Testing and Clustering with Gauss linear assumption for a household data

Valentina Fedorova,
Ilia Nouretdinov,
Alex Gammerman
Computer Learning Research Centre
Royal Holloway, University of London
{name}@cs.rhul.ac.uk

Evgeny Ivin
Moscow School of Economics
Lomonosov Moscow State University

Abstract

The paper considers an application of martingales for testing the hypothesis that the data are generated by the Gauss linear model. In addition, the martingales are also used for suggesting some possible clustering of the data. A Russian household dataset is taken to show results of this approach. In particular, with the help of martingales the households were split into groups with a similar structure of expenditure. It was assumed that within each group dependence of the household total consumption on its income can be described by the Gauss linear model. The structure of expenditure was tested for the households within the groups.

Keywords: testing assumption, martingales, conformal prediction, clustering.

1 Introduction

In many applications an assumption is made that the data are generated by the Gauss linear model - that our response variables y are linearly dependent on the vectors of explanatory variables $\mathbf{x}$ with the normally distributed random noise ξ:

$$y = \beta_1 + \beta_2 \cdot \mathbf{x} + \xi$$

where β_1 is unknown coefficient and β_2 is a vector of coefficients of the same length as $\mathbf{x}$. Our aim is to test this assumption for a special class of predictors, called conformal predictors (Vovk et al., 2005). Conformal predictors allow us to estimate confidence of the predictions and their main property is validity: for a given significance level ϵ the number of errors would not exceed ϵ. It has been proved that this property of validity holds for many different models including i.i.d. model[1], Gauss linear model and some other models. The underlying models (assumptions) can be tested using martingales: they allow us to accumulate evidence against the assumptions. That is if the martingale value is large then we reject the assumption. In this paper we use the martingale testing for the Gauss linear assumption and present results of the testing for a Russian household dataset.

[1] independent and identically distributed data

The martingales can also help to split the data into certain clusters when a martingale changes its behavior. In the work Ho (2005) a similar problem of splitting was considered for the i.i.d. assumption: how the data sequence may be split by a martingale into several i.i.d. clusters. Within each of them data examples followed a random order. However, the Gauss linear assumption does not require anything of the order of examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, so we can sort them in a way convenient for testing our assumption. In this work we use ranking by $\mathbf{x}_i$ (that is one-dimensional). As we will see, this allows us to find deviations from the assumption effectively and to clusterize the data by the deviations so that a cluster is a group of households with a similar level of income and a structure of expenditure. This direction of study was mostly motivated by the work Aivazian and Kruglyak (2010) where authors describe the data as a mixture of several clusters with the normal distributions.

The rest of the paper is organized as following: in section 2 we briefly describe the data and used model, in section 3 we outline the concepts of martingale testing, in section 4 we explain the ideas of our clustering, in section 5 we present results of analysis of the data and finally we summarize the work in section 6.

2 Data

We will study the part of the Russian households dataset (`http://www.micro-data.ru`) collected during the first quarter of 2003 year. It consists of 53149 households, each household is described by several attributes. The attributes could be numerical (i.e. income, number of members of family etc.) or categorized (i.e. householder's profession). To split the dataset into clusters we consider only links between the income budget of household and its total consumption. We use the following Gauss linear model:

$$\log_{10}(\text{consumption}) = \beta_1 + \beta_2 \cdot \log_{10}(\text{income}) + \xi, \tag{1}$$

where $\beta = (\beta_1, \beta_2)$ is a real number vector of coefficients and $\xi \sim N(0, \sigma)$ is independent and normally distributed noise.

Figure 1 plots the data on the logarithmic scale of income and consumption. The dark dashed line represents function $\log_{10}(\text{consumption}) = \log_{10}(\text{income})$.

The paper will study households with the income budget more than 500 rubles during the first quarter of 2003 year. We suppose that for the rest of households studied time period does not give enough information on their income. The dark solid vertical line in figure 1 shows the income level of $500 = 10^{2.7}$ rubles. We can see that the points that lie to the left of the line look like outliers. From this considerations for our analysis we will use the dataset of 53082 households with income greater than 500 rubles during the quarter.

3 Testing

In this section we describe the testing of the hypothesis that data agree with the Gauss linear model. It proceeds in two steps. The first one is a calculation of p-values as an output of conformal predictors. And the second one is a calculation of martingales that reflect the "correctness" of the assumption.

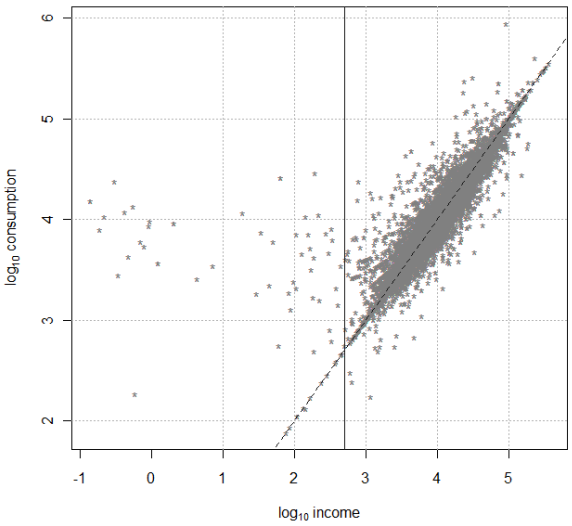

Figure 1: The Russian household data is given for the first quarter of 2003 year: income and total consumption are on the logarithmic scale. The dark dashed line is for $\log_{10}(\text{consumption}) = \log_{10}(\text{income})$ and the dark solid vertical line is for income level of 500 rubles ($\log_{10}(500) = 2.7$ in the scale).

3.1 Conformal predictors

We start testing by obtaining a sequence of p-values using conformal predictor for the Gauss linear assumption. Recalling the model (1) we can denote i-th example of the data as (x_i, y_i), where $x_i = \big(1, \log_{10}(\text{income})\big)$ is a vector of $p = 2$ attributes and $y_i = \log_{10}(\text{consumption})$ is a label. To describe the process of p-value calculation we will use the following notation: $\hat{y}_i^l$ is the Least Square prediction for x_i, based on examples $\Big\{(x_1, y_1), \ldots, (x_l, y_l)\Big\}$; $\hat{y}_n$ is shorthand for $\hat{y}_n^{n-1}$; $s_{n-1}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n-1}(y_i - \hat{y}_i^{n-1})^2$ is the standard estimation of variance of ξ from the first l examples; X_l is a matrix consists of rows $x_1, \ldots, x_l$; Y_l is a column-vector of labels $y_1, \ldots y_n$; symbol $'$ denotes operation of transposition.

To calculate p-values by a conformal predictor we specify a "non-conformity" measure of an example (x_i, y_i) with respect to the given examples $\Big\{(x_1, y_1), \ldots, (x_l, y_l)\Big\}$ as

$$\alpha_i = |y_i - \hat{y}_i^l|. \tag{2}$$

The "non-conformity" measure is a way of scoring how i-th example is different from the rest of given examples. It is known that the random variable:

$$t_i = \frac{y_i - \hat{y}_i}{s_{i-1}\sqrt{1 + x_i'(X_{i-1}'X_{i-1})^{-1}x_i}}$$

has a t-distribution with $i - p - 1$ degrees of freedom (see, eg. Ryan, 1997, p. 17). Consider the value t_i provided by the given data and denote $t_{i-p-1}^{|t_i|}$ for the $|t_i|$-th quintile of t-distribution with $i - p - 1$ degrees of freedom. Then p-value is estimated by the equation

$$p_i = 2 \cdot (1 - t_{i-p-1}^{|t_i|}).$$

This equation for p-value agrees with the equation (2) for "non-conformity" measure (for details see Vovk et al., 2005, pp. 202 - 203). A p-value is a number from $[0, 1]$. It is close to 1 if the pair (x_i, y_i) fits well to the model (1) with the parameters estimated by the given examples $\Big\{(x_1, y_1), \ldots, (x_l, y_l)\Big\}$.

To obtain a sequence of p-values we use conformal predictor in on-line mode (examples arrive one after another) and the algorithm 1 summarizes this process.

Algorithm 1 Generating p-values

Input: $\Big\{(x_1, y_1), \ldots, (x_n, y_n)\Big\}$ – data, sorted by x_i

Output: $\Big(p_1, \ldots, p_n\Big)$ – sequence of p-values

for $i = p + 1$ **to** n **do**

$X_{i-1} = \begin{pmatrix} x_1 \\ \vdots \\ x_{i-1} \end{pmatrix}$

$Y_{i-1} = (y_1 \ldots y_{i-1})'$

(x_i, y_i) is a new example

$\hat{\beta} = (X'_{i-1}X_{i-1})^{-1}X'_{i-1}Y_{i-1}$

$\hat{y}_i = x_i \cdot \hat{\beta}$

$s^2_{i-1} = \frac{1}{i-p-1}\sum_{j=1}^{i-1}(y_j - \hat{y}_j^{i-1})^2$

$t_i = \frac{y_i - \hat{y}_i}{s_{i-1}\sqrt{1+x'_i(X'_{i-1}X_{i-1})^{-1}x_i}}$

find quintile of t-distribution $t^{|t_i|}_{i-p-1}$

$p_i = 2 \cdot (1 - t^{|t_i|}_{i-p-1})$

end for

In the paper we assume only the Gauss linear model (1) for dependence between x_i and y_i. It allows us to put data in the order relevant for the problem (sort the households by the value of income). Output of the algorithm 1 is a sequence of p-values: that is the p-value p_i that corresponds to the data point (x_i, y_i). It was proved that the p-values distribute independently and uniformly in [0,1] if the assumption of conformal predictor is satisfied (see Vovk et al., 2005, Theorem 8.1).

3.2 Martingales

The main tool for testing our assumption is martingales. Denote $z_i = (x_i, y_i), i = 1, \ldots n$ and assume that each z_i is generated by the probability distribution $\mathbf{P}$. Then a sequence of non-negative random variables $M_0, M_1, \ldots$ is a *martingale* if each M_n is a measurable function of $z_1, \ldots, z_n$ and

$$M_n = \mathbf{E}(M_{n+1}|M_1, \ldots, M_n),$$

where $\mathbf{E}$ refers to the expected value in the probability space $\mathbf{P}$. We set $M_0 = 1$. After observing a new example, the martingale value reflects the strength of evidence against the assumption that probability distribution $\mathbf{P}$ generates the data.

According to Ville's inequality (see Ville, 1939, p. 100)

$$\mathbf{P}\Big\{\exists n : M_n \geq C\Big\} \leq 1/C, \forall C > 0.$$

it is unlikely that any M_n would have a large value.

3.2.1 Martingale based on p-values

In this paper we use the so called sleepy jumper martingale introduced in Vovk et al. (2003) as a generalization of power martingales.

A family of *power martingales* is defined as

$$M_n^{\varepsilon} = \prod_{i=1}^{n} (\varepsilon p_i^{\varepsilon - 1}), \tag{3}$$

where $\varepsilon \in [0, 1]$ and p_i-s are the p-values output by algorithm 1. If the p-values are distributed independently and uniformly then the power martingales keep value close to 0.

The sleepy jumper martingale approximates the performance of the best power martingale (martingale from the family of power martingales with the largest final value). For a sequence $\widetilde{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$, the equation (3) is generalized as

$$M_n^{\widetilde{\varepsilon}} = \prod_{i=1}^{n} (\varepsilon_i p_i^{\varepsilon_i - 1}),$$

where $\varepsilon_i \in [0, 1], i = 1, \ldots n$. Then *sleepy jumper martingale* is defined as

$$M_n = \int_{[0,1]^{\infty}} M_n^{\widetilde{\varepsilon}} \mu d\widetilde{\varepsilon}, \tag{4}$$

where μ is a probability distribution in $[0, 1]^{\infty}$. The sleepy jumper distribution is specified by two parameters: R for probability of awake/asleep and S for probability of change current ε_i to another value from $[0, 1]$. To obtain the distribution μ sleepy jumper creates sequences of states. When the state is (s, j) the new state is generated from the Cartesian product $\{awake, asleep\} \times [0, 1]$ with respect to the following conditions:

- if s is asleep then the new state is (asleep, j) with the probability R or it is (awake, j) with the probability $1 - R$;
- if s is awake then the new state is $(\overline{s}, \overline{j})$: $\overline{s}$ is awake with the probability $1 - R$ or it is asleep with the probability R, $\overline{j}$ takes random value from $[0, 1]$ with the probability J or it stays the same j with the probability $1 - J$.

Sleepy jumper starts from the state $(s_1, j_1) =$ (asleep, 1) and further generates sequences $(s_1, j_1), (s_2, j_2), \ldots$. From a sequence of states corresponding sequence $\widetilde{\varepsilon}$ for equation (4) is defined as

$$\varepsilon_i = \begin{cases} j_i, & \text{if } s_i = \text{ awake;} \\ 1, & \text{otherwise.} \end{cases}$$

This process gives the distribution μ and we can calculate the sleepy jumper martingale (4).

4 Clustering

By studying a martingale performance we can split the data into clusters. Since the data are sorted by values x_i the martingale tests whether y_i can be reliably predicted from x_i and previous data according to the Gauss linear assumption. We assume that the points where the martingale performance becomes different (for example by visual inspection) can reflect changes in the structure of expenditure. We split the data by those points and then each cluster is analyzed separately.

We will present results and additional comments about clustering for the data in the section 5.

5 Results

We study the data on the logarithmic scale and sorted it by value of income.

5.1 Testing and clustering

Figure 2 shows sleepy jumper martingale performance for the household data. The x-axis is for the number of a household (as sorted by income) and y-axis is for the value of the martingale.

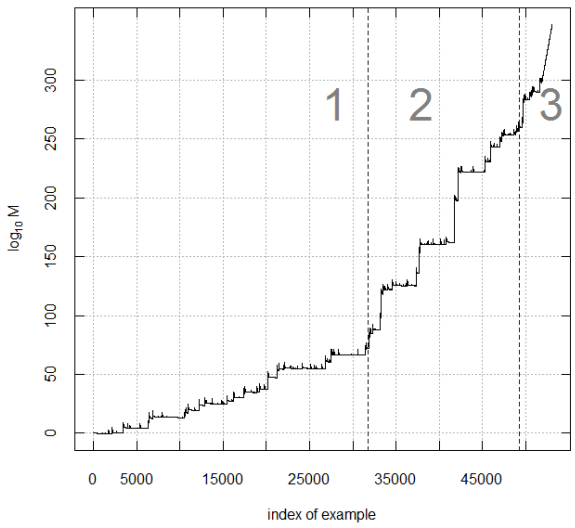

Figure 2: Sleepy jumper martingale is performing on the household data. The final value of the martingale is higher than 10^{300}. The dark dashed lines roughly split the household data according to the different growth rate of the martingale.

The large final value of the martingale means that the Gauss linear assumption is falsified for the data.

According to the martingale performance in figure 2 we split the data into three clusters:

- in the first cluster the martingale grows insignificantly except of several points where p-values occasionally appear to be very close to zero;

- in the second cluster the martingale tends to grow faster;
- in the third cluster the growth of the martingale becomes very fast.

To illustrate the splitting on the data the dashed vertical lines in figure 3 show the borders that we see in the figure 2 for the martingale performance. But unlike figure 2 this picture is plotted in terms of income value itself (not the number of the household). There we can see that, for example, in the first cluster (the income of households less than 14000 rubles) the whole income is used for the consumption (up to the noise). For the households with higher income the consumption is usually smaller than the income. It is clear from the figure where the diagonal line corresponds to the equality between the income and the consumption.

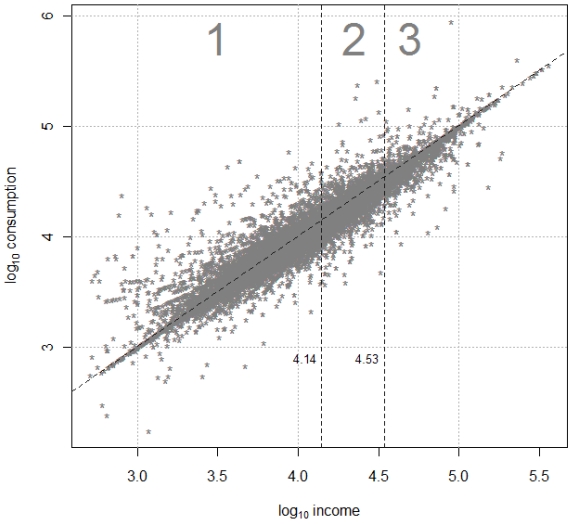

Figure 3: The household data are split according to the clusters found in the figure 2.

5.2 Expenditure structure for clusters

The martingale performance for the whole household data showed that the Gauss linear assumption is falsified. But the martingale values were changing differently. Apparently for each cluster the assumption is violated in a different way. The clusters can be useful if they correspond to a certain expenditure structure. We will check within each cluster whether the spending on food, non-food goods, services and alcohol is proportional to the total consumption.

For the testing, each example is represented by the total consumption as an attribute and a category of spending as a label. We use the conformal predictor for the Gauss linear assumption to generate the p-values and finally calculate performance of the martingale for each cluster.

Figure 4 shows the analysis of the expenditure structure inside the three clusters. A large martingale value shows that the linear dependence between the category of spending and the total consumption is falsified or normality of the noise is broken.

The results in figure 4 show that most of martingales take large values. Only in the second cluster most of the martingale values do not increase. There hypotheses about the expenditure structure can work. An exception is the spending on alcohol, it seems to be less predictable than other spendings.

5.3 Conformal prediction and validity for clusters

Let us now compare the clusters by applicability of the Gauss linear model for predictions. We use the on-line conformal predictor for the Gauss linear assumption and we check whether its predictions are valid.

If the assumption were satisfied for data then predictions would be automatically valid, i.e. the error rate does not exceed beforehand chosen significance level (for details see Vovk et al., 2005, section 8.5). The question is whether the predictions are valid within the clusters.

The error rate for the significance level of 5% and median accuracy of obtained predictions are plotted in figure 5. The dashed grey line on the graphs of errors (left column in figure 5) shows the allowed level of 5% for errors.

So the Gauss linear model is practically applicable inside the first, second and most of the third clusters.

6 Discussion and conclusions

In the paper we use the martingale approach for testing the Gauss linear assumption and clustering data according to the martingale performance. From what we presented here the following results can be concluded:

- The Gauss linear assumption for the whole dataset is falsified.
- According to the martingale behavior we split the data into the clusters. Because the dataset was sorted by income the clusters correspond to some interval of household income values. We have seen that different clusters show different consumption behavior. Within each of the clusters conformal predictor for the Gauss linear assumption produce practically valid predictions.

For further research we can try to split clusters into smaller ones taking into account more attributes.

Acknowledgements:

This work has been supported by EraSysBio+ grant funds from the European Union, BBSRC and BMBF "Living with uninvited guests: comparing plant and animal responses to endocytic invasions" to the Salmonella Host Interactions Project European Consortium, SHIPREC; Animal Health and Veterinary Laboratories Agency of Department for Environment, Food and Rural Affairs on Machine learning algorithms for analysis of large veterinary datasets.

We are grateful to S.A.Aivazian and M.V.Kruglyak for providing the data and useful information for this work. We also would like to thank Royal Holloway University of London for support and funding.

References

S. A. Aivazian and M. V. Kruglyak. *Typology of the Russian Household's Consumption Behaviour (Metodology of Investigation, Informational Maintenance and Experimental Approbation) (Rus)*. CEMI Russian Academy of Science, Moscow, 2010.

S.-S. Ho. A martingale framework for concept change detection in time-varying data streams. In *Proc. Int. Conf. on Machine Learning (ICML 2005)*, 2005.

T. P. Ryan. *Modern Regression Methods*. John Willey & Sons, New York, 1997.

`http://www.micro-data.ru`.

J. Ville. *Etude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

V. Vovk, I. Nouretdinov, and A. Gammerman. Testing exchangeability on-line. In *Proceedings of the 20th International Conference on Machine Learning (ICML 2003)*, pages 768–775, Washington, DC, 2003.

V. Vovk, A. Gammerman, and G. Shafer. *Algoritmic learning in a random world*. Springer, New York, 2005.

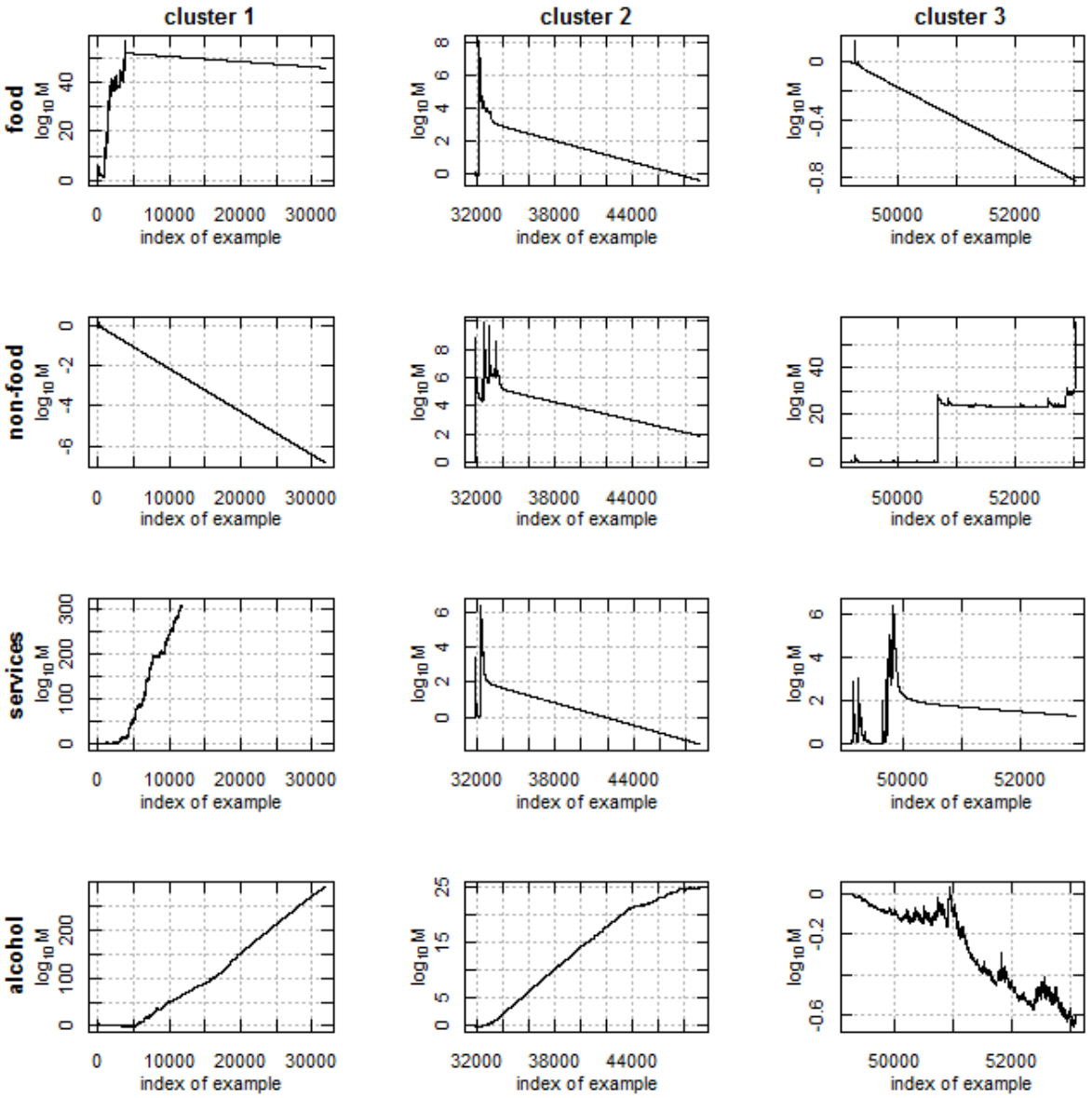

Figure 4: Martingale is plotted for testing structure of expenditure. The columns from left to right are for the first, second and third clusters respectively. The rows from top to bottom are for different categories of spending. If a spending takes fixed share of the total consumption (up to normally distributed noise) martingale should not grow.

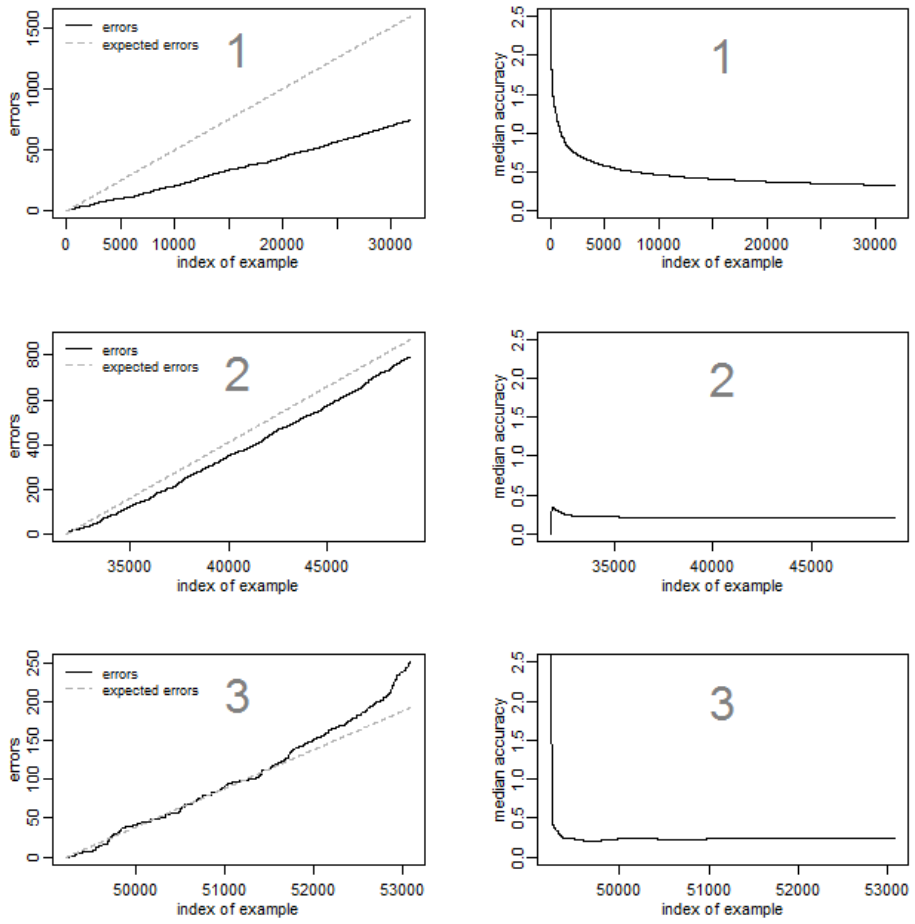

Figure 5: Error rate and median accuracy for on-line conformal prediction are plotted for the significance level of 5% for the first, second and third clusters from top row to bottom. Obtained predictions are practically valid with the only exception of the end of the third cluster.