Portfolio of Automated Trading Systems: Complexity and Learning Set Size Issues

Sarunas Raudys

***Abstract*— In this paper, we consider using profit/loss histories of multiple automated trading systems (ATSs) as N input variables in portfolio management. By means of multivariate statistical analysis and simulation studies, we analyze the influences of sample size (L) and input dimensionality on the accuracy of determining the portfolio weights. We find that degradation in portfolio performance due to inexact estimation of N means and $N(N-1)/2$ correlations is proportional to N/L; however, estimation of N variances does not worsen the result. To reduce unhelpful sample size/dimensionality effects, we perform a clustering of N time series and split them into a small number of blocks. Each block is composed of mutually correlated ATSs. It generates an expert trading agent based on a nontrainable $1/N$ portfolio rule. To increase the diversity of the expert agents, we use training sets of different lengths for clustering. In the output of the portfolio management system, the regularized mean-variance framework-based fusion agent is developed in each walk-forward step of an out-of-sample portfolio validation experiment. Experiments with the real financial data (2003–2012) confirm the effectiveness of the suggested approach.**

***Index Terms*— Complexity, efficient-market hypothesis, investments, Markowitz, multiagent systems, optimization, portfolios, regularization, sample size.**

I. Introduction

THE TASK of portfolio management (PM) is one of the most important research topics in financial engineering (seminal works of Markowitz [1], [2], and reviews of thousands of research papers aimed at improving Markowitz's solution [3]–[6]). PM methods are significant because, besides finance, they can also be applied in economics, industry sector allocation, computer hardware and software technologies, web design, teaching, planning health promotion programs, and other disciplines (see [7], [8]).

In portfolio design, we search for a multidimensional weight vector $\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_N)$ that determines "optimal" investment proportions for N assets [in this paper, the automated trading systems (ATSs) or trading robots]. In this paper's one day sell and buy analysis, we assume that the coefficients $\boldsymbol{w}_j$ fulfill the following requirements:

$$\boldsymbol{w}_j \geq 0, \qquad \sum_{j=1}^{N} \boldsymbol{w}_j = 1. \tag{1}$$

Manuscript received May 25, 2012; revised September 20, 2012; accepted November 23, 2012. Date of publication January 9, 2013; date of current version January 30, 2013. This work was supported by the Research Council of Lithuania under Grant MIP-043/2011 and Grant MIP-018/2012.

The author is with the Department of Informatics, Vilnius University, Vilnius LT-03225, Lithuania (e-mail: sarunas.raudys@mif.vu.lt).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2012.2230405

Usually, to find the weights one uses L days (time moments in a more general problem setting) history of investment profit/loss (the learning set), x_{ji} ($i = 1, 2, \ldots, L$; $j = 1, 2, \ldots, N$). Then the ith day portfolio profit/loss is defined as

$$x_{Pi} = \sum_{j=1}^{N} \boldsymbol{w}_j x_{ji} = \boldsymbol{X}_i \boldsymbol{w}^T \tag{2}$$

where $\boldsymbol{X}_i = (x_{1i}, x_{2i}, \ldots, x_{Ni})$, and "$T$" denotes a transpose operation.

A standard approach to solve the PM task is the mean-variance framework [1], [2]. Here, one maximizes the sample mean (mean) and standard deviation (std) ratio (a modified Sharpe ratio with no risk-free rate)

$$Sh_{\text{in}} = \frac{\text{mean}(x_{\text{Pi}})}{\text{std}(x_{\text{Pi}})} \tag{3}$$

where mean $(x_{\text{Pi}}) = \overline{\boldsymbol{X}}\boldsymbol{w}^T$, std$(x_{\text{Pi}}) = \sqrt{\boldsymbol{w}\boldsymbol{S}\boldsymbol{w}^T}$, $\overline{\boldsymbol{X}}$ is an N-dimensional (N-D) vector row, a sample estimate of mean of returns, $\boldsymbol{\mu}$ and

$$\boldsymbol{S} = \frac{1}{L-1} \sum\nolimits_{i=1}^{L} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T (\boldsymbol{X}_i - \overline{\boldsymbol{X}}) \tag{4}$$

is the sample estimate of the $N \times N$ dimensional covariance matrix (CM) $\boldsymbol{\Sigma}$. In practice, instead of (3) an annualized number, Sh, is used

$$Sh = \left[\frac{\text{mean}(x_{\text{Pi}})}{\text{std}(x_{\text{Pi}})}\right] \times \sqrt{252}. \tag{5}$$

Such a solution is optimal if x_{Pi} follows the Gaussian distribution, and the mean and CM are known exactly. In real situations, we deal with non-Gaussian variables, training vectors are correlated, we have finite learning sets, and the data characteristics are changing all the time. Thus, such a solution is approximate. Employment of the mean-variance framework has two important advantageous features.

1) Provided the number of inputs (assets or automated trading systems) is sufficiently large, distribution of the sum (2) can move toward Gaussian (for strict definitions, see [4]).
2) In case of Gaussian returns and *a priori* fixed probability Prob ($x_{\text{Pi}} < P_{\max}$), maximization of (3) means that the mean value of returns is maximized, where $P_{\max}$ is a chosen risk level. Optimality of the mean-variance framework is that the maximization result does not depend on the choice of $P_{\max}$.

In situations where the number of assets is small, however, we come across asymmetry and excess of portfolio profit

distribution. In such circumstances, the variance is not an appropriate risk measure; it counts positive fluctuations above the expected returns (also called upside volatility) as a part of the risk. For this reason, the downside risk (the volatility below the expected returns only) has been introduced. New Sharpe-ratio-related portfolio optimization schemes based on downside risk have been developed [2]–[4], [6], [9], [10]. Ignoring observations with positive fluctuations can become useful in highly asymmetric cases. However, one needs to bear in mind that the quantity of data used to evaluate the risk is reduced by half. For that reason, paying no attention to positive fluctuations can become harmful for a Gaussian portfolio.

In contrast to numerous studies with rigorous asymmetry and excess, in this paper we focus on tactic lines of attack to approach normality, i.e., an increase in the number of inputs, N. An encouraging feature is also found in the fact that, in many practical tasks, a good Gaussian fit is obtained when the number of inputs N exceeds 100 even in the case of correlated inputs and non-Gaussian univariate distributions of x_{ji}. Along these lines, the mean-variance framework can be useful since it does not require normality of distribution for each single output. One needs the distribution of the weighted output $x_{Pi} = \mathbf{X}_i \mathbf{w}^T$, to be close to a Gaussian.

Positive effects of approaching normality are lessened by sample size/dimensionality problems. Accuracy of determination of the portfolio vector depends on the accuracy of the estimation of the mean returns $\overline{\boldsymbol{X}}$ and the covariance matrix $\boldsymbol{S}$. If the sample size L is small, and the number of inputs N is large, instabilities and a reduction in portfolio performance arise. Inexact estimation of the $N \times (N+1)/2$ elements of the covariance matrix can be reduced by regularizing constraints, shrinkage estimators, or penalty terms that are incorporated in the optimization procedure [11]–[13]. In the shrinkage approach, the off-diagonal elements of the matrix are moderated (or shrunk) compared to the typically large off-diagonal elements of the sample matrix. The variance elements in the diagonal are left untouched. In hard regularization, we can reduce the number of parameters estimated from the learning set up to N variances.

Use of assumptions about a block-diagonal structure of the covariance matrix can dramatically reduce the number of parameters to be estimated. In pattern classification, this model has already been used for four decades [14], [15]. It has now been successfully applied to PM [16]–[19]. A remarkable reduction in covariance matrix complexity can also be achieved while using a tree-type dependence between the variables $x_{a1}, x_{a2}, \ldots, x_{aN}$ [20]–[22].

There is some interest in finding an estimator of the CM that performs substantially better than the standard learning-set-based matrix (4). The authors of paper [17] concluded that, from the experiments, it is actually impossible to claim that one of the estimators is better than the others.

In an estimation of the N-D mean vector of returns $\overline{\boldsymbol{X}}$, we also face a small learning set problem. The only known way to simplify the solution is by the use of the nontrainable $1/N$ portfolio trading rule (equally weighted portfolio) with ($\boldsymbol{w}_1 = \boldsymbol{w}_2 = \cdots = \boldsymbol{w}_N = 1/N$). DeMiguel *et al.* [11] also compared various advanced methods consisting of Bayesian estimation, shrinkage, robust allocation, etc., and found that none of the 14 models they implemented could consistently outperform the $1/N$ portfolio. Therefore, the $1/N$ portfolio was suggested as a benchmark method.

For an efficient search of optimal portfolio calculation methods in out-of-sample (i.e., cross-validation, where the validation set follows the training set in time) regime, one needs a deep theoretical analysis [23], [24]. Examination of small learning sample questions is also very important in situations in which environments are changing very rapidly. In the modern world affected by fast technological, political, and economic changes, financial situations vary very often and unpredictably. Therefore, lengthy time series are often unsuitable for precise portfolio calculation. One needs to employ shorter training histories [6], [25], [26]. In certain papers [6], [27], a supervised learning decision system that adapts portfolio weights to changed environments has been suggested.

The above literature review shows that knowledge about performance, sample size, and input dimensionality relationships is vital to the portfolio design. Unlike statistical pattern recognition, these questions have not been investigated sufficiently in this research field. In this paper, we use a double asymptotic analysis of learning set size and input dimensionality effects on the decision-making accuracy. In this examination, both the sample size and the dimensionality are increasing [15], [21], [28], [29]. Earlier, this methodology was successfully applied in pattern classification.

In Section II, we suggest a direct formula for portfolio weights calculation. We use it to derive asymptotic relationships between L, N, and the success in portfolio optimization performance. In Section III, we consider the sample size and dimensionality issues in high-frequency trading, where, instead of assets, a large number of ATSs (robots) are used in a two-stage multiagent PM system. Section IV describes the results of the experimental performance analysis. Section V concludes this paper.

II. Learning Set Size and Complexity Issues

A. Finding the Portfolio Weights

In standard mean-variance framework, finding vector $\boldsymbol{w}$ is performed as a calculation of an efficient frontier for F return values, $\boldsymbol{w}_1\overline{\boldsymbol{X}}^T = q_1, \boldsymbol{w}_2\overline{\boldsymbol{X}}^T = q_2, \ldots, \boldsymbol{w}_F\overline{\boldsymbol{X}}^T = q_F$ [1]. For each *a priori* fixed value q_i, a minimum of $\mathrm{std}(x_{\mathrm{Pi}}) = \sqrt{\boldsymbol{w}_t \boldsymbol{S} \boldsymbol{w}_t^T}$ is found. Afterward, F values of the Sharpe ratio are calculated, and the value with the largest ratio is selected. In standard procedure, optimization is based on Lagrange multipliers that allows fulfilling the constraints (1) and $\boldsymbol{w}_t\overline{\boldsymbol{X}}^T = q_t$, $t = 1, 2, \ldots, F$. To find the portfolio weight vector $\boldsymbol{w}_t$, we will use L days N-dimensional time series, i.e., the $N \times L$ data matrix. To speed up calculations, instead of exact solution, i.e., minimization of standard deviations by using the Lagrange multiplier, we will use approximate analytic solution. Like in the perceptron training, we minimize the cost function, where

constraints (1) are incorporated into the cost function

$$\text{Cost}(q_t) = \boldsymbol{w}_t \boldsymbol{S}\ \boldsymbol{w}_t^T + \Lambda_q \left(\boldsymbol{w}_t \overline{\boldsymbol{X}}^T - q_t\right)^2 + \Lambda_1 \left(1\boldsymbol{w}_t^T - \mathbf{1}\right)^2. \tag{6}$$

In (6) $\mathbf{1} = [1\ \ 1, \ldots, 1]$ stands for a row vector composed of N "1s." Scalars Λ_q and Λ_1 control the constraint violations. Values of Λ_q, Λ_1 should be sufficiently large to ensure that the terms $(\boldsymbol{w}_t\overline{\boldsymbol{X}}^T - q_t)^2$ and $(\mathbf{1}\boldsymbol{w}_t^T - 1)^2$ onverge to 0. If they approach infinity, both methods result in identical solutions. Parameters Λ_q, Λ_x control the minimization accuracy. In finite-learning-set situations, lower accuracy can play a positive role of regularization and improve accuracy of solution [22]. The origin of this phenomenon is similar to overtraining in neural networks learning, or finding an optimal value of regularization parameter (see [11]–[13], [15], [21], [22], also Section II-D). Our experiments with 169 time series of financial data ($N = 169$) have shown that, for this data, values $\Lambda_q = \Lambda_1 = \Lambda$ between 10^8 and 10^6 is a good choice [22]. Too large Λ values, however, result in a slightly worse result. Given q_t and Λ, the minimum of (6) is found by equating the derivative to zero and solving the equation obtained

$$\frac{\partial \text{Cost}(q_t)}{\partial \boldsymbol{w}_t} = 2\boldsymbol{w}_t\boldsymbol{S} + 2\Lambda\left(\boldsymbol{w}_t\overline{\boldsymbol{X}}^T - q_t\right)\overline{\boldsymbol{X}} + 2\Lambda\ \left(\mathbf{1}\boldsymbol{w}_t^T - 1\right)\mathbf{1} = 0.$$

Then

$$\boldsymbol{w}_t = (q_t\overline{\boldsymbol{X}} + \mathbf{1})(\boldsymbol{S}/\Lambda\ \overline{\boldsymbol{X}}^T\overline{\boldsymbol{X}} + \mathbf{1}^T\mathbf{1})^{-1}. \tag{7}$$

To satisfy constraint (1), we brought negative weights to naught if $\boldsymbol{w}_{tj} \leq 0$. Afterward, we normalized $\boldsymbol{w}_t$ to meet constraint (1). In experimental comparison of diverse portfolio design methods, we used calculations based on (7) and the standard optimization procedure, the "frontcon" function realized in MATLAB financial toolbox.

Lagrange multipliers provide a strategy for finding the minima of a function subject to equality constraints. They require calculating two or more gradient vectors. In our approach, we use single modified gradient vector and take into account that the sample size is finite. For that reason, we reduce accuracy of minimization and gain both in accuracy and speed. Simulations show that, after proper regularization, (7) most often outperforms the MATLAB answer. Moreover, (7) was roughly 30 times faster. It is very important in complex multiagent system (MAS) design: simple analytical expression and high calculation speed are very important when we have to generate a large number of agents differing in subsets of trading robots and training parameters values. Furthermore, the simplicity of (7) allows rough analytical examination of the sample size–dimensionality issues.

B. In-Sample and Out-of-Sample Standard Portfolio

To theoretically examine the influence of estimation accuracy of the N-D mean vector $\overline{\boldsymbol{X}}$ and the $N \times N$ dimensional covariance matrix $\boldsymbol{S}$, we will use representation (7). While examining the ATSs, data, we dealt with the return histories where components of mean vector $\overline{\boldsymbol{X}}$ exceeded 1 notably. Standard deviation in diagonal elements of matrix $\boldsymbol{S}$ markedly exceeded the components of vector $\overline{\boldsymbol{X}}$ as well. We performed numerical examination of magnitudes of all three components in (7). For a rough examination of factors that affect portfolio accuracy, we ignored the two smallest terms in (7). So, the weights are expressed as

$$\boldsymbol{w}\ = \overline{\boldsymbol{X}}\boldsymbol{S}^{-1}. \tag{8}$$

We employed this rough approximation for sample size/complexity analysis in this paper. Use of representation (8) in (2) shows that the in-sample (training set-based) mean value of weighted sum (2) can be expressed as

$$\boldsymbol{w}\overline{\boldsymbol{X}}^T = \overline{\boldsymbol{X}}\boldsymbol{S}^{-1}\overline{\boldsymbol{X}}^T. \tag{9}$$

The variance of in-sample returns

$$\boldsymbol{w}\boldsymbol{S}\boldsymbol{w}^T = \overline{\boldsymbol{X}}\boldsymbol{S}^{-1}\boldsymbol{S}^{-1}\overline{\boldsymbol{X}}^T = \overline{\boldsymbol{X}}\mathbf{S}^{-1}\overline{\boldsymbol{X}}^T. \tag{10}$$

Therefore, in-sample Sharpe ratio (3) of the N-dimensional portfolio roughly can be expressed as

$$S\hat{h}_{in} \approx \sqrt{\overline{\boldsymbol{X}}\boldsymbol{S}^{-1}\overline{\boldsymbol{X}}^T}. \tag{11}$$

Equation (11) depends on the random vector $\overline{\boldsymbol{X}}$ and matrix $\boldsymbol{S}$. Thus, it is a random variable, too. We will use standard transformations traditionally used in multivariate statistical analysis [30]–[32] $\boldsymbol{Y} = \boldsymbol{X}\mathbf{F}$ (where the $N \times N$ matrix $\mathbf{F}$ is such that $\mathbf{F}^{-1}\Sigma\mathbf{F} = \mathbf{I}_N$ ($\mathbf{I}_N$ stands for the $N \times N$ identity matrix), $\mathbf{F} = \mathbf{G}\mathbf{d}^{-1}\mathbf{T}$, $\mathbf{d}^2$ stands for diagonal matrix of eigenvalues of matrix $\boldsymbol{\Sigma}$, $\mathbf{G}$ is orthogonal matrix of eigenvectors of $\boldsymbol{\Sigma}$ and $\mathbf{T}$ is an orthogonal matrix such that $\boldsymbol{\mu}\mathbf{G}\mathbf{d}^{-1}\mathbf{T} = \boldsymbol{\Delta} = (\delta, 0, 0, \ldots, 0)$. It can be shown that representation

$$S\hat{h}_{in}^2 = \overline{\boldsymbol{X}}\boldsymbol{S}^{-1}\overline{\boldsymbol{X}}^T = \overline{\boldsymbol{X}}\mathbf{F}\mathbf{F}^{-1}\boldsymbol{S}^{-1}\mathbf{F}\mathbf{F}^{-1}\overline{\boldsymbol{X}}^T = \overline{\boldsymbol{X}}\mathbf{F}(\boldsymbol{S}_{\boldsymbol{Y}})^{-1}\mathbf{F}^{-1}\overline{\boldsymbol{X}}^T$$

can be converted into

$$(\boldsymbol{\Delta} + \overline{\boldsymbol{\Theta}})S^{-1}(\boldsymbol{\Delta} + \overline{\boldsymbol{\Theta}})^T = (\boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}} + \overline{\boldsymbol{\Theta}\boldsymbol{\Theta}}^{\mathrm{T}})s^{11} \tag{12}$$

where δ stands for asymptotic value of ratio (3) $S\hat{h}_{in}$, when sample size $L \to \infty$, components of the N-D vector $\overline{\boldsymbol{\Theta}} = \overline{\boldsymbol{X}} - \boldsymbol{\mu})\mathbf{F}\ = (\boldsymbol{\theta}_1, \ldots, \theta_N)$ are Gaussian N(0, $1/L$) random variables, covariance matrix $\boldsymbol{S}_Y = \mathbf{F}^{-1}\boldsymbol{S}\boldsymbol{F}$ is distributed as the Wishart matrix with matrix parameter $\mathbf{I}_N$ and L degrees of freedom, and s^{11} is upper left element of matrix $(\boldsymbol{S}_Y)^{-1}$: $\mathrm{s}^{11} \sim 1/\chi^2_{L-N^2}$.

Consequently, $S\hat{h}_{in}^2$ distributed as a scaled ratio of the noncentral and central chi-square random variables (a scaled noncentral F random variable)

$$S\hat{h}_{in}^2 \sim \chi^2_{L\delta^2,N} / \chi^2_{L-N}. \tag{13}$$

A double asymptotic approach when both the sample size and dimensionality are increasing [15], [21], [28], [29] advocates that asymptotic variances of both the numerator and denominator in (13) approach zero. Therefore, the mean values can be approximated by their asymptotic limit values

$$\chi^2_{\delta^2,N}\ \ \to (\delta^2 + N),\ \text{and}\ \chi^2_{L-N} \to (L - N).$$

Consequently, expectation of the in-sample estimate

$$S\hat{h}_{in} \to \sqrt{(\delta^2 + N/L)/(1 + N/(L-N))} = \delta \times \sqrt{(1 + N/(L \times \delta^2))(1 + N/(L-N))}. \quad (14)$$

Use of (8) and the standard orthogonal transformations shows that distribution of the out-of-sample (cross-validation) Sharpe ratio can be represented as

$$S\hat{h}^{=}_{\text{out}} \frac{\boldsymbol{\mu}\boldsymbol{w}^T}{\sqrt{\boldsymbol{w}\mathbf{I}^{-1}\boldsymbol{w}^T}} \approx \frac{\boldsymbol{\Delta}\mathbf{S}^{-1}(\boldsymbol{\Delta} + \overline{\boldsymbol{\Theta}})^T}{\sqrt{(\boldsymbol{\Delta} + \overline{\boldsymbol{\Theta}})\mathbf{S}^{-1}\mathbf{I}^{-1}\mathbf{S}^{-1}(\boldsymbol{\Delta} + \overline{\boldsymbol{\Theta}})^T}} = \frac{\delta^2 \boldsymbol{s}^{11}}{\sqrt{\boldsymbol{\Delta}(\boldsymbol{s}^{1*}(\boldsymbol{s}^{1*})^T\boldsymbol{\Delta}^T + \sum_{i,j=1}^{N} \boldsymbol{\Theta}_i\boldsymbol{\Theta}_j \sum_{\alpha=1}^{N} \boldsymbol{s}^{i\alpha}\boldsymbol{s}^{j\alpha T}}}$$

where $\mathbf{s}^{1*}$ is the first row of the inverse matrix $\boldsymbol{S}^{-1} = ((s^{ij}))$.

We recall that $\boldsymbol{\Delta} = (\delta, 0, 0, \ldots, 0)$. Thus, $\boldsymbol{\Delta}\boldsymbol{s}^{1*}(\boldsymbol{s}^{1*})^T\boldsymbol{\Delta}^{\mathrm{T}} = (\delta \times \boldsymbol{s}^{11})^2$. In analysis we need expectations $\mathrm{E}(s^{ii}) = L/(L–N)$, $E(\boldsymbol{s}^{ii})^2 = L^2/(L-N)^2$, $E(s^{ij}) = 0$, $E(s^{ij})^2 = L^2/(L-N)^3$ if $j \neq i$ [31], [32].

After a little simple but tedious algebra, we get that mean value of out-of-sample Sharpe ratio

$$S\hat{h}_{\text{out}} \to \delta/\sqrt{(1 + N/(L \times \delta^2))(1 + N/(L-N))}. \quad (15)$$

The term $T_X = 1 + N/(L \times \delta^2)$ in (14) and (15) is responsible for inaccuracies that arise while estimating the mean of portfolio returns. The term $T_S = 1 + N/(L-N)$ is responsible for inaccuracies that arise while estimating the covariance matrix.

Both terms increase the in-sample profit, but reduce the out-of-sample one. In portfolio optimization tasks, the value of δ [the limiting value of the ratio mean (x_{Pi})/std(x_{Pi})] usually is less than 1. Hence, in relatively large learning set size situations, the term T_X can become larger than the term T_S. It means that, in such situations, the profit decreases mainly due to inexact estimation of the means.

In most situations, especially in rapidly changing environments, the sample size is small. In ATSs [33], [34], the number of activities (autonomous trading robots, agents) N often is close or even exceeds learning set size L. Equations (14) and (15) show that effect of learning set size can be very influential if the dimensionality N is close to L. It means that, in very small learning set situations, proper estimation of the covariance matrix becomes a critical task. The situation can become even dangerous: a small increase in dimensionality can ruin the portfolio completely.

Equations (14) and (15) are approximate. To check the diminution in accuracy caused by ignoring two terms in (7), we performed simulation experiments with 169-D Gaussian data. Characteristics of the data (mean vector $\boldsymbol{\mu}$, and CM $\boldsymbol{\Sigma}$) were calculated from two-thousand 169-D vectors of financial data used in one of ATS experiments considered in Sections III and IV. Graphs 1 and 5 in Fig. 1(a), (b) were computed theoretically for $\delta = \boldsymbol{\mu}\boldsymbol{w}^T/\sqrt{\boldsymbol{w}\boldsymbol{\Sigma}^{-1}\boldsymbol{w}^T} = 0.717$, where $\boldsymbol{w}$ was calculated according (7) for a known $\boldsymbol{\mu}$ value $\boldsymbol{\Sigma}$.

Straight lines 3 in Fig. 1(a), (b) characterize asymptotic gain, $\delta \times 15.8745 = 11.38$ (annualized ratio). In simulation

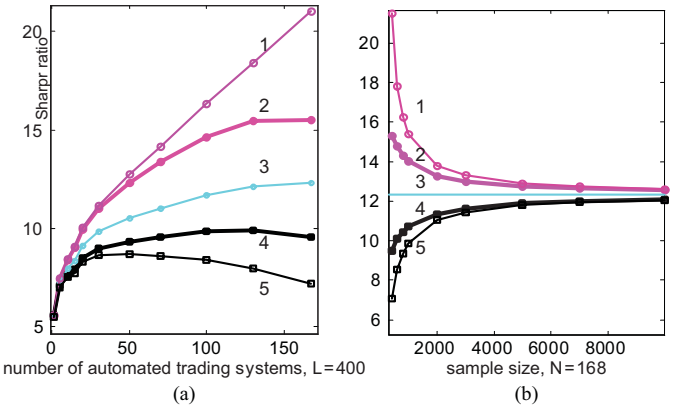

Fig. 1. Sharpe ratio as a function of the dimensionality (a) learning set size. (b) 1: "in-sample" analytical formula, 2: in-sample experiment with Gaussian data, 3: asymptotic analytically calculated value, 4: out-of-sample experiment, 5: "out-of-sample" analytical formula; $N = 168$.

with Gaussian data, we performed 100 independent experiments with learning set sizes $L = 600$, 800, 1000, 2000, 5000, and 10 000. For each randomly generated set of L vectors, we found a vector of means CM and calculated the portfolio vector. Curves 2 and 4 show the means of 100 "in-sample" and "out-of-sample" values, $S\hat{h}^{=}_{\text{in}} \overline{\boldsymbol{X}}\boldsymbol{w}^T/\sqrt{\boldsymbol{w}S^{-1}\boldsymbol{w}^T}$ and $S\hat{h}_{\text{out}} = \boldsymbol{\mu}\boldsymbol{w}^T/\sqrt{\boldsymbol{w}\boldsymbol{\Sigma}^{-1}\boldsymbol{w}^T}$.

The right part of Fig. 1(b) approves the symmetry of the in-sample and out-of-sample performance of empirical and theoretical curves, 2 versus 4 and 1 versus 5, with respect to the asymptotic straight line 3. The symmetry between the in-sample and out-of-sample means of the performance curves was already taken into account earlier in pattern recognition [15], [21].

Comparison of experimental curves (2 and 4) with the theoretically calculated ones (1 and 5) points out the lower accuracy of theoretical approximation in situations where the sample size is very small. In large-sample situations, the accuracy is higher. Nevertheless, theoretical and empirical curves demonstrate that, in the case of high dimensions ($N \approx 160$), the sample size necessary to find portfolio weights must be at least 1000, approximately. In small-sample situations, one cannot trust in-sample evaluations of portfolio efficacy, and obligatorily must use out-of-sample datasets.

In Fig. 1(a), we present an analogous family of five curves obtained for situation when the learning set size is fixed ($L = 400$) and the dimensionality is changing. Such type of dependencies heavily depends on the input variable (in our situation ATSs) ranking. In this experiment, we ranked ATSs according to their individual Sharpe ratio values, and paid no attention to covariance matrix. Therefore, our ranking was far from ideal. Nevertheless, we observed an obvious peaking effect. In both theoretical and empirical graphs, we see that the out-of-sample Sharpe ratio increases with increase in the number of trading agents. Later, the Sharpe ratio starts diminishing. The approximate theoretical equation (15) and the experiment give the same qualitative result: i.e., the necessity to select a small number of most profitable agents. It also shows the obligation to regularize the covariance matrix or use

other means to stabilize portfolio solutions. The latter question will be considered in Sections II-C and II-D.

C. Block-Diagonal Covariance Matrix

Term $T_S = 1 + N/(L - N)$ in (15) explains the reason for the dramatic diminishing of portfolio performance when $N \to L$ if full sample covariance matrix (4) is used. In the standard estimate (4), we take into account N variances and $N(N-1)/2$ correlations. Use of assumptions about the block-diagonal structure of covariance matrix can dramatically reduce the number of parameters to be estimated.

To simplify analytical formulas, suppose we will postulate that the N ATSs are split into R mutually independent $N_{\rm B}$-dimensional parts (blocks) where $N_{\rm B} = N/R$ is an integer. According to the block independence assumption, we have to estimate R block covariance matrices $\boldsymbol{S}_{\rm BL}$, totally $N(N-R)/(2R)$ nonzero correlations. In this case, we have to estimate approximately R times smaller number of unknown parameters. Like in the full covariance matrix case, one can construct orthogonal transformations, such that

$$\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \ldots, \boldsymbol{\Delta}_R),\ \boldsymbol{\Delta}_\alpha = (\delta_{\alpha 1}, 0, 0, \ldots, 0),\ (\alpha = 1, \ldots R).$$

Thus, the out-of-sample Sharpe ratio can be represented as

$$\frac{\boldsymbol{\Delta}\mathbf{S}_{\rm BL}^{-1}(\boldsymbol{\Delta} + \overline{\boldsymbol{\Theta}})^T}{\sqrt{(\boldsymbol{\Delta} + \overline{\boldsymbol{\Theta}})\boldsymbol{S}_{\rm BL}^{-1}\mathbf{I}^{-1}\boldsymbol{S}_{\rm BL}^{-1}(\boldsymbol{\Delta} + \overline{\boldsymbol{\Theta}})^T}} = \frac{\sum_{\alpha=1}^{R}\delta_\alpha^2 s_*^{11}}{\sqrt{\sum_{\alpha=1}^{R}\delta_\alpha (s_*^{11})^2 + \sum_{i=1}^{N}\Theta_i^2\sum_{j=1}^{N}(s_*^{ij})^2}}$$

where s_α^{ij} is the ijth element of $N_{\rm B} \times N_{\rm B}$-dimensional inverse matrix $\boldsymbol{S}_{\alpha BL}^{-1}$, the symbol "*" means "any of $1, 2, \ldots, R$," and according to standard theory [31], [32], the expected values

$$E(s_*^{ii})^2 = \frac{L^2}{(L - N/R)^2} \text{ and } E(s_*^{ij})^2 = \frac{L^2}{(L - N/R)^3},\ \text{if } j \neq i.$$

After some algebra, we obtain that the mean value is

$$S\hat{h}_{\rm out}^{\rm BD} \to \delta/\sqrt{(1 + N/(L \times \delta^2))(1 + N/(L - N/R))}. \quad (16)$$

Equation (16) shows that the small-sample properties of the portfolio improve especially if N is close to L. Unfortunately, the noticeable influence of inexact estimation of the mean returns does not change.

D. Regularized Portfolio

One more possible way of enhancing the portfolio solution is to use a shrinkage estimate where the off-diagonal elements of the covariance matrix are moderated. The variance elements in the diagonal are kept untouched

$$\boldsymbol{S}_{\rm reg1} = \boldsymbol{S} \times (1 - \lambda) + \boldsymbol{D} \times \lambda \quad (17)$$

where $\boldsymbol{D}$ is a diagonal $N \times N$ matrix of variances found from matrix $\mathbf{S}$, and λ is regularization parameter ($0 \le \lambda \le 1$).

When $\lambda = 0$, we have no regularization; when $\lambda = 1$, we have hard regularization. In latter case, we use only the diagonal elements (variances) of matrix $\mathbf{S}$. An alternative regularization formula

$$\boldsymbol{S}_{\rm reg2} = \boldsymbol{S} + \mathbf{I} \times \lambda,\ \text{where } (0 \le \lambda \le \infty) \quad (18)$$

is often used in pattern recognition. Like in (17), for $\lambda = 0$, we have no regularization; however, when $\lambda \to \infty$, we start to ignore covariance matrix. In principle, such a method can be applied to portfolio optimization, too. Below we will consider the case of hard regularization where only diagonal elements, matrix $\boldsymbol{D}$, are used in portfolio optimization.

Like in (8) we consider the rough approximation

$$\boldsymbol{w}_D = \overline{\boldsymbol{X}}\boldsymbol{D}^{-1}\mathbf{d}. \quad (19)$$

To simplify the final analytical formula for expected mean value of the out-of-sample Sharpe ratio, in this paper we suppose that input return values are uncorrelated. Performing simple normalizing of the variances, we conclude that we have to consider the following model:

1) the distribution of the N-dimensional vector $\boldsymbol{X}$ is Gaussian with mean $\boldsymbol{\Delta} = (\delta_1, \delta_1, \ldots, \delta_N)$ and $N \times N$ CM $\mathbf{I_N}$;
2) $\overline{\boldsymbol{X}} = (\boldsymbol{\Delta} + \overline{\boldsymbol{\Theta}})$, where covariance matrix of zero mean vector $\overline{\boldsymbol{\Theta}}$ is $\mathbf{I}_N \times 1/L$;
3) the N elements, $d_1, d_2, \ldots, d_N$ in the diagonal sample-based matrix $\boldsymbol{D}$ are independent and distributed as χ_L^2/L. When $L \to \infty$, its variance approaches zero.

Use of estimate (19) results in that the distribution of the out-of-sample Sharpe ratio can be represented as

$$\frac{\boldsymbol{\Delta}\boldsymbol{D}^{-1}(\boldsymbol{\Delta} + \overline{\boldsymbol{\Theta}})^T}{\sqrt{(\boldsymbol{\Delta} + \overline{\boldsymbol{\Theta}})\boldsymbol{D}^{-1}\mathbf{I}_N\boldsymbol{D}^{-1}(\boldsymbol{\Delta} + \overline{\boldsymbol{\Theta}})^T}} = \frac{\boldsymbol{\Delta}\boldsymbol{D}^{-1}\boldsymbol{\Delta}^T}{\sqrt{(\boldsymbol{\Delta}\boldsymbol{D}^{-2}\boldsymbol{\Delta}\overline{\boldsymbol{\Theta}}\boldsymbol{D}^{-2}\overline{\boldsymbol{\Theta}}^T)}}. \quad (20)$$

The diagonal matrix $\boldsymbol{D}^{-1}$is composed of N independent elements L/χ_L^2 in its diagonal, and the matrix $\boldsymbol{D}^{-2}$ is composed of N independent elements $(L/\chi_L^2)^{-2}$ in its diagonal. The mean values of the diagonal elements of matrices $\mathbf{D}^{-1}$ and $\mathbf{D}^{-2}$ are $(L-1)/(L-3)$ and $(L-1)^2/[(L-3)(L-5)]$. They tend to 1 asymptotically when both $N \to \infty$ and $L \to \infty$.

After some algebra we obtain mean value of $Sh_{\rm Dout}^2$ as

$$S\hat{h}_{\rm out}^{\rm D} \to \delta/\sqrt{1 + N/(L \times \delta^2)}. \quad (21)$$

Equation (20) shows that the hard regularization portfolio degradation factor depends only on inaccuracies that arise while estimating the mean of portfolio returns. Inaccuracies in the estimation of variances do not affect asymptotically. At the same time, we are obliged to have in mind that, in this heavy regularization case, we are ignoring all correlations.

Equation (21) advocates that regularization (17) is more useful than regularization (18). In (17), asymptotically (when $\lambda \to 1$) we do not ignore the variances. In (18), asymptotically (when $\lambda \to \infty$) we ignore both the correlations and the variances. Both regularization methods, however, display similar properties when λ is small.

In Fig. 2, we present out-of-sample Sharpe ratios (means of 100 independent experiments) as functions of the learning set

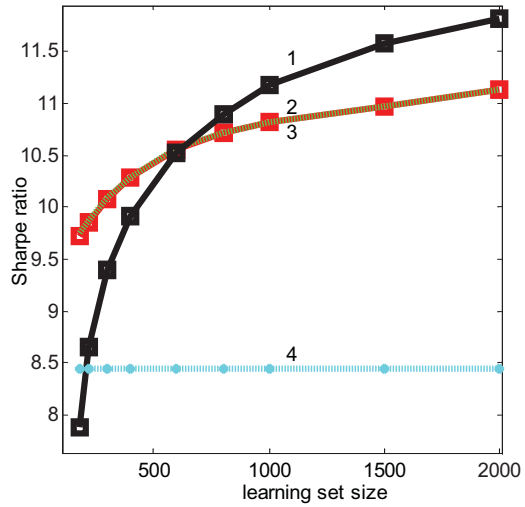

Fig. 2. Scissors effect in the portfolio design. Out-of-sample Sharpe ratio as a function of learning set size in experiments with 169-D Gaussian data: 1: full sample covariance matrix (4) was used to calculate the weights, 2: diagonal elements of the matrix were used, 3: true values of diagonal elements were used, 4: $1/N$ portfolio; means of 100 independent experiments.

size in experiments with 169-D Gaussian data considered in this paper. Curve 1 corresponds to portfolio weights when full sample covariance matrix (4) was used in (7). We see that the method becomes ineffective when sample size is below 400.

As an alternative to full covariance matrix, we considered hardly regularized CM where only the diagonal elements of matrix $\boldsymbol{S}$ were used in (7) (Curve 2). As predicted by theory, [i.e., (21)], use of exact values of the diagonal elements (in this simulation we knew the covariance matrix $\boldsymbol{\Sigma}$ exactly) resulted in the same portfolio accuracy: curve 3 matches curve 2 exactly. In this distinct series of experiments, simple $1/N$ portfolio rule results in much lower profit (cyan curve, 4).

The pair of curves 1 and 2 in Fig. 2 reminds the scissors shape and advocates that in small-sample situations one needs to apply simple decision-making rules (use only diagonal elements of matrix $\boldsymbol{S}$). In large-sample situations, it is preferable to use more complex rules. This outcome was called "a scissors effect." In pattern recognition, it was proved in [35], see also [21], [36]). Now we demonstrate this behavior in portfolio optimization. Another observation that follows from theoretical and simulation studies is that estimation of the variances asymptotically do not affect portfolio accuracy; mostly, estimation of correlations is responsible for term T_S in (15).

III. Experimental Analysis of Sample Size Effects

A. Data Used and Experiment Design

New information communication technologies allow converting dissimilar trading strategies into ATSs (trading robots). Simultaneous analysis of a large number of traders' actions allows us to take into account experience and knowledge of competing investors and integrate this knowledge into one's own decision-making process [22]. Moreover, weighted summation of a multitude of such systems moves distribution of the portfolio returns toward a Gaussian.

To calculate the weights and estimate the performances of the portfolio optimization schemes, we investigated a 169-D profit/loss time series representing the simulated track records of real-life trading systems in the period 2003–2012. The automated trading firm that provided the data used a two-level trading design. In the first level, the ATSs were used. In the second level, the robot outputs were fused by the $1/N$ strategy. The firm trades in U.S. and European futures: stock index, energies, metals, commodities, energies, interest rate products, FX products, etc.

The original portfolio was composed of high-frequency and short-term ATSs. Most systems exit their position within 24 h but some can hold positions longer. We already mentioned above that most of the time the ATSs tend not to trade but wait for the right opportunity. As a result, we were dealing with very sparse data, mainly filled with zeros, the days when the systems were not trading. After inspection of the data over several years, we observed that for different ATSs the Sharpe ratio (5) varied between 0 and 4.

The first-level systems were continually renewed: each day a new set of ATSs was selected. More profitable and less correlated systems were most wanted. Therefore, after several months the collections of ATSs usually were almost entirely new. In our research we carried out experiments with seven different ATS datasets acquired in 3–4 months intervals.

All comparative experiments were performed in the out-of-sample regime using a walk forward methodology. One trains the model with L N-dimensional vectors of profit/loss series and tests on the subsequent $L_{\text{valid}} = 100$ vectors. Next, one walks L_{valid} days forward: one shifts the training and validation sets' windows by L_{valid} days. We used $M = 14$ walk-forward steps with 100 day intervals. So, for each size of training set, we trained expert and fusion agents 14 times. In diverse experiments, we trained expert agents associated with $L = 100 - 1000$ days, preceding each testing interval.

B. Multiagent ATSs

To demonstrate the usefulness of the above theoretical issues in real-world PM tasks, we considered the high-frequency trading where, in principle, it is possible to probe hundreds or even thousands of investments [10], [33]. Our portfolio optimization strategy is based on the following two theoretically established facts:

1) if one uses a large number of input variables (automated trading systems) and calculates the portfolio weights for each of them, then according to the central limit theorem in probability theory, the distribution of the weighted sum $x_{\text{Pi}} = \boldsymbol{X}_i \boldsymbol{w}_T$, becomes close to Gaussian; subsequently, one has more arguments to apply the mean-variance paradigm;
2) to satisfy sample size/dimensionality requirements, one is obliged to use all means possible to use additional information and simplify the calculation of portfolio weights.

The exclusive feature of the profit/loss series consists in a large number of days where the ATSs refrain from active investments: a very large part of data consists of zeros.

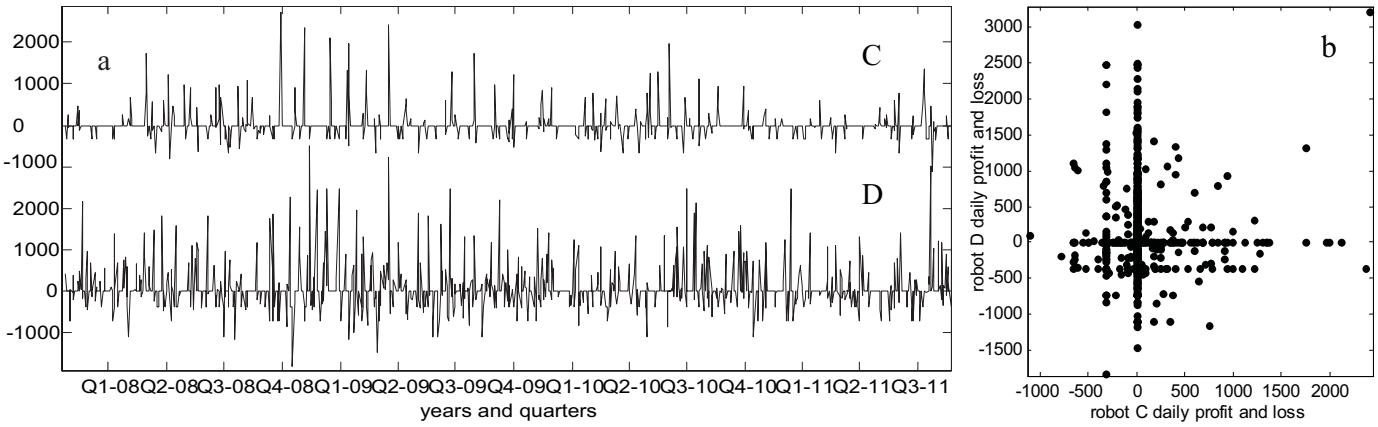

Fig. 3. (a) Fluctuations of robots' (ATSs') C and D profits and losses during 1000 working days. (b) Bivariate scatter diagram of their profit and losses.

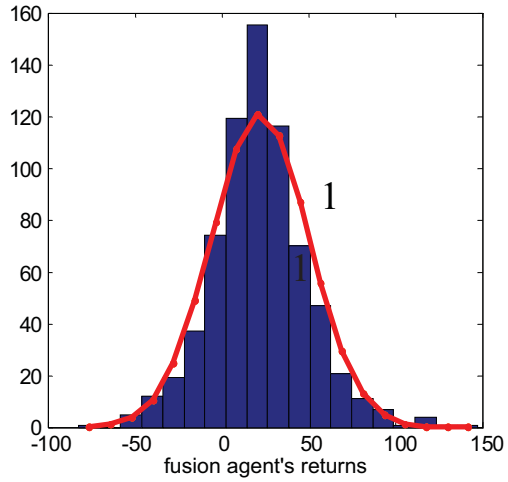

Fig. 4. Histogram of outputs of the fusion agent placed in the output of portfolio weights finding MAS (700 working days data). Red graph (1) is its Gaussian approximation.

In Fig. 3(a), we see an example where ~70% of days have zero profit and loss. A bivariate distribution is depicted in Fig. 3(b). The scatter diagram shows unambiguously that distributions of profits are far from Gaussian. Their density functions advocate that we have a mixture of three densities at least.

One component of the mixture is composed of zero days with zero profit. In normal trading days we have moderate fluctuations. Their mean value exceeds zero marginally. The third component is composed of outliers with large fluctuations. It is very advantageous that the profits of diverse trading robots are almost statistically independent [Fig. 3(b)].

Situation with nonnormality of portfolio returns' distribution is improved when we fuse outputs of a large number of ATSs. In Fig. 4 we see a histogram of distribution of weighted sum $x_{\mathrm{Pi}} = \boldsymbol{X}_i \mathbf{w}^T$, calculated by the MAS shown in Fig. 5. Here, $L_{\mathrm{Fusion}} = 700$ days was used to calculate $N = 169$ portfolio weights. In spite of the obvious nonnormality of the ATSs, the distribution of the weighted sum is notably closer to a Gaussian curve. While comparing with the Gaussian approximation performed from the histogram data, however, we see a sharp peak in the middle of the histogram. It means that we have a mixture of two densities and that there is scope for further improving the portfolio design algorithms.

In the first experiments of ATS-based portfolio optimization strategy [22], we were considering a single-stage schema with $N = 142$ ATS outputs fused linearly: $x_{\mathrm{Pi}} = \boldsymbol{X}_i \boldsymbol{w}^T$. We used the standard Markowitz portfolio optimization algorithm "frontcon" rule implemented in the MATLAB Financial Toolbox. In this series of experiments, it outperformed the $1/N$ portfolio.

The analysis confirmed a necessity to shorten the amount of training history: a 500-day data history produced the highest Sharpe ratio [22]. In an attempt to improve the learning set size/dimensionality properties of the weight calculation algorithm, we introduced regularization and used the constrained covariance matrix $\boldsymbol{S}$ where only $2N-1$ elements of the inverse matrix $\boldsymbol{S}^{-1}$ participated in the weights calculation (the first-order tree dependence model [20], [21]). It was found that the portfolio performance notably depended on the regularization parameter and learning set size.

In [34], the objective was directed toward developing an evolvable MAS aimed at finding the proper values of parameters just mentioned. In this approach, a multitude of "expert trading agents" based on N_A input ATSs were designed and compared on the basis of training set information ($N_A < N$). The agents differed in: 1) sets of N_A ATSs selected randomly; 2) values of regularization parameter reg used to estimate covariance matrix; and 3) levels of risk used to determine days when expert agents refused from investments. Finally, outputs of the R best agents were fused by the R-dimensional portfolio rule based on Markowitz's "frontcon" principle. In the out-of sample regime, this approach allowed reaching or even surpassing the performance of the best expert agent made known after concluding the experiment.

With further objective to simplify the portfolio weights finding rule to meet the small learning set size requirements, we made efforts to reduce dimensionality of expert agents. In Fig. 5 we present the multilayer-perceptron-like feedforward information flow schema. In this schema, the expert agents are based on a small number of distinct ATSs (trading robots) that mostly operate on one or several assets.

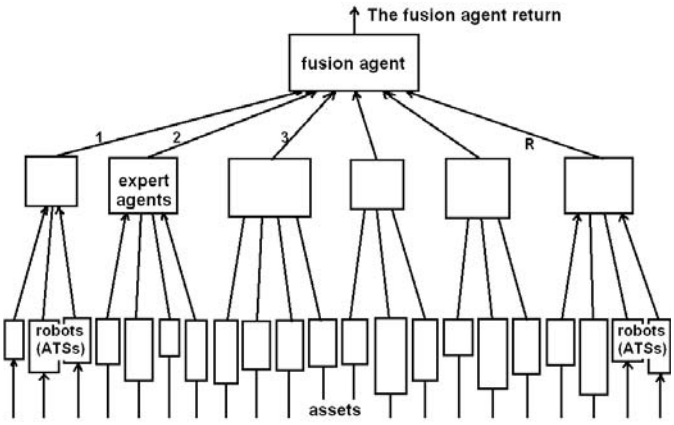

Fig. 5. Feed forward flow of information in two-layer portfolio calculation MAS.

TABLE I
BLOCK-DIAGONAL STRUCTURE OF THE CORRELATION MATRIX

12	**39**	**37**	**32**	**39**	**23**	**33**	-1	1	3	-2	2	1	0	1	3	-1	-1	-1	-1
0	**10**	**34**	**32**	**5**	**15**	**5**	-3	1	0	9	6	9	0	1	1	-1	-3	-4	-1
	0	**20**	**26**	**27**	**12**	**10**	0	-3	-1	-1	0	3	1	1	-2	2	-6	-5	-2
		0	**83**	**50**	**33**	**55**	0	2	2	8	11	3	7	0	2	0	-3	-3	-1
			0	**48**	**35**	**65**	0	1	1	7	11	3	7	0	2	0	-4	-4	-1
				0	**36**	**29**	1	3	1	4	6	2	0	3	3	-1	-2	-1	-1
					0	**31**	-7	2	0	6	3	7	-11	10	2	-1	-6	-7	0
						0	0	1	1	2	-1	0	0	-1	1	1	-3	-2	0
							0	**25**	**22**	**14**	**25**	**20**	**85**	**-11**	**17**	-2	-4	-1	-5
								0	**83**	**22**	**40**	**44**	**20**	**19**	**79**	-1	3	4	0
									0	**21**	**37**	**46**	**20**	**22**	**83**	-1	6	8	0
										0	**17**	**28**	**22**	**7**	**26**	0	10	12	3
											0	**25**	**29**	**8**	**36**	-4	-1	-5	-4
												0	**19**	**28**	**52**	2	4	3	1
													0	**-11**	**19**	0	-4	1	-3
														0	**28**	3	0	0	-1
															0	0	3	3	0
																0	**4**	**12**	**10**
																	0	**63**	**25**
																		0	**32**

C. Means to Simplify Decision-Making Schema

Learning set size and complexity issues considered in Section II emphasized that trustful extra information about the covariance matrix can help to improve the portfolio weights calculation. In preliminary experiments [22], we found that a 500-day data history produced the highest Sharpe ratio (5) somewhere about 10. So, the ratio mean $(x_{\mathrm{Pi}})/\mathrm{std}(x_{\mathrm{Pi}}) = \delta \approx 2/3$, and $\delta^2 \approx 0.5$. In the ATSs data, only 30% of days are active. One can hope that a number of efficient days L_{eff}, which actually determines small-sample properties of sample estimates of the covariance matrix, is notably smaller. For a preliminary assessment suppose, $L_{\mathrm{eff}} = 200$. Then the term $T_X = 1 + N/(L \times \delta^2) = 1 + 169/200/0.5 \approx 2.5$, but term $T_S = 1 + N/(L - N) = 1 + 169/(200 - 169) \approx 6.5$. It means that conventional CM estimate (4) does not fit in such sample size/dimensionality situations. Calculations with $L_{\mathrm{eff}} = 500$ also result in large values of the terms considered. To find extra ways to introduce additional information into the portfolio calculation procedure, we studied block-diagonal covariance matrix assumptions.

For the block-diagonal covariance matrix structure term $T^{\mathrm{BD}}_{\mathrm{Sout}} = 1 + N/(L - N/R)$. For $R = 25$, assuming $N_B = L/R \approx 7$ we find $1 + 169/(200 - 7) \approx 1.8$, i.e., much smaller value than $T_S = 6.5$.

In order to check if the covariance matrix of multitude ATSs can be converted into the block-diagonal structure, we performed analysis of empirical 169-D data. To split the ATSs into groups (blocks), we applied the k-means clustering algorithm implemented in MATLAB Statistics toolbox. Earlier, in Section II-D we showed that estimation of variances (diagonal elements of matrix $\mathbf{S}$) does not affect portfolio accuracy when $L \to \infty$ and $N \to \infty$ asymptotically. Thus, after data normalization according to the standard deviations, instead of the covariance matrix we can consider correlation matrix $\mathbf{K}$. For that reason, we used the $|1 - \mathrm{correlation}|$ between the robot outputs as a dissimilarity criterion. In an example experiment, we formed $R = 25$ blocks (clusters) of robots out of 169 ATSs.

In Table I we present a part of correlation matrix (three blocks out of $R = 25$) obtained in one of the experiments with sample size $L = 400$. To get the correlation coefficients, the numbers in the Table I should be divided by 100 (in the diagonal of the matrix, instead of 1 we printed 0).

Limitation of the cluster analysis procedure resulted in an imperfect block-diagonal structure. We see inside each single block, correlations are notably larger than correlations outside the block (first block: mean of correlations $\rho_m = 0.3$, standard deviation $\rho_{sd} = 0.21$; second block: $\rho_m = 0.3$, $\rho_{sd} = 0.23$; between both blocks: $\rho_m = 0.03$, $\rho_{sd} = 0.04$).

Among two dozen portfolio calculation schemas aimed at introducing additional information into weights calculation algorithm, we considered a number of block-diagonal structures. Experiments with complete estimation of all coefficients in each block and estimation of N values of mean returns did not lead to notable increase in portfolio performance. Theoretical analysis of terms T_X and $T^{\mathrm{BD}}_{\mathrm{Sout}}$ also confirms excessive values of product terms T_X and $T^{\mathrm{BD}}_{\mathrm{Sout}}$.

In a further attempt to introduce additional guesses into the parameter estimation process, we noticed that similar outputs of the ATSs that belong to the single cluster are notably correlated. So, it is reasonable to make a forceful assumption that correlations ρ_{xy} between the outputs of N_{Bt} robots of the rth group are equal, i.e., $\rho_{xy} = \rho_r$. If this assumption is correct, the correlation matrix of the rth block would be

$$\mathbf{K}_{\mathrm{BD}r} = \mathbf{I} \times (1 - \rho_r) + \mathbf{1}^T\mathbf{1} \times \rho_r$$

where the matrix $\mathbf{I}$ and vector $\mathbf{1}$ are as defined in Section II.

In the multiagent PM system, each expert agent is fed by outputs of the ATSs that belong to the single cluster of dimensionality N_{Br}. In (15), (16), and (21), we showed that the decrease in portfolio efficacy due to inexact estimation of the mean of portfolio returns $\overline{X}$ is proportional to $T_X = 1 + N/(L \times \delta^2)$. After making assumption that all $\rho_{xy} = \rho_r$, we see that the term T_X can turn out to be larger than term $T^{\mathrm{BD}}_{\mathrm{S}} = 1 + N/(L - N/R)$. To keep away from estimation of the

means vectors of the returns, assume that $\overline{\boldsymbol{X}}_r \mathbf{D}^{-1/2=}\overline{\boldsymbol{Y}}_r = \mathbf{1}$. Inserting the assumed vectors into (7) we obtain

$$\boldsymbol{w}_{BDr} = (q_s\mathbf{1} + \mathbf{1})(\boldsymbol{K}_{BDr}/\Lambda + \times\mathbf{1}^T\mathbf{1})^{-1}. \tag{22}$$

Matrix algebra shows that solution of (22) is straightforward: $\boldsymbol{w}_{BDr} = \gamma \times \mathbf{1}$, where γ is a positive scalar. Without a loss of generality, it can be normalized to meet requirement (1). Hence, the above deduction gives arguments that we can try for using $1/N_{Br}$ portfolio for any ρ_r and Λ values.

Consequently, theoretical considerations support easily intuitive understandable guesses that if:

1) the variances of certain group of the ATSs are equalized;
2) mean returns of N_B robots are approximately equal;
3) ATSs are equally correlated;
4) a sum of the ATSs output is approximately normally distributed, the benchmark $1/N_B$ portfolio rule is the optimal solution.

We employed this deduction to justify the architecture of portfolio managing MAS (Fig. 5) where the correlated input ATSs were grouped into the blocks. To reduce the scornful small-sample problems, we ignored scatter of the correlations and scatter in mean values inside each single block, and applied nontrainable $1/N_B$ rules (equally weighted portfolio). After averaging, the distributions of outputs of the blocks (expert agents) are closer to Gaussian. For that reason, we have more arguments to combine the experts' outputs by the regularized linear fusion rule based on the mean-variance framework. To increase diversity of the expert agents, we used varied training sets to perform cluster analysis in each walk-forward step of the validation experiment. The training sets differed in their lengths.

IV. Results With MAS

The objective of experimental study was to compare the novel multiagent learning system with two benchmark portfolio weight determining rules: 1) the simple $1/N$ rule where all N weights were equal to $1/N$, and 2) the standard MATLAB Markowitz efficient frontier finding rule MatMar, both applied to outputs of N trading systems.

A. Experiment Design

In the benchmark $1/N$ portfolio, we have no training. In the MATLAB Markowitz (MatMar) efficient frontier finding rule, we used all previous training vectors up to beginning of each 100-day testing interval. In the novel MAS-based system, to ensure diversity:

1) the agents differ in subsets of clustered ATSs and dimensionality $N_{\mathrm{B}r}(\sum_{r=1}^{R} N_{\mathrm{B}r} = N)$;
2) the lengths of learning sets used to perform clustering were either $L_1 = 200$, $L_2 = 300$, or $L_3 = 400$.

Thus, the total number of agents was $3R$. In the final decision-making stage (fusion of $3R$ expert agents' outputs), we used (7) and the efficient frontier approach. For estimating the mean vector of returns and covariance matrix of the $3R$ agent returns, we used $L_{\mathrm{Fusion}} = 600$ days history.

Use of MatMar algorithm to calculate portfolio weights in the MAS resulted in a slightly lower out-of-sample Sharpe ratio values (approximately by 0.5) in most cases. The possible reason is a standard regularization algorithm incorporated into the weights calculation rule. In the novel algorithm (7), we use the accuracy constant Λ that acts as an additional regularization factor. Besides, we brought negative weights to naught and normalized the weight magnitudes to meet constraints (1). This procedure simplifies and partially regularizes the solution as well.

The increase in variety of expert agents differing in regularization parameter and/or in the learning set size assisted in obtaining higher Sharpe ratios. It is worth mentioning that period 2009–2012 is very unhandy to perform comparison of financial engineering algorithms. Results of out-of-sample evaluations during the 100-day validation interval depend on exact positions of the start and end of learning and validation sets. Thus, 0.5 inaccuracies in determining 14 intervals' averages of the Sharpe ratio (5) estimation cannot be avoided.

B. Experiments With Block-Diagonal Matrix Structures

In experiments with the block-diagonal structure, we also investigated many variants. To design expert agents, at first we used the conventional mean-variance paradigm with full estimation of the mean returns, different ways to define matrix structures (numbers of blocks, all correlations in a single block are the same), the $1/N_{\mathrm{B}}$ portfolio rule. The experiments showed that, in the crisis period, learning sample sizes should be rather small. While performing the experiments with empirical data till the end of 2011, we found the best structure of the MAS and its parameters. The best results were obtained with $R = 25$ blocks, the number of replicates in clustering 30, and exploitation of $1/N_{\mathrm{B}}$ Portfolio rule to design 3×25 expert agents. For fusion of 75 expert outputs, we used notably regularized (reg $= 0.8$) 75×75 covariance matrix and $L_{\mathrm{Fusion}} = 600$ days training sequences. We would like to draw the reader's attention to the fact that the best results were obtained when rather short training histories were used to form the first decision level expert agents: $L_1 = 200$, $L_2 = 300$, and $L_3 = 400$. It means that the financial data structure changed remarkably during the short period. This fact has also been observed earlier by our partners who renewed the sets of the ATSs almost every day.

The global parameters of the portfolio weights calculation system (reg, R, L_{Fusion}, etc.) were established from separate dataset recorded up to the end of 2011. The performance of novel portfolio optimization system was tested with a new collection of $N = 154$ ATSs recorded in period 2003–May 2012. In Fig. 6 we present the last five years' dynamics of the Sharpe ratio (5) in fourteen validation intervals. Time is indicated for the last day of each validation interval.

Because of the relatively short validation intervals, we were estimating means and standard deviations with notable errors. Hence, we were unable to avoid substantial fluctuations. In this experiment, the Matlab frontcon rule MatMar was trained in its standard conditions (all training data available was used to estimate $\overline{\boldsymbol{X}}$ and $\boldsymbol{S}$). Therefore, this rule failed to conquer MAS when training history became too lengthy. The novel MAS was designed specifically to adapt to changes. It used

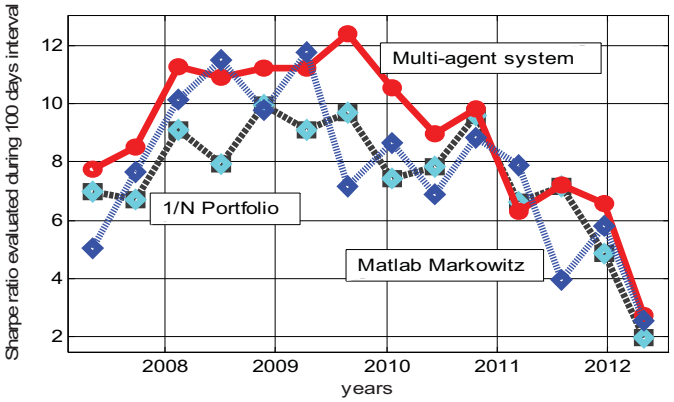

Fig. 6. Four years dynamics of the Sharpe ratio in test intervals; $N = 169$ robots; $3 \times 25 = 75$ expert agents trained $L = 200$, 300, or 400 days, average of 20 experiments.

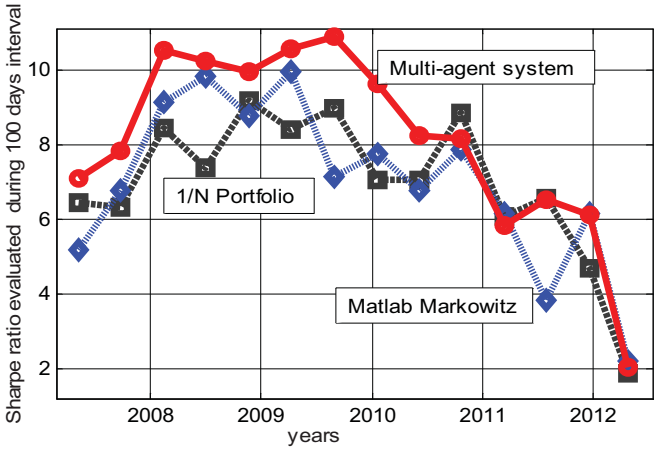

Fig. 7. Five years dynamics of the Sharpe ratio; N=100 robots selected randomly from 154 ones; means of 200 independent experiments with training set sizes $L = 200$, 300, and 400 days.

short learning sequences to group the ATSs into blocks. For that reason, it outperformed both benchmark rules almost in all validation intervals. At the very end (last months of 2012), because of the chaotic behavior of the financial market due to harsh political and economic events, all methods became almost ineffective.

To compare performance more profoundly, we carried out three series of 200 additional experiments with 200 diverse 100-D ATSs collections randomly formed from the Data No. 7. In Fig. 7 we see average values, which confirms the conclusions obtained in the experiment with the original 154-D ATSs data set.

In Fig. 8 we see a scatter diagram of the Sharpe ratio obtained in 200 independent experiments with diverse datasets. The triangles in Fig. 8 show the novel MAS portfolio (y-axis) versus MatMar mean-variance approach (x-axis). In all 200 validation experiments, the novel MAS-based scheme outperformed the nontrainable $1/N$ portfolio rule as well.

Figs. 7 and 8 show that the PM's success is heavy influenced by the variations of the financial market. The profit increases up to 2009 but falls later. To check the influence of learning set sizes, we repeated experiments with shorter and lengthier

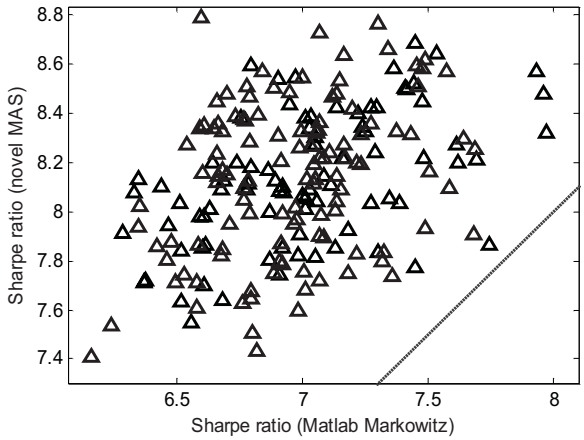

Fig. 8. Scatter diagrams of bi-variate distributions of the Sharpe ratio in 200 independent experiments.

TABLE II
SHARPE RATIO OF MAS AND TWO BENCHMARK METHODS

Method Average of	MAS $L_1 = 200$	MAS $L_1 = 100$	MAS $L_1 = 300$	$1/N$	MatMar
14 periods	8.13 (0.29)	8.12	**8.18**	6.97	6.98
First 7 periods	9.38 (0.42)	9.39	**9.40**	7.71	8.29
Last 6 periods	6.17 (0.36)	6.17	**6.27**	5.87	5.51
Last 3 periods	4.92 (0.43)	4.84	**5.05**	4.38	4.08

training histories: $L_1 = 100$, $L_2 = 200$, $L_3 = 300$, and $L_1 = 300$, $L_2 = 450$, $L_3 = 600$. However, we did not notice any significant differences between the results.

In Table II, we present mean Sharpe values calculated for all 1400-, the first 700-, and the last 600- and 300-day periods. Inside brackets we present the standard deviations. In the last two columns of the Table II, we show the performance of two benchmark methods. Novel hierarchical MAS visibly outperform the benchmark methods: the differences in the Sharpe ratio values exceed 1 (e.g., 8.13 versus 6.98).

V. CONCLUSION

By means of multivariate statistical analysis and simulation studies, we analyzed the influence of sample size and input dimensionality effect on the accuracy of determination of N portfolio weights. We performed theoretical and simulation studies of small-sample properties of diverse weight calculation schemes. We found that degradation in portfolio performance due to inexact estimation of N means and $N(N-1)/2$ correlations is proportional to two product terms:

1) term $T_X = 1 + N/(L \times \delta^2)$ is responsible for the inexact determination of the mean returns, vector $\overline{\boldsymbol{X}}$;
2) term $T_S = 1 + N/(L - N)$ is responsible for the inaccuracies that arise when estimating the correlations.

Both terms increase the in-sample profit, but they reduce the out-of-sample one. In certain situations, T_X may even exceed T_S. The term T_S clarifies why portfolio performance

diminishes dramatically if the sample size is close to the number of inputs. The drawbacks of the increased dimensionality of the portfolio weight vector can be reduced by a number of techniques considered in multivariate statistical analysis: dimensionality reduction, regularization, use of special structures of the covariance matrix, etc. An interesting deduction that followed from analytical and simulation studies is that asymptotic estimation of N variances does not worsen the result.

To diminish unhelpful sample size dimensionality effects, we suggested a special MAS architecture where the number of trading agents is reduced by splitting the ATSs into separate blocks. We showed that the nontrainable $1/N$ portfolio rule can become optimal if the means of ATSs' returns in the single block are equal, and they are equally correlated. We recommend to approach these conditions by performing $|1 - \text{correlation}|$ based cluster analysis of the profit and loss time series and having a large number of trading robots in the block. To increase the diversity of expert agents, we used training sets of different lengths in the clustering procedure. Considerable regularization of the covariance matrix of the expert outputs allowed obtaining comparatively reliable fusion rule of outputs of the MAS.

Comparison of the novel MAS system-based PM rule with two standard benchmark rules was performed in out-of-sample regime with financial data of 2003–2012. A novel PM system was renewed in each walk-forward step (100 working days) in out-of sample portfolio validation experiments. In the experiments, the novel decision-making schema resulted in a 12% higher Sharpe ratio than both benchmark PM methods applied to ATSs.

It is worth mentioning that the cluster analysis procedure used in this paper is far from perfect. We used the k-means algorithm and $|1 - \text{correlation}|$ as the dissimilarity measure. Improving the clustering criterion is an important subject for the future study. Novel approaches aimed at preprocessing truly high-dimensional input data to low-dimensional representations combined with regularization [37], [38] can facilitate enhanced expert agent design. Among the topics for future research could be taking into account the fact that, in finite ATSs situations, the weighted sums of expert agents produce mixtures of two densities (see Fig. 4). We need to analyze much larger datasets with tens of thousands of trading robots. We need to develop more flexible evolving MASs capable of adapting expert agents and their learning parameters (learning set sizes, regularization parameters, levels of risk, degree of collaboration with other agents, etc.) to continuously changing environments.

We ought to bear in mind that many economic theories explain the efficient-market hypothesis (EMH) by regarding the stock market as a game in which a large number of competing players are trying to maximize their winnings by buying low and selling high. The EMH declares that the market is effective and that, whenever new information comes up, the market absorbs it by correcting itself [39]–[42]. The generation of artificial financial market data and evolving multiagent data analysis systems suggest a variety of models where, in an explicit way, it is assumed that asset prices in the future depend on the prediction and actions of all its participants. Hence, ATSs and the hierarchical MAS-based approach are sources of specific information, and can suggest theoretical arguments for an explanation of EMH and financial market control.

REFERENCES

[1] H. M. Markowitz, "Portfolio selection," *J. Finan.*, vol. 7, no. 1, pp. 77–91, 1952.

[2] H. M. Markowitz, "Portfolio selection," in *Efficient Diversification of Investments*. New York: Wiley, 1959, p. 344.

[3] R. C. Grinold and R. N. Kahn, *Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Selecting Superior Returns and Controlling Risk*, 2nd ed. New York: McGraw-Hill, 1999.

[4] J. Voit, *The Statistical Mechanics of Financial Markets*. Berlin, Germany: Springer-Verlag, 2001.

[5] P. D. McNelis, *Neural Networks in Finance: Gaining Predictive Edge in the Market*. New York: Academic, 2005.

[6] K. K. Hung, Y. M. Cheung, and L. Xu, "An extended ASLD trading system to enhance portfolio management," *IEEE Trans. Neural Netw.*, vol. 14, no. 2, pp. 413–425, Mar. 2003.

[7] M. Jeffery and I. Leliveld, "Best practices in IT portfolio," *MIT Sloan Manag. Rev.*, vol. 45, no. 3, pp. 41–49, 2004.

[8] D. M. Berwick, "Disseminating innovations in health care," *J. Amer. Med. Assoc.*, vol. 289, no. 15, pp. 1969–1975, 2003.

[9] C. Fishburn, "Mean-risk analysis with risk associated with below target returns," *Amer. Econ. Rev.*, vol. 67, no. 2, pp. 116–126, 1977.

[10] K. K. Hung, C. C. Cheung, and L. Xu, "New sharpe-ratio-related methods for portfolio selection," in *Proc. Comput. Intell. Finan. Eng.*, 2000, pp. 34–37.

[11] V. D. DeMiguel, L. Garlappi, and R. Upppal, "Optimal versus naïve diversification: How inefficient is the 1/N portfolio strategy?" *Rev. Finan. Studies*, vol. 22, no. 5, pp. 1915–1953, 2009.

[12] R. Kan and G. Zhou, "Optimal portfolio choice with parameter uncertainty," *J. Finan. Quant. Anal.*, vol. 42, no. 3, pp. 621–656, 2007.

[13] D. J. Disatnik and S. Benninga, "Shrinking the covariance matrix-simpler is better," *J. Portfolio Manage.*, vol. 33, no. 4, pp. 56–63, 2007.

[14] S. V. Goldin and N. N. Poplavskij, "Methods to increase a robustness of discriminant function," in *Proc. Math. Methods Oil Geol. Geophysis.*, vol. 36. 1970, pp. 129–55.

[15] S. Raudys and D. Young, "Results in statistical discriminant analysis: A review of the former Soviet Union literature," *J. Multivar. Anal.*, vol. 89, no. 1, pp. 1–35, 2004.

[16] G. Papp, S. Pafka, M. A. Nowak, and I. Kondor, "Random matrix filtering in portfolio optimization," *Acta Phys. Polon. B*, vol. 36, pp. 2757–2765, Sep. 2005.

[17] M. Bonato, M. Caporin, and A. Ranaldo, "Forecasting realized (co)variances with a block structure Wishart autoregressive model," Swiss Banking Institute, Univ. Zurich, Zurich, Switzerland, Tech. Rep., 2008.

[18] D. Disatnik, "Portfolio optimization using a block structure for the covariance matrix estimating," *J. Bus. Finan. Account.*, vol. 39, nos. 5–6, pp. 806–843, 2012.

[19] N. Hautsch, L. M. Kyj, and R. C. A. Oomen, "A blocking and regularization approach to high dimensional realized covariance estimation," Quantitative Products Lab., Berlin, Germany, Tech. Rep., 2009.

[20] S. Raudys and A. Saudargiene, "First-order tree-type dependence between variables and classification performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 1324–1328, Feb. 2001.

[21] S. Raudys, *Statistical and Neural Classifiers: An Integrated Approach to Design*. NY: Springer-Verlag, 2001.

[22] S. Raudys and A. Raudys, "High frequency trading portfolio optimisation: Integration of financial and human factors," in *Proc. 11th Int. Conf. Intell. Syst. Design Appl.*, Nov. 2011, pp. 696–701.

[23] Z. Bai, H. Liu, and W.-K. Wong, "Enhancement of the applicability of Markowitz's portfolio optimization by utilizing random matrix theory," *Math. Finan.*, vol. 19, no. 4, pp. 639–667, 2009.

[24] A. Gandy and L. A. M. Veraart, "The effect of estimation in high–dimensional portfolios," *Math. Finan.*, to be published.

[25] S. Raudys and I. Zliobaite, *Prediction of Commodity Prices in Rapidly Changing Environments* (Lecture Notes in Computer Science), vol. 3686, Berlin, Germany: Springer-Verlag, 2005, pp. 154–163.

[26] S. Raudys and A. Mitasiunas, *Multi-Agent System Approach to React to Sudden Environmental Changes* (Lecture Notes in Artificial Intelligence), vol. 4571. Berlin, Germany: Springer-Verlag, 2007, pp. 810–823.
[27] F. D. Freitas, A. F. De Souza, and A. R. De Almeida, "Prediction-based portfolio optimization model using neural networks," *Neurocomputing*, vol. 72, nos. 10–12, pp. 2155–2170, 2009.
[28] S. Raudys, "On determining training sample size of linear classifier," *Acad. Sci. USSR*, vol. 28, pp. 79–87, Mar. 1967.
[29] A. Zollanvari, U. M. Braga-Neto, and E. R. Dougherty, "Analytic study of performance of error estimators for linear discriminant analysis," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4238–4255, Sep. 2011.
[30] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Hoboken, NJ: Wiley, 2003.
[31] A. A. Bowker, *A Represetation of Hotelling's T^2 and Anderson's Clasifications Statistics W in Terms of Single Statistics*. Stanford, CA: Stanford Univ. Press, 1963.
[32] S. Raudys, "On the amount of a priori information in designing the classification algorithm," *Proc. Acad. Sci. USSR*, vol. 4, pp. 168–174, Aug. 1972.
[33] T. Hendershott and R. Riordan, "High frequency trading and price discovery," UC Berkeley, Berkeley, Tech. Rep., 2011.
[34] S. Raudys and A. Raudys, "Three decision making levels in portfolio management," in *Proc. IEEE Conf. Comput. Intell. Finan. Eng. Econ.*, Mar. 2012, pp. 197–204.
[35] S. Raudys, "On the problems of sample size in pattern recognition," in *Proc. 2nd All-Union Conf. Stat. Methods Control Theory*, vol. 2. 1970, pp. 64–76.
[36] S. Raudys, "Evolution and generalization of a single neurone. II. Complexity of statistical classifiers and sample size considerations," *Neural Netw.*, vol. 11, pp. 297–313, Mar. 1998.
[37] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition and verification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 526–534, Mar. 2012.
[38] A. Stuhlsatz, J. Lippel, and T. Zielke, "Feature extraction with deep neural networks by a neneralized discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 4, pp. 596–608, Apr. 2012.
[39] E. Fama, "Effcient capital markets: A review of theory and empirical work," *J. Finan.*, vol. 25, no. 2, pp. 383–417, 1970.
[40] E. F. Fama, "Random walks in stock market price," *Finan. Anal. J.*, vol. 21, no. 5, pp. 55–59, 1965.
[41] B. G. Malkiel, "The efficient market hypothesis and its critics," *J. Econ. Perspect.*, vol. 17, no. 1, pp. 59–82, 2003.
[42] J. Mockus and A. Raudys, "On the effcient-market hypothesis and stock exchange game model," *Expert Syst. Appl.*, vol. 37, no. 8, pp. 5673–5681, 2010.

Sarunas Raudys received the Masters and Ph.D. degrees in computer science from the Kaunas University of Technology, Kaunas, Lithuania, and the U.S.S.R. Doctor of Science (Habil.) degree from the Riga Institute of Electronics and Computer Science, Riga, Latvia, in 1978.

He is currently a Head Researcher with the Department of Informatics, Faculty of Mathematics and informatics, Vilnius University, Vilnius, Lithuania. His current research interests include multivariate analysis, statistical pattern recognition, data mining, artificial neural networks, deep learning, evolvable MASs, artificial economics, and artificial life.