Akademie věd České republiky
Ústav teorie informace a automatizace

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Petr Nedoma, Miroslav Kárný, Roman Kytka

ANALYSIS OF FINANCIAL MARKET DATA

No.: 2135 December 20, 2005

ÚTIA AVČR, P.O.Box 18, 182 08 Prague,
Czech Republic

Fax: (+420)286890378, http://www.utia.cas.cz,
E-mail: utia@utia.cas.cz

Contents

1 Aims of the study

Generally, properties of the software system Jobmain [1] estimating (finite probabilistic) mixtures (putting together the software basis [2] and reflecting the theory [**?**]) are tested on non-trivial financial data. Specific aims of this study are:

1. inspection of influence of the upper bound on the orders of auto-regressions used in mixture components as well as influence of forgetting factors used in mixture estimation;
2. judgment whether the obtained mixtures combined with a simple one-stage-ahead optimizing marketing could be profitable;
3. evaluation of all experiments and recommendation of the most suitable models and/or marketing strategies.

2 Description of the study

In this study we are trying to find out optimal upper bound on order of auto-regression (ORD) and forgetting factor (FRG) while other parameters determining mixture estimation are unchanged. Upper bound on **order of auto-regression** determines upper bound on memory length used for respective predictions. Taking into account publicly recommended memory about 40, we processed the learning data with $ORD = 10$, 20 and 30. The results make us to stop there.

All processed data are used in estimation of structure and parameters of the mixture. By changing the **forgetting factor**, we can suppress influence of very old data on parameter estimates and thus adapt the mixture to recent "average" behavior of the modelled process. The experiments with $FRG = 0.999999$ (practically no forgetting), 0.999 and 0.99 were run. The lower the factor is, the shorter time period is taken into consideration for structure and parameter estimation. Effective window over data is $1/(1 - FRG) \in \{1e6, 1e3, 1e2\}$

3 Data

We processed the data from four financial markets, dated from year 1985 (or from opening date of the market) until today. This encounters approx-

imately about 5000 records for each market. Four different markets were tried. The respective data files are called exp1a, exp2a, exp3a, exp4a.

Data items (called channels in the used software environment) that are available and used in modelling are given in Table 1

Channel no.	Channel name	Channel description
1	DATE	date of making record
3	OPEN	opening exchange rate for this day
5	LOW	lowest exchange rate for this day
6	CLOSE	closing exchange rate for this day
9	CASH	the price on the spot market
10	INCREMENT	CLOSE(t+1) - CLOSE(t)

Table 1: Channels description

In all examples, we set channels 1, 3, 5, 6, 9 as channels in condition of the resulting predictor and channel 10 as the predicted channel. Channels in condition were selected during previous experiments, according to their explanatory ability with respect to the predicted channel.

4 Processing

First we processed data with changing value of the upper bound on order of auto-regression and no forgetting ($FRG = 0.999999$). The good bound found in this case is conjectured as suitable one for other forgetting factors, too. This organization of experiments allowed us to reduce otherwise exponentially increasing evaluation time. As stated in Section 2, we processed data with values $ORD = 10, 20, 30$.

4.1 Experiments with different orders of auto-regression

4.1.1 ORD=10: Computation time: about 2 hours

The results obtained for all considered markets are summarized in Table 2. For comparison, a mixture with single component was estimated, too.

The complete exploitation of the upper bound on the model order (seen in structure variables "Job.Mix.Facs{:}.str") made us to continue processing with higher ORD value. Moreover, the presented validation test was not

Data	mixll	Rel. SE	Valid. test	mixll	Rel. SE	Valid. test
–	Mixture identification			Single com. identification		
exp1a	1.594e005	0.0136	0	-4.450e004	0.0138	0
exp2a	1.394e005	0.0132	0	NaN	NaN	1
exp3a	-9.082e003	0.0146	0	NaN	NaN	1
exp4a	8.534e004	0.0138	0	-3.173e005	0.0140	0

Table 2: Mixture identification; mixll is the value of the log-likelihood for the final model variant: maximum is search for; Rel. SE is ratio of standard deviation of prediction error to that of data; Valid. test is output of a model-validation test with meaning 1= model is OK, 0= model is BAD; NaN means not-a-number

passed as well as the newer validation test expressed during Jobmain run graphically.

4.1.2 ORD=20: Computation time: about 5 hours.

The results obtained for all considered markets are summarized in Table 3.

Data	mixll	Rel. SE	Valid. test	mixll	Rel. SE	Valid. test
–	Mixture identification			Single com. identification		
exp1a	1.595e005	0.0135	1	NaN	NaN	1
exp2a	5.573e004	0.0135	0	NaN	NaN	1
exp3a	6.247e003	0.0725	0	-5.170e005	0.0146	0
exp4a	7.721e004	0.0145	0	NaN	NaN	1

Table 3: Mixture identification; mixll is the value of the log-likelihood for the final model variant: maximum is search for; Rel. SE is ratio of standard deviation of prediction error to that of data; Valid. test is output of a model-validation test with meaning 1= model is OK, 0= model is BAD; NaN means not-a-number

For exp1a, the result is better and even validation test was passed successfully. Also for exp3a, mixll increased significantly but the validation test indicates still model insufficiency. For exp2a, exp4a, the results are poorer.

The complete exploitation of the upper bound on the model order (seen in structure variables "Job.Mix.Facs{:}.str") made us to continue processing with higher ORD value.

4.1.3 ORD=30: Computation time: about 54 hours.

The results obtained for all considered markets are summarized in Table 4.

Data	mixll	Rel. SE	Valid. test	mixll	Rel. SE	Valid. test
–	Mixture identification			Single com. identification		
exp1a	1.385e005	0.0136	0	-4.450e004	0.0138	0
exp2a	1.422e005	0.0132	0	NaN	NaN	1
exp3a	0	0.0146	0	NaN	NaN	1
exp4a	0	0.0138	0	-3.173e005	0.0140	0

Table 4: Mixture identification; mixll is the value of the log-likelihood for the final model variant: maximum is search for; Rel. SE is ratio of standard deviation of prediction error to that of data; Valid. test is output of a model-validation test with meaning 1= model is OK, 0= model is BAD; NaN means not-a-number.

The validation for files exp3a, exp4a was unfinished as the computation was broken. Even the incomplete results indicate that the prediction quality is not significantly better than that with $ORD = 20$.

4.2 Experiments with different forgetting factor

The results without forgetting indicate that different markets require estimation with different ORD. In this exploratory phase of the research, we took ORD =20 as "reasonable" one and performed experiments with forgetting for a fixed common structure obtained in the case without forgetting, i.e. just estimation and validation were run. (steps = [0 0 0 1 1 0 0 0] in Jobmain).

4.2.1 FRG=0.999

The results obtained for all considered markets are repeatedly in Table 5.

These results are rather poor. Full combined run of structure estimation and forgetting is almost surely necessary.

4.2.2 FRG=0.99

The results obtained for all considered markets are summarized in Table 6.

Data	mixll	Rel. SE	Valid. test	mixll	Rel. SE	Valid. test
–	Mixture identification			Single com. identification		
exp1a	-6.019e005	15386.9	0.5	8.711e004	0.0279	0.5
exp2a	-5.997e005	3253.62	0.5	7.343e003	0.0710	0.5
exp3a	-5.331e005	642.355	0.5	-3.575e004	0.0217	0.5
exp4a	-6.027e005	1182	0.5	7.131e004	0.4566	0.5

Table 5: Mixture identification; mixll is the value of the log-likelihood for the final model variant: maximum is search for; Rel. SE is ratio of standard deviation of prediction error to that of data; Valid. test is output of a model-validation test with meaning 1= model is OK, 0= model is BAD; NaN means not-a-number.

Data	mixll	Rel. SE	Valid. test	mixll	Rel. SE	Valid. test
–	Mixture identification			Single com. identification		
exp1a	9.203e004	0.0247	0.5	9.216e004	0.0247	0.5
exp2a	2.196e004	0.0888	0.5	7.343e003	0.0910	0.5
exp3a	-2.251e004	0.0213	0.5	-3.575e004	0.0217	0.5
exp4a	2.332e004	0.0161	0.5	7.131e003	0.4566	0.5

Table 6: Mixture identification; mixll is the value of the log-likelihood for the final model variant: maximum is search for; Rel. SE is ratio of standard deviation of prediction error to that of data; Valid. test is output of a model-validation test with meaning 1= model is OK, 0= model is BAD; NaN means not-a-number.

These results are more promising. They are relatively close to the case without forgetting and can be improved by the joint run initialization and forgetting. Even then, a bit worse result can be expected in learning phase but there is a chance for an overall gain when using adaptive marketing strategy.

5 Results

In this section we will provide results in a summarized view. Only variables important for our decisions and summarizing will be displayed here. Further technical and more detailed information you can find in chapter 7.

5.1 Numerical evaluation of estimation

The numerical evaluation of estimation and prediction results were presented above. Here, we just stress that mixll is the most important variable. The higher this value is, the better result it means. The values of Rel. SE help us in a finer selection among variants that have similar mixlls. The evaluation of the value Validity test $\in [0, 1]$ $\equiv$[invalid,fully valid] was found unreliable (probably both theoretical and implementation problems) so that we take it as an auxiliary indicator only.

5.2 Numerical evaluation of a preliminary marketing strategy

In order to judge whether the prediction quality could be sufficient for successful marketing, a simple, one-stage-ahead marketing strategy its non-adaptive version was implemented as script invest4.m and its adaptive version in invest5.m. Various runs differing in mixtures used are judged with the help of following indicators.

succa – the number of successful actions. For example, our prediction was "buy" while real data, unknown in the moment of prediction, support the same decision ("buy") as appropriate one.

succapr – relative version of the succa variable, i.e., $\text{succapr} = \frac{\text{succa}}{\text{number of data}} * 100$.

sumcp – the final value of cumulated gains with the tested strategy.

sumcpid – the final value of cumulated gains in the ideal situation when the future change of prices is known.

reldiff – the ratio of the real and ideal gains whose relative difference of ideal-real gains. This number describes the relation between possible gains in ideal and real situations. Its best value is 1. Positive values indicate that the applied strategy brings a profit.

Data	succa	succapr	reldiff	sumcp	sumcpid
Invest4					
exp1a	3760	73.39	0.0289	0.2888	9.9826
exp2a	2245	44.07	0.1036	186.18	1.7970e003
exp3a	2182	48.66	0.0523	571.39	1.0934e004
exp4a	2307	45.10	0.0982	145.38	1.4801e003
Invest5					
exp1a	3735	72.91	-0.0142	-0.1417	9.9826
exp2a	2415	47.41	0.0294	52.820	1.7970e003
exp3a	2173	48.46	0.0441	481.78	1.0934e004
exp4a	2293	44.83	0.0675	99.943	1.4801e003
exp1a_FRG99	3341	65.22	-0.8673	-8.6582	9.9826
exp2a_FRG99	2485	48.59	0.0240	43.050	1.7970e003
exp3a_FRG99	2187	48.77	-0.0058	-63.390	1.0934e004
exp4a_FRG99	2404	46.999	-0.0199	-29.439	1.4801e003
exp1a_FRG999	2974	58.05	-1.5892	-15.8648	9.9826
exp2a_FRG999	2480	48.68	0.0282	50.6200	1.7970e003
exp3a_FRG999	2196	48.97	0.0045	48.3100	1.0934e004
exp4a_FRG999	2414	47.19	-0.0133	-19.6691	1.4801e003

Note that the above results are only qualitative: use of real unit price and real transaction costs may influence it substantially.

5.3 Graphical representation of results

In this section you will find graphical comparison of reached results. It will be depicted in the following types of graphs:

Histogram shows histogram of prediction error. This is output of Jobcontrol.

Segment validation shows probability of the model validity. At present, this is the most reliable test available. This is output of Jobcontrol.

Prediction of increment shows prediction of increment in time. This is output of invest*n*.m, n=4,5.

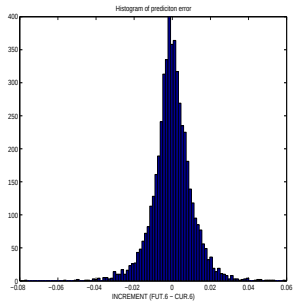
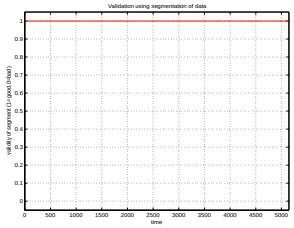
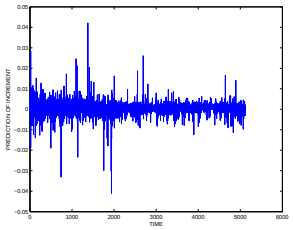
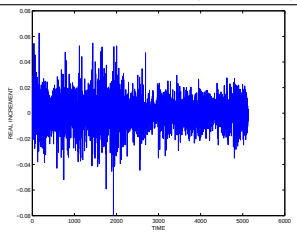
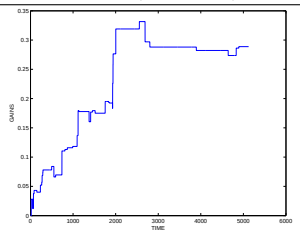
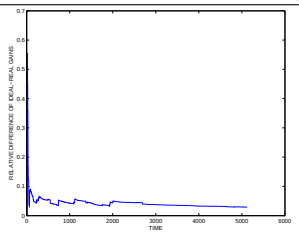
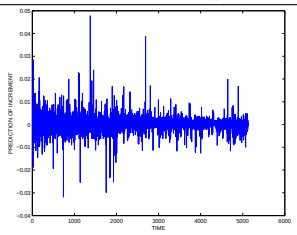
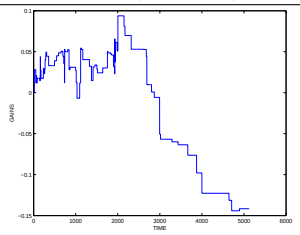
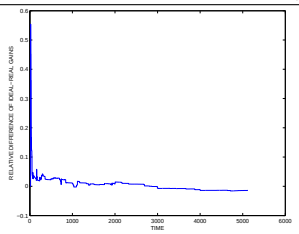
Real increment shows real increment in time. This is output of invest*n*.m.

Gains shows gains reached with marketing strategy. This is output of invest*n*.m, n=4,5.

Ratio of real and ideal gains - shows relative difference between real

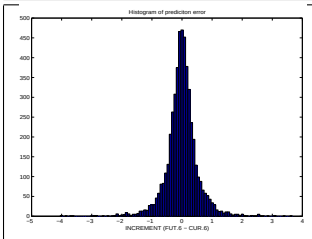
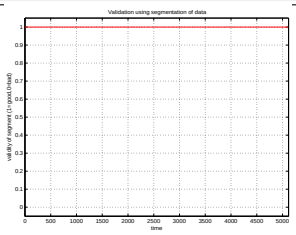
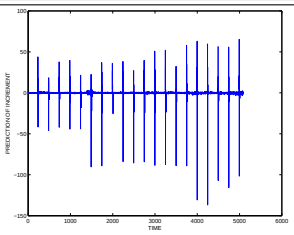
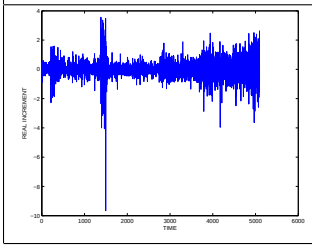
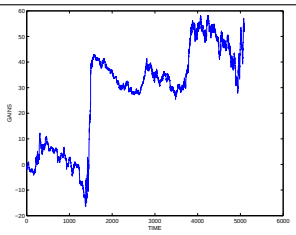
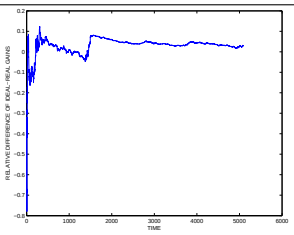
(predicted) and ideal (data-based best possible) gains.

5.3.1 Experiment exp1a, ORD=20, FRG=0.999999

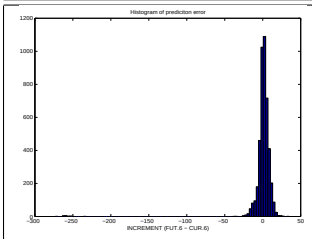
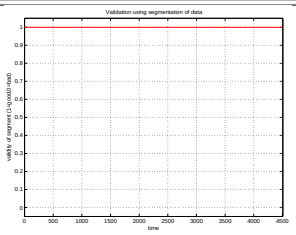
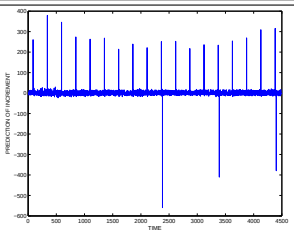
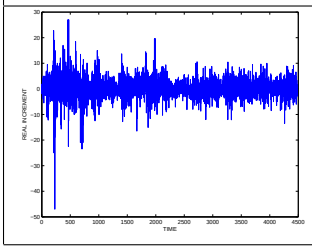
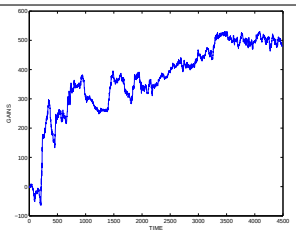
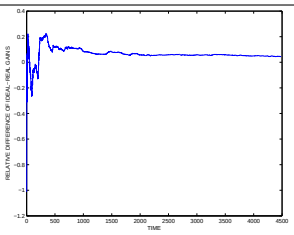
Histogram	Segm.validation	Pred.of increment(invest4)
		
Real increment(invest4)	Gains(invest4)	Relative diff.of ideal-real gains(invest4)
		
Pred.of increment(invest5)	Gains(invest5)	Relative diff.of ideal-real gains(invest5)
		

5.3.2 Experiment exp2a, ORD=20, FRG=0.999999

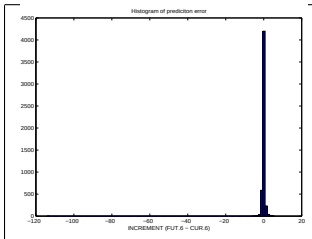
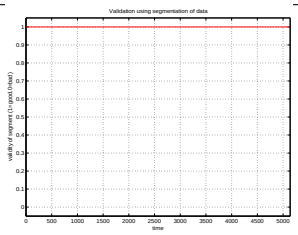
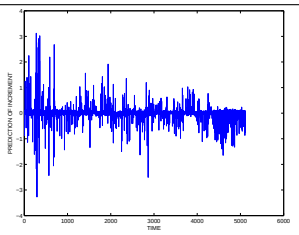
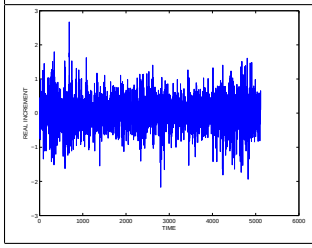
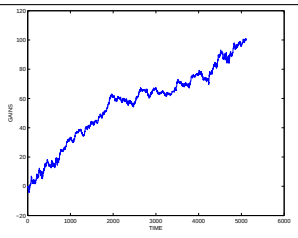
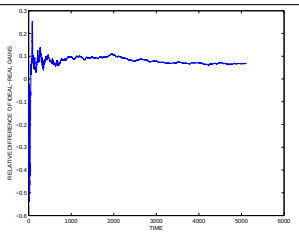
From now on, only invest5.m is used.

Histogram	Segm.validation	Pred.of increment
		
Real increment	**Gains**	**Relative diff.of ideal-real gains**
		

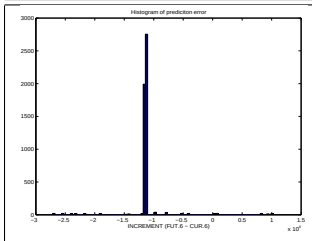
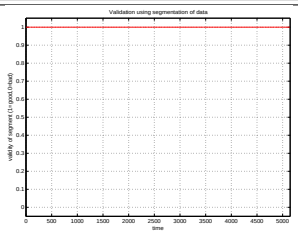
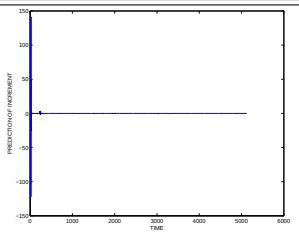
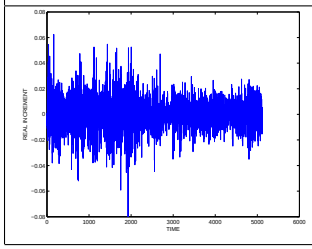
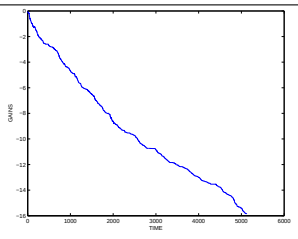
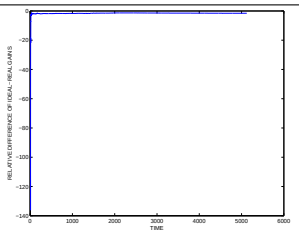
5.3.3 Experiment exp3a, ORD=20, FRG=0.999999

Histogram	Segm.validation	Pred.of increment
		
Real increment	**Gains**	**Relative diff.of ideal-real gains**
		

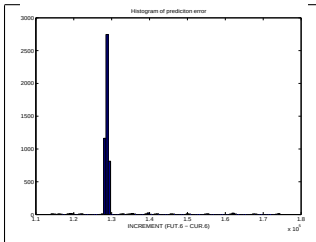
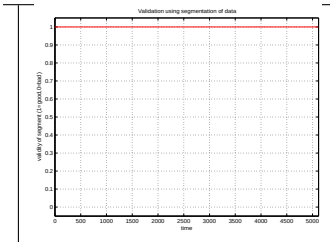
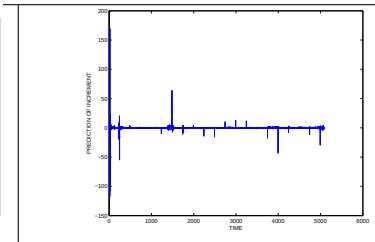
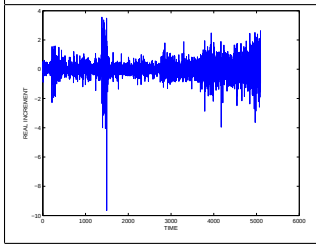
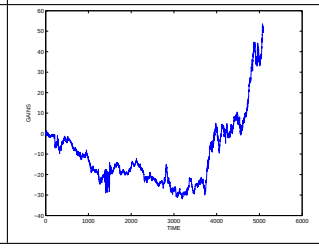
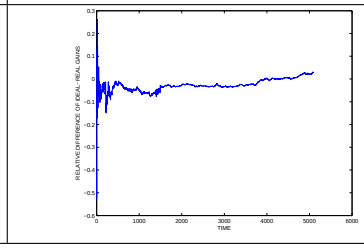
5.3.4 Experiment exp4a, ORD=20, FRG=0.999999

Histogram	Segm.validation	Pred.of increment
		
Real increment	**Gains**	**Relative diff.of ideal-real gains**
		

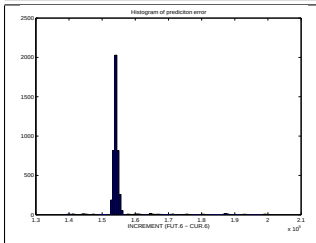
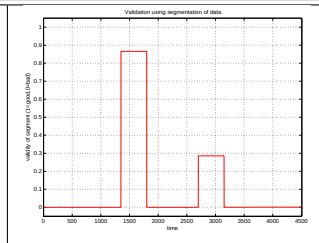
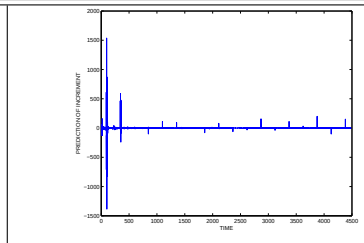
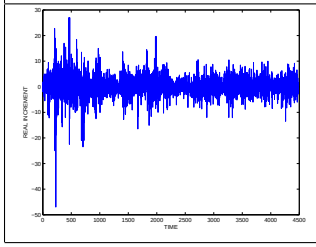
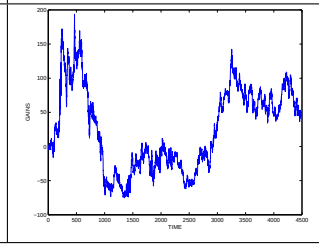
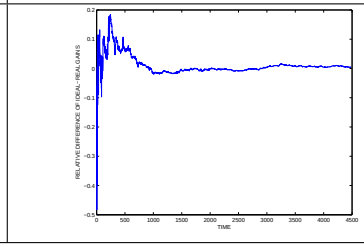
5.3.5 Experiment exp1a, ORD=20, FRG=0.999

Histogram	Segm.validation	Pred.of increment
		
Real increment	**Gains**	**Relative diff.of ideal-real gains**
		

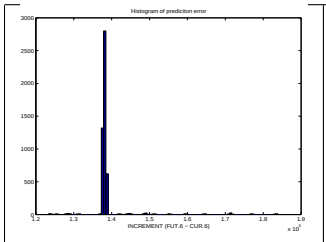
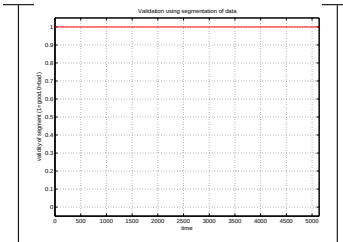
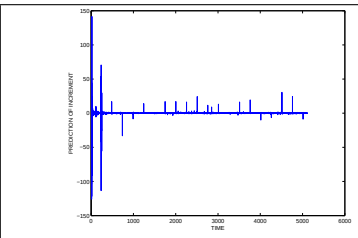
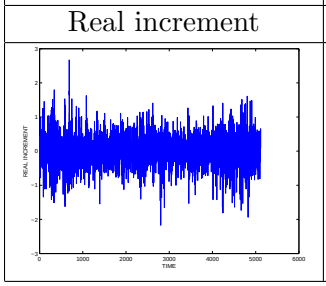
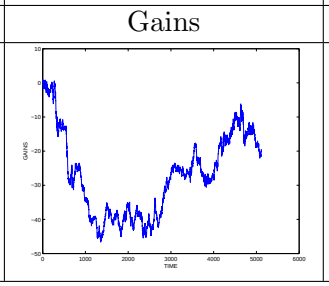
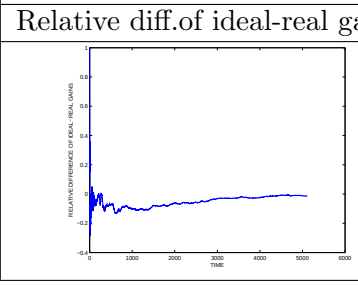
5.3.6 Experiment exp2a, ORD=20, FRG=0.999

Histogram	Segm.validation	Pred.of increment
		
Real increment	**Gains**	**Relative diff.of ideal-real gains**
		

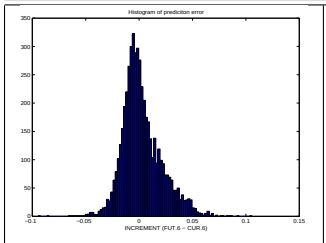
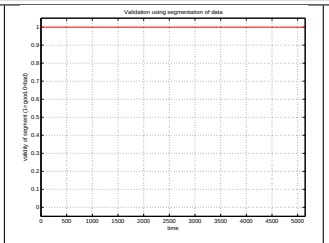
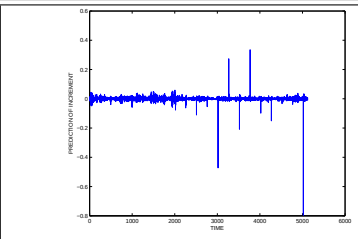
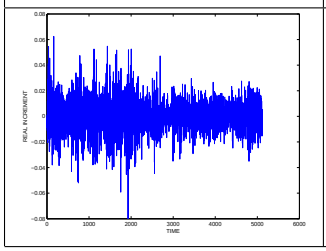
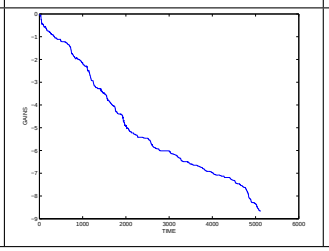
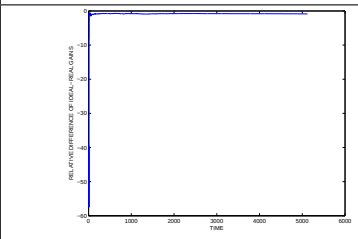
5.3.7 Experiment exp3a, ORD=20, FRG=0.999

Histogram	Segm.validation	Pred.of increment
		
Real increment	**Gains**	**Relative diff.of ideal-real gains**
		

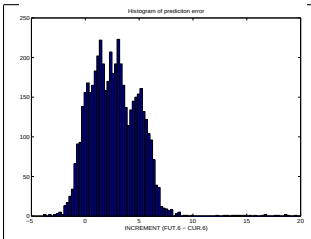
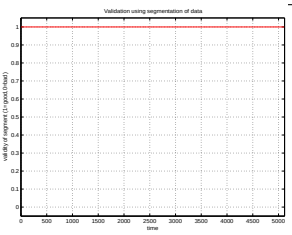
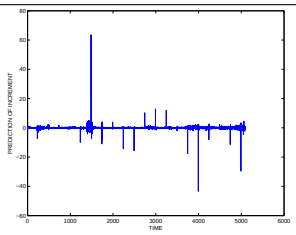
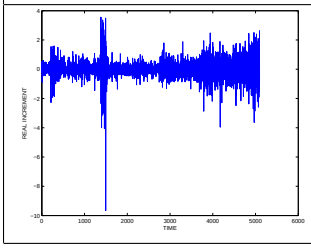
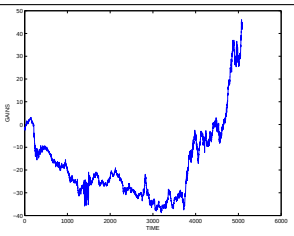
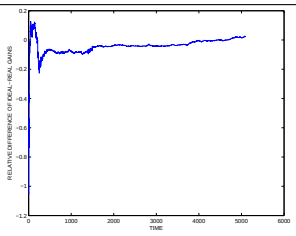
5.3.8 Experiment exp4a, ORD=20, FRG=0.999

Histogram	Segm.validation	Pred.of increment
		
Real increment	**Gains**	**Relative diff.of ideal-real gains**
		

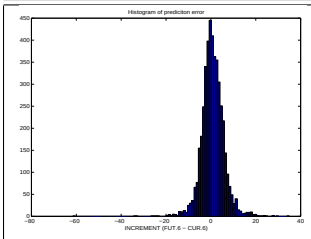
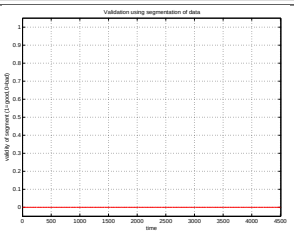
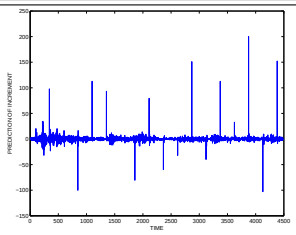
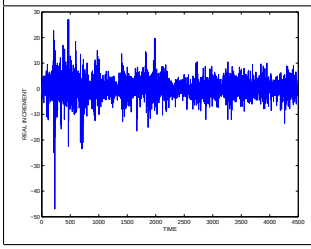
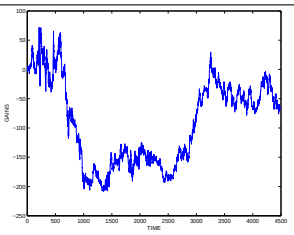
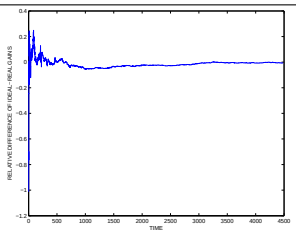
5.3.9 Experiment exp1a, ORD=20, FRG=0.99

Histogram	Segm.validation	Pred.of increment
		
Real increment	**Gains**	**Relative diff.of ideal-real gains**
		

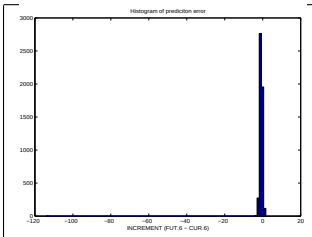
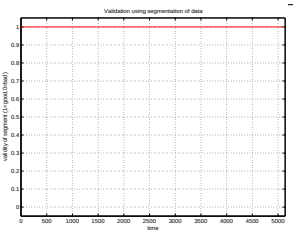
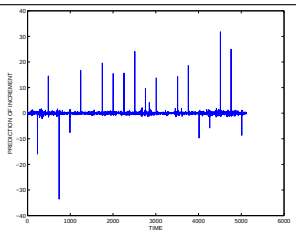
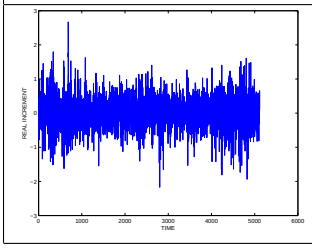
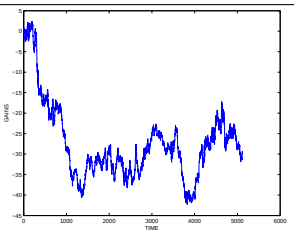
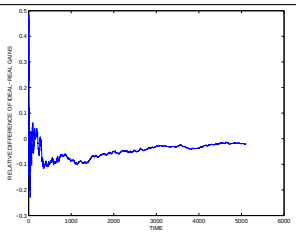
5.3.10 Experiment exp2a, ORD=20, FRG=0.99

Histogram	Segm.validation	Pred.of increment
		
Real increment	**Gains**	**Relative diff.of ideal-real gains**
		

5.3.11 Experiment exp3a, ORD=20, FRG=0.99

Histogram	Segm.validation	Pred.of increment
		
Real increment	**Gains**	**Relative diff.of ideal-real gains**
		

5.3.12 Experiment exp4a, ORD=20, FRG=0.99

Histogram	Segm.validation	Pred.of increment
		
Real increment	**Gains**	**Relative diff.of ideal-real gains**
		

6 Conclusions

The obtained preliminary results lead to the following temporary conclusions:

- The results are promising to such an extent that a further development makes sense.
- Mixture models of the order several tens seems to be suitable for sufficiently precise short term prediction.
- Estimation results are not stable enough and the model structure has to be selected specifically for each market.
- Use of forgetting and consequently the adaptive version do not fulfil general promises, i.e., it call for focus on the fixed version in near future.
- One-stage-ahead strategy has to be generalized to multi-step-ahead one, which will allow to "keep" positions open for several days and thus to decrease transaction costs.

Concerning to Jobmain as universal tool for solving similar tasks, the overall impression is positive. Some small bugs have been already removed and other ones at least indicated (like validation test or form of the automatically generated output).

7 Additional technical information

More detailed technical information are available at our SVN server, in folder: business/archiv/reporty/ver1-3/detail.

Acknowledgement

This research was partially supported by AV ČR S1075351 and by the research center DAR, grant MŠMT ČR 1M6798555601.

References

[1] Ludvík Tesař, Petr Nedoma, and Miroslav Novák, *Mixture learning script Jobcontrol (Program)*, ÚTIA AV ČR, Praha, 2004.

[2] P. Nedoma, J. Böhm, T.V. Guy, L. Jirsa, M. Kárný, I. Nagy, L. Tesař, and J. Andrýsek, "Mixtools: User's Guide", Tech. Rep. 2060, ÚTIA AV ČR, Praha, 2002.