to integrate to $1 = \int_0^1 \rho(y|x)dy$. Each diagram contains 10 example predictions, vertical lines show the actual values $(y^i, \rho(y^i|x^i))$: the higher the better prediction, without prediction all would have height 1. Companies were chosen to present prediction examples of various evaluation levels. The best ones predict mainly narrow unimodal distributions in line with the actual values, although weaker ones can usually only predict wide often multimodal distributions. We can see rapid growths at the ends - they are likely artefacts of using polynomials, their additional removal might

improve prediction. The bottom part presents their sorted predicted densities in the actual values $\{\rho(y^i|x^i)\}_i$, with marked gray $\rho = 1$ line of using no prediction and green exp(log-likelihood) line corresponding to average improvement over no prediction. The points are of different colours denoting one of 10 rounds of 10-fold cross-validation.

Integration required for normalization is relatively costly to compute, especially in higher dimensions, hence for efficient calculation the predicted polynomial $\tilde{\rho}$ was discretized here into 100 values on a $((i-0.5)/100)_{i=1,\ldots,100}$ lattice, which corresponds to approximating the density with a piecewise constant function on length 1/100 subranges. Then $\max(\cdot, 0.03)$ was applied, and division by the sum for normalization. Finally, the density in discretized $\lceil 100y^i \rceil/100$ position was used as $\rho(y^i|x^i)$ in the log-likelihood evaluation.

In Figure 4 the results of cross-validation are presented. Model '123' denotes using the three basic variables: where '1' denotes the closing price (P), '2' volume (V), and '3' the difference between high and low price normalized by dividing by the closing price: $(H-L)/P$. The last column presents the averaged evaluation for using common model for all data. We can also see that there are large differences between companies, hence we will mostly focus on building individual models for each company. The three lowest dots correspond to predicting from single variable, then evaluation grows when adding information from succeeding variables.

Copulas are a general, well-established method of modelling multivariate distribution. In higher dimensions r-vines are a flexible class of mulivariate distributions. This type of copulas allows for flexible modelling of asymmetric and nonlinear dependence patterns Gurgul and Machno (2016). For comparison purposes we estimated such models and it turns out that on average log-likelihoods for individual model from copulas were smaller than from HCR. In Figure 4 points denoted by "123vc" correspond to results from r-vines. On average, log-likelihood for individual HCR models was 0.603, while for vine-copulas it was 0.366, getting better representation of complex behaviour thanks of allowing for high parametric models. HCR also has much less expensive estimation (least squares regression of moments), and interpretation of the found parameters as moment dependencies.

While the optimal choice of the basis seems a difficult open problem, an exhaustive search over all subsets is rather impractically costly, Figure 5 presents some heuristic approaches. The $\mathscr{B}$ family seems generally a good start, e.g. to successively modify some its parameter by one as long as improvement is observed. In this Figure we can see a large improvement while the number of predicted moments rises up to ≈ 7, which suggests that the complexity of the conditional distributions for this problem requires this degree of polynomial in order to be described properly. This Figure also contains trials of using different orders of some first mixed moments. The selective removal, which is presented there, seems a reasonable optimization: for each mixed moment from B_X calculate the evaluation when it is removed, finally remove the one that leads to the best evaluation, and so on as long as the evaluation improves.

Examples of β matrix are visualized in Fig. 6 for $|B_X| = 53$, $|B_Y| = 1+8$. Trying to split all companies into subsets of similar behaviour, as visualized in tree Fig. 7, splitting into two subsets we obtain the comL and comR models - correspondingly for the left (DPW, BEI, HNK3, FME, SAP, DB1, RWE, FRE, HEI, DTE, IFX) and right (DAI, SIE, TKA, CON, MRC, LHA, VOW3, MUV2, ALV, BMW, DBK) subtree of this tree. Then individual

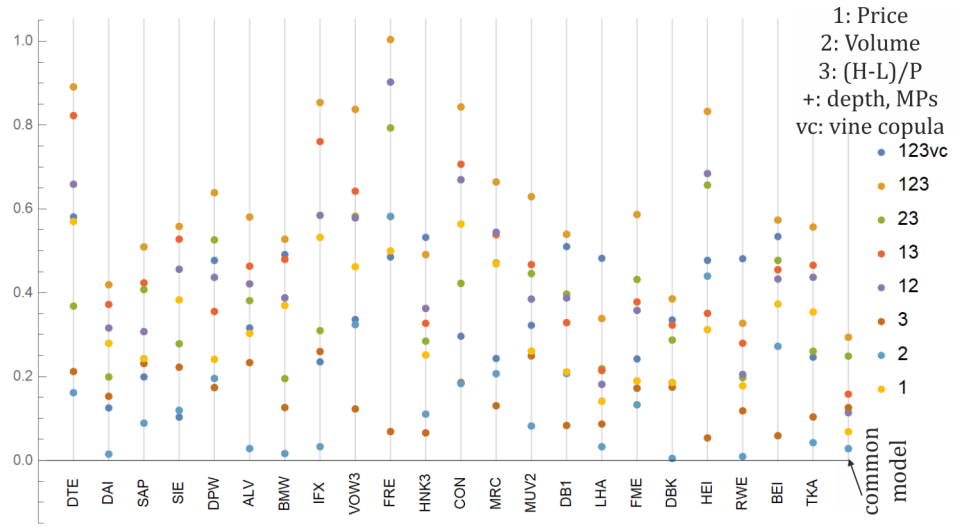

Figure 4: Log-likelihoods from 10-fold cross-validation for individual models for companies using various types of information. We can see individual behaviour of companies and growth of prediction evaluation while adding information from succeeding variables. The "123vc" points correspond to vine copulas using the same evaluation: for HCR average log-likelihood for individual models was 0.603, while for vine-copulas it was 0.366, additionally requiring $\approx 100\times$ more computational time.

models for 5 selected companies were presented. The rows correspond to the predicted moments of Y, as linear combinations of mixed moments of X corresponding to columns. Row zero has always only 000 nonzero coefficient equal to 1 for normalization. The next row describes the prediction of the expected value, the next one of variance and so on. In the top model, common for all companies, we can, e.g. see large positive $001 \to 1$ coefficient: the spread increases with the growth of $H - L$, negative $010 \to 1$: the spread decreases with growth of V, and negative $011 \to 2$: variance of spread decreases for correlated V and $H - L$. Blue $100 \to 3$ for FRE denotes a reduction in skewness of spread with growth of price. Generally, we can see rather individual behaviour for different companies, starting with $100 \to 1$ analogous to the price-spread correlation, which seems the main dividing factor between comL and comR companies.

4.6. Individual vs common models, universality

A natural question is how helpful for prediction a given variable is - Fig. 4 presents some answers by calculating the log-likelihood also for models using only some of the variables. We can see different companies can have very different behaviour here, e.g. for some V is helpful (volume and spread are correlated), for some it is not. Fig. 6 shows that they can even display the opposite behaviour: e.g. for $100 \to 1$ dependence on price.

It is a general lesson that while we would like predictors to be nice simple formulas, the reality might be much more complicated - the models found here are the results of the

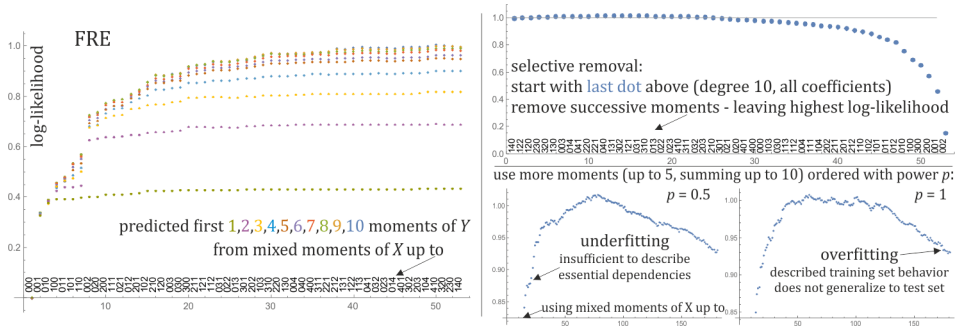

Figure 5: Left: Optimizing the basis and model size using the example of the company FRE and $B_X = \mathscr{B}((4,4,4),5,3)$ size 53 basis of mixed moments from '123' model. Log-likelihoods for predicting the first $1\ldots10$ moments (denoted by colours) using some first of mixed moments (sorted lexicographically) of 3 X variables: $P,V,(H-L)/P$. We can see that we should predict ≈ 8 moments, higher moments are necessary to represent more complex distributions. Top right: selective removal of successive mixed moments to maximize log-likelihood - we can see that we can slightly improve evaluation this way, additionally reducing the model size. However, it requires individual optimization for each company. Bottom right: analogously as top, but using size 181 larger $B_X = \mathscr{B}((5,5,5),10,3)$, also trying different orders of mixed moments: accordingly to $\sum_i (j_i)^p$. While using all such mixed moments clearly leads to overfitting, selectively using some of the first ones can lead to slightly improved evaluation.

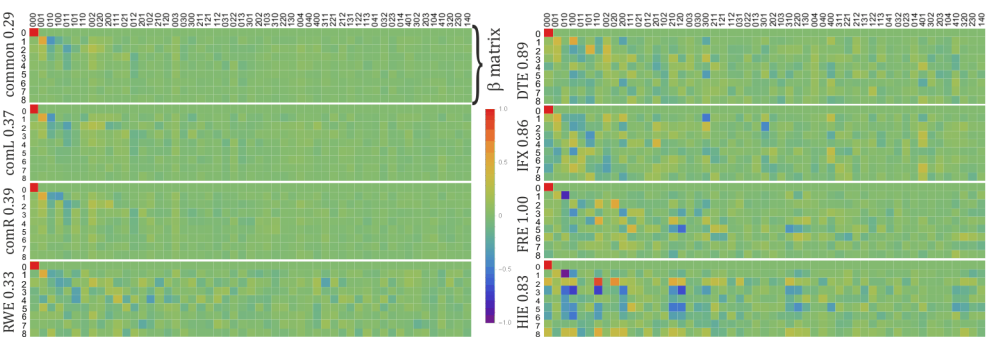

Figure 6: Visualized coefficients of '123' models (9×53 matrix β for $\rho(y|x) = \sum_j f_j(y) \sum_k \beta_{jk} f_k(x)$) for $(P,V,H-L)$ variables, the numbers above the names are log-likelihoods. The 'common' is the model built for all the data combined - it presents general trends. The 'comL' and 'comR' models are for the left (DPW, BEI, HNK3, FME, SAP, DB1, RWE, FRE, HEI, DTE, IFX) and right (DAI, SIE, TKA, CON, MRC, LHA, VOW3, MUV2, ALV, BMW, DBK) subtree in Fig. 7 - we can see that these subsets of companies mainly differ by $100 \to 1$ coefficient corresponding to correlation between price and spread.

cultures of traders of the stocks of individual companies, which can essentially vary between companies.

Therefore, to obtain the most accurate predictions we should build individual models

for each company. Furthermore, a specific behaviour of a given company can additionally evolve in time - which could be exploited, e.g. by building separate models for shorter time periods, or using adaptive least-squares linear regression Duda (2019), and this is planned for future investigation.

However, building such models requires training data, which in the case of variables like bid-ask spread might be difficult to access. Hence, it is also important to search for universality - e.g. try to guess a model for a company for which we lack such data, on the basis of the available information for other companies. This generally seems a very difficult problem, Fig. 7 shows that even having all the data, using the common model for multiple companies we should expect a large evaluation drop. For example, we can see that the behaviour of DTE completely disagrees with the common model for all.

As we can see in this tree Figure, the use of common model situation improves if we can cluster companies into groups of similar behaviour - results are also presented for splitting companies into just two groups with separate models (comL, comR in Fig. 6), also visually leading to slightly better predictions as we can see comparing the 3rd and 4th column in Fig. 1. The heights of the names show the evaluation of using an individual model for a given company, orange dots show the successive reduction of log-likelihood for a given company while using common models for subsets that grow according to this tree. The lowest dots correspond to the use of one common model for all (common in Fig. 6) we can see that it is worse than zero only for DTE (we get zero when using no prediction at all). Splitting companies into a left and right subtree and using separate two models for them (comL and comR in Fig. 6), we essentially obtain a better prediction (one dot up). The tree structure was calculated by combining subsets to maximize (log-likelihood of common model / average log-likelihood of individual models) - grouping companies into pairs and then further, up to a single common model for all. The positions of lines represent such grouped companies: a light-gray line their averaged log-likelihoods of individual models, dark-gray line their log-likelihood for a common model. The difference between these two lines represent a loss while using the common model. The common models are fixed hence there is no cross-validation (CV) used, which artificially improves performance, for example for the first dot of FME corresponding to the common model with HNK - making it above CV individual model, generally suggesting large time inhomogeneities - to be included in future adaptive models.

5. Conclusions and further work

A general methodology has been presented for extracting and exploiting complex statistical dependencies between multiple variables in an inexpensive and interpretable way for predicting conditional probability distributions, using the example of the difficult problem of predicting bid-ask spreads from more accessible information. This expands the approach of Duda and Szulc (2018) by inferring from mixed moments, and searching for a basis in large spaces of possibilities.

Figure 1 presents a comparison between it and standard methods when using only the expected value from such predicted conditional density. A perfect predictor would lead to diagonal scatter plot, standard methods provide rather a noise instead, while the predictions

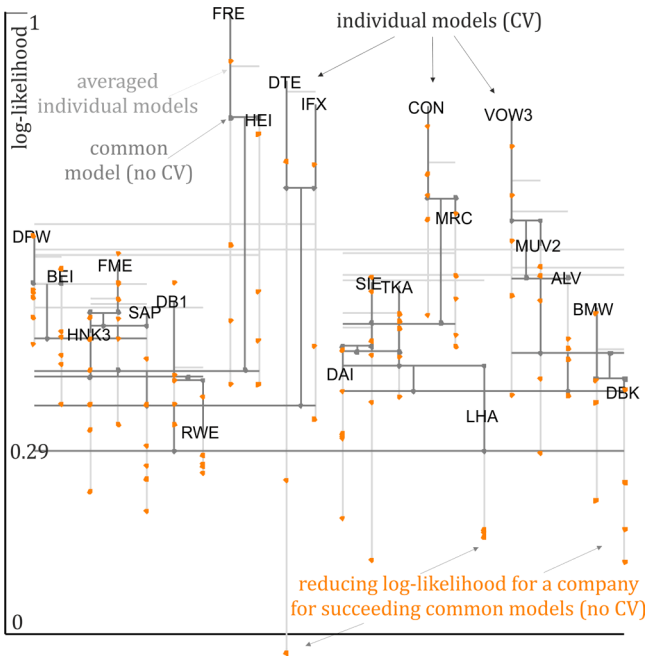

Figure 7: Visualization of optimized hierarchical grouping and evaluation loss while using common models for multiple companies, the height denotes log-likelihoods. It was constructed by starting with individual models, then successively joining subsets of companies leading to lowest loss of evaluation while using a common model for them.

from the approaches discussed indeed often resemble diagonal plot, especially when using individual models. The predicted conditional probability density provides much more information than the value alone: e.g. it allows one to additionally estimate the uncertainty of such a prediction as value, or provide prediction for multimodal densities, or it allows random values to be generated, e.g. for Monte-Carlo simulations, or just provides the entire density for accurate considerations especially if transforming such random variables through some further nonlinear functions.

There are many directions for further development of this relatively new general methodology, for example:

- Optimal choice of the basis is a difficult problem, which should be automatized especially for a larger number of variables - selecting from the basis of orthonormal polynomials discussed, or maybe automatically optimizing a completely different basis on the basis of a data set.
- Large differences between the behaviours of individual companies have been observed - raising difficult questions regarding how to optimize for common behaviour, optimize models on the basis of an incomplete information, etc. Additionally, such behaviour has probably also time inhomogeneity - the models should evolve in time,

requiring adaptive models to improve performance, where the problem of data availability becomes even more crucial.

- These models rapidly grow with the number of variables, which requires some modifications for exploiting high dimensional information - like extracting features from these variables, e.g. as averages, dimensionality reduction like PCA, etc.
- We have predicted the conditional distributions for one-dimensional variables, but the methodology was introduced to be more general: predicting for multidimensional Y should be just a matter of using proper B_Y, which is planned to be tested in the future.
- The densities predicted as polynomials often have rapid growths at the ends of $[0, 1]$ - their removal might improve performance.
- A linear relation was assumed between moments with least-squares optimization, which is inexpensive and has good interpretability, but is not necessarily optimal - one could consider, e.g. using neural networks instead, and optimizing criteria closer to the log-likelihood of final predictions.
- In the light of the Epps effect we can see the dependence of stock return cross-correlations on the data sampling frequency, i.e. for high-resolution data the cross-correlations are significantly smaller than their asymptotic value as observed for daily data. One should check the performance of HCR with respect to the data sampling frequency.
- The share of algorithmic trading in the market is growing. The HCR method may be helpful in the forecast of quoted and effective bid-ask spread regressed on the share of algorithmic trading in the market.
- A comparison of the results of bid-ask spread modelling and forecasting using HCR methodology with respect to the microstructure of stock markets in particular countries, their size and the level of development.

Acknowledgement

We would like to thank the Editors of this journal, and anonymous Referees for their valuable comments on earlier versions of the paper.

Henryk Gurgul thanks Professor Roland Mestel for providing the bid-ask data from the data bank ”Finance Research Graz Data Services” and Professor Erik Theissen and Stefan Scharnowski from Mannheim for providing data from the ”Market Microstructure Database”. Henryk Gurgul was financed by AGH University of Science and Technology in Krakow (institutional subsidy for maintaining Research Capacity Grant 16.16.200/396.)

References

ABDI, F., RANALDO, A., (2017). A simple estimation of bid-ask spreads from daily close, high, and low prices. The Review of Financial Studies 30(12), pp. 4437–4480

AMIHUD, Y., (2002). Illiquidity and stock returns: cross-section and time-series effects. Journal of Financial Markets 5(1), pp. 31–56

ARMITAGE, S, BRZESZCZYŃSKI, J., SERDYUK, A., (2014). Liquidity measures and cost of trading in an illiquid market. Journal of Emerging Market Finance 13(2), pp. 155–196

BĘDOWSKA-SÓJKA, B., ECHAUST, K., (2019). Commonality in Liquidity Indices: The Emerging European Stock Markets. Systems 7(2), pp. 7–24

BESSEMBINDER, H., (2003). Issues in assessing trade execution costs. Journal of Financial Markets 6(3), pp. 233–257

BUTLER, A. W., GRULLON, G., WESTON, J. P., (2005). Stock market liquidity and the cost of issuing equity. Journal of Financial and Quantitative Analysis 40(2), pp. 331–348

CHAN, L. K., LAKONISHOK, J., (1993). Institutional trades and intraday stock price behavior. Journal of Financial Economics 33(2), pp. 173–199

CHEN, X., LINTON, O., YI, Y., (2017). Semiparametric identification of the bid–ask spread in extended Roll models. Journal of Econometrics 200(2), pp. 312–325

CHRISTIE, W. G., SCHULTZ, P. H., (1994). Why do NASDAQ market makers avoid odd-eighth quotes? The Journal of Finance 49(5), pp. 1813–1840

CORWIN, S. A., SCHULTZ, P., (2012). A simple way to estimate bid-ask spreads from daily high and low prices. The Journal of Finance 67(2), pp. 719–760

DUDA, J., (2018). Exploiting statistical dependencies of time series with hierarchical correlation reconstruction. arXiv preprint arXiv:180704119

DUDA, J., (2019). Parametric context adaptive Laplace distribution for multimedia compression. arXiv preprint arXiv:190603238

DUDA, J., SZULC, A., (2018). Credibility evaluation of income data with hierarchical correlation reconstruction. arXiv preprint arXiv:181208040

DUDA, J., SYREK, R., GURGUL, H., (2019). Modelling bid-ask spread conditional distributions using hierarchical correlation reconstruction. arXiv preprint arXiv:191102361

FONG, K. Y., HOLDEN, C. W., TRZCINKA, C. A., (2017). What are the best liquidity proxies for global research? Review of Finance 21(4), pp. 1355–1401

GAJEWSKI, J. F., GRESSE, C., (2007). Centralised order books versus hybrid order books: A paired comparison of trading costs on NSC (Euronext Paris) and SETS (London Stock Exchange). Journal of Banking & Finance 31(9), pp. 2906–2924

GURGUL, H., MACHNO, A., (2016). Modeling dependence structure among European markets and among Asian-Pacific markets: a regime switching regular vine copula approach. Central European Journal of Operations Research 24(3), pp. 763–786

GURGUL, H., MACHNO, A., (2017). The impact of asynchronous trading on Epps effect on Warsaw Stock Exchange. Central European Journal of Operations Research 25(2), pp. 287–301

GURGUL, H., SYREK, R., (2019). Dependence Structure of Volatility and Illiquidity on Vienna and Warsaw Stock Exchanges. Finance a Uver: Czech Journal of Economics & Finance 69(3), pp. 298–321

HANSCH, O., NAIK, N. Y., VISWANATHAN, S., (1999). Preferencing, internalization, best execution, and dealer profits. The Journal of Finance 54(5), pp. 1799–1828

HUANG, R. D., STOLL, H. R., (1996). Dealer versus auction markets: A paired comparison of execution costs on NASDAQ and the NYSE. Journal of Financial Economics 41(3), pp. 313–357

LESMOND, D. A., OGDEN, J. P., TRZCINKA, C. A., (1999). A new estimate of transaction costs. The Review of Financial Studies 12(5), pp. 1113–1141

MESTEL, R., MURG, M., THEISSEN, E., (2018). Algorithmic trading and liquidity: Long term evidence from Austria. Finance Research Letters 26, pp. 198–203

NAIK, N. Y., YADAV, P. K., (2003). Trading costs of public investors with obligatory and voluntary market-making: Evidence from market reforms. In: EFA 2003 Annual Conference Paper, 408

REISS, P. C., WERNER, I. M., (1996). Transaction costs in dealer markets: Evidence from the London Stock Exchange. In: The Industrial Organization and Regulation of the Securities Industry, University of Chicago Press

ROLL, R., (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. The Journal of Finance 39(4), pp. 1127–1139

STOLL, H. R., (2000) Presidential address: friction. The Journal of Finance 55(4), pp. 1479–1514