My first scientific paper

Week 7

Error analysis

Vadim Strijov

Moscow Institute of Physics and Technology

2021

Significant increase in complexity and modest increase in accuracy

	train	test	out-of-time	# parameters
Logistic regression	53,08%	55,18%	57,50%	= 12
Neural network	59,85%	57,04%	58,27%	~ 240
Regression forest	61,85%	57,01%	59,61%	> 1000
Gradient boosting	**63,58%**	**58,31%**	**59,50%**	> 10,000

Model selection is an important problem!

... it was a banking credit scoring model

Задача нахождения наиболее правдоподобных параметров

Задана выборка $D = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I}$, функция ошибки модели S и модель — параметрическое семейство функций $f(\mathbf{w}, \mathbf{x})$. Требуется найти такие параметры $\mathbf{w}$ модели, которые бы доставляли минимум функции ошибки

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{W}} S(\mathbf{w}|D, f). \tag{1}$$

Функция ошибки, определенная посредством логарифмической функции правдоподобия

$$S(\mathbf{w}) = -\ln\big(p(D|\mathbf{w}, f)\big),$$

обеспечивает максимизацию правдоподобия параметров. Параметры, найденные минимизацией этой функции ошибок, будут называться наиболее правдоподобными.

Примеры функции ошибки в регрессии и классификации

Регрессия

Гипотеза порождения данных: $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \mathbf{I})$.
Функция ошибки:

$$S(\mathbf{w}) = \|\mathbf{y} - \mathbf{f}\|_2^2.$$

Классификация

Гипотеза порождения данных: $\mathbf{y} \sim \mathcal{B}(f, 1 - f)$.
Функция ошибки:

$$S(\mathbf{w}) = \sum_{i \in \mathcal{I}} y_i \ln f(\mathbf{w}^\mathsf{T}\mathbf{x})_i + (1 - y_i) \ln\bigl(1 - f(\mathbf{w}^\mathsf{T}\mathbf{x})_i\bigr).$$

Задача выбора оптимального набора признаков

- Задана выборка $D = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I}$.
- Задано случайное разбиение множество индексов элементов выборки $\mathcal{I} = \mathcal{L} \sqcup \mathcal{C}$.
- Множество независимых переменных $\mathbf{x} = [x_1, \ldots, x_j, \ldots, x_n]$ проиндексировано $j \in \mathcal{J} = \{1, \ldots, n\}$.
- Задано множество моделей-претендентов $\mathfrak{F} = \{f(\mathbf{w}, \mathbf{x})\}$.
- Модель — параметрическое семейство функций $f(\mathbf{w}, \mathbf{x}) = \mu(\mathbf{w}^\mathsf{T}\mathbf{x})$, где μ — функция связи (в случае регрессии $\mu = \mathrm{id}$, в случае классификации $\mu = \frac{1}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x})}$.
- Структура модели $f_\mathcal{A}$ задана множеством индексов $\mathcal{A} \subseteq \mathcal{J}$ и означает включение переменных $\mathbf{x}_\mathcal{A}$. Иначе, используются только признаки-столбцы матрицы $\mathbf{X}$ с индексами из множества $\mathcal{A}$.
- Задана функция ошибки S.

Среднее значение и стандартное отклонение ошибки

Задана выборка $\mathfrak{D} = \{\mathbf{x}_i, y_i\}$. Ее элементы проиндексированы:

$$i \in \mathcal{I} = \{1, \ldots, m\}.$$

Разобьем выборку равномерно случайно, на две равномощные подвыборки, обучение и контроль, K раз:

$$\mathcal{I} \longrightarrow \mathcal{L}_k \bigsqcup \mathcal{C}_k, \qquad k \in \{1, \ldots, K\}.$$

Задана модель $f(\mathbf{w}, \mathbf{w})$ и функция ошибки $S(\mathbf{w}|\mathfrak{D})$. Параметры модели оптимизированы на обучении $\mathfrak{D}_{\mathcal{L}}$ как

$$\hat{\mathbf{w}} = \arg\min S(\mathbf{w}|f, \mathfrak{D}_{\mathcal{L}}).$$

Для каждого из K разбиений вычисляем ошибку на обучении и на контроле. Получаем два набора ошибок:

$$\{S_k(\hat{\mathbf{w}}_k|f, \mathfrak{D}_{\mathcal{L}k})\}, \qquad \{S_k(\hat{\mathbf{w}}_k|f, \mathfrak{D}_{\mathcal{C}k})\}, \qquad k \in \{1, \ldots, K\}.$$

Задача выбора оптимального набора признаков

Требуется найти такое подмножество индексов $\mathcal{A} \subseteq \mathcal{J}$, которое бы доставляло минимум функции

$$\mathcal{A}^* = \arg\min_{\mathcal{A} \subseteq \mathcal{J}} S(f_{\mathcal{A}} | \mathbf{w}^*, D_{\mathcal{C}})$$

на разбиении выборки D, определенном множеством индексов $\mathcal{C}$.

При этом параметры $\mathbf{w}^*$ модели должны доставлять минимум функции

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{W}} S(\mathbf{w} | D_{\mathcal{L}}, f_{\mathcal{A}})$$

на разбиении выборки, определенном множеством $\mathcal{L}$.

Зависимость среднего значения ошибки от объема выборки

Для двух наборов ошибок вычислим среднее значение и поправленное стандартное отклонение:

$$\bar{S} = \frac{1}{K}\sum_{k=1}^{K} S_k, \qquad \sigma = \frac{1}{K-1}\sqrt{\sum_{k=1}^{K}(\bar{S} - S_k)^2}.$$

Повторим процедуру на ограниченном объеме выборки, например:

$$m = \overline{1, M},$$

где M — наибольший объем доступной выборки.
Построим график зависимости ошибки и стандартного отклонения от объема выборки.
Домашнее задание 3 находится по адресу http://bit.ly/16UIlQH

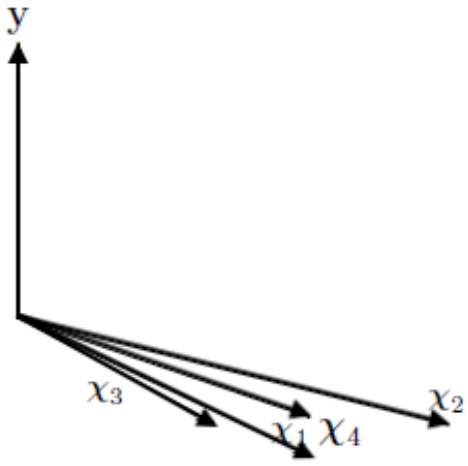
Некторые задачи машинного обучения

- Задача оценки параметров модели,
- задача выбора признаков или объектов выборки,
- задача выбора модели оптимальной сложности,
- задача построения и выбора структуры модели,
- задача проверки гипотезы порождения данных.

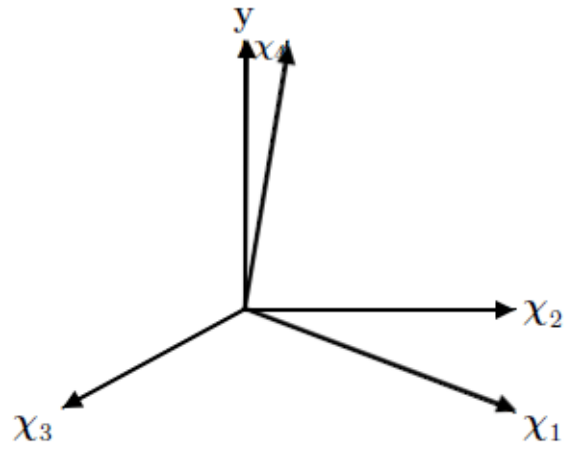
Предполагается, что функция ошибки $S(\mathbf{w}|D, f)$ задана исходя из

- гипотезы порождения данных,
- либо из практических соображений.

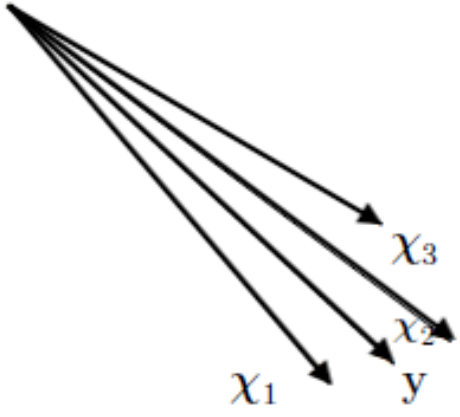
Configurations of feature space



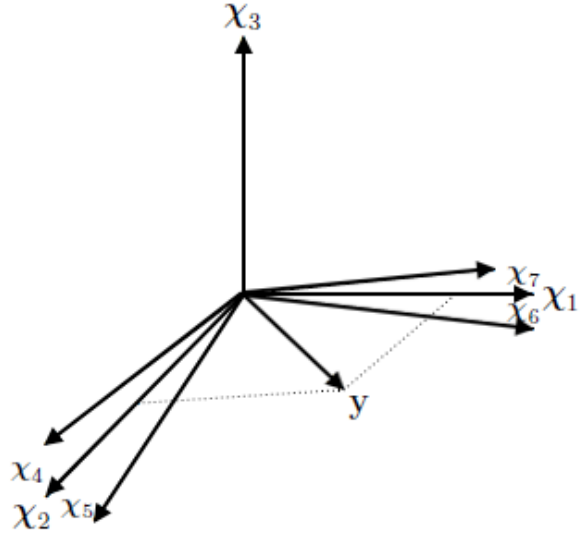
Non-adequate correlated



Adequate random



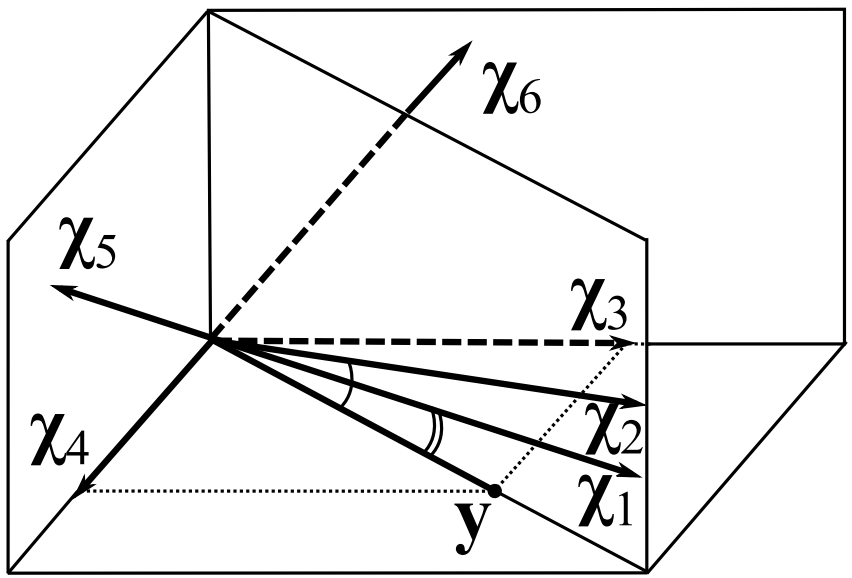
Adequate redundant



Adequate correlated

Выбор устойчивой и точной модели

Выборка содержит мультикоррелирующие χ_1, χ_2 и устойчивые χ_5, χ_6 признаки — столбцы матрицы «объект-признак» $\mathbf{X}$. Требуется выбрать два признака из шести.



Точность и устойчивость при заданной сложности

Решение: χ_3, χ_4 — набор ортогональных признаков с наименьшим значением функции ошибки.

Minimize number of similar and maximize number of relevant features

The model is defined by a ~~vertex~~ point in the n-dimensional cube.

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

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

Number of correlated features Sim $\to$ min, number of correlated to the target Rel $\to$ max, where matrix $\mathbb{Q} \in \mathbb{R}^{n\times n}$ of pairwise similarities of features χ_i and χ_j is

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

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

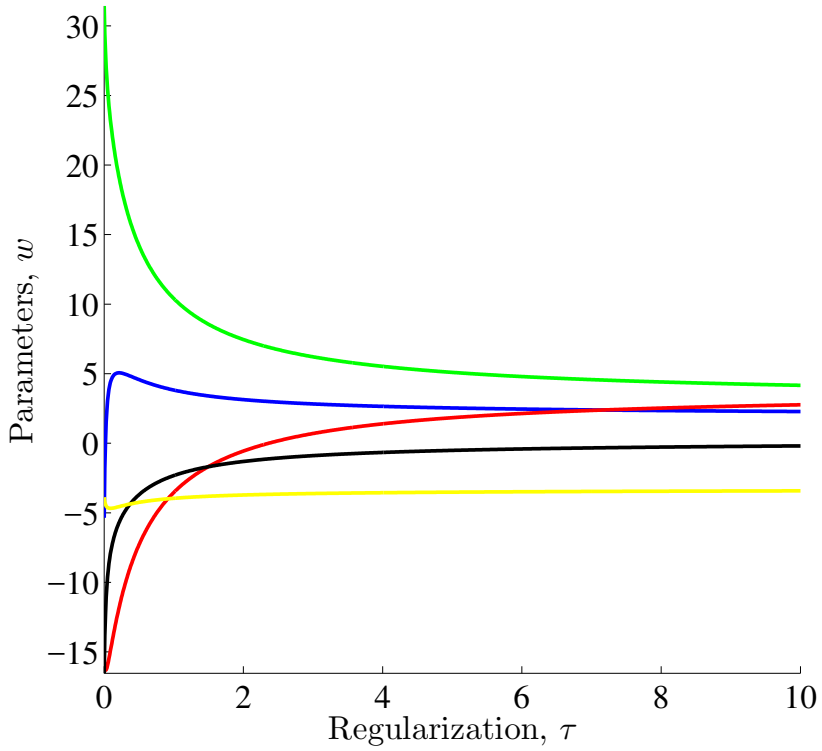
$$\mathbf{b} = [b_i] = \mathrm{Rel}(\chi_i),$$

elements $b_i = \|\mathrm{Corr}(\chi_i, \mathbf{y})\|$.

Katrutsa, Strijov. 2017. Comprehensive study of feature selection methods to solve multicollinearity problem // Expert Systems with Applications

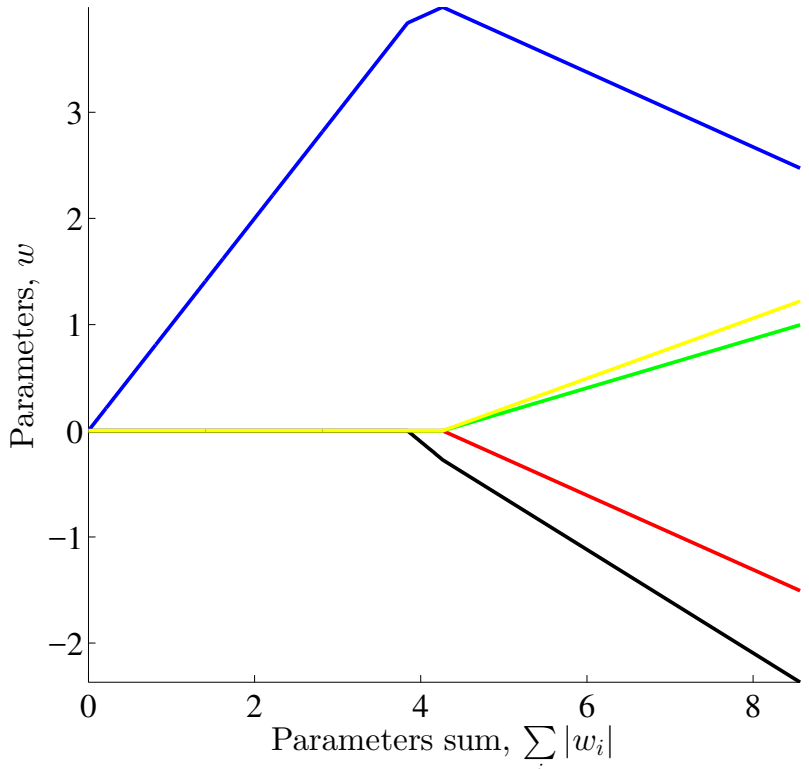
Model parameters with regularization

Ridge regression



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

Lasso



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

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:

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

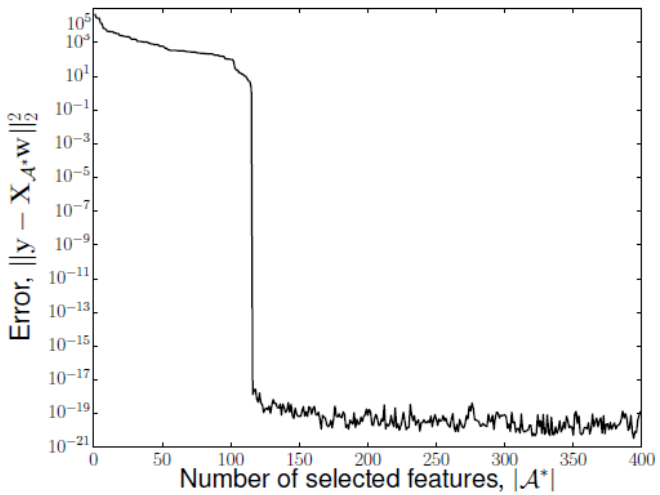
$$[a_{q,1}, \ldots, a_{q,\nu}, a_{p,\nu+1}, \ldots, a_{p,n}] \to \mathbf{a}'_q;$$

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.

Evaluation criteria for the NIR spectra data set

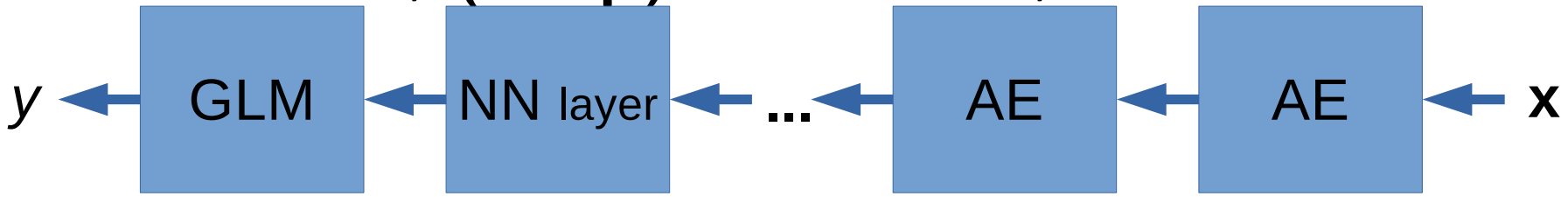
Method	C_p	RSS	$\ln \frac{\lambda_1}{\lambda_n}$ SVD	VIF	BIC
QP ($\tau = 10^{-9}$)	-110	$1.37 \cdot 10^{-18}$	-25.7	$6.43 \cdot 10^{6}$	548.38
Genetic	-110.88	$7.68 \cdot 10^{-30}$	-24	$8.13 \cdot 10^{5}$	534.19
LARS	$3.22 \cdot 10^{21}$	$2.07 \cdot 10^{-7}$	-28.3	$7.94 \cdot 10^{7}$	529.47
Lasso	$2.5 \cdot 10^{28}$	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. Stresstest procedure for feature selection algorithms // Chemometrics and Intelligent Laboratory Systems, 2015, 142 : 172-183.

Linear model, (deep) neural net, and autoencoder



$$f = \sigma_k \circ \underset{1\times 1}{\mathbf{w}_k^{\mathsf{T}}}\boldsymbol{\sigma}_{k-1} \circ \mathbf{W}_{k-1}\boldsymbol{\sigma}_{k-2} \circ \cdots \circ \underbrace{\underset{n_2\times 1}{\mathbf{W}_2}\boldsymbol{\sigma}_1 \circ \underset{n_1\times n}{\mathbf{W}_1}\underset{n\times 1}{\mathbf{x}}}_{E_{\mathbf{x}}=\sum_{\mathbf{x}_i\in\mathfrak{D}}\|\mathbf{x}_i-\mathbf{r}(\mathbf{x}_i)\|_2^2} \in \mathfrak{D}$$

$$E_D=\sum_{(\mathbf{x}_i,y_i)\in\mathfrak{D}}\Big(y_i-f(\mathbf{x}_i)\Big)^2$$

$$S = \lambda_1 E_D + \lambda_2 E_{\mathbf{x}} + \lambda_3 E_{\mathbf{w}} = \boldsymbol{\lambda}^{\mathsf{T}}\mathbf{s}$$

$E_{\mathbf{w}}$ is some regularisation error, for

principal component analysis: $\mathbf{W}^{\mathsf{T}}\mathbf{W} = \mathbf{I}_n$,

skip block: $\mathbf{W} = \mathbf{I}_n$, $\sigma = \mathrm{id}$,

classification: $\sigma \in \{\text{logistic}, \text{softmax}, \text{ReLu}, \ldots\}$.

... including LM, LR, PCA, AE, SAE, 2NN, DLL, CNN, etc.

Probabilistic model selection

Bayesian inference delivers the error function $S(\mathbf{w})$

Posterior Likelihood Prior

$$p(\mathbf{w}|\mathcal{D},\mathbf{A},\mathbf{B},\mathbf{f}) = \frac{p(\mathcal{D}|\mathbf{w},\mathbf{B},\mathbf{f})p(\mathbf{w}|\mathbf{A},\mathbf{f})}{p(\mathcal{D}|\mathbf{A},\mathbf{B},\mathbf{f})}.$$

Evidence (to select a model)

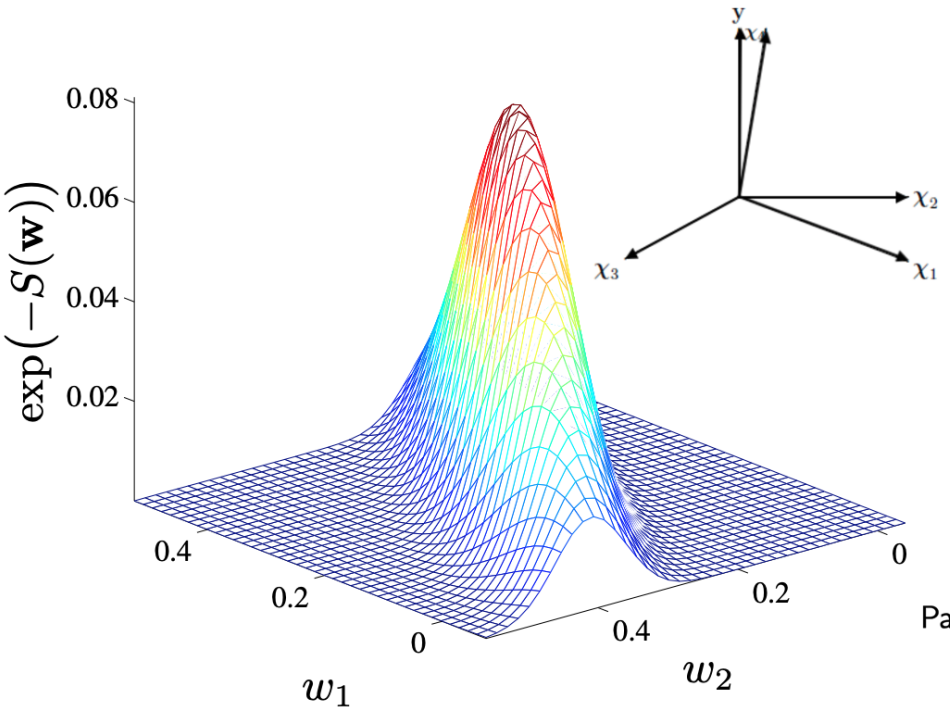
Write the error function given hyperparameters $\mathbf{A}, \mathbf{B}$

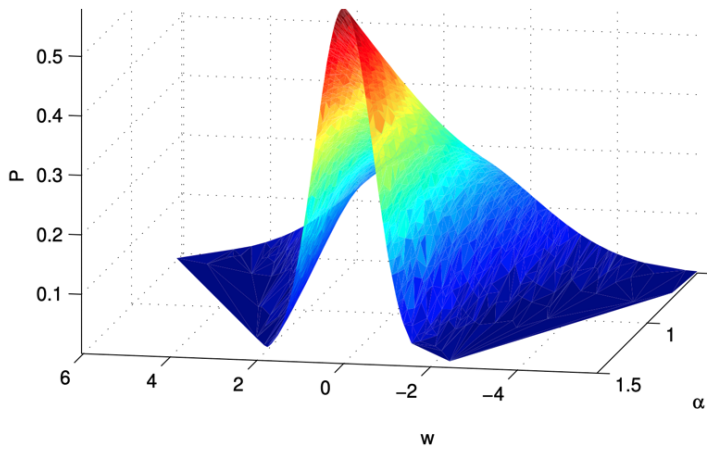
$$S(\mathbf{w}) = \underbrace{\frac{1}{2}(\mathbf{y}-\mathbf{f})^{\mathsf{T}}\mathbf{B}(\mathbf{y}-\mathbf{f})}_{\text{approximation error}} + \underbrace{\frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^{\mathsf{T}}\mathbf{A}(\mathbf{w}-\hat{\mathbf{w}})}_{\text{regularisation error}},$$

$$S = E_D + E_{\mathbf{w}} = \boldsymbol{\lambda}^{\mathsf{T}}\mathbf{s}, \qquad \text{metaparameters } \lambda = \frac{1}{2}.$$

Empirical distribution of model parameters

The value of error function $S(\mathbf{w}|\mathfrak{D}, f)$ depends on parameters.





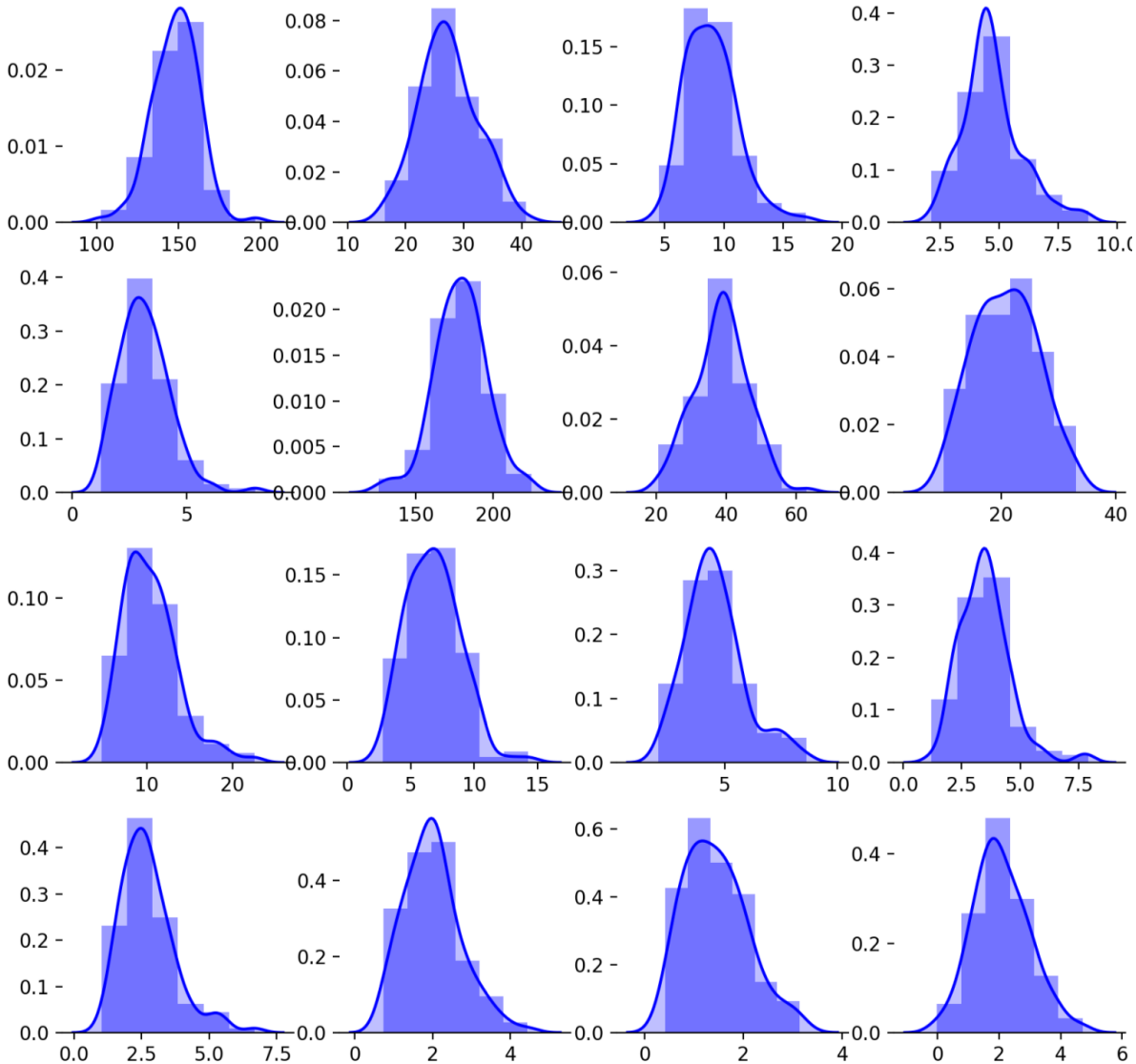
Parameters $\mathbf{w}$, variance α^{-2}, and $p(\mathbf{w}|\mathfrak{D}, \alpha)$ is the evidence.

x-axis and y-axis: parameters $\mathbf{w}$, z-axis: $\exp(-S(\mathbf{w}))$

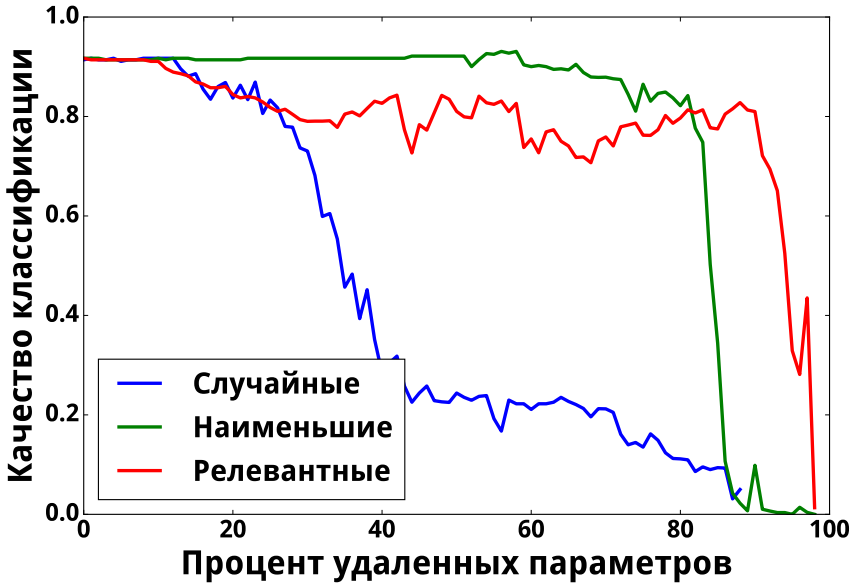
Kuznetsov, Tokmakova, Strijov. 2016. Analytic methods of structure parameter // Informatica

How to check the i.i.d hypothesis

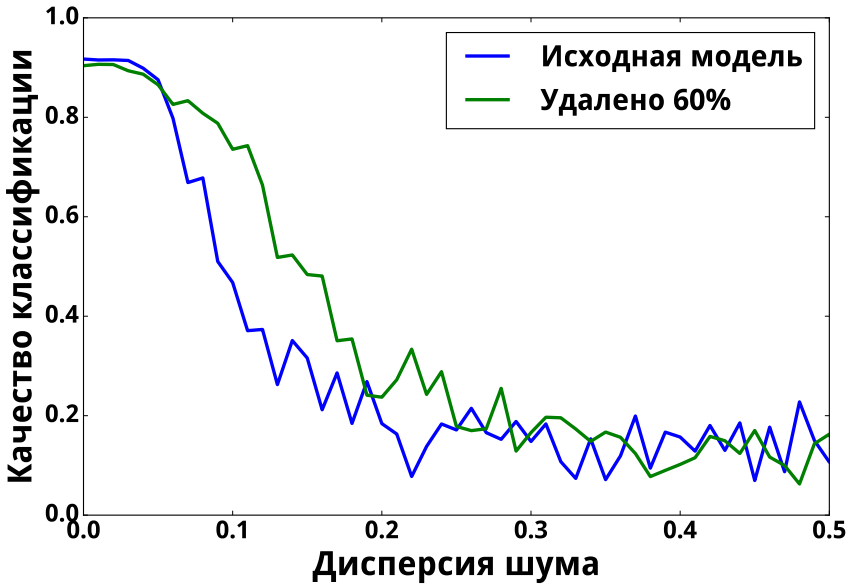
T-test) $E\varepsilon = 0, D\varepsilon = const$, as well as the spectrum analysis



Правдоподобие моделей с избыточным числом параметров не изменяется значимо при их удалении



Избыточность параметров модели



Устойчивость модели

Глубокое обучение предполагает оптимизацию моделей с заведомо избыточной сложностью.

Bakhteev, Strijov. 2019. Comprehensive analysis of gradient-based hyperparameter optimization algorithms // Annals of Operations Research

Двоичное представление структуры модели

Модель f выбирается из множества моделей-претендентов $\mathfrak{F}$ путем оптимизации двоичного вектора $\mathbf{a} \in \mathbb{B}^n$,

$$\hat{y} = f(\mathbf{w}, \mathbf{x}) = a_1 w_1 x_1 + \cdots + a_n w_n x_n$$

для линейной модели $f(\mathbf{w}, \mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{w}$
и для нейронной сети

$$\mathbf{f}(\mathbf{w}, \mathbf{x}) = \frac{\exp\big(\mathbf{h}(\mathbf{x})\big)}{\sum_k \exp(h_k(\mathbf{x}))}, \quad \mathbf{h}(\mathbf{x}) = \mathbf{W}_2^\mathsf{T}\mathbf{tanh}(\mathbf{W}_1^\mathsf{T}\mathbf{x}), \quad \mathbf{w} = \mathbf{vec}(\mathbf{W}_1 \vdots \mathbf{W}_2)$$

путем зануления соовтествующего параметра

$$w_j = 0$$

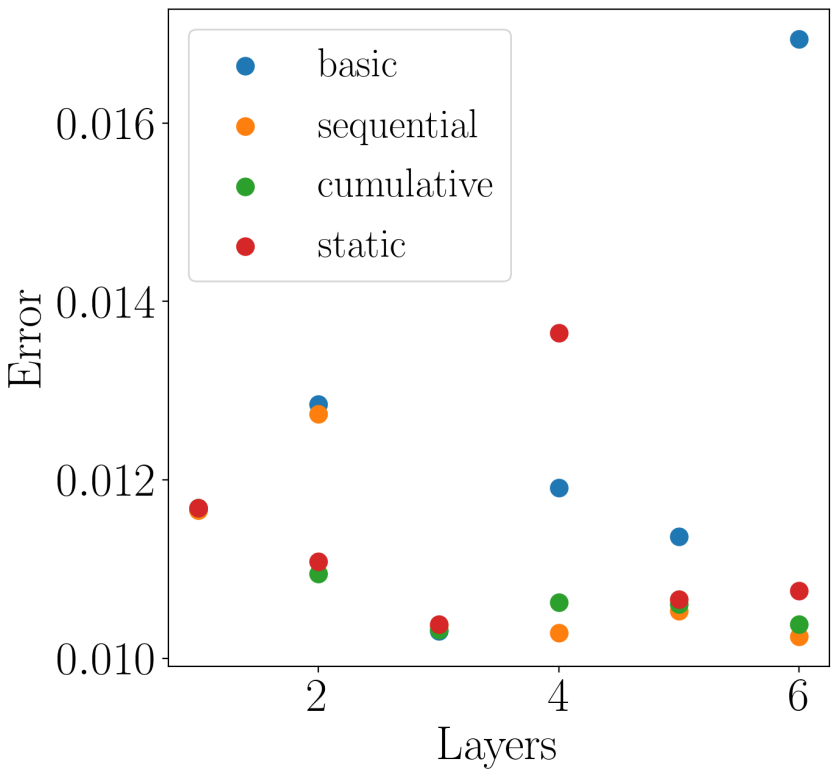
или согласно методу оптимального прореживания

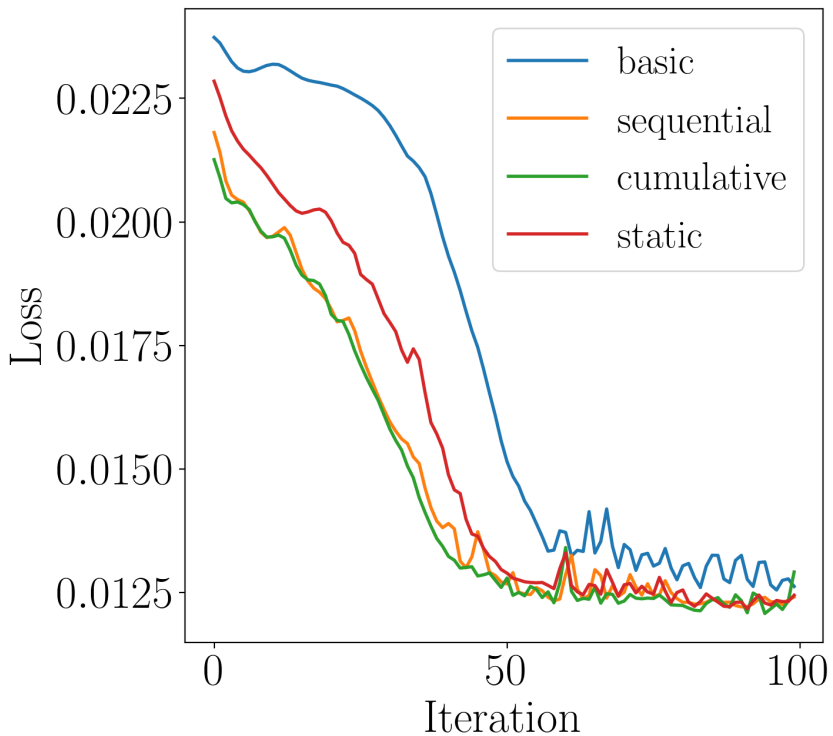
$$\mathbf{e}_j^\mathsf{T}\Delta\mathbf{w} + w_j = 0$$

где j-й элемент вектора $\mathbf{e}$ равен 1, прочие равны 0.
Модель задана вершиной двоичного n-мерного куба.

Процедура оптимизации параметров

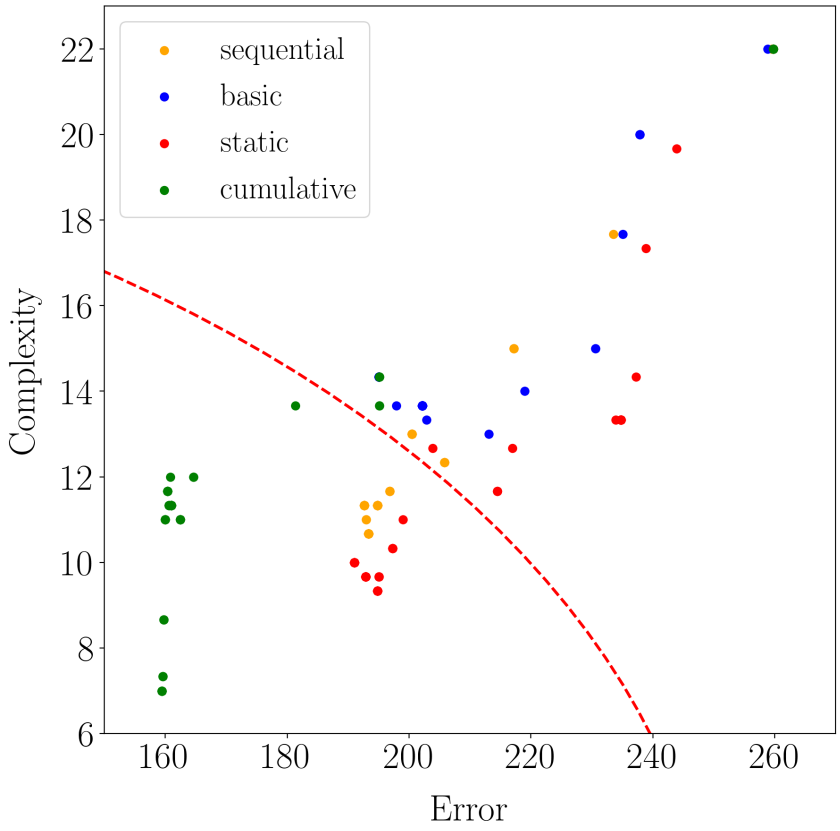
Результаты работы алгоритма выбора расписания оптимизации. Слева — зависимость ошибки нейросети на тестовой выборке от используемого метода задания весов регуляризации и количества слоев. Справа — ошибка на валидационной выборке во время обучения.

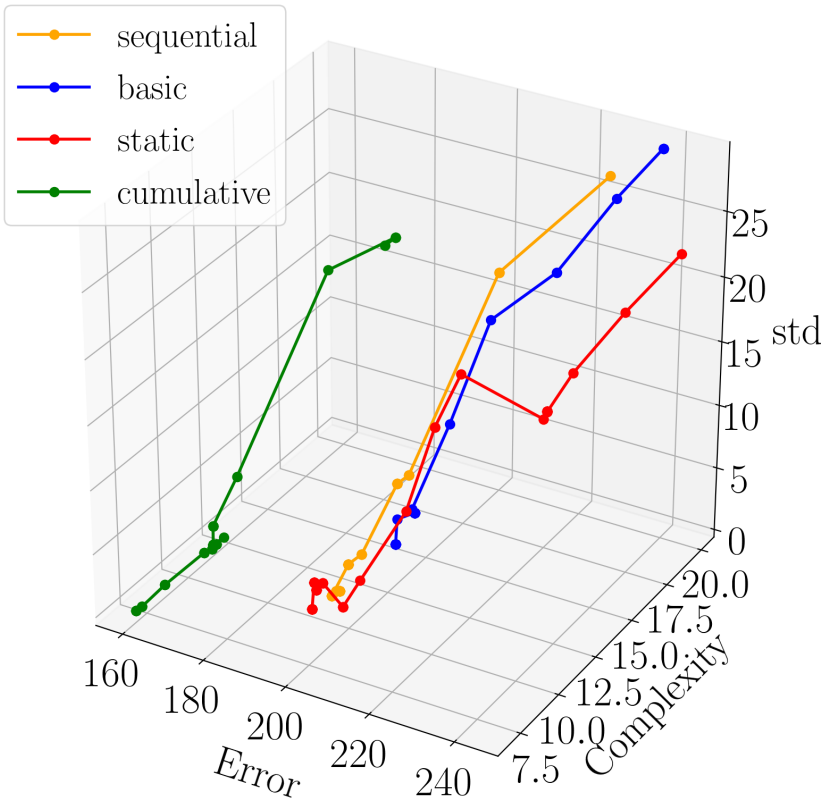




Использование регуляризации приводит к более быстрой сходимости и меньшей ошибке.

Процедура оптимизации структуры

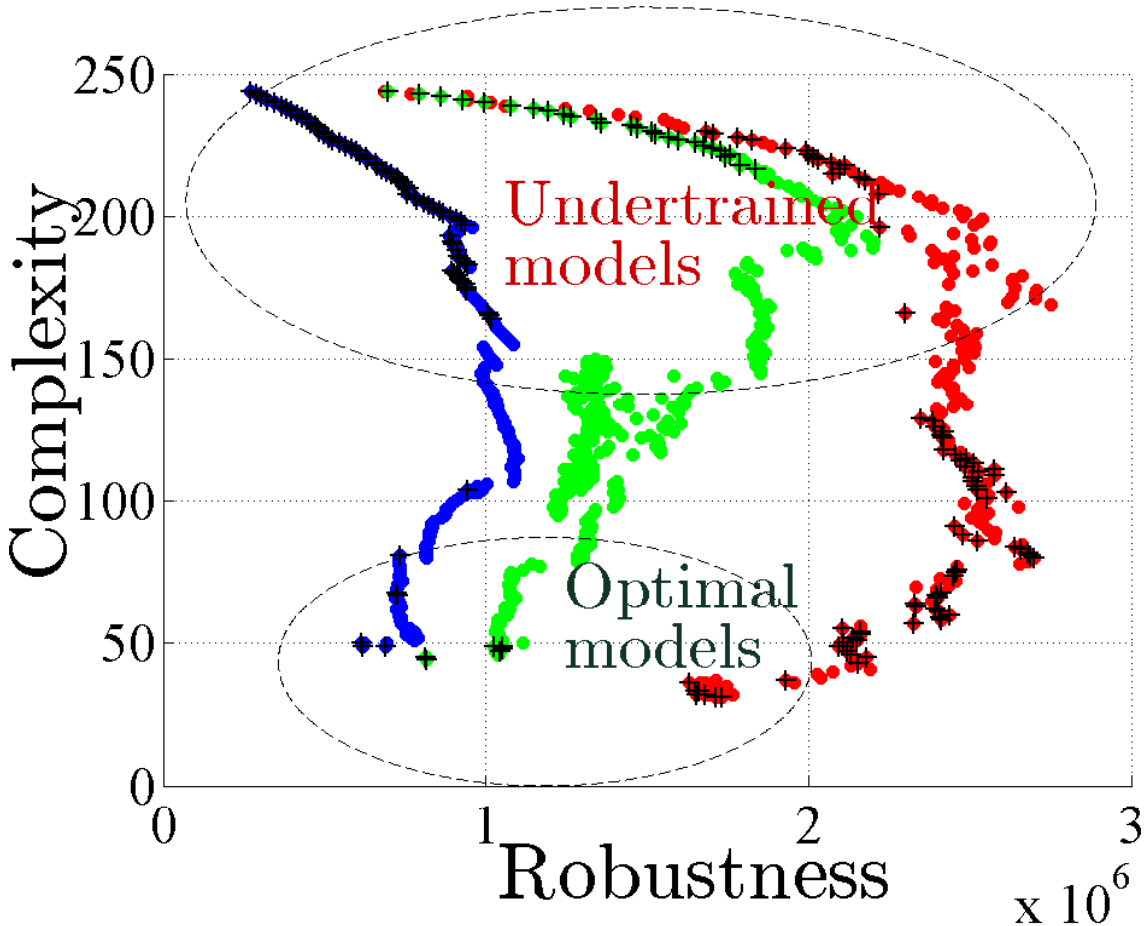




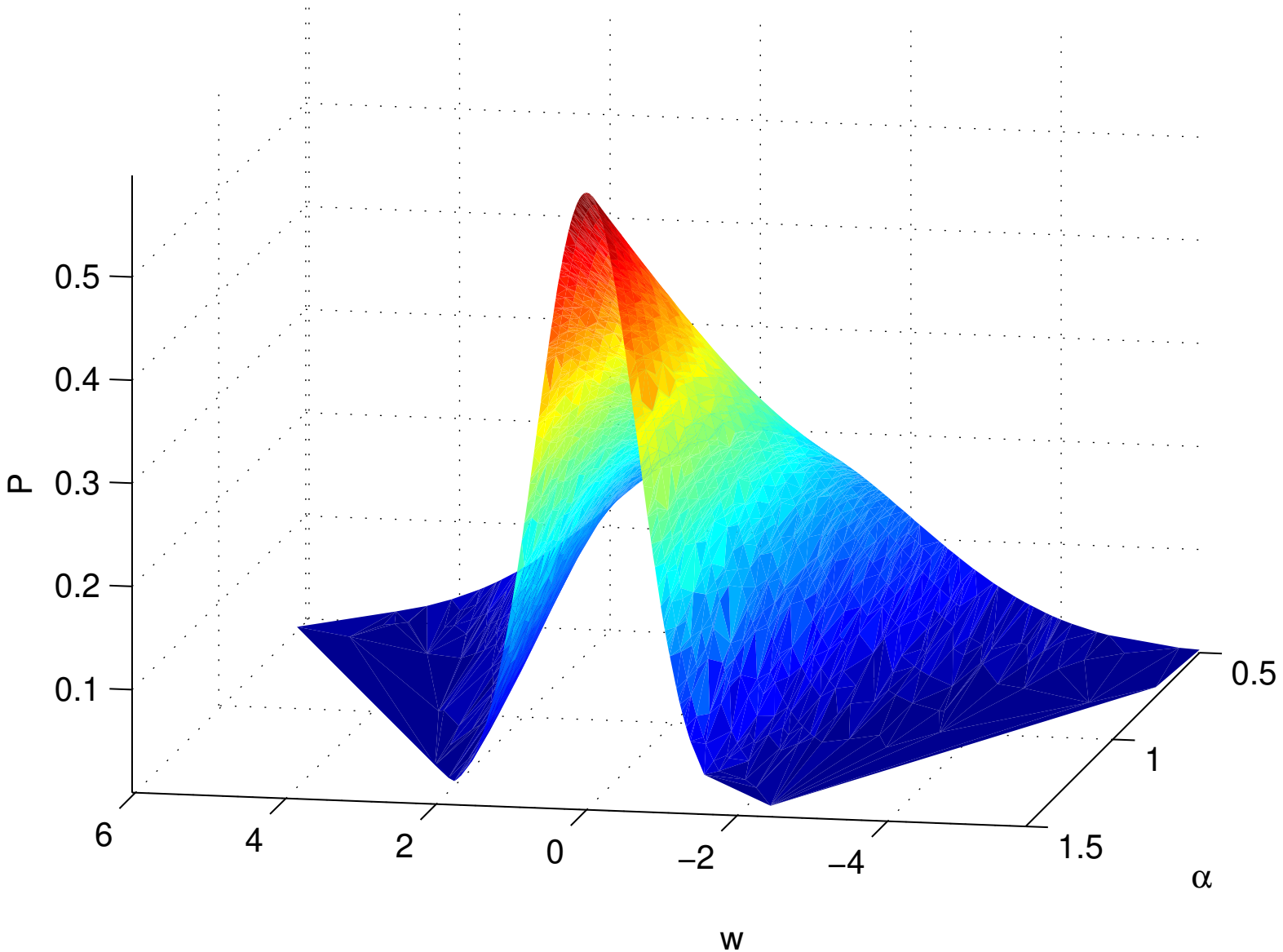
Изменение точности, сложности и устойчивости моделей при итерациях генетического алгоритма

Последовательный выбор моделей:

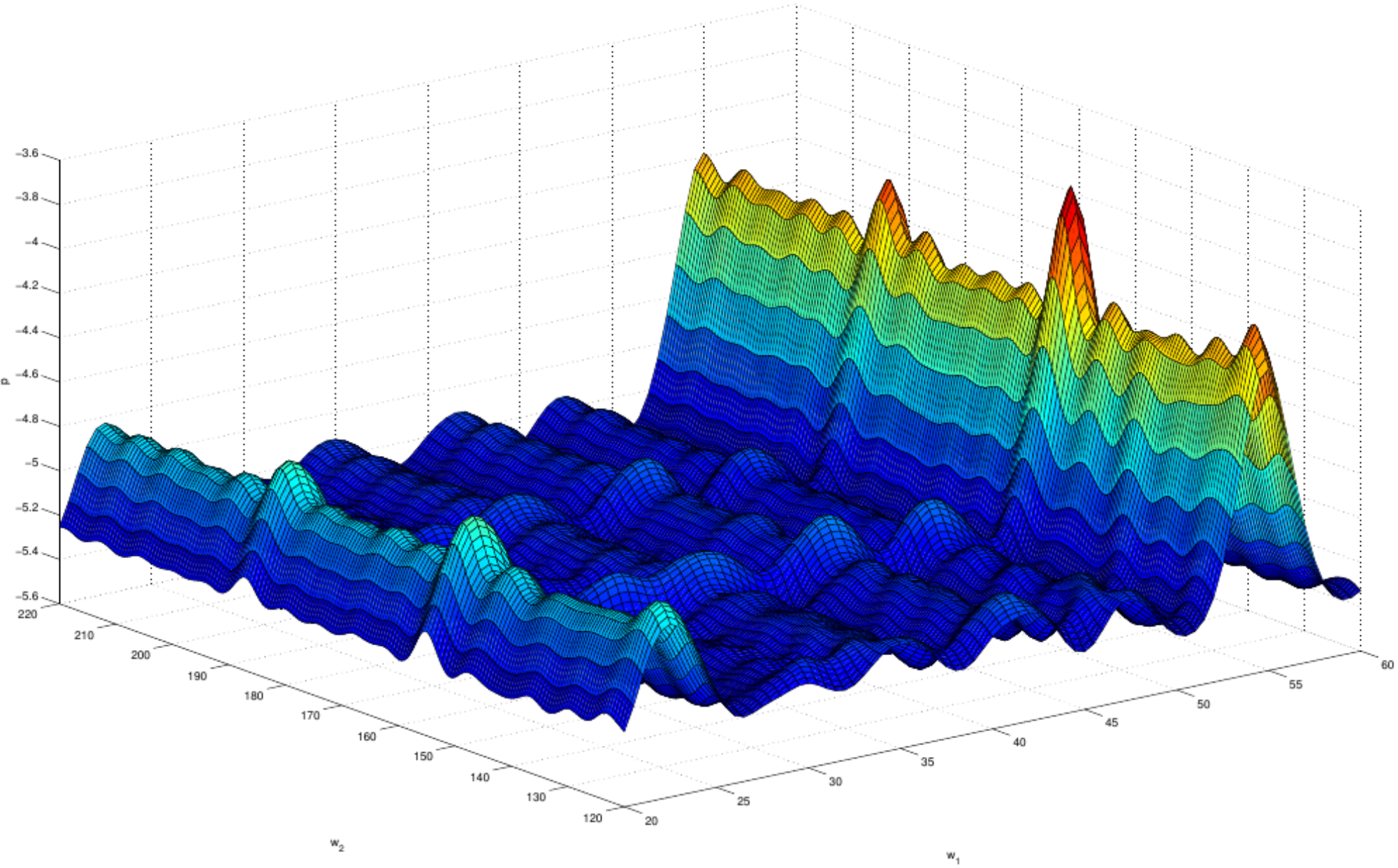
точность, сложность, устойчивость



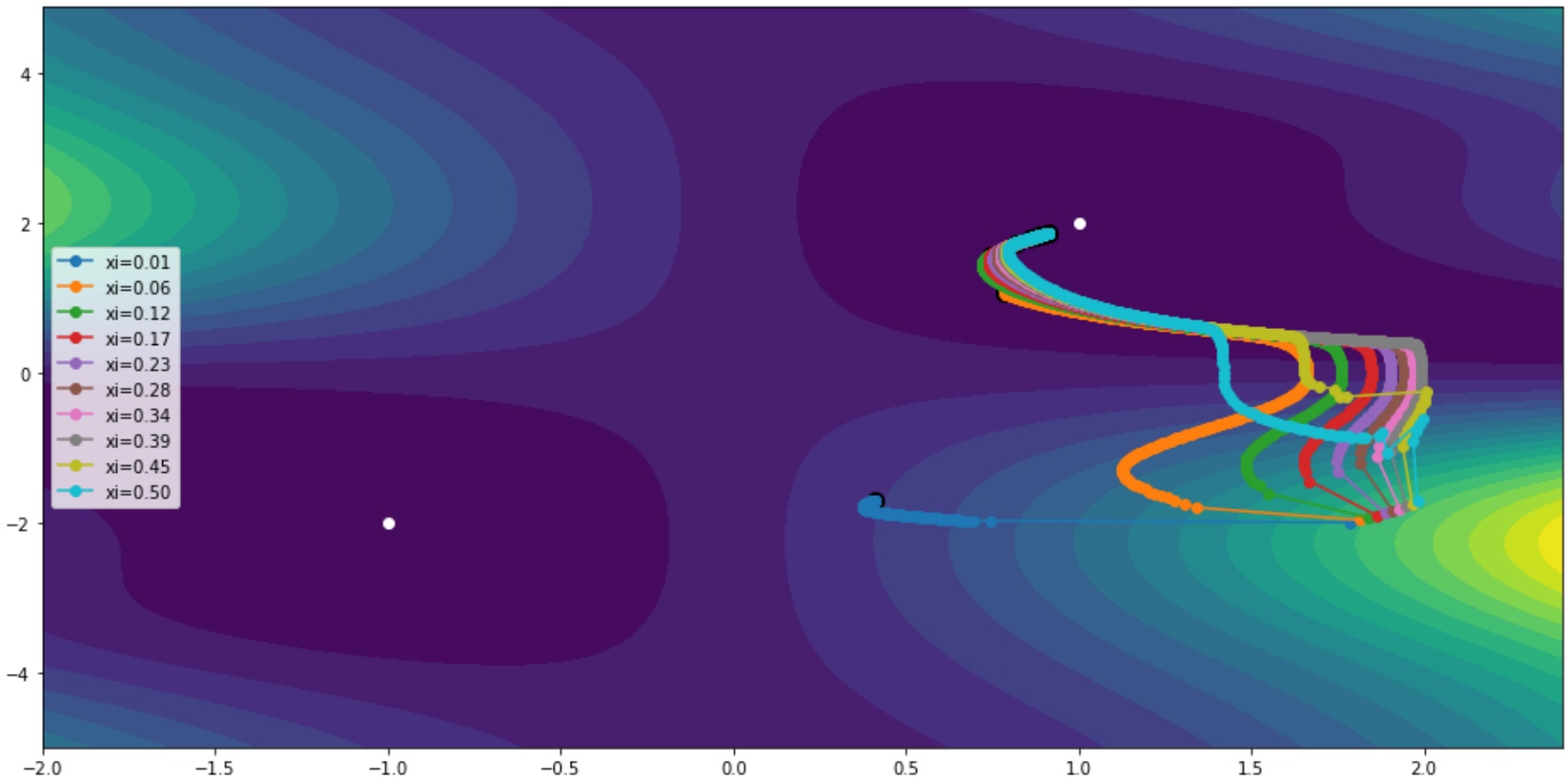
Точность или устойчивость



Многоэкстремальность функции ошибки



Многоэкстремальность, сходимость и мультистарт



$$\mathbf{w} \in \mathbb{R}^2$$

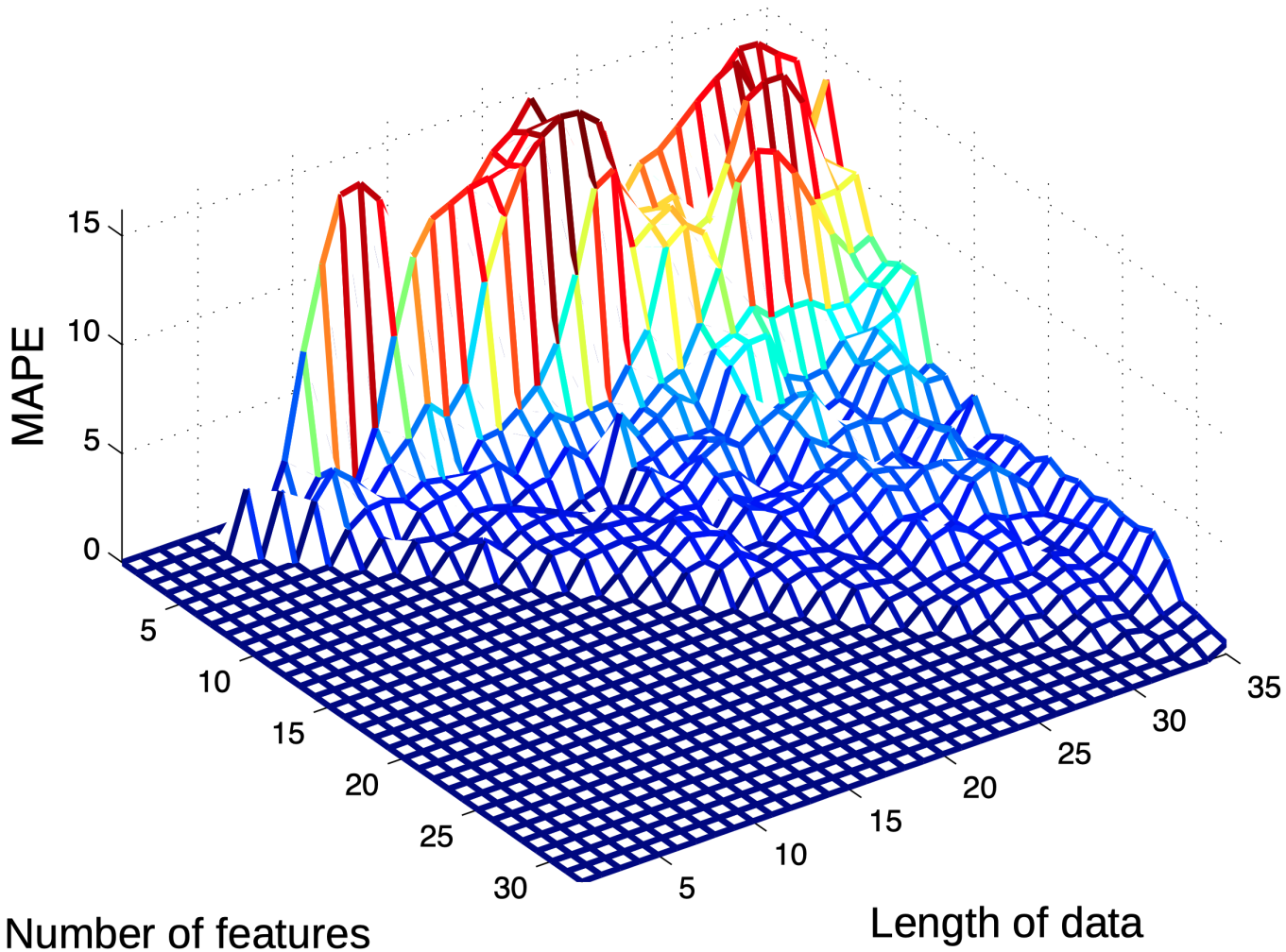
Стохастический градиент и ковариация параметров

Стохастический градиент и ковариация параметров

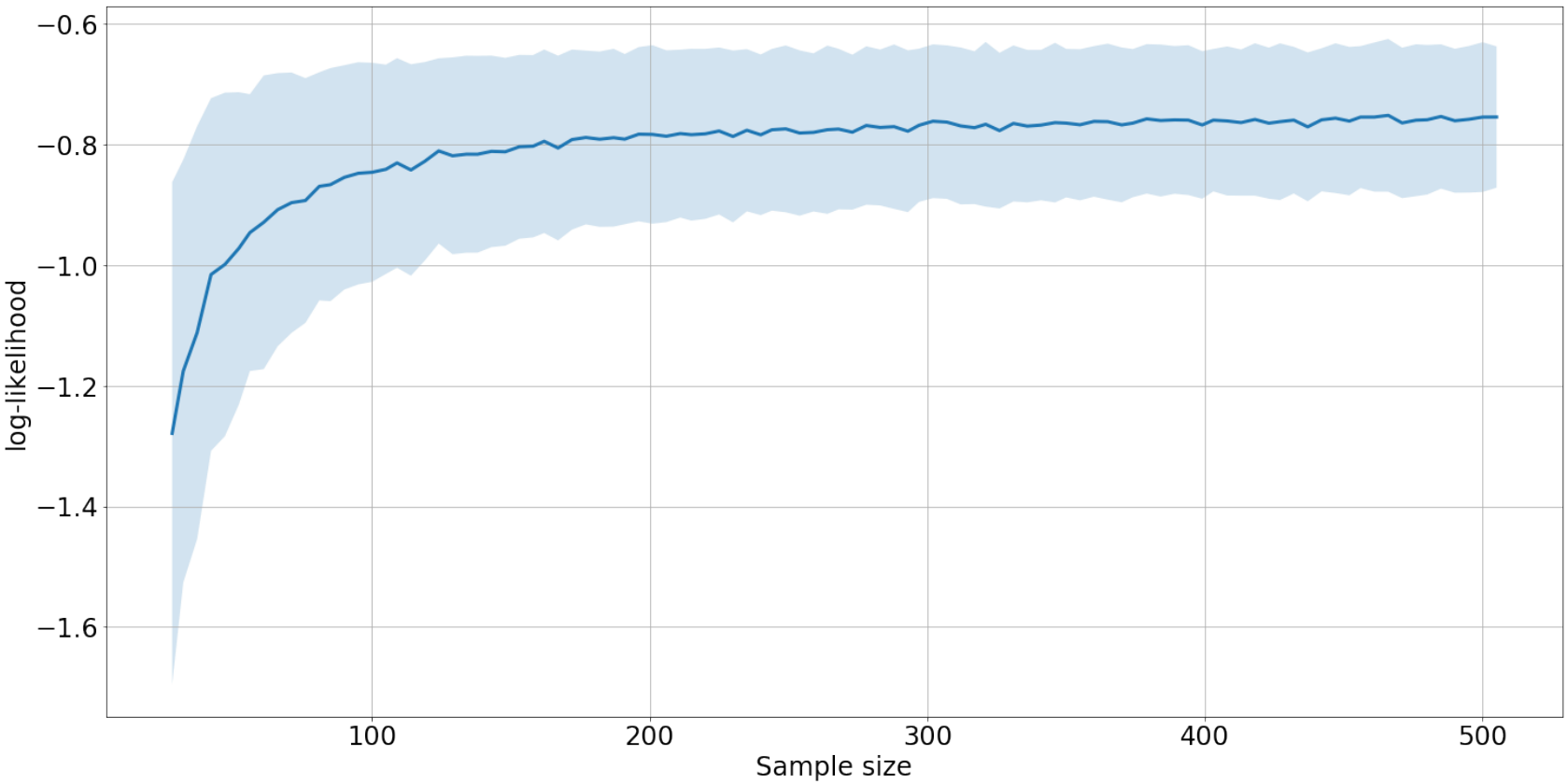
Стохастический градиент и ковариация параметров

Стохастический градиент и ковариация параметров

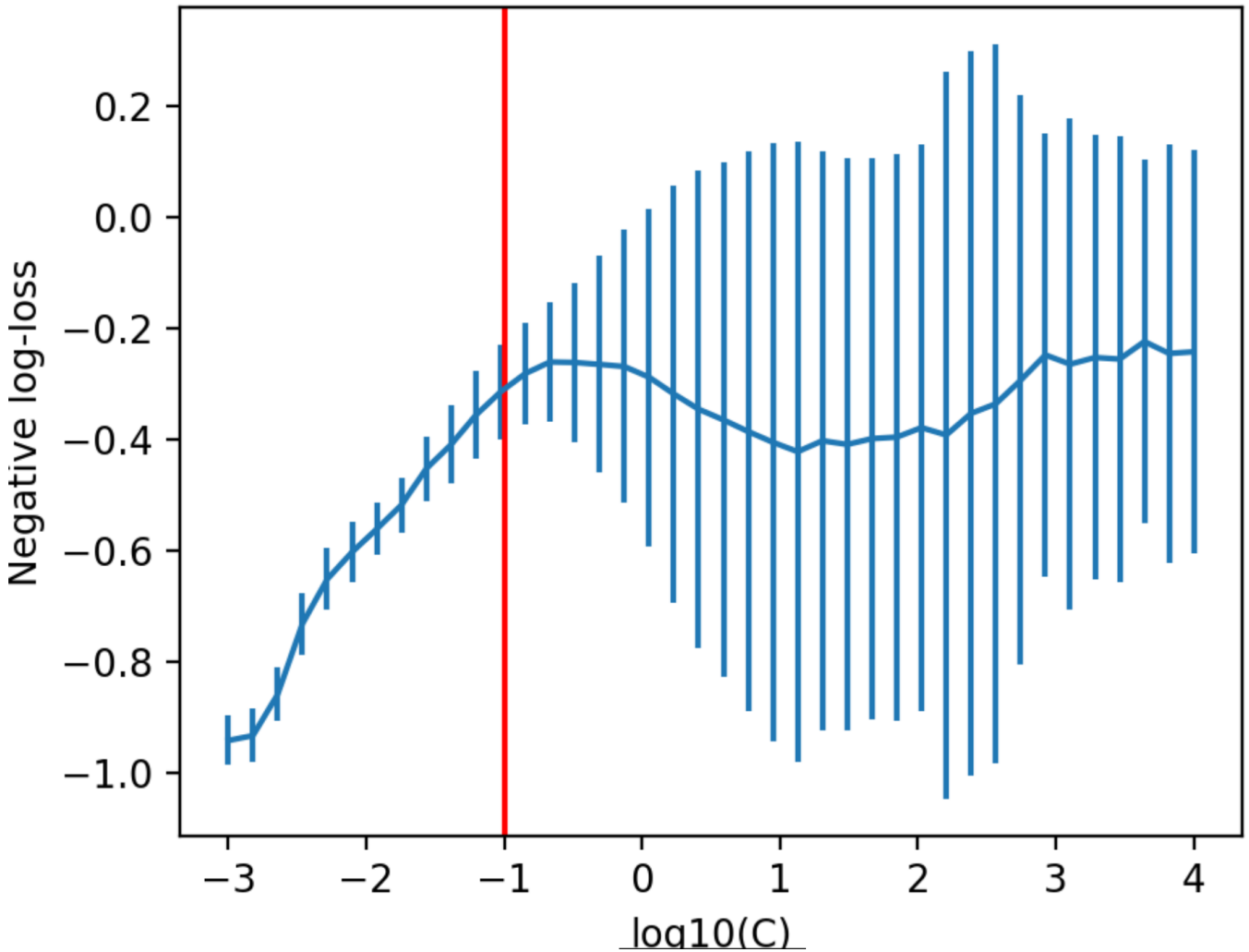
Ошибка (переобученной!) модели



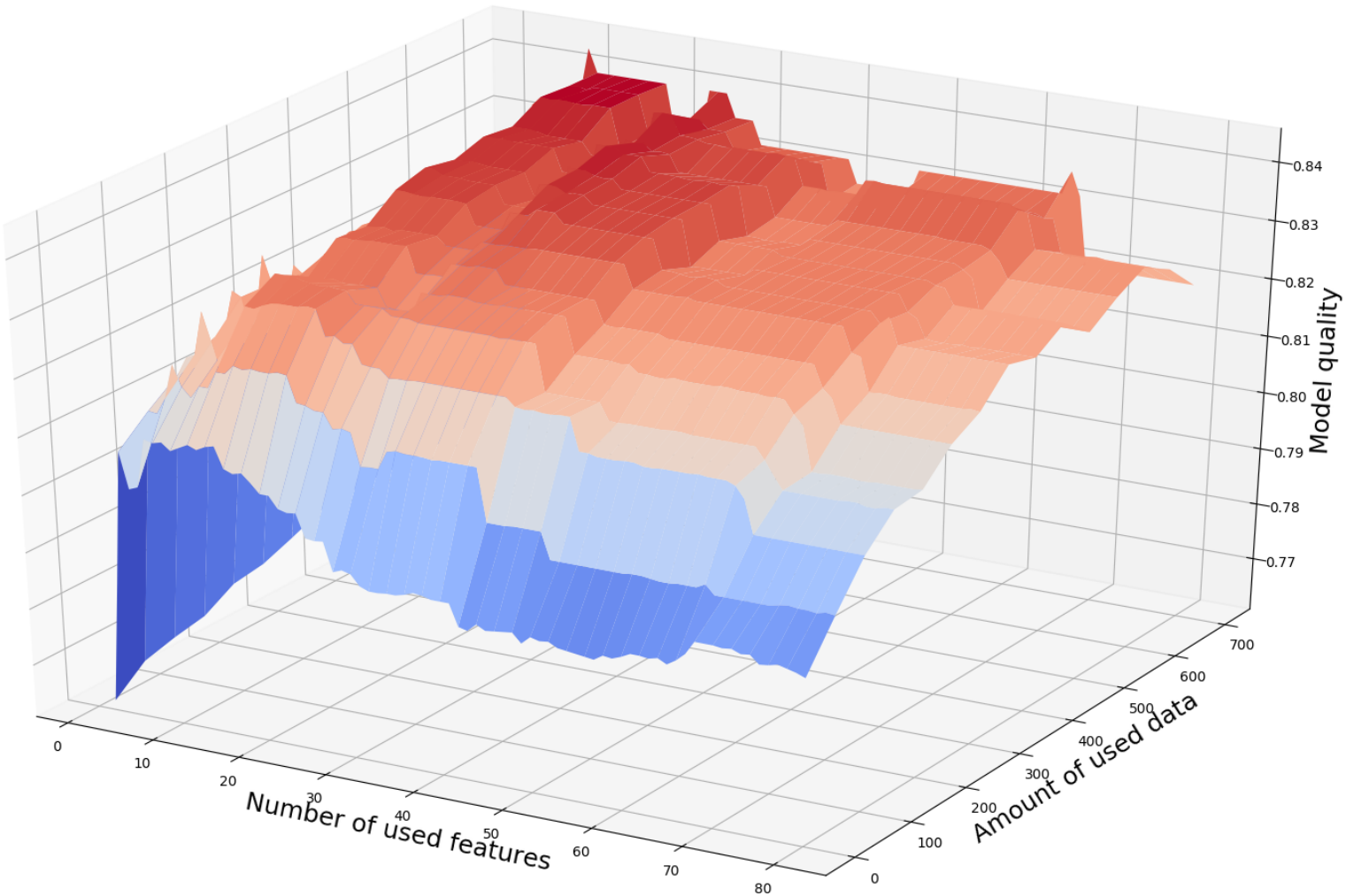
– Ошибка и её дисперсия при пополнении выборки



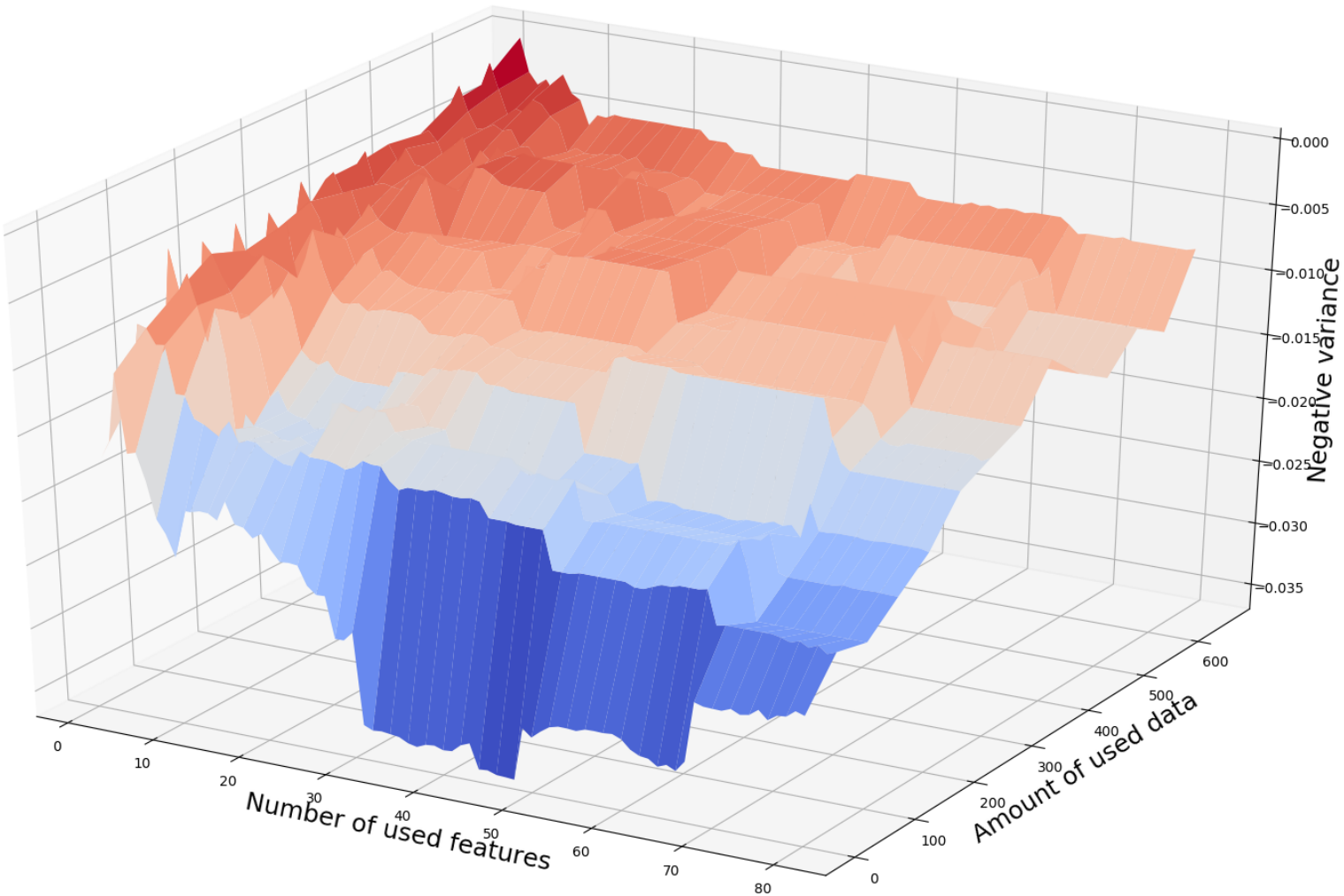
Дисперсия ошибки при повышении сложности модели



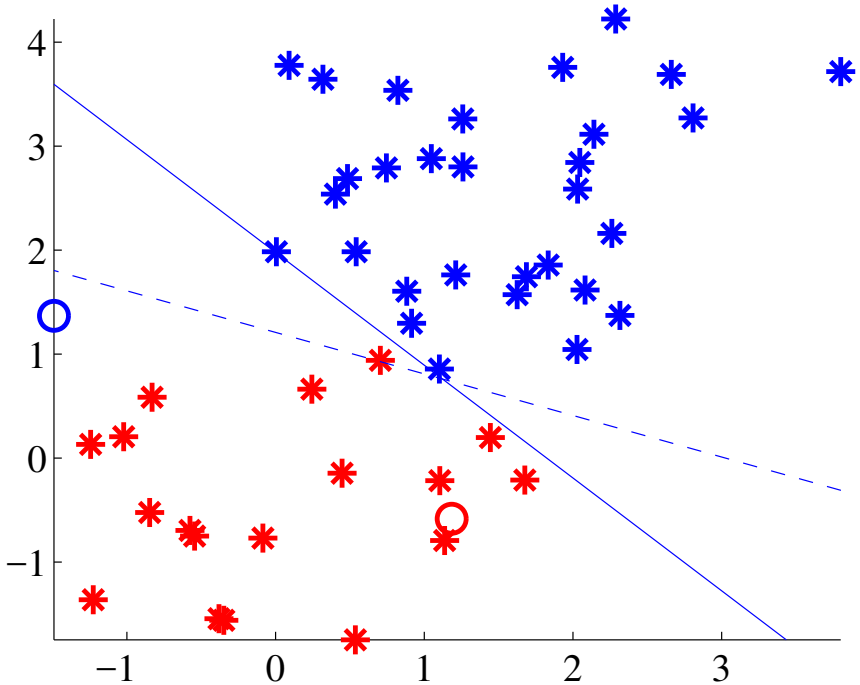
– Ошибка при различных объемах выборки

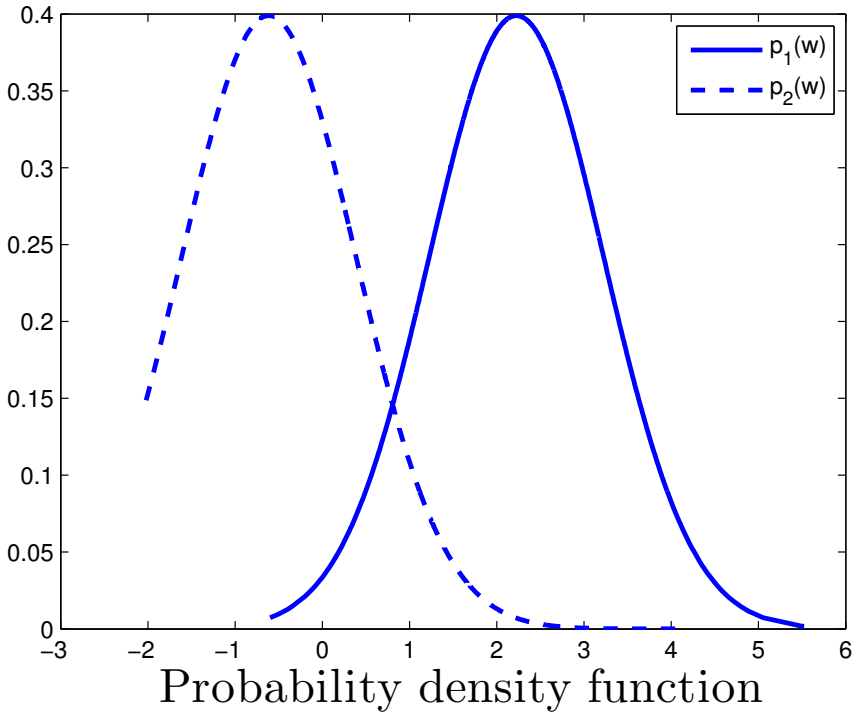


– Дисперсия ошибки при различных объемах выборки



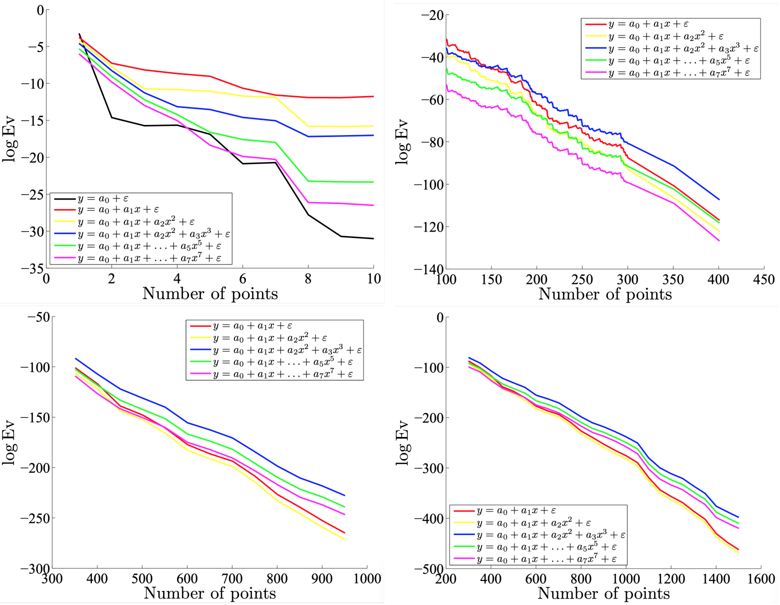
Изменение эмпирического распределения параметров



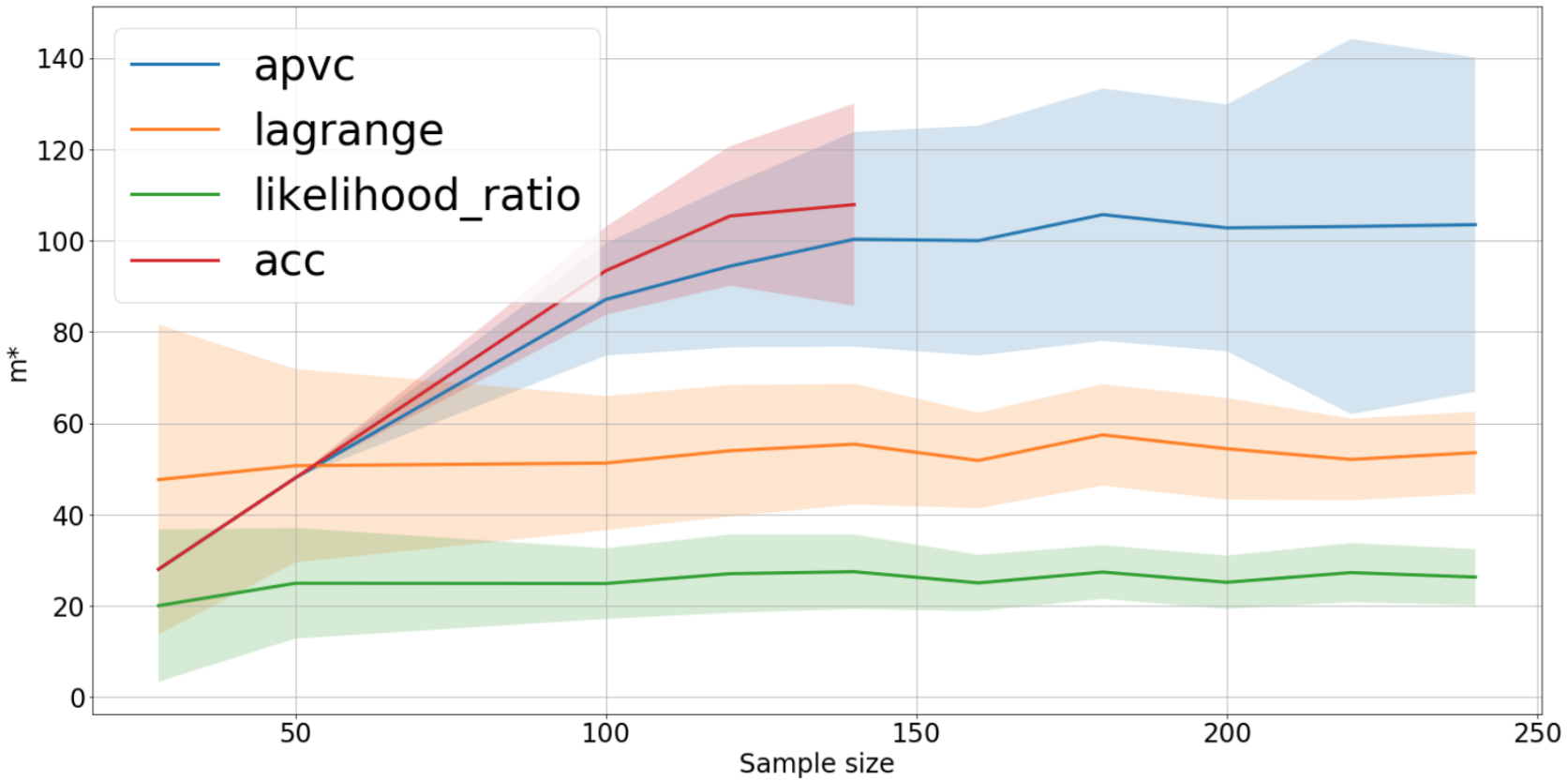


Объем выборки m^* из распределения P достаточен, если выборки X_1, X_2 размера $m > m^*$ из P схожи согласно функции сходства $D(\hat{P}_1, \hat{P}_2)$ между эмпирическими распределениями, полученными на этих выборках.

Правдоподобные модели при разных объемах выборки



Объем выборки, спрогнозированной на раннем этапе сбора данных



Имея выборку объема m требуется спрогнозировать оптимальный объем m^*.