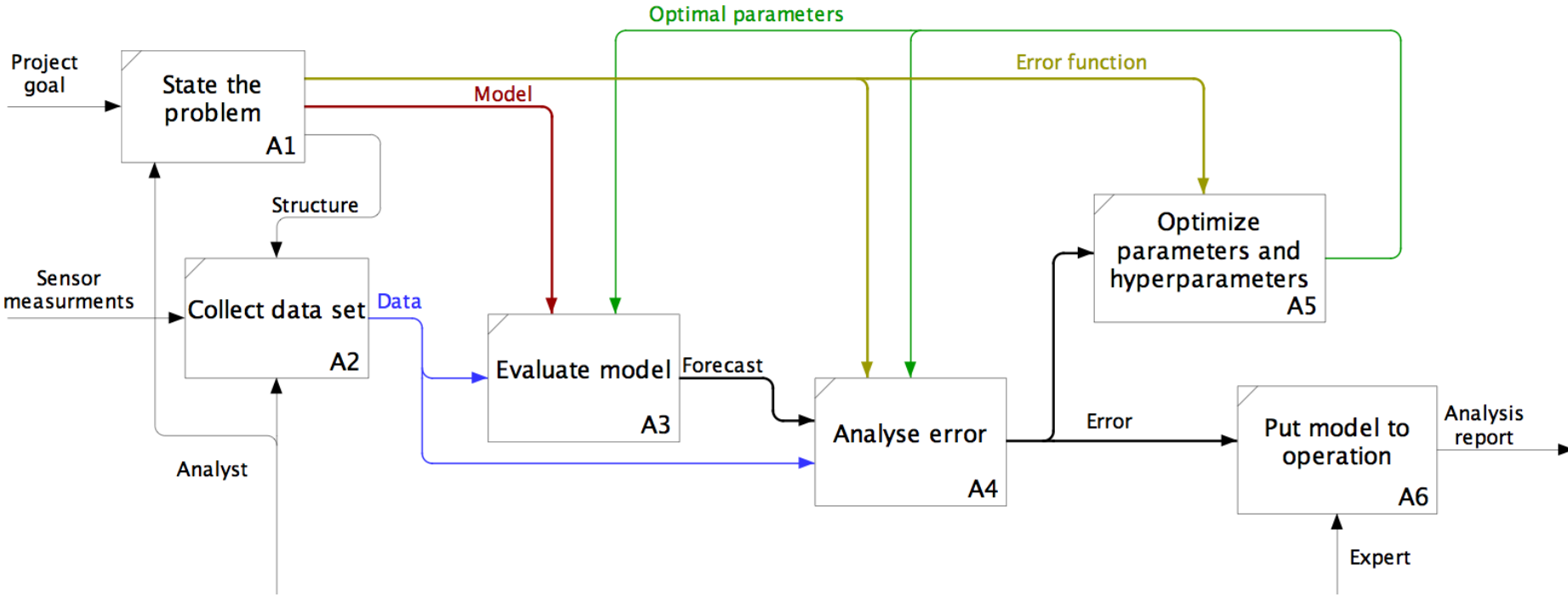
Постановка задач и выбор моделей в машинном обучении

Вадим Викторович Стрижов

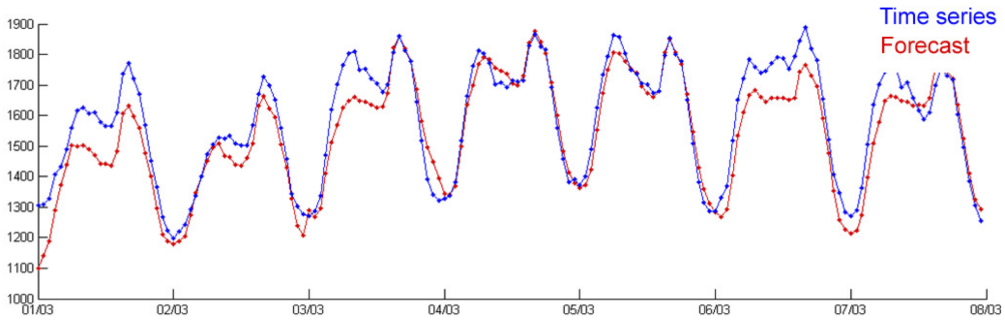
Московский физико-технический институт

Осенний семестр 2019

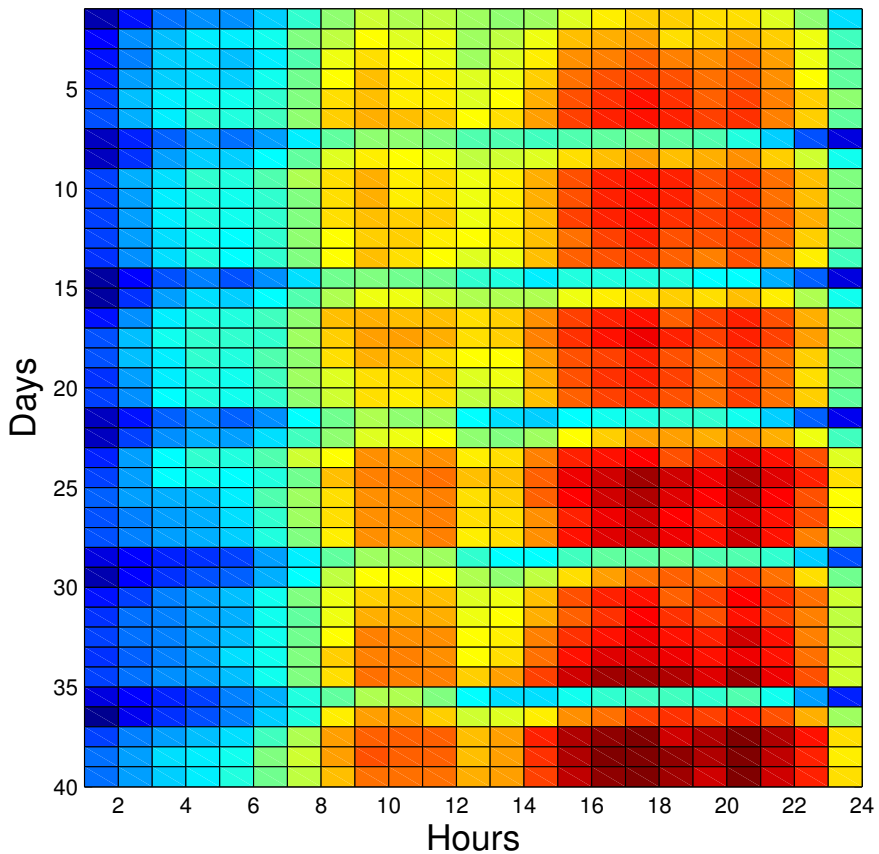
Analyst creates a model for expert to put it to operation



The autoregressive design matrix and the model



$$\underset{(m+1)\times(n+1)}{\mathbf{X}^*} =$$



$$= \left[\begin{array}{c|c} \underset{1\times 1}{\hat{s}_T} & \underset{1\times n}{\mathbf{x}_{m+1}} \\ \hline \underset{m\times 1}{\mathbf{y}} & \underset{m\times n}{\mathbf{X}} \end{array}\right].$$

In terms of regression: $\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \mathbf{w}) = \mathbf{X}\mathbf{w}$, $\hat{y}_{m+1} = \hat{s}_T = \langle \mathbf{x}_{m+1}, \hat{\mathbf{w}} \rangle$.

Model generation

Introduce a set of the primitive functions $\mathfrak{G} = \{g_1, \ldots, g_r\}$, for example $g_1 = 1$, $g_2 = \sqrt{x}$, $g_3 = x$, $g_4 = x\sqrt{x}$, etc.

The generated set of features $\mathbf{X} =$

$$\left(\begin{array}{cccc|c|ccc}
g_1 \circ s_{T-1} & \ldots & g_r \circ s_{T-1} & \ldots & g_1 \circ s_{T-\kappa+1} & \ldots & g_r \circ s_{T-\kappa+1} \\
\hline
g_1 \circ s_{(m-1)\kappa-1} & \ldots & g_r \circ s_{(m-1)\kappa-1} & \ldots & g_1 \circ s_{(m-2)\kappa+1} & \ldots & g_r \circ s_{(m-2)\kappa+1} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
g_1 \circ s_{n\kappa-1} & \ldots & g_r \circ s_{n\kappa-1} & \ldots & g_1 \circ s_{n(\kappa-1)+1} & \ldots & g_r \circ s_{n(\kappa-1)+1} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
g_1 \circ s_{\kappa-1} & \ldots & g_r \circ s_{\kappa-1} & \ldots & g_1 \circ s_1 & \ldots & g_r \circ s_1
\end{array}\right).$$

Kolmogorov-Gabor polynomial as a variant for model generation

$$y = w_0 + \sum_{i=1}^{UV} w_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} x_i x_j + \cdots + \sum_{i=1}^{n} \ldots \sum_{z=1}^{n} w_{i\ldots z} x_i \ldots x_z,$$

where the coefficients

$$\mathbf{w} = \left(w_0, w_i, w_{ij}, \ldots, w_{i\ldots z}\right)_{i,j,\ldots,z=1,\ldots,n}.$$

Examples of nonparametric transformation functions

- Univariate

Formula	Output dimension
$\sqrt{x}$	1
$x\sqrt{x}$	1
$\arctan x$	1
$\ln x$	1
$x \ln x$	1

- Bivariate

Plus	$x_1 + x_2$
Minus	$x_1 - x_2$
Product	$x_1 \cdot x_2$
Division	$\frac{x_1}{x_2}$
	$x_1\sqrt{x_2}$
	$x_1 \ln x_2$

Nonparametric aggregation: sample statistics

Nonparametric transformations include basic data statistics:

- Sum or average value of each row $\mathbf{x}_i$, $i = 1, \ldots, m$:

$$\phi_i = \sum_{j=1}^{n} x_{ij}, \text{ or } \phi_i' = \frac{1}{n} \sum_{j=1}^{n} x_{ij}.$$

- Min and max values: $\phi_i = \min_j x_{ij}$, $\phi_i' = \max_j x_{ij}$.
- Standard deviation:

$$\phi_i = \frac{1}{n-1} \sqrt{\sum_{j=1}^{n} (x_{ij} - \text{mean}(\mathbf{x}_i))^2}.$$

- Data quantiles: $\phi_i = [X_1, \ldots, X_K]$, where

$$\sum_{j=1}^{n} [X_{k-1} < x_{ij} \leq X_k] = \frac{1}{K}, \text{ for } k = 1, \ldots, K.$$

Nonparametric transformations: Haar's transform

Applying Haar's transform produces multiscale representations of the same data.

Assume that $n = 2^K$ and init $\phi_{i,j}^{(0)} = \phi_{i,j}'^{(0)} = x_{ij}$ for $j = 1, \ldots, n$.

To obtain coarse-graining and fine-graining of the input feature vector $\mathbf{x}_i$, for $k = 1, \ldots, K$ repeat:

- data averaging step

$$\phi_{i,j}^{(k)} = \frac{\phi_{i,2j-1}^{(k-1)} + \phi_{i,2j}^{(k-1)}}{2}, \quad j = 1, \ldots, \frac{n}{2^k},$$

- and data differencing step

$$\phi_{i,j}'^{(k)} = \frac{\phi_{i,2j}'^{(k-1)} - \phi_{i,2j-1}'^{(k-1)}}{2}, \quad j = 1, \ldots, \frac{n}{2^k}.$$

The resulting multiscale feature vectors are $\phi_i = [\phi_i^{(1)}, \ldots, \phi_i^{(K)}]$ and $\phi_i' = [\phi_i'^{(1)}, \ldots, \phi_i'^{(K)}]$.

Examples of parametric transformation functions

Function name	Formula	Output dim.	Num. of args	Num. of pars
Add constant	$x + w$	1	1	1
Quadratic	$w_2 x^2 + w_1 x + w_0$	1	1	3
Cubic	$w_3 x^3 + w_2 x^2 + w_1 x + w_0$	1	1	4
Logarithmic sigmoid	$1/(w_0 + \exp(-w_1 x))$	1	1	2
Exponent	$\exp x$	1	1	0
Normal	$\frac{1}{w_1\sqrt{2\pi}} \exp\left(\frac{(x-w_2)^2}{2w_1^2}\right)$	1	1	2
Multiply by constant	$x \cdot w$	1	1	1
Monomial	$w_1 x^{w_2}$	1	1	2
Weibull-2	$w_1 w_2 x^{w_2-1} \exp -w_1 x^{w_2}$	1	1	2
Weibull-3	$w_1 w_2 x^{w_2-1} \exp -w_1 (x - w_3)^{w_2}$	1	1	3
...	...	...	...	...

Ill-conditioned matrix, or curse of dimensionality

Assume we have hourly data on price/consumption for three years. Then the matrix $\underset{(m+1)\times(n+1)}{\mathbf{X}^*}$ is

$$156 \times 168, \text{ in details: } 52\text{w} \cdot 3\text{y} \times 24\text{h} \cdot 7\text{d};$$

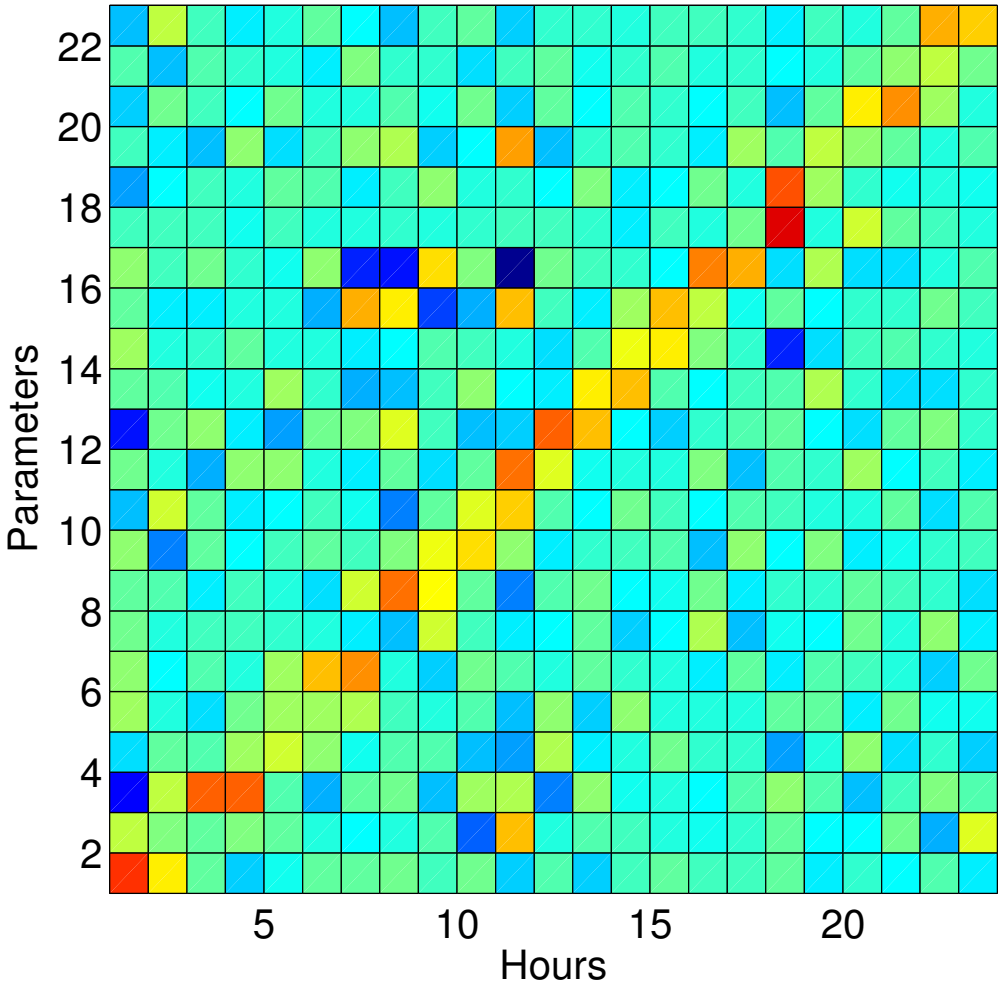
- for 6 time series the matrix $\mathbf{X}$ is 156×1008,
- for 4 primitive functions it is 156×4032,

$$m << n.$$

The autoregressive matrix could be considered as *ill-conditioned* and *multi-correlated*. The model selection procedure is required.

How many parameters must be used to forecast?

The color shows the value of a parameter for each hour.



Estimate parameters $\mathbf{w}(\tau) = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$, then calculate the sample $s(\tau) = \mathbf{w}^\mathsf{T}(\tau)\mathbf{x}_{m+1}$ for each τ of the next $(m+1\text{-th})$ period.

произвольный вектор $\mathbf{Xv}$, ортогональный вектору регрессионных остатков $\mathbf{Xw} - \mathbf{y}$:

$$(\mathbf{Xv})^\intercal(\mathbf{Xw} - \mathbf{y}) = \mathbf{v}^\intercal(\mathbf{X}^\intercal\mathbf{Xw} - \mathbf{X}^\intercal\mathbf{y}) = 0.$$

Так как это равенство должно быть справедливо для произвольного вектора $\mathbf{v}$, то $\mathbf{X}^\intercal\mathbf{Xw} - \mathbf{X}^\intercal\mathbf{y} = 0$, см. рис. 22. Если столбцы матрицы $\mathbf{X}$ линейно независимы, то матрица $\mathbf{X}^\intercal\mathbf{X}$ обратима и уравнение имеет единственное решение относительно параметров

$$\mathbf{w} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}. \tag{38}$$

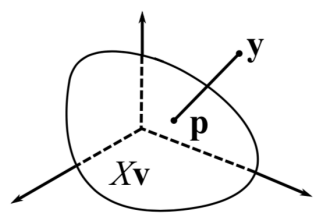
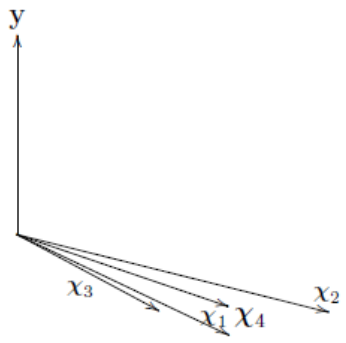


Рис. 6. Проекция вектора зависимой переменной на пространство столбцов матрицы плана.

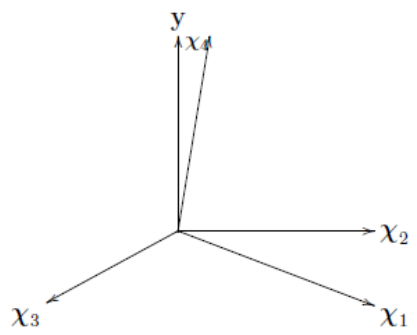
Проекция вектора $\mathbf{y}$ на пространство столбцов матрицы $\mathbf{X}$ имеет вид

$$\mathbf{p} = \mathbf{Xw} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} = \mathbf{Py}.$$

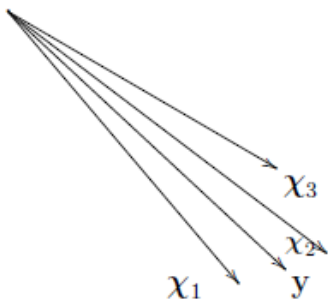
Multicollinear features to forecast: possible configurations



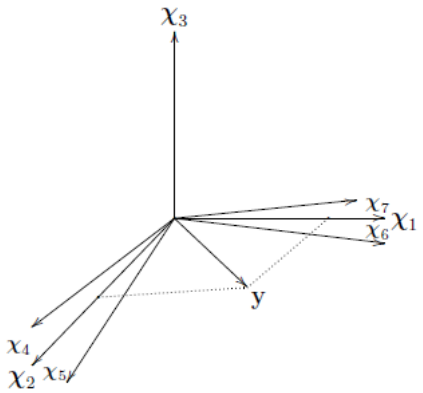
Inadequate and correlated



Adequate and random



Adequate and redundant

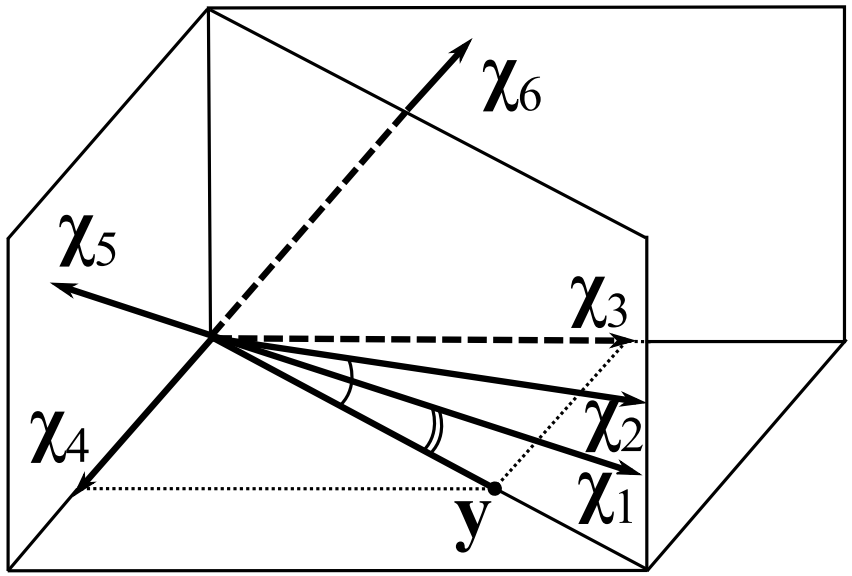


Adequate and correlated

Katrutsa A.M., Strijov V.V. Stresstest procedure for feature selection algorithms // Chemometrics and Intelligent Laboratory Systems, 2015, 142 : 172-183.

Selection of a stable set of features of restricted size

The sample contains multicollinear χ_1, χ_2 and noisy χ_5, χ_6 features, columns of the design matrix $\mathbf{X}$. We want to select two features from six.



Stability and accuracy for a fixed complexity

The solution: $\chi_3, \chi_4$is an orthogonal set of features minimizing the error function.

Add-del, Exhaustive, Genetic, FOS, LARS, Stepwise, Stagewise, Ridge, Lasso, QPFS

Exhaustive search algorithm

The basic linear model includes all independent variables

$$y = w_0 + \alpha_1 w_1 x_1 + \alpha_2 w_2 x_2 + \ldots + \alpha_n w_n x_n.$$

structure

The hyperparameter $\alpha \in \{0, 1\}$ is included for the model. The exhaustive search

α_1	α_2	$\ldots$	α_n
1	0	$\ldots$	0
0	1	$\ldots$	0
$\ldots$	$\ldots$	$\ldots$	$\ldots$
1	1	$\ldots$	1

$A \subseteq Y = \{1, \ldots, n\}$

Denote by vector α the indicator function.

$$S(w|D_\alpha) = \|y - f_\alpha\|_2^2 \to \min_\alpha$$

Add (append a feature)

Step 0.
The active set $\mathcal{A}_0 = \emptyset$, and $\mathcal{W}$ is the set if feature indices, $P = |\mathcal{W}|$.
Step $k = 1, \ldots,$
Select the next best feature index

$$\hat{j} = \arg \min_{j \in P \setminus \mathcal{A}_k} \min_{\mathbf{w} \in \mathbb{W}_k} \|(X_{\mathcal{A}_k} \vdots \mathbf{x}_j)\mathbf{w} - \mathbf{y}\|_2^2,$$

then

$$\mathcal{A}_{k+1} = \mathcal{A}_k \cup \hat{j}.$$

Assume the following

The column vectors

$$\mathbf{x}^j = \{x_i^j | i \in 1, \ldots, \ell\} \quad \text{and} \quad \mathbf{y} = \{y_i | i \in 1, \ldots, \ell\}.$$

The model

$$\mathbf{y} = w_1 \mathbf{x}^1 + \ldots + w_P \mathbf{x}^n + \varepsilon,$$

in the other words,

$$\mathbf{y} = X\mathbf{w} + \varepsilon.$$

Assume for all $j \in N$

$$\|\mathbf{x}^j\|_1 = 0, \quad \|\mathbf{x}^j\|_2 = 1 \quad \text{and} \quad \|\mathbf{y}\|_1 = 0, \quad \|\mathbf{y}\|_2 = 1.$$

For all $j, k \in N$, $j \neq k$ the vectors $\mathbf{x}^j, \mathbf{x}^k$ are linear independent.
Then the vector of correlation coefficients

$$\mathbf{c} = X^T \mathbf{y}.$$

Fast orthogonal search

Step 0.
The residuals $\varepsilon_0 = \mathbf{0}$, the active set $\mathcal{A}_0 = \emptyset$.

Step $k = 1, \ldots, P$.

$$\mathcal{A}_k = \mathcal{A}_{k-1} \cup \hat{j},$$

where $\hat{j}$ — feature, which has maximum correlation with ε_k:

$$\hat{j} = \arg \max_{j \in \{N \setminus \mathcal{A}_k\}} \frac{\langle \mathbf{w}, \mathbf{x}^j \rangle}{\|\mathbf{x}\| \|\varepsilon_k\|},$$

and

$$\varepsilon_k = X_{\mathcal{A}} \mathbf{w}_{\mathcal{A}} - \varepsilon_{k-1}.$$

Least angle regression, LARS

Denote $\mu = X\mathbf{w}$.

Step 0.
$\mu_0 = \mathbf{0}$, residual vector $\varepsilon_0 = \mathbf{y} - \mu_0$.

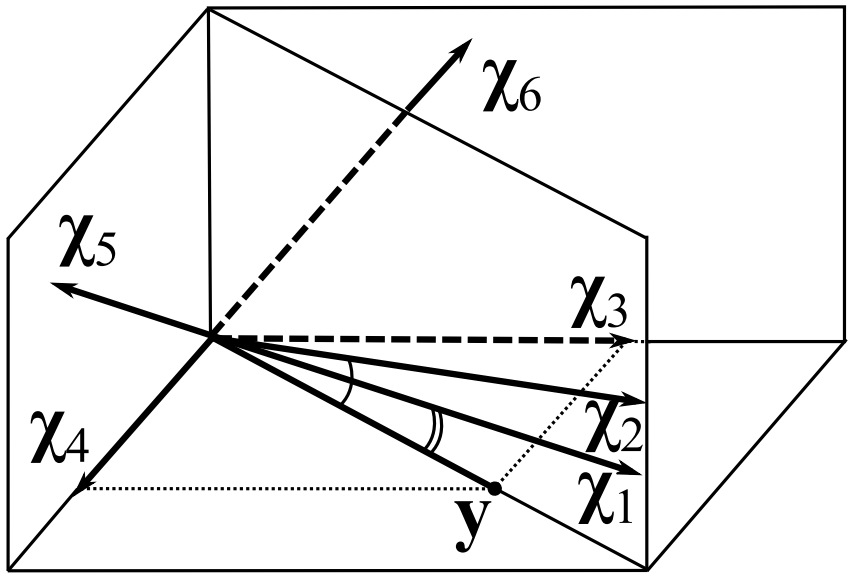
Step 1.
Let $\mathbf{y}$ has greater correlation with $\mathbf{x}^1$ than with $\mathbf{x}^2$. Then the new value of $\mu_1 = \mu_0 + w_1\mathbf{x}^1$, where w_1 is chosen so, that the vector $\mathbf{y}_2 - \mu$ — is a bisector for the vectors $\mathbf{x}^1, \mathbf{x}^2$.

Step 2.
For the unit bisector $\mathbf{u}_2$ calculate w_2:

$$\mu_2 = \mu_1 + w_2\mathbf{u}_2 = \mathbf{y}_2 \quad \text{for } n=2.$$

Selection of a stable set of features of restricted size

The sample contains multicollinear χ_1, χ_2 and noisy χ_5, χ_6 features, columns of the design matrix $\mathbf{X}$. We want to select two features from six.



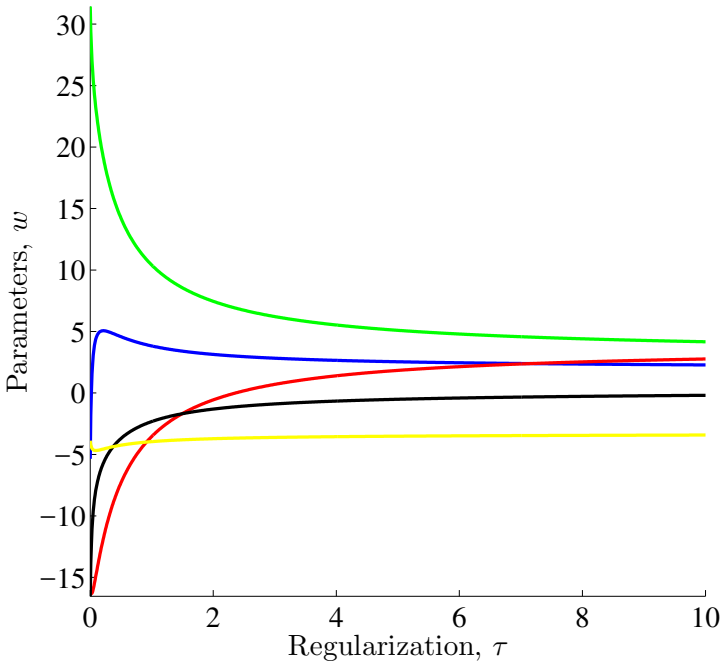
Stability and accuracy for a fixed complexity

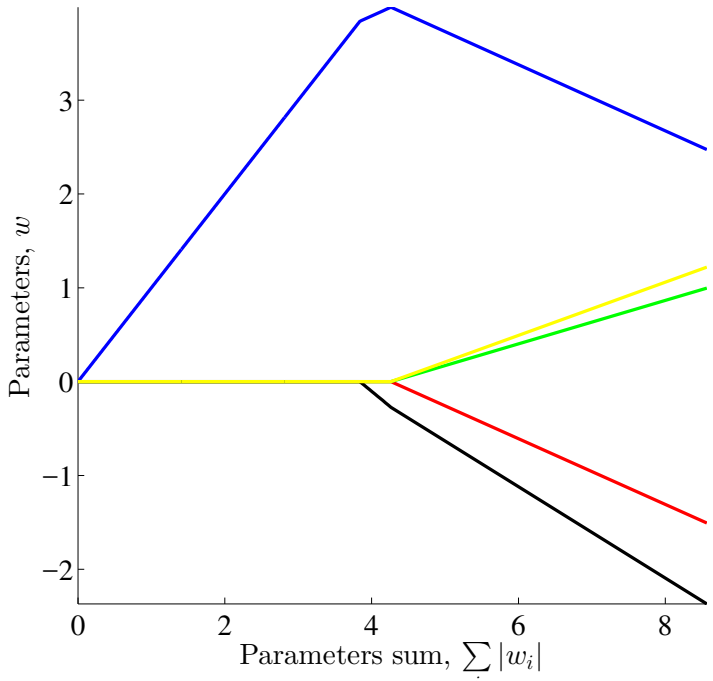
The solution: $\chi_3, \chi_4$is an orthogonal set of features minimizing the error function.

Add-del, Exhaustive, Genetic, FOS, LARS, Stepwise, Stagewise, Ridge, Lasso, QPFS

Model parameter values with regularization

Vector-function $\mathbf{f} = \mathbf{f}(\mathbf{w}, \mathbf{X}) = [f(\mathbf{w}, \mathbf{x}_1), \ldots, f(\mathbf{w}, \mathbf{x}_m)]^\mathsf{T} \in \mathbb{Y}^m$.





$$S(\mathbf{w}) = \|\mathbf{f}(\mathbf{w}, \mathbf{X}) - \mathbf{y}\|^2 + \gamma^2 \|\mathbf{w}\|^2$$

$$S(\mathbf{w}) = \|\mathbf{f}(\mathbf{w}, \mathbf{X}) - \mathbf{y}\|^2, \quad T(\mathbf{w}) \leqslant \tau$$

$+\delta^2|w|$ elastic

Minimize number of similar and maximize number of relevant features

Introduce a feature selection method QP(Sim, Rel) to solve the optimization problem

$$\mathbf{a}^* = \arg\min_{\mathbf{a}\in\mathbb{B}^n} \mathbf{a}^\top \mathbf{Q}\mathbf{a} - \mathbf{b}^\top \mathbf{a},$$

where matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ of pairwise similarities of features χ_i and χ_j is

$$\mathbf{Q} = [q_{ij}] = \mathrm{Sim}(\chi_i, \chi_j) = \left| \frac{\mathrm{Cov}(\chi_i, \chi_j)}{\sqrt{\mathrm{Var}(\chi_i)\mathrm{Var}(\chi_j)}} \right|$$

and vector $\mathbf{b} \in \mathbb{R}^n$ of feature relevances to the target is

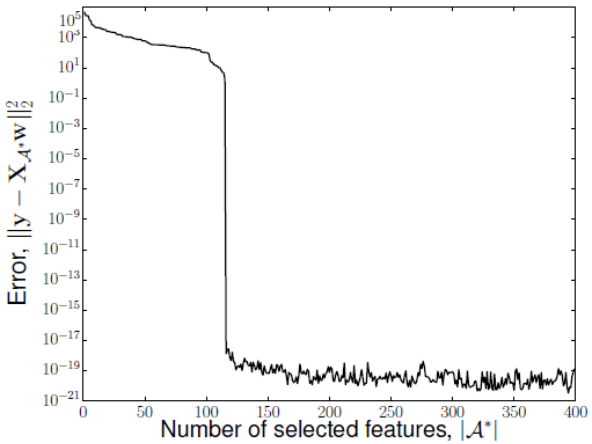
$$\mathbf{b} = [b_i] = \mathrm{Rel}(\chi_i),$$

where elements b_i equal absolute values of the sample correlation coefficient between feature χ_i and the target vector $\mathbf{y}$.

Number of correlated features Sim $\to$ min, number of correlated to the target Rel $\to$ max.

Evaluation criteria for the NIR spectra data set

Method	C_p	RSS	ln $\frac{\lambda_1}{\lambda_n}$ SVD	VIF	BIC
QP ($\tau = 10^{-9}$)	-110	$1.37 \cdot 10^{-18}$	-25.7	$6.43 \cdot 10^{6}$	548.38
Genetic	-110.88	$7.68 \cdot 10^{-30}$	-24	$8.13 \cdot 10^{5}$	534.19
LARS	$3.22 \cdot 10^{21}$	$2.07 \cdot 10^{-7}$	-28.3	$7.94 \cdot 10^{7}$	529.47
Lasso	$2.5 \cdot 10^{28}$	1.61	-27.72	$1.03 \cdot 10^{21}$	1712.92
ElasticNet	$2.51 \cdot 10^{28}$	1.61	-27.72	$1.03 \cdot 10^{21}$	1712.92
Stepwise	$3.66 \cdot 10^{29}$	23.56	-36.78	$1.94 \cdot 10^{22}$	1919.23
Ridge	$1.59 \cdot 10^{28}$	1.02	-36.22	$1.07 \cdot 10^{22}$	$1.79 \cdot 10^{3}$



Dependence of residual norm on the number of selected features QP(Sim, Rel).

Katrutsa A.M., Strijov V.V. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria // Expert Systems with Applications, 2017, 76 : 1-11.

Discrete genetic algorithm for feature selection (simple ver.)

1. There are set of binary vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_P\}$, $\mathbf{a} \in \{0,1\}^n$;
2. get two vectors $\mathbf{a}_p, \mathbf{a}_q$, $p, q \in \{1, \ldots, P\}$;
3. chose random number $\nu \in \{1, \ldots, n-1\}$;
4. split both vectors and change their parts:

$$[a_{p,1}, \ldots, a_{p,\nu}, a_{q,\nu+1}, \ldots, a_{q,n}] \to \mathbf{a'}_p,$$

$$[a_{q,1}, \ldots, a_{q,\nu}, a_{p,\nu+1}, \ldots, a_{p,n}] \to \mathbf{a'}_q;$$

5. choose random numbers $\eta_1, \ldots, \eta_Q \in \{1, \ldots, n\}$;
6. invert positions $\eta_1, \ldots, \eta_Q$ of the vectors $\mathbf{a'}_p, \mathbf{a'}_q$;
7. repeat items 2-6 $P/2$ times;
8. evaluate the obtained models.

Repeat R times; here P, Q, R are the parameters of the algorithm and n is the number of the corresponding model features.

Discrete genetic algorithm for grouping

1. There are set of binary vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_P\}$, $\mathbf{a} \in \{1, \ldots, k\}^n$;
2. get two vectors $\mathbf{a}_p, \mathbf{a}_q$, $p, q \in \{1, \ldots, P\}$;
3. chose random number $\nu \in \{1, \ldots, n-1\}$;
4. split both vectors and change their parts:

$$[a_{p,1}, \ldots, a_{p,\nu}, a_{q,\nu+1}, \ldots, a_{q,n}] \to \mathbf{a'}_p,$$

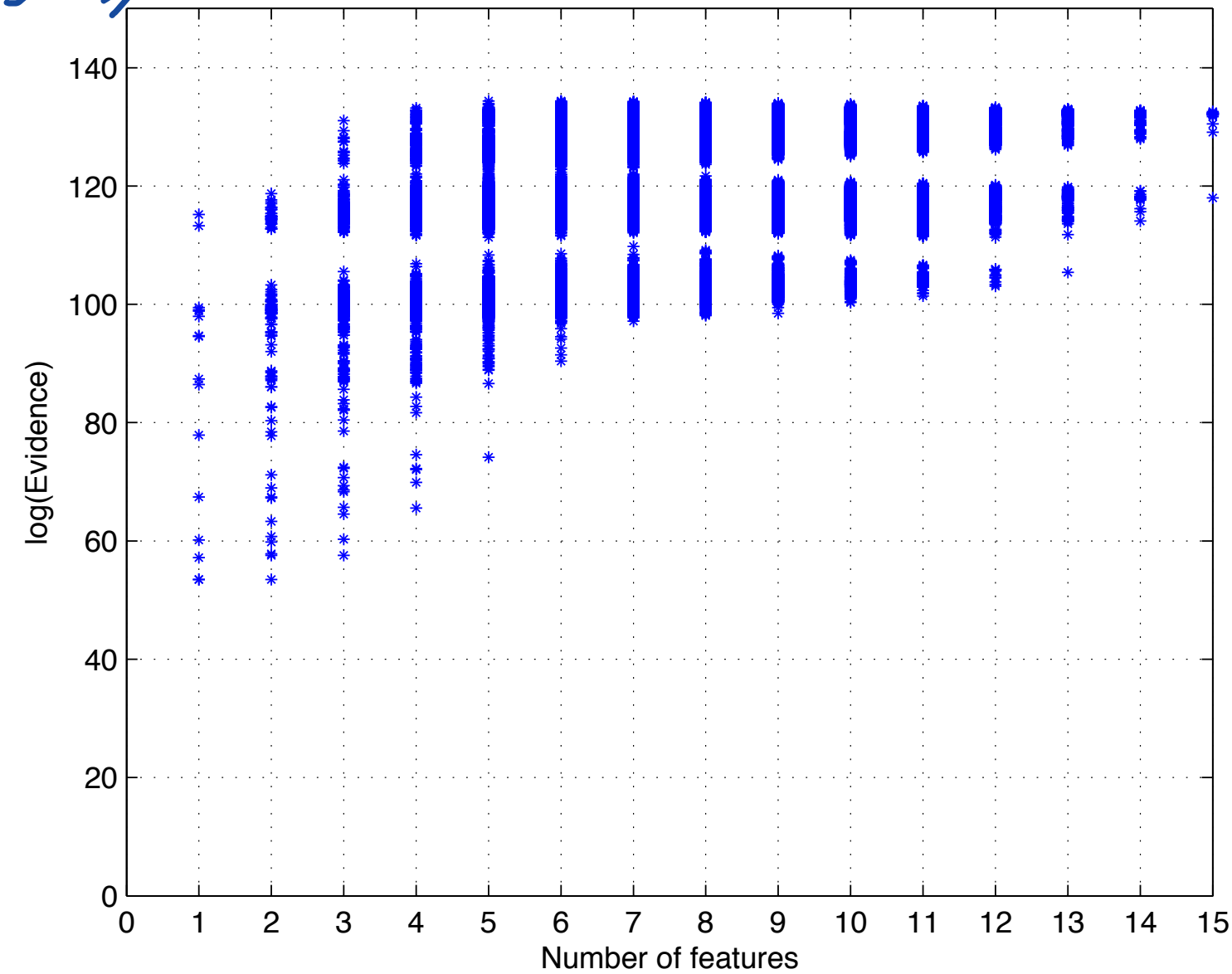
$$[a_{q,1}, \ldots, a_{q,\nu}, a_{p,\nu+1}, \ldots, a_{p,n}] \to \mathbf{a'}_q;$$

5. choose random numbers $\eta_1, \ldots, \eta_Q \in \{1, \ldots, n\}$;
6. replace values in positions $\eta_1, \ldots, \eta_Q$ of the vectors $\mathbf{a'}_p, \mathbf{a'}_q$ for random values from $\{1, \ldots, k\}$;
7. repeat items 2-6 $P/2$ times;
8. evaluate the obtained models.

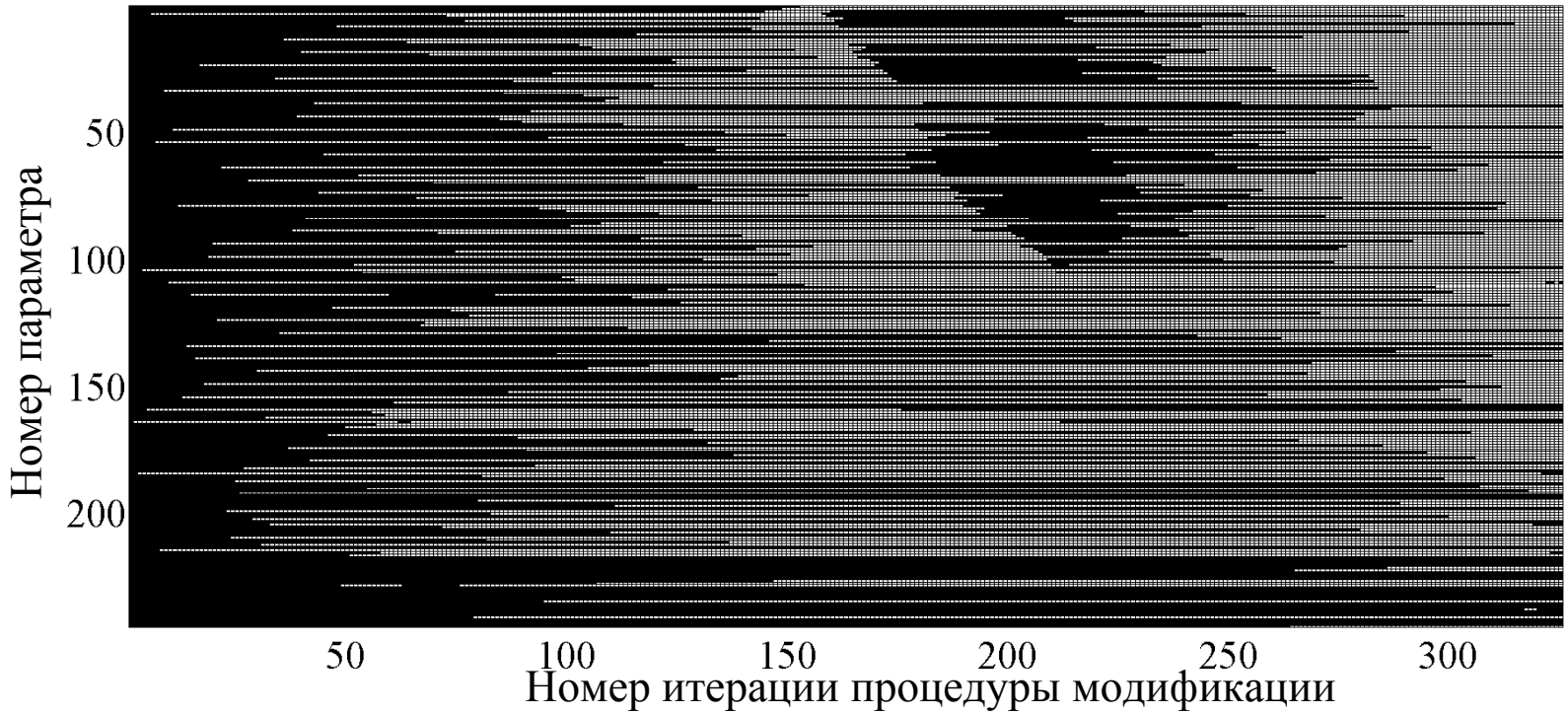
Repeat R times; here P, Q, R are the parameters of the algorithm and k is desired number of categories.

Change of likelihood at the arbitrary modification

$\exp(-S(w))$

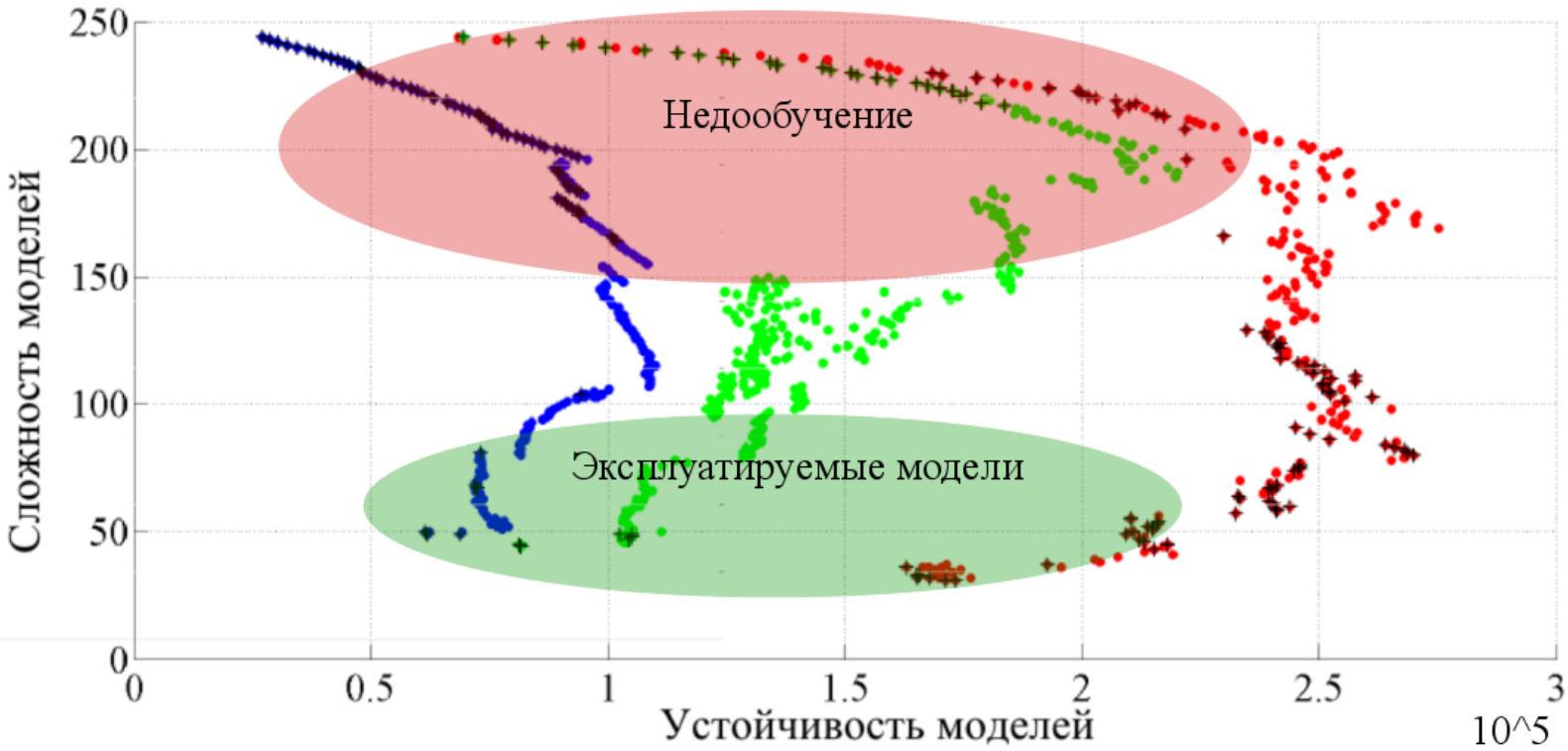


Процедура модификации как путь в кубе



- каждый столбец описывает структуру модели на данной итерации;
- белая клетка – параметр неактивен;
- черная клетка – параметр активен.

Эксперимент: результаты



Стратегия	C	S	η
Оптимальное прореж-ние	50	877	$2.2 \cdot 10^5$
Последовательное прореж-ние	50	870	$1.2 \cdot 10^5$
Устойчивое прореж-ние	50	866	$6 \cdot 10^4$