Yield Curve Arbitrage – a hierarchical correlation reconstruction approach

Małgorzata Snarska[1], **Przemysław Jaśko**[2], **Jarek Duda**[3]

[1] Financial Markets Department,
Cracow University of Economics, 27 Rakowicka St., Krakow, POLAND
Email: snarskam@uek.krakow.pl
[2] Computational Systems Department,
Cracow University of Economics, 27 Rakowicka St., Krakow, POLAND
Email: jaskop@uek.krakow.pl
[3] Institute of Computer Science,
Jagiellonian University, 6 Łojasiewicza St., Krakow, POLAND
Email: jaroslaw.duda@uj.edu.pl

Abstract

We consider an yield curve arbitrage as a kind of a statistical arbitrage strategy involving portfolios of fixed-income instruments with specific maturity, which are rich or cheap points on the yield curve, and for which convergence is anticipated by the constructed statistical model for a yield curve.

We recall mathematical definintion of a statistical arbitrage, involving specific characteristics of a stochastic process representing dynamics of a strategy.

As a main contribution, we propose a statistical model for a yield curve functional data, which we use for predictions of a yield curve distribution. Using forecasts of a proposed model, we construct a strategy of portfolios for a fixed-income instrument. Then, *ex post* we statisticaly test conformance of employed strategy with a statistical arbitrage one.

Key words: yield curve arbitrage, statistical arbitrage, stochastic processes, functional data, algotrading

1 Introduction

Yield Curve Arbitrage strategy consists of portfolios involving long and short positions at different points along the yield curve of a fixed-income instrument with different maturities.

We consider an yield curve arbitrage as a kind of statistical arbitrage strategy involving portfolios of fixed-income instruments with specific maturity, which are rich or cheap points on the yield curve, and for which convergence is anticipated by the constructed statistical model for a yield curve.

In the section 2, we recall mathematical definintion [18] of a statistical arbitrage, involving specific characteristics of a stochastic process representing dynamics of a strategy. We also present a statistical test introduced by [18], for an assesement of conformance of considered strategy, with a statistical arbitrage one. Mentioned test is based on strategy value time series data. Then we briefly discuss how a statistical arbitrage differs from a standard arbitrage, and that latter can be considered as a specific case of a former. We also recall statistical test wich can be used to check if strategy is a statistical arbitrage. The statistical test is based on time series, which is a realization ofconsidered strategy value stochastic process.

In section 3, we give a general understanding of a yield curve arbitrage strategy, and consider it as a case of a statistical arbitrage strategy for portfolios of fixed-income instruments.

In sections from 4 to 6, we present our original proposition of a statistical model for a yield curve functional data.

During the construction of our model, we treat yield curves as random functions, and employ for them a Karhunen-Loève expansion. This enables us to represent yield curve random functions as linear combinations of deterministic functions (e.g. specific kind of polynomial functions), and coefficients of considered combinations are random variables.

Based on predictions of our model for a yield curve distribution (more precisely for a distribution of its decompostion coefficients), we construct a strategy of portfolios for a fixed-income instrument. Then, *ex post* we statisticaly test conformance of employed strategy with a statistical arbitrage one.

2 A concept of a statistical arbitrage strategy

In this section we recall formal definitions of concepts related to a statistical arbitrage strategy. The concepts will be later used for a construction of a yield curve arbitrage, which we consider as a variant of a statistical arbitrage strategy involving fixed-income portoflios.

2.1 Preliminary concepts related to a statistical arbitrage strategy

We consider *multi-step single-instrument* market model containing two underlying securites:

- The *risk-free asset* (money-market account), described by deterministic function:

$$A(t) = (1+R)^t, \tag{1}$$

  where $R > -1$ is the risk-free rate,

- the *risky instrument*, which is thought as a fixed-income instrument of a specific maturity, or a portfolio of fixed-income instrument with different maturities.

Following definitions are based on the work [3].

[Strategy] A *strategy* is sequence of pairs of random variables showing, for $t = 1, 2, \ldots, T$, the number of units $x(t)$ of an instrument and the number $y(t)$ of money market account units chosen at time $t-1$ and held until time t, and such that $(x(t), y(t))$ is $\mathcal{F}_{t-1}$. In general, a sequence of random variables satisfying this measurability requirement is called *predicatble*.

[Predictability of a process] A process $X = (X(t))_{t\geq 1}$ is *predicatble* relative to given filtration $(\mathcal{F}_t)_{t\geq 0}$ if for every $t \geq 1$, $X(t)$ is $\mathcal{F}_{t-1}$-measurable.

[Value process of strategy] The *value process* of a strategy is a sequence defined for $n = 1, \ldots, N$ by

$$V_{(x,y)}(t) = x(t)S(t) + y(t)A(t), \tag{2}$$

together with the initial investement

$$V_{(x,y)}(0) = x(1)S(t) + y(1)A(t), \tag{3}$$

[Self-financing strategy] A strategy is *self-financing* if

$$V_{(x,y)} = x(t+1)S(t) + y(t+1)A(t), \tag{4}$$

for all $t = 1, \ldots, T-1$.

As we said above, in a multi-step single-instrument model, portfolio of risky instruments can be treated as a single underlying instrument. But in a case when we want to explicitly consider each risky instrument in a strategy portolios, it will be more convenient to use *multi-step multiple instrument model*, for more details see [3].

In the following definitions, we have $d = 1$ for a multi-step single instrument model and $d > 1$ for a multi-step multiple instrument model.

[Portfolio] A *portfolio* is the vector $(\mathbf{x}, y) = (x_1, \ldots, x_d, y)$ with coordinates representing positions in the corresponding securities.

[Strategy] A *strategy* is a sequence of portfolios

$$(\mathbf{x}(t), y(t)) = (x_1(t), \ldots, x_d(t), y(t)),\ t = 1, \ldots, T. \tag{5}$$

These are R^{d+1}-valued functions (i.e. random vectors), except for the initial portolio, which is determined at time 0 and so is deterministic. We shall write $(\mathbf{x}, y)$ for this sequence.

[Strategy value process] The *value* of a strategy a time t is $\mathrm{V}_{(\mathbf{x},y)}(t) = \sum_{j=1}^{d} x_j(t)S_j(t) + y(t)A(t), for t = 1, \ldots, T.$
$V_{(\mathbf{x},y)}(0) = \sum_{j=1}^{d} x_j(1)S_j(0) + y(1)A(0).$

2.2 Statistical arbitrage vs. standard arbitrage strategy

Below we recall definitions of a standard arbitrage and a statistical arbitrage, and later emphasize the relation between them.

[Standard arbitrage] A strategy $(\mathbf{x}, y)$ is an (standard) *arbitrage opportunity* in the underlying market if its value process satisfies $V_{(\mathbf{x},y)}(0) = 0$, $V_{(\mathbf{x},y)}(t) \geq 0$ for all n and for some t there is an ω such that $V_{(\mathbf{x},y)}(t, \omega) > 0$.

Here $\omega \in \Omega$ is an event of a probability space $(\Omega, \mathcal{F}, P)$, over which considered stochastic processes are defined.

Below for a sake of a notation simplicity we omit a subscript $(\mathbf{x}, y)$ from a $V_{(\mathbf{x},y)}$.

[Statistical arbitrage [18]] A statistical arbitrage is a zero initial cost, self-financing trading strategy with cumulative discounted trading profits $V(n)$ such that:

1. $V(0) = 0$,
2. $\lim_{t\to\infty} E^P[V(t)] > 0$,
3. $\lim_{t\to\infty} P(V(t) < 0) = 0$,
4. $\lim_{n\to\infty} Var[\Delta V(t) | \Delta V(t) < 0] = 0$.

If a strategy is explicitly stated as a mathematical model of a stochastic process, we can directly check if a strategy value stochastic process meets a formal definition of *statistical arbitrage* strategy, which is stated above. This approach was used in [14] and [15] among others.

Another possibility is to test empirically *ex post*, if a strategy is a statistical arbitrage one, using realization of a considered strategy value stochastic process.

Jarrow et al. [18] assumed that underlying stochastic process for strategy value can be stated as an *Arithemtic Brownian Motion*.

For an Arithmetic Brownian Motion Jarrow et al. [18] stated subset of a parameter space, for which such a strategy value process meets the definition of a statistical arbitrage strategy.

It is assumed that a following Arithmetic Brownian Motion called *unconstrained mean model* (UM), represents incremetal trading profits:

$$\Delta V_i = \mu i^{\theta} + \sigma i^{\lambda} z_i, \tag{6}$$

where z_i are $ii\mathcal{N}(0,1)$ random variables (but frequetly to meet empirical facts normality and independence assumptions are relaxed). The initial quantities z_0 and ΔV_0 are both zero by definition.

For the *UM model*, the following restrictions have to be *satisfied simultaneously* for a *statistical arbitrage opportunity* to exist [18]:

1. $R_1 : \mu > 0$,
2. $R_2 : -\lambda > 0$ lub $\theta - \lambda > 0$,
3. $R_3 : \theta - \lambda + (1/2) > 0$ oraz
4. $R_4 : \theta + 1 > 0$.

Thus, statistical arbitrage is defined by an **intersection of sub-hypotheses**.

Using DeMorgan's laws, the no statistical arbitrage null hypothesis can be also stated, with use of a following alternitve which is a negation of above stated conjunction for parameter conditions:

1. $R_1^c : \mu \leq 0$ lub
2. $R_2^c : -\lambda \leq \lambda$ oraz $\theta - \lambda \leq 0$ lub
3. $R_3^c : \theta - \lambda + (1/2) \leq 0$ lub
4. $R_4^c : \theta + 1 \leq 0$.

We can estimate parameters of an UM model using empirical realization of a considered strategy value process.

For a UM model and no statistical arbitrage null hypothesis (stated as logical alternative), Jarrow et al. proposed a statistical test with a Min-t statistics defined as follows:

$$Min - t = \min\{t(\hat{\mu}), t(\hat{\theta} - \hat{\lambda} + 0.5), t(\hat{\theta} + 1), \max\{t(-\hat{\lambda}), t(\hat{\theta} - \hat{\lambda})\}\}, \tag{7}$$

where $t(\hat{\mu}), t(-\hat{\lambda}), t(\hat{\theta} - \hat{\lambda} + 0.5), t(\hat{\theta} + 1)$ are respective t statistics for logical alternative constituent hypotheses, and $\hat{\mu}, \hat{\theta}, \hat{\lambda}$ are MLE (Maximum Likelihood Estimators) of an UM model parameters.

Taking into account Jarrow et al. definition of a statistical arbitratge strategy, when the probability of loss becomes zero in finite time T , i.e. $P(v(t) < 0) = 0$ for all $t \geq T$, this implies the existence of a standard arbitrage opportunity.

So, it was showed [15] that a standard arbitrage opportunity is a special case of statistical arbitrage.

[Standard arbitrage as a special case of a statistical arbitrage] For a standard arbitrage strategy V (self-financing) there exists a finite time T_m such that $P(V_{(\mathbf{x},y)}(t) > 0) > 0$ and $P(V_{(\mathbf{x},y)}(t) \geq 0) = 1$ for all $t \geq T_m$ and the proceeds of this profit can be deposited into money market account for the rest of the infinite time horizon.

It should be kept in mind that a statistical arbitrage is not a risk free profit generating strategy in contrast to a standard arbitrage strategy. A statistical arbitrage strategy in its assumption, tries to minimise a market volatility impact on a portfolio (market neutral strategy), but in practice it cannot fully reduce a portfolio market risk, as is in a case of a standard arbitrage.

3 Yield curve arbitrage as a case of statistical arbitrage strategy

In our work we consider an yield curve arbitrage as a kind of statistical arbitrage strategy involving portfolios of fixed-income instruments with specific maturity, which are rich or cheap points on the yield curve, for which convergence is anticipated by a statistical model for a yield curve.

In finance, arbitrage strategies imply positive returns in different markets regardless a market is in bull or bear state and are typically hedged in some way. This means that such strategies are meant to carry relatively low volatility or are neutral to changes in the key market variables. Fixed income arbitrage styles are fairly complex and usually reduce competition in trading. Less competition leads to higher risk-adjusted returns and weakens the performance [1].

Yield curve arbitrage is a relative value trading strategy within government debt or related interest rates. Hence, the strategy is about identifying overtly rich and cheap points on the yield curve with the assumption that these mispricings convergence in the near future, so that they can be traded profitably. This means, that short rate follows a stochastic mean-reverting process [6].

The idea behind studying the yield curve arbitrage as a trading strategy comes from the notion that some points of the term structure of interest rates may not at all times be in sync with each other. The yields for different maturities are not determined independently; they are all linked across the yield curve. As pointed out in earlier research by [6], trading of the rates that are out of the line with each other can result in highly attractive return profiles. No-arbitrage should guarantee that the yield curve is internally consistent, i.e. that all the forward rates are unique, and no 'textbook arbitrage' is possible. Nevertheless, the 'mispricings' on the yield curve arise from the nearby bond maturities trading at prices dissimilar enough. In other words, arbitrage opportunities related to the bond risk premia can exist despite the uniqueness of all forward rates [6].

Although changes in term structure of interest rates are relatively easy to interpret they are however very difficult to model and forecast due to no proper economic theory underlying such events. Yield curves are usually represented by multivariate yet quite sparse time series ie. at any point in time infinite dimensional curve is portrayed via relatively few points in a multivariate space of data and as a consequence multimodal statistical dependencies behind these curves are relatively hard to extract and forecast via typical multivariate statistical methods. We propose to model yield curves via reconstruction of joint probability distribution of parameters in functional space as a high degree polynomial. Thanks to adoption of an orthonormal basis, the MSE estimation of coefficients of a given function is just an average over a data sample in the space of functions. Since such polynomial coefficients are independent and have cumulant-like

interpretation: ie. they describe corresponding perturbation from an uniform joint distribution, our approach can also be extended to any d-dimensional space of yield curve parameters (also in neighboring times) due to controllable accuracy. We believe that this approach to modeling of local behavior of a sparse multivariate curved time series can complement prediction from standard models like ARIMA, that are using long range dependencies, but provide only inaccurate prediction of probability distribution, often as just Gaussian with constant width.

4 Statistical model for a yield curve forecasting

4.1 The concept of yield curve forecasting

As successful forecasting of financial time series could be often turned into profit, it makes it difficult to get a better prediction for the following value that just the previous value. However, above self-regulatory mechanism of the market does not restrict prediction of probability distribution of values, what is crucial for example for risk evaluation or Monte Carlo simulations.

Standard approaches to predict probability distribution of values like ARIMA usually models this distribution as Gaussian, often of constant width: predicts some value and its inaccuracy (standard deviation). In contrast, we will model this probability distribution using large number of independent coefficients describing joint distribution as polynomial - what turns out leading to very different and more complex distribution than standardly assumed Gaussian - for example multimodal in the discussed example.

Specifically, as it is difficult to obtain a better prediction than just the previous value, we will focus on sequence of differences between two succeeding values. In discussed example it will be 3-dimensional space of parameters of Yield Curves of Diebold-Li model [5] (for fixed $\lambda = 0.0609$), which dimensionality can be further increased for improved prediction by operating on time window: use a few previous values as context for prediction.

For convenience of fitting polynomial, we will first normalize each variable to nearly uniform distribution on $[0, 1]$. It can be done by transforming variables with CDF (cumulative probability distribution) of approximated distribution of this variable, for which we will use Laplace distributions as it agrees well with empirical CDF (Fig. 2)

Taking d such normalized variables, e.g. for different parameters in given or neighboring times, if uncorrelated they would come from nearly uniform $\rho \approx 1$ distribution on $[0, 1]^d$. We will model perturbation from this uniform density as linear combination of orthonormal polynomials $\rho(\mathbf{x}) = \sum_{\mathbf{j}} a_{\mathbf{j}} f_{\mathbf{j}}(\mathbf{x})$. It makes MSE optimal estimation very inexpensive [10]: $a_{\mathbf{j}} = \frac{1}{|X|} \sum_{\mathbf{x} \in X} f_{\mathbf{j}}(\mathbf{x})$ is just average over sample X. Coefficients for different $\mathbf{j}$ are independent and have multivariate cumulant-like specific interpretation, can be used for describing statistical dependencies between tested variables.

This article extends methodology from [7] for 1D variable (on example of Dow Jones Industrial Averages time series) into the case of multidimensional random variables.

4.2 Normalization to nearly uniform density

We will discuss on example of time series of 6470 (from 1993 to 2018) daily Yield Curve $\beta_1, \beta_2, \beta_3$ parameters $\{\beta_1, \beta_2, \beta_3\}_{t=1..n_0}$ for $n_0 = 6470$.

Time series are usually normalized for example to allow assumption of stationary process: such that joint probability distribution does not change while shifting position.

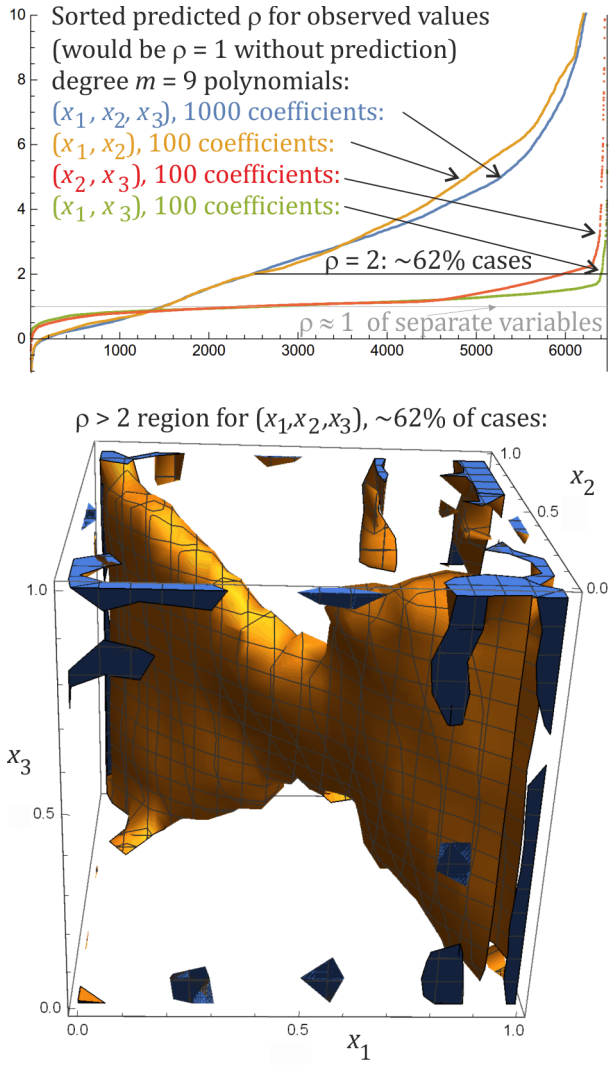

Figure 1: Each variable is normalized to have nearly uniform $\rho = 1$ density (PDF) on $[0, 1]$ range. Top: sorted predicted ρ for observed values is usually much higher than base $\rho = 1$ thanks to exploiting joint distribution. Four graphs correspond to joint distribution of parameters reconstructed with degree 9 polynomial for all three variables, or all their pairs. Surprisingly, we see that (x_1, x_2) gives even better predictions than for all 3 variables here. Bottom: region of predicted $\rho > 2$ for all 3 variables. From above plot we can read that observed values were there in $\approx 62\%$ of cases. In contrast to usually assumed Gaussian, distribution obtained from the real data turns out multimodal here. Density focused near diagonal for (x_1, x_2) means they are anti-correlated.

The standard approach, especially for Gaussian distribution, is to subtract mean value, then divide by the standard deviation. However, such normalization does not exploit local dependencies between values, what we are interested in.

Hence we will work on sequence of differences (errors, residues) from current value to its prediction based on previous values, which can be taken for example from ARIMA-like models. For simplicity we will use here the previous value as predictor: operate on $\beta_i(t+1)-\beta_i(t)$ sequence for $t = 1 \ldots n_1$ where $n_1 = n_0 - 1$. In practical applications $\beta_i(t)$ can be replaced with a more sophisticated predictior, for example exploiting long-range dependencies.

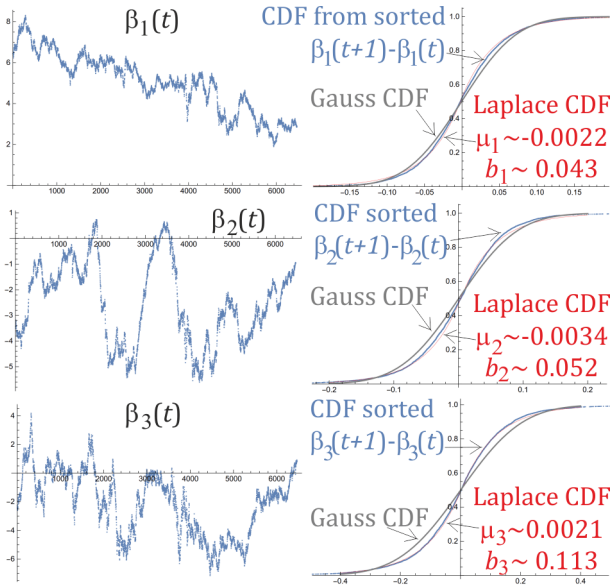

Figure 2: Left: time series of 6470 (from 1993 to 2018) daily Yield Curve $\beta_1, \beta_2, \beta_3$ parameters (Diebold-Li model [5]) fitted using $\lambda = 0.0609$ standard assumption. We will work on $x_i(t) := CDF_{Laplace(\mu_i,b_i)}(\beta_i(t+1) - \beta_i(t))$ time series: normalized to nearly uniform distribution on $[0, 1]$. Right: comparison of empirical CDF obtained from sorted values with CDF of Laplace and Gaussian distribution with estimated parameters - we will use Laplace as it has better agreement.

As shown in Fig. 2, such sequences of differences from predictor turns out to have nearly Laplace distribution: $\mathrm{g(y)}=\frac{1}{2b\exp\left(-\frac{|y-\mu|}{b}\right)} where maximum likelihood estimation of parameters is just: \mu =$ median of y, $b =$ mean of $|y - \mu|$.

For simplicity, we use Laplace distributions here to normalize variables to nearly uniform in $[0, 1]$, with separate parameters for different variables: $\mathrm{x}_i(t) := G_i(\beta_i(t+1) - \beta_i(t)) where\ \mathrm{G(y)}=\int_{-\infty}^{y} g(y')\,dy'$ is CDF of used distribution (Laplace here).

We will search for $\rho_X(x)$ density. To remove transformation (4.2) to retrieve the final density of $(\beta_1, \beta_2, \beta_3)$, observe that $P(y' = G^{-1}(x) \leq y) = P(x \leq G(y))$. Differentiating over y, we get $\rho_Y(y) = \rho_X(G(y)) \cdot g(y)$.

4.3 Hierarchical correlation reconstruction

After normalization, we have $\{x_1(t), x_2(t), x_3(t)\}$ time series with nearly uniform density of separate variables. Taking its d values: as different coordinates or neighboring in time, if uncorrelated they would come from nearly uniform distribution in $[0, 1]^d$ - difference from uniform distribution describes statistical dependencies in our time series. We will use a polynomial to describe this difference: estimate joint density for d neighboring values of x.

Assuming we have $\{\mathbf{x}^t\}_{t=1,\ldots,n} \subset [0, 1]^d$ ie. a vector sequence of neighboring values (we will discuss various possibilities later), we would like to model density of such vectors as polynomial. It turns out [10] that using orthonormal basis, which for multidimensional case can be products of 1D orthonormal polynomials, mean square (MSE, L^2)

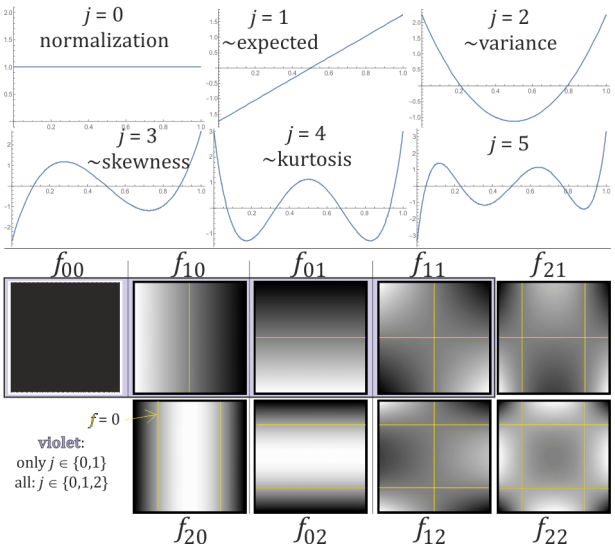

Figure 3: Top: the first 6 of used 1D orthonormal basis of polynomials ($\langle f,g\rangle = \int_0^1 fg\,dx$): $j=0$ coefficient guards normalization, the remaining functions integrate to 0, and their coefficients describe perturbation from uniform distribution. These coefficients have similar interpretation as cumulants, but are much more convenient for reconstruction of density. Bottom: 2D product basis $f_{\mathbf{j}}(\mathbf{x}) = f_{j_1}(x_1)f_{j_2}(x_2)$ for $m=2$: $j\in\{0,1,2\}$. The $j=0$ coordinates do not modify the corresponding variable - generally, the given coefficient describes statistical dependencies between coordinates having nonzero index.

optimization leads to elementary formula for estimated coefficients:

$$\rho(\mathbf{x}) = \sum_{\mathbf{j}\in\{0,m\}^d} a_{\mathbf{j}} f_{\mathbf{j}}(\mathbf{x}) = \sum_{j_1\dots j_d=0}^{m} a_{\mathbf{j}}\, f_{j_1}(x_1)\cdot\ldots\cdot f_{j_d}(x_d)$$

with estimated coefficients: $\mathrm{a}_{\mathbf{j}} = \frac{1}{n}\sum_{t=1}^{n} f_{\mathbf{j}}(\mathbf{x}^t)$

The basis used this way has $|B| = (m+1)^d$ functions. Besides, inexpensive calculation, this simple approach has also the very convenient property of coefficients being independent, giving each $\mathbf{j}$ unique value and interpretation. Independence also allows for flexibility of considered basis - instead of considering all $\mathbf{j}$, we can focus on more promising ones: for example with larger absolute value of coefficient, replacing negligible $a_{\mathbf{j}}$. Instead of MSE optimization, we can use often preferred: likelihood maximization [8], but it requires additional iterative optimization and introduces dependencies between coefficients.

Above f_j 1D polynomials are orthonormal in $[0,1]$: $\int_0^1 f_j(x)f_k(x)dx = \delta_{jk}$, getting (rescaled Legendre): $f_0 = 1$ and for $j = 1,2,3,4,5$ correspondingly:

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

$$3(70x^4-140x^3+90x^2-20x+1),$$

$$\sqrt{11}(252x^5-630x^4+560x^3-210x^2+30x-1).$$

They are plotted in the top of Fig. 3. f_0 corresponds to normalization. The $j=1$ coefficient decides about reducing or increasing the mean - have similar interpretation as expected value. Analogously $j=2$ coefficient decides about focusing or spreading given variable, similarly as variance. And so on: further f_j have similar interpretation

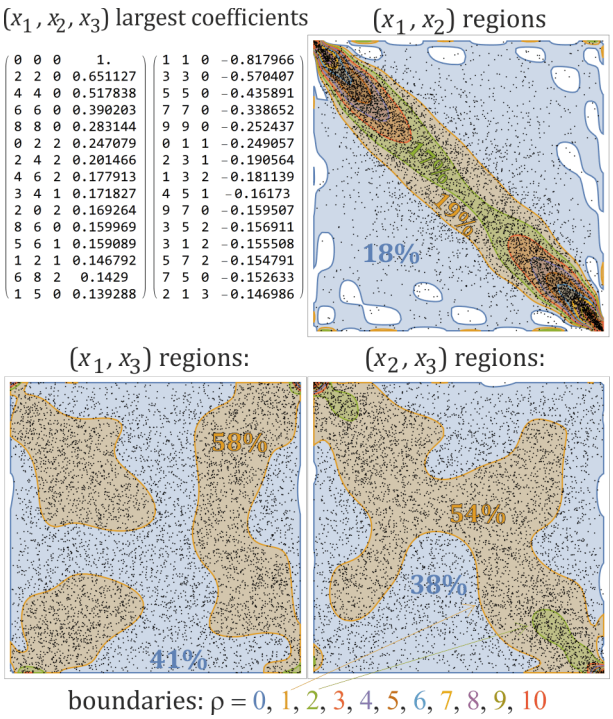

Figure 4: Modelling joint probability distribution of (x_1, x_2, x_3) variables, each normalized to nearly uniform distribution on $[0, 1]$. Top left: largest positive and negative obtained coefficients $a_{\mathbf{j}}$ of polynomial used as density estimation: corresponding to $a_{\mathbf{j}} f_{j_1}(x_1) f_{j_2}(x_2) f_{j_3}(x_3)$ correction from uniform distribution on $[0, 1]^3$. The remaining 3 region plots show the actual values (6469 tiny black points) and region plots of obtained density as degree $m = 9$ polynomial for all 3 pairs of variables, presenting non-uniformity of their joint distribution, especially for the (x_1, x_2) pair (top right). If these variables would be uncorrelated ($\rho \approx 1$), probability of a region would be proportional to its area. In contrast, the blue region here corresponding to estimated density $1 \leq \rho \leq 2$ has more than $1/2$ of area, but only $\approx 18\%$ of probability, which is mostly concentrated in the diagonal, near its edges. It can be seen in the largest coefficients: negative 110 gives anti-correlation, positive 220 increases probability of extreme values. The third variable appears much further in coefficient list, what means weaker statistical dependency.

as cumulants, however, while reconstructing density from moments is a difficult moment problem, presented description is directly coefficients of polynomial estimating the density.

For multiple variables, $a_{\mathbf{j}}$ describes only correlations between $C = \{i : j_i > 0\}$ coordinates, does not affect $j_i = 0$ coordinates, as we can see in the bottom of Fig. 3. Each coefficient has also a specific interpretations here, for example a_{11} decides between increase and decrease of second variable with increase of the first, a_{12} analogously decides focus or spread of the second variable.

Errors of such estimated coefficients come from approximately Gaussian distribution: $\tilde{a}_{\mathbf{j}} - a_{\mathbf{j}} \sim \mathcal{N}\left(0, \frac{1}{\sqrt{n}} \sqrt{\int (f_{\mathbf{j}} - a_{\mathbf{j}})^2 \rho \, d\mathbf{x}}\right)$ $For \rho = 1$ the integral has value 1, getting $\sigma = 1/\sqrt{n} \approx 0.013$ in our case. As we can see in Fig. 4, many coefficients are more that tenfold larger here: can be considered as essential, not a result of a noise.

5 Context-free modeling

We will work analogously to Markov modelling: model probability distribution of a new value basing on one or a few previous values - referred as context for the prediction. In contrast to standard Markov situation, values we are using are continuous: from $[0,1]$ or $[0,1]^2$ or $[0,1]^3$ here. The number of previous values considered for prediction is referred as the order of model. Order 0 or context-free models values as independent random variables: all from the same probability density. Order 1 is standard Markov process: uses context as one previous value to model probability distribution of the current value. Analogously for higher-order models: using a few previous values as context for prediction.

Let us start with basic context-free approach: just model joint distribution of observed values, not looking for statistical dependencies with neighboring values. In discussed example $d = 3$ dimension and $\mathbf{x}^t = (x_1(t), x_2(t), x_3(t))$ for $t = 1 \ldots n = n_1$.

We could also use only a subset of such variables, e.g. for $d = 2$ we have three possible pairs here: (x_1, x_2), (x_1, x_3) and (x_2, x_3). Reducing to $d = 1$ should lead to nearly uniform density due to the used normalization. Imperfection of e.g. assumed Laplace distribution used for this purpose will be corrected while fitting polynomial - in multidimensional case by $\mathbf{j}$ coefficients with $j_i = 0$ for all but a given coordinate.

Top of Fig. 1 contains evaluation for fitting $m = 9$ degree polynomial (for each variable) and 4 cases: $d = 3$ and all 3 pairs $d = 2$. It shows sorted ρ for predictions of actually observed values - the higher it is, the better prediction. Beside the used model, efficiency of such prediction strongly relies on objective statistical dependencies between these variables - for example will fail if they are uncorrelated. Surprisingly, we can see from this plot that using just first two variables gives better prediction than for all three here - the third variable is weakly correlated, more strongly with x_2 (red plot) than with x_1 (green).

Such plots evaluating prediction also allow to calibrate density plots, including interpretation for negative densities being artifact of the presented method: we can see that $\approx 3\%$ of cases here got negative density, hence $\rho > 0$ region is expected to indicate the proper value in $\approx 97\%$ of cases.

Analogously we can interpret further lines of these graphs, for example we can see that the blue and orange lines in Fig. 1 have $\rho > 2$ in $\approx 62\%$ of cases. Lower 3D density plot shows this $\rho > 2$ region in $[0,1]^3$: while it contains only $\approx 14\%$ of the volume, what would be its probability for uncorrelated variables, it contains here much more: $\approx 62\%$ of points from the sample.

2D regions for multiple isolines of constant ρ for all 3 pairs are presented in Fig. 4, which alternatively can be obtained as marginalization of 3D density. It gives better visualization of strong statistical dependance between x_1 and x_2, and much weaker with x_3. Percentages indicate probability of cases observed in a given region, for example for $0 < \rho < 1$ for light blue regions. The missing probability is localized in further regions. Tiny black points are the actual 6469 data points - presented density region plots MSE fit degree $m = 9$ polynomial to sum of Dirac deltas in all points of the sample.

Figure 4 also contains the largest positive coefficients (left, always starts with 1 for normalization), and negative (right). They provide unique independent cumulant-like description of statistical dependencies in modelled sample. For example largest positive is $a_{220} \approx 0.65$, what corresponds to parabola in first and second variable: statistical avoidance of being both near the center. Largest negative is $a_{110} \approx -0.82$, saying that with growth of the first variable, there comes reduction of the second.

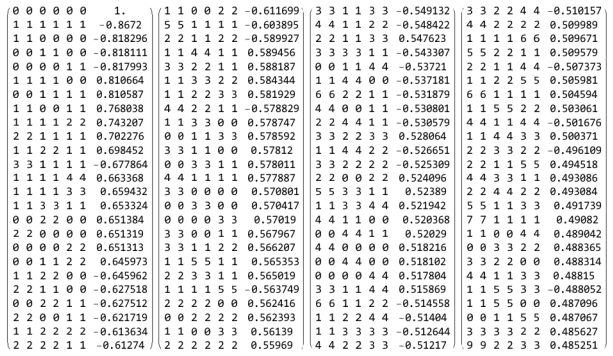

Figure 5: The most significant statistical dependencies: 100 largest absolute value coefficients for the million coefficient model: $m = 9$, $d = 6$ modeling three neighboring pairs. The corresponding 6 coordinates are: $(x_1(t), x_2(t), x_1(t-1), x_2(t-1), x_1(t-2), x_2(t-2))$. The list obviously starts with, $a_{000000} = 1$ corresponding to normalization (the remaining functions integrate to 0). Then we have "11" pairs as already seen in Fig. 4, this time in all 3 positions with nearly identical coefficient (tiny differences come from occurrences at the beginning and the end). Then we see large $a_{111100} \approx a_{001111} \approx 0.81$ positive coefficient describing dependency between neighboring pairs: saying e.g. that with growth of the first 3 variables, the fourth is also likely to grow. While using up to $m = 9$ order, we see that the "9" index appears only in the last: 100th position here - dominant statistical dependencies are described by relatively low order polynomials here. Assuming uniform density on $[0,1]^6$, these coefficients should come for a Gaussian distribution centered in zero with $\sigma = 1/\sqrt{n} \approx 0.012$, hence the above coefficients $> 40\sigma$ can be seen as statistical significant: should not be interpreted as a result of noise.

6 Context-dependent modelling

The next step is trying to exploit statistical dependencies between values neighboring in time: based on context representing the history, for example a few previous values, or extracted crucial information about the past for example in some dimensionality reduction method like PCA: corresponding to the largest eigenvalues of covariance matrix.

For simplicity and reducing dimension we will work on (x_1, x_2) pairs as x_3 has much weaker correlation. We have considered one previous pair as the context ($d = 4$): $\mathbf{x}^t = (x_1(t), x_2(t), x_1(t-1), x_2(t-1))$, or two previous pairs ($d = 6$): $\mathbf{x}^t = (x_1(t), x_2(t), x_1(t-1), x_2(t-1), x_1(t-2), x_2(t-2))$ for $t = 1 \dots n$ which is $n_1 - 1$ or $n_1 - 2$ correspondingly.

The most significant 100 coefficients for the largest considered model ($d = 6$, $m = 9$, 10^6 coefficients) are presented in Fig. 5. Each is independent and has a specific meaning: correction $a_{\mathbf{j}} \prod_{i=1}^{d} f_{j_i}(x_i)$ to initially uniform density on $[0,1]^6$ - providing unique description of statistical dependencies in the observed data sample. For certainty that they are not just a result of random noise, $\sigma \approx 0.012$ here for $\rho = 1$ (uniform density on $[0,1]^d$), which is exceeded a few dozens of times in this sample.

The results of this $m = 9$ order 2 (right, 10^6 coefficients) and analogously order 1 (left, 10^4 coefficients) are presented in Fig. 6. Especially, the order 2 model gives nearly perfect agreement: in $\approx 80\%$ of cases, the actually observed value is in the smallest predicted region (red boundary for $\rho = 10$). However, this is fitting a million coefficient model to just 6467 data points - polynomial approaching spikes in data points.

The proper prediction evaluation should test generalization capabilities, what is presented in Fig. 7. These tests of 27 models first randomly split data sample into two disjoint subsets, use the first one to calculate coefficients, and test on the second subset. We see that the million coefficient model ($d = 6, m = 9$) in $\approx 25\%$ of cases gives negative

density - has strong overfitting. However, focusing on predicted high density regions, it most frequently gives the proper prediction.

Finally, we see that the choice of the most appropriate model is a difficult question, it might be worth to consider a few models and somehow mix their predictions.

7 Conclusion and further perspectives

While there is usually assumed Gaussian distribution for financial data, in reality it is often much more complicated, including multimodal distributions. There was presented basics of systematic approach for modelling such joint distribution with a polynomial - what allows to effectively find and work with parametrisation using thousands of unique and independent cumulant-like coefficients, each one having a specific interpretation, and being inexpensive to calculate.

The used example applied basic methodology for educative reasons, we plan to investigate its extensions in the future, for example:

- Selective choice of basis: we have used complete basis of polynomials, what makes its $(m+1)^d$ size impractically large especially for high dimensions. However, usually only a small percentage of coefficients is above noise - we can selectively choose and use a sparse basis of significant values instead - describing real statistical dependencies. Alternatively, we can selectively reduce polynomial degree for some of variables.

- Adaptive choice of coefficients: we have assumed that coefficients are constant in time, what corresponds to stationarity of time series. However, in practice it is often non-stationary, what can be modelled using coefficients being not average of all values of a given function like here, but some local averages instead, for example with exponentially decaying weight [8].

- Long-range value prediction: combination with state-of-art prediction models exploiting long-range dependencies, for example using a more sophisticated (than just the previous value) predictor of the current value.

- Improving information content of context used for prediction: instead of using a few previous values as the context, we can use some features e.g. describing long-range behavior like average over a time window, or for example obtained from dimensionality reduction methods like PCA (principal component analysis).

While the approach used here was analogous to Markov modelling, an alternative approach to consider in the future is using time as one of coordinates, e.g. fit polynomial to $(x_1(t), x_2(t), t)$ triples in a moving time window. It would require much lower dimension, allowing to model longer correlations directly. It also allows working with continuous time.

References

[1] AGARWAL, Vikas, et al. Risk and return in convertible arbitrage: Evidence from the convertible bond market. Journal of Empirical Finance, 2011, 18.2: 175-194.

[2] BURGESS, Andrew Neil, et al. A computational methodology for modelling the dynamics of statistical arbitrage. 2000. PhD Thesis. University of London.

[3] CAPIŃSKI, Marek; KOPP, Ekkehard. Discrete models of financial markets. Cambridge University Press, 2012.

[4] DAVIS, Richard A.; LII, Keh-Shin; POLITIS, Dimitris N. Remarks on some nonparametric estimates of a density function. In: Selected Works of Murray Rosenblatt. Springer, New York, NY, 2011. p. 95-100.

[5] DIEBOLD, Francis X.; LI, Canlin. Forecasting the term structure of government bond yields. Journal of econometrics, 2006, 130.2: 337-364.

[6] DUARTE, Jefferson; LONGSTAFF, Francis A.; YU, Fan. Risk and return in fixed-income arbitrage: Nickels in front of a steamroller?. The Review of Financial Studies, 2007, 20.3: 769-811.

[7] DUDA, Jarek. Exploiting statistical dependencies of time series with hierarchical correlation reconstruction. arXiv preprint arXiv:1807.04119, 2018.

[8] DUDA, Jarek. Hierarchical correlation reconstruction with missing data, for example for biology-inspired neuron. arXiv preprint arXiv:1804.06218, 2018.

[9] DUDA, Jarek. Normalized rotation shape descriptors and lossy compression of molecular shape. arXiv preprint arXiv:1509.09211, 2015.

[10] DUDA, Jarek. Rapid parametric density estimation. arXiv preprint arXiv:1702.02144, 2017.

[11] DUDA, Jarek; SNARSKA, Małgorzata. Modeling joint probability distribution of yield curve parameters. arXiv preprint arXiv:1807.11743, 2018.

[12] EDGEWORTH, Francis Y. On the probable errors of frequency-constants (contd.). Journal of the Royal Statistical Society, 1908, 71.4: 651-678

[13] GALLANT, A. Ronald. SNP: nonparametric time series analysis. In: Macroeconometrics and Time Series Analysis. Palgrave Macmillan, London, 2010. p. 245-249.

[14] GÖNCÜ, Ahmet; AKYILDIRIM, Erdinç. Statistical arbitrage with pairs trading. International Review of Finance, 2016, 16.2: 307-319.

[15] GÖNCÜ, Ahmet. Statistical arbitrage in the Black–Scholes framework. Quantitative Finance, 2015, 15.9: 1489-1499.

[16] HOGAN, Steve, et al. Testing market efficiency using statistical arbitrage with applications to momentum and value strategies. Journal of Financial economics, 2004, 73.3: 525-565.

[17] HYVÄRINEN, Aapo; OJA, Erkki. Independent component analysis: algorithms and applications. Neural networks, 2000, 13.4-5: 411-430.

[18] JARROW, Robert, et al. An improved test for statistical arbitrage. Journal of Financial Markets, 2012, 15.1: 47-80.

[19] LAZZARINO, Marco, et al. What is statistical arbitrage?. Theoretical Economics Letters, 2018, 8.05: 888.

[20] MANDELBROT, Benoit. The Pareto-Levy law and the distribution of income. International economic review, 1960, 1.2: 79-106.

[21] PARZEN, Emanuel. On estimation of a probability density function and mode. The annals of mathematical statistics, 1962, 33.3: 1065-1076.

[22] PEDERSEN, Lasse Heje. Efficiently inefficient. Princeton University Press, 2015.

[23] SHOHAT, James Alexander; TAMARKIN, Jacob David. The problem of moments. American Mathematical Society (RI), 1950.

[24] VARANASI, Mahesh K.; AAZHANG, Behnaam. Parametric generalized Gaussian density estimation. The Journal of the Acoustical Society of America, 1989, 86.4: 1404-1415.

[25] WEINBERGER, Marcelo J.; SEROUSSI, Gadiel; SAPIRO, Guillermo. The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS. IEEE Transactions on Image processing, 2000, 9.8: 1309-1324.

[26] WILKS, Samuel S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. The annals of mathematical statistics, 1938, 9.1: 60-62.

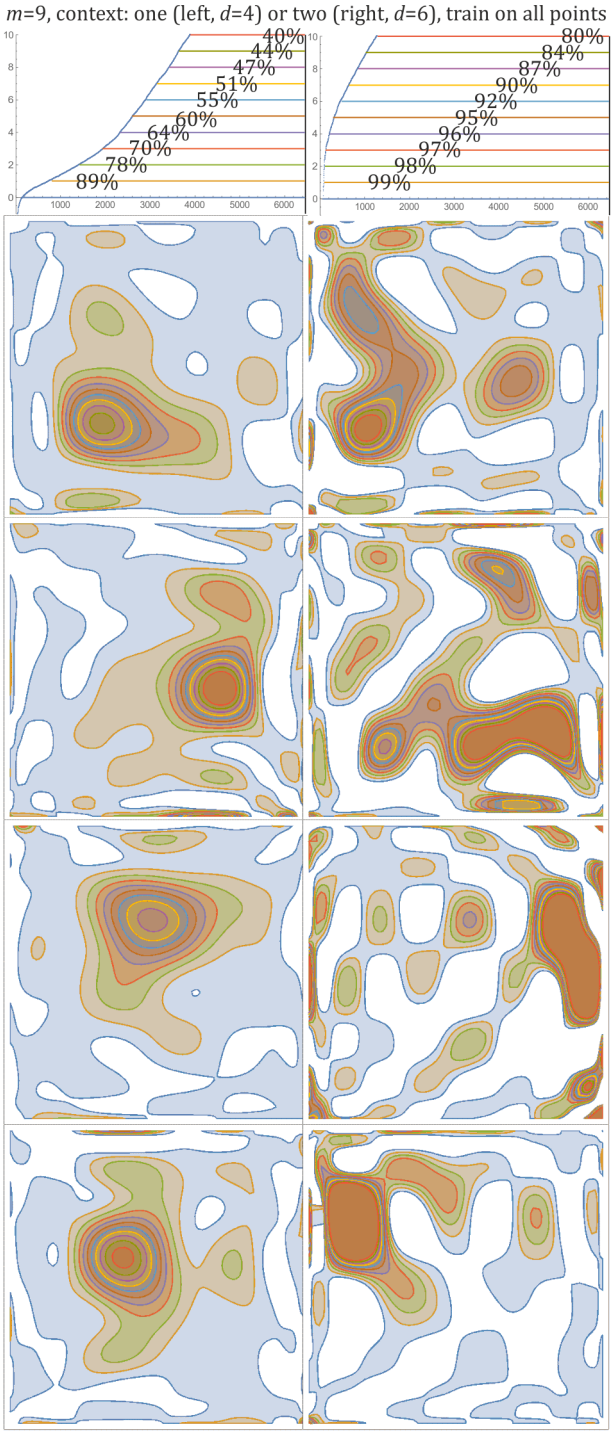

Figure 6: Top: sorted predicted densities for the actually observed values for two degree $m = 9$ models: using one (left, 10^4 coefficients) or two (right, 10^6 coefficients) previous (x_1, x_2) pairs as the context. It contains percentages of cases when density was above $\rho = 0, 1, \ldots, 10$ thresholds, drawn below in region plot. Bottom: region plots for predicted densities in some four random points in time, the same for both models. We can see overfitting, especially in the right column, with large white regions denoting predicted $\rho < 0$. This model fits million coefficients to size 6467 sample - approaching density as polynomial with spikes in the used points. The proper model evaluation should test its generalization capabilities instead: estimate coefficients on a subset of sample, and test on the remaining points - its results are presented in Fig. 7.

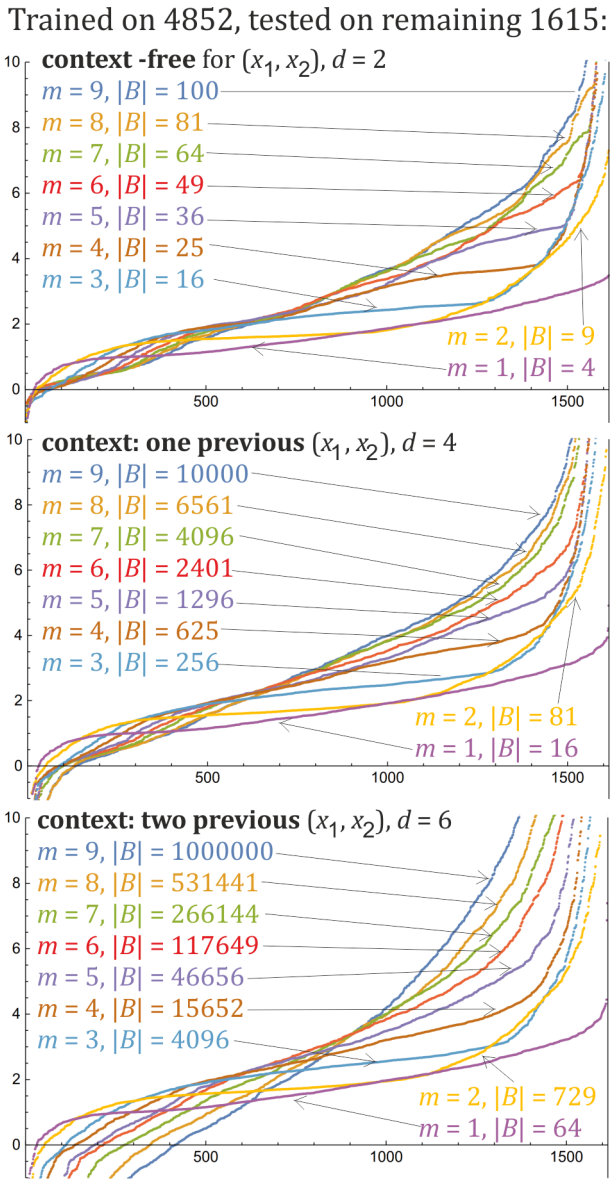

Figure 7: The proper evaluation of 27 models: sorted predicted densities for the actually observed values (the higher, the better prediction) in randomly chosen 25% of data points, using the remaining 75% of points to train the model (estimate coefficients). There were used context-free ($d = 2$), order 1 ($d = 4$) and order 2 ($d = 6$) models for (x_1, x_2) and all degrees $m = 1, \ldots, 9$. The highest (blue) plots are analogous as in Fig. 6, but this time with disjoint training and test sets to prevent overfitting.