Оценка оптимального объема выборки и выбор моделей машинного обучения

Вадим Викторович Стрижов

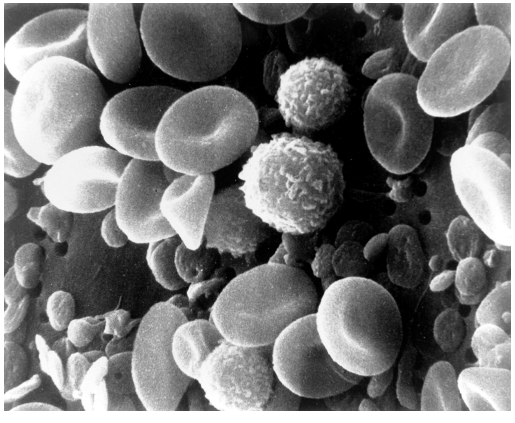
Московский физико-технический институт

2019

Классификация пациентов с нарушениями работы сердечно-сосудистой системы

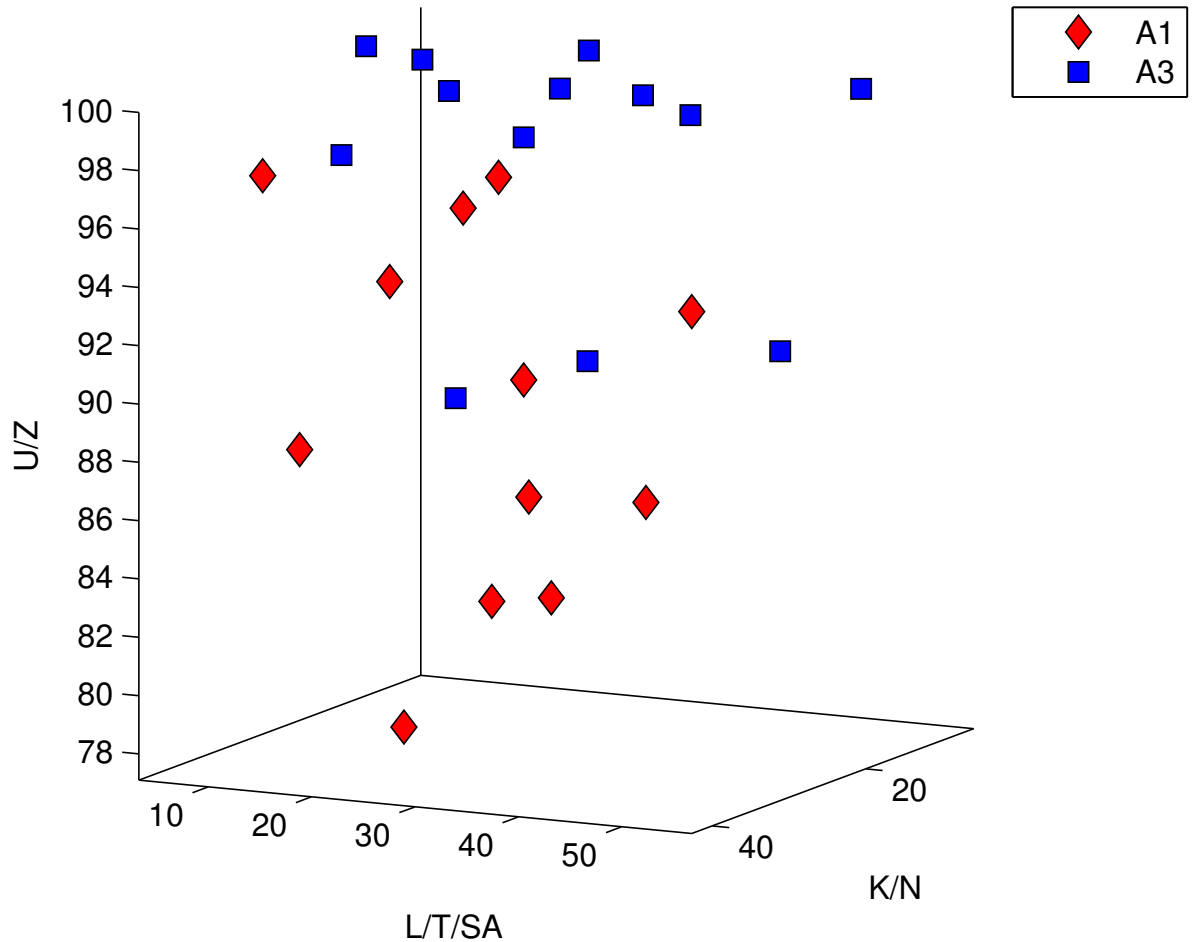
Классы:	A1 прооперированы и A3 в зоне риска
Объекты:	12 пациентов в группе A1 и 13 в A3
Признаки:	20 биомаркеров K, L, K/M, L/M, ...

- Критерий качества: число неверно классифицированных
- Модель: обобщенно-линейная
- Гипотеза порождения данных: простая (i.i.d.) выборка



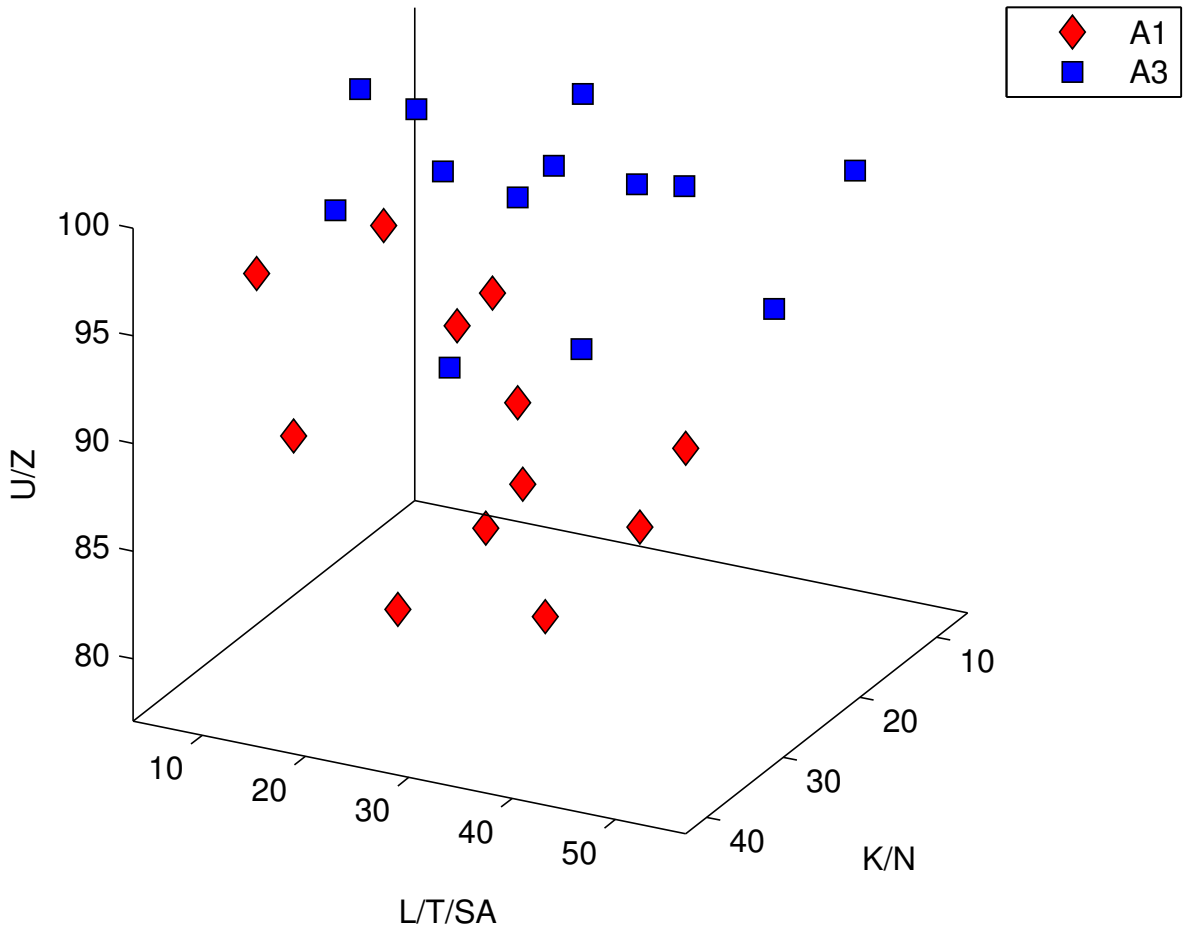
Class	Patient	K	L	K/M	L/M	
A1	P001	58.3	16.7	0.52	0.00	
A1	P004	40.2	6.0	NaN	NaN	
A1	P005	54.3	13.1	NaN	NaN	
A1	P008	48.7	9.8	0.05	0.02	etc.
A3	P023	46.6	21.2	0.40	0.08	
A3	P026	50.7	26.2	0.12	0.00	
A3	A007	45.3	24.5	0.05	0.02	
A3	P039	46.3	13.1	1.23	0.13	
				etc.		

Выбраны три признака из 20, линейная модель



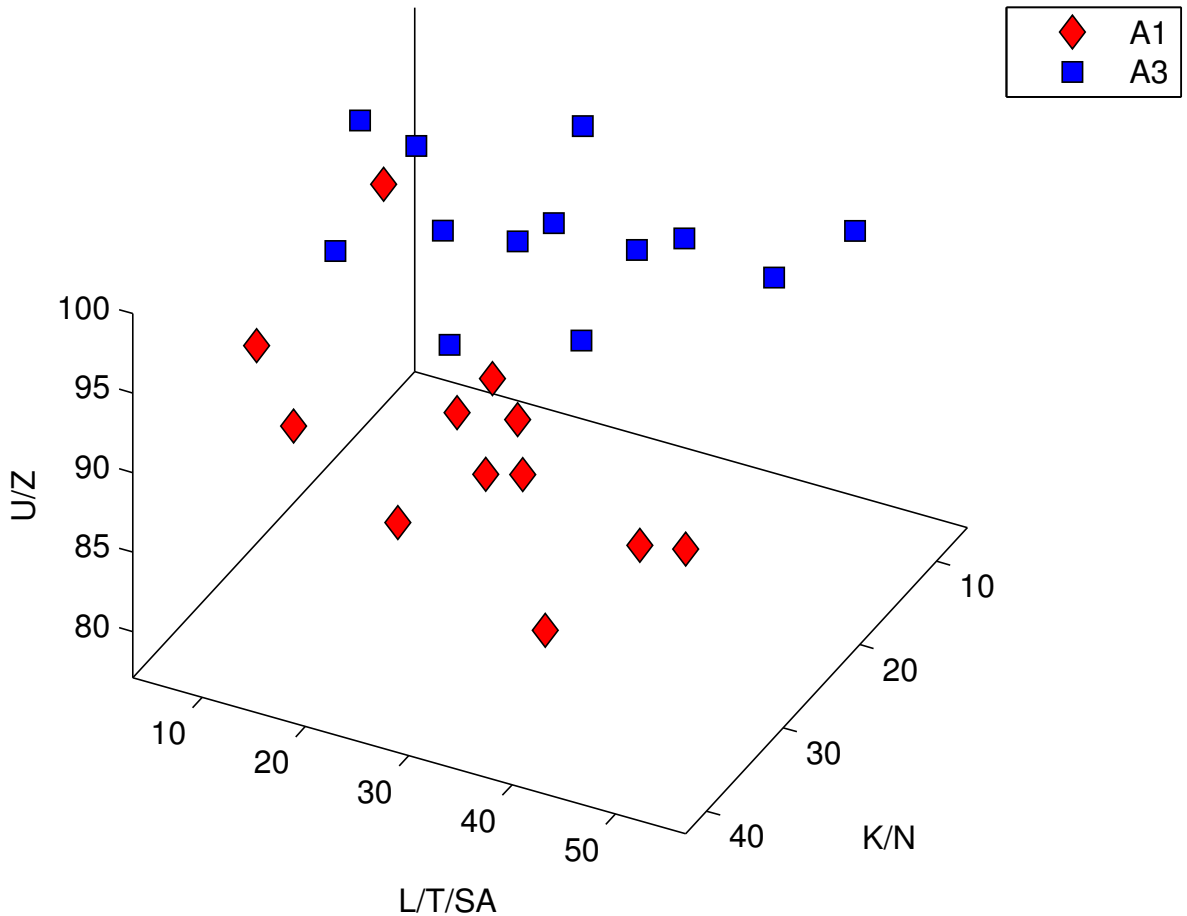
$$\hat{y} = f(\mathbf{w}, \mathbf{x}) = \text{sign}(\mathbf{w}^\mathsf{T}\mathbf{x} - b) = \text{sign}([0.35, 0.72, 0.29]^\mathsf{T}\mathbf{x} - 34.16)$$

Выбраны три признака из 20, линейная модель



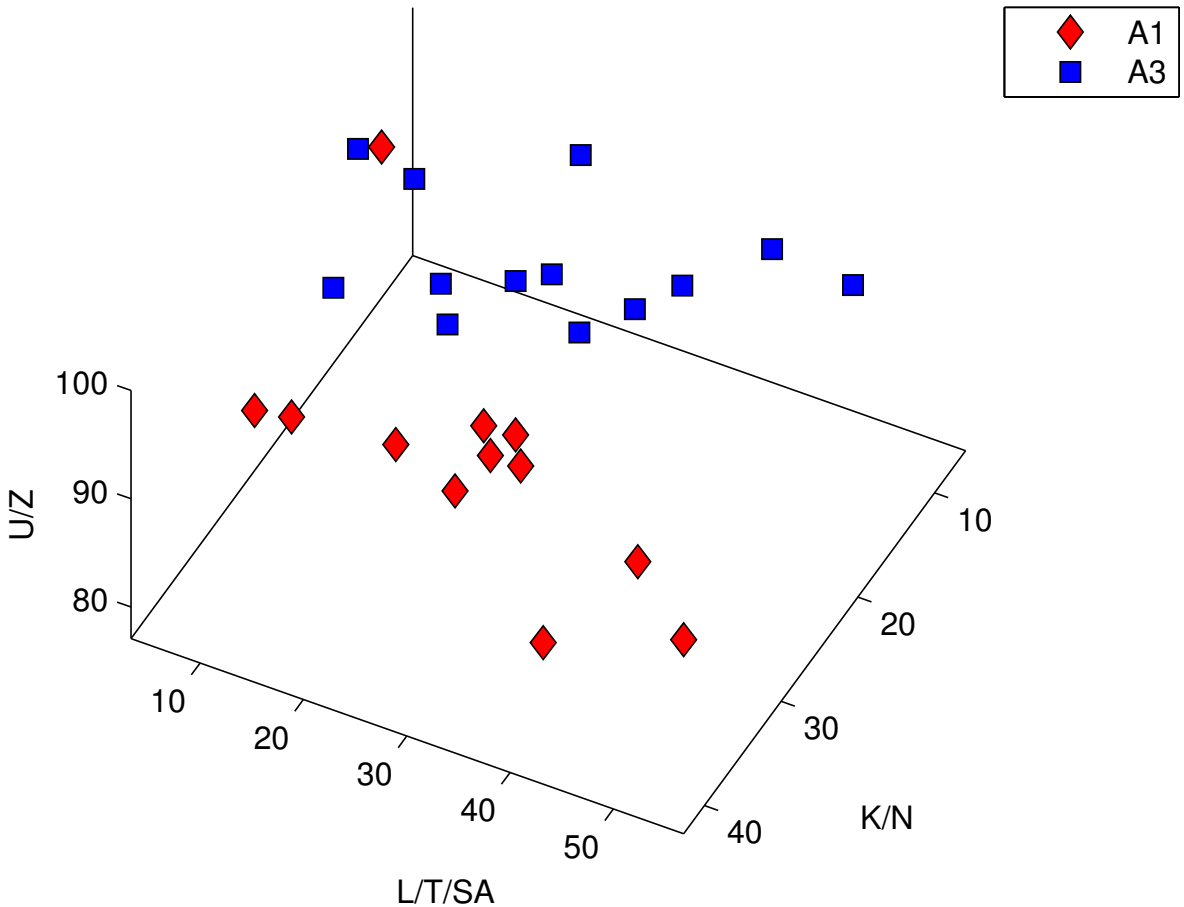
$$\hat{y} = f(\mathbf{w}, \mathbf{x}) = \text{sign}(\mathbf{w}^{\mathsf{T}}\mathbf{x} - b) = \text{sign}([0.35, 0.72, 0.29]^{\mathsf{T}}\mathbf{x} - 34.16)$$

Выбраны три признака из 20, линейная модель



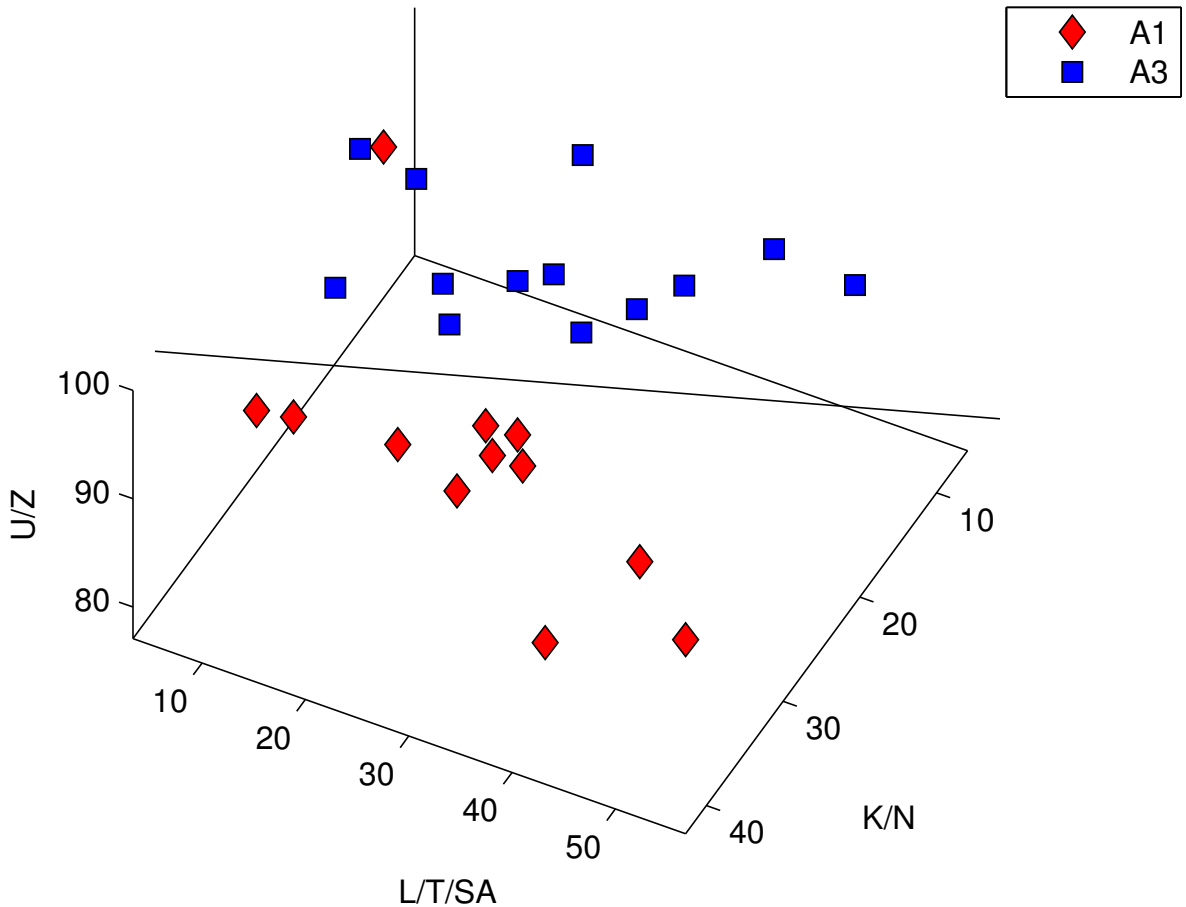
$$\hat{y} = f(\mathbf{w}, \mathbf{x}) = \text{sign}(\mathbf{w}^{\mathsf{T}}\mathbf{x} - b) = \text{sign}([0.35, 0.72, 0.29]^{\mathsf{T}}\mathbf{x} - 34.16)$$

Выбраны три признака из 20, линейная модель



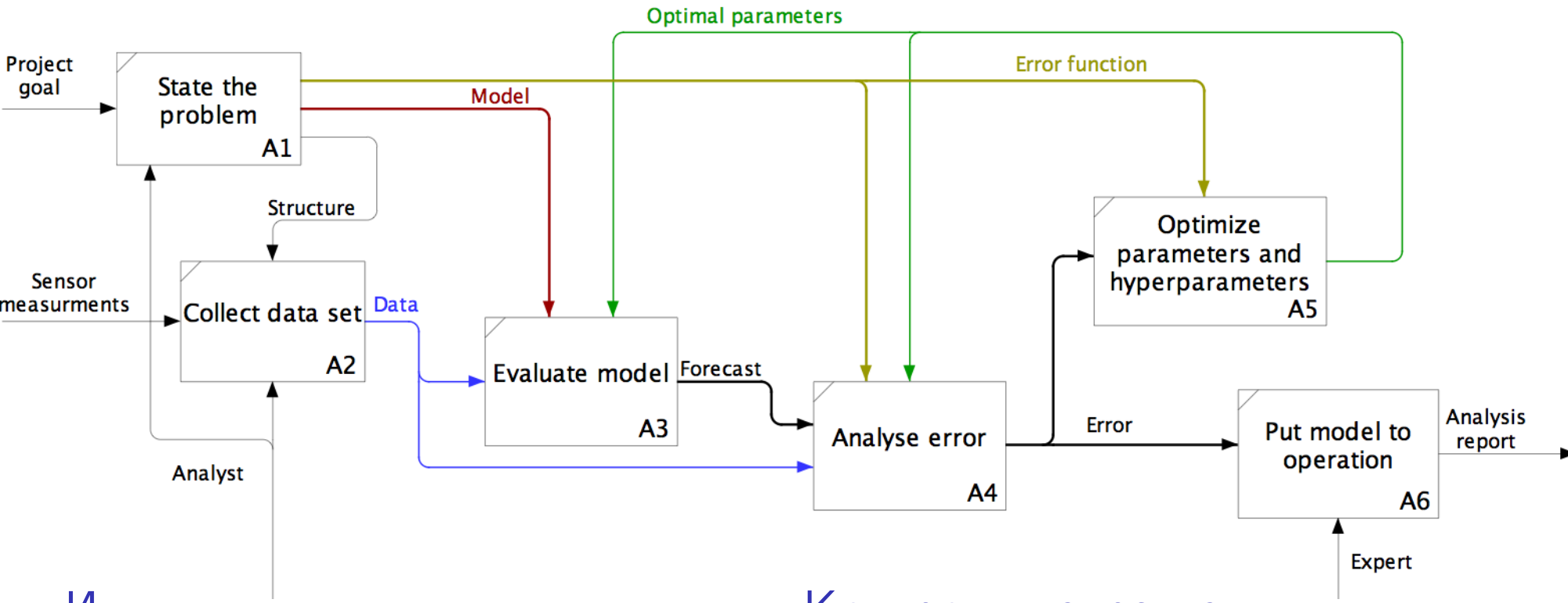
$$\hat{y} = f(\mathbf{w}, \mathbf{x}) = \text{sign}(\mathbf{w}^{\mathsf{T}}\mathbf{x} - b) = \text{sign}([0.35, 0.72, 0.29]^{\mathsf{T}}\mathbf{x} - 34.16)$$

Выбраны три признака из 20, линейная модель



$$\hat{y} = f(\mathbf{w}, \mathbf{x}) = \text{sign}(\mathbf{w}^{\mathsf{T}}\mathbf{x} - b) = \text{sign}([0.35, 0.72, 0.29]^{\mathsf{T}}\mathbf{x} - 34.16)$$

Создание модели для её эксплуатации



Источники критериев качества

1. Эксплуатация: доход, число отказов
2. Теория: статистические гипотезы порождения данных
3. Технология: критерий удобен для оптимизации

Критерии качества

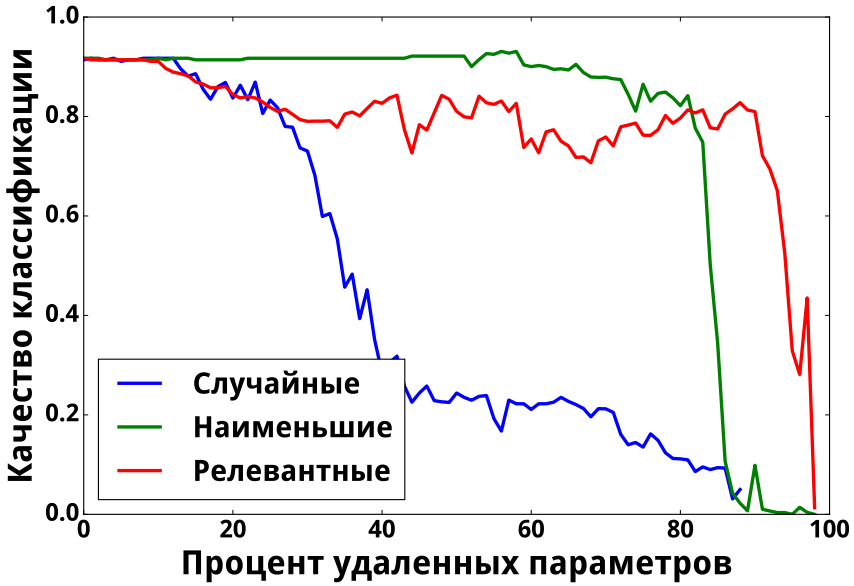
- Точность: MAPE, AUC, F1 score
- Устойчивость: дисперсия прогноза, ошибки, ковариация параметров
- Сложность: число параметров, Колмогоровская

Значительное повышение сложности и скромный прирост точности

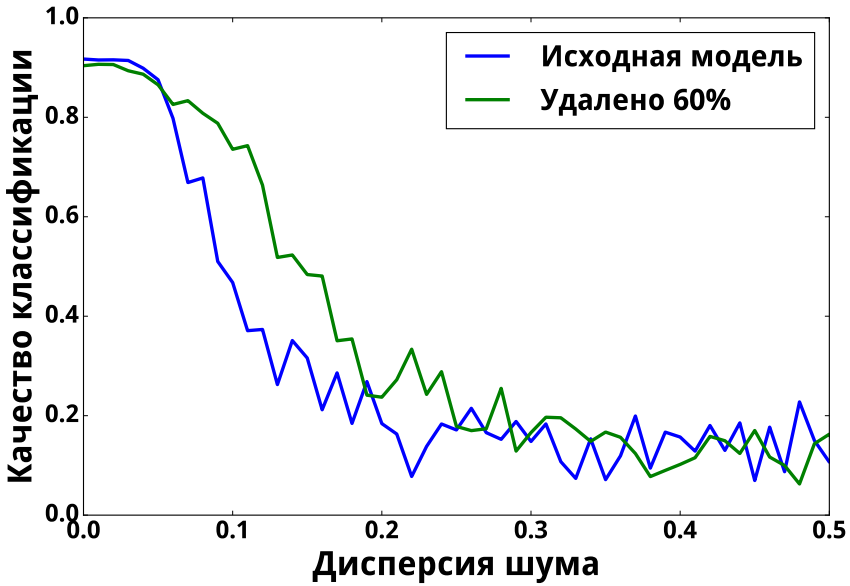
	train	test	Number of parameters
Логистическая регрессия	53,08%	55,18%	= 12
Нейронная сеть	59,85%	57,04%	~ 240
Случайный лес	61,85%	57,01%	> 1000
Градиентный бустинг	**63,58%**	**58,31%**	> 10 000

... это был скоринг

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



Избыточность параметров модели

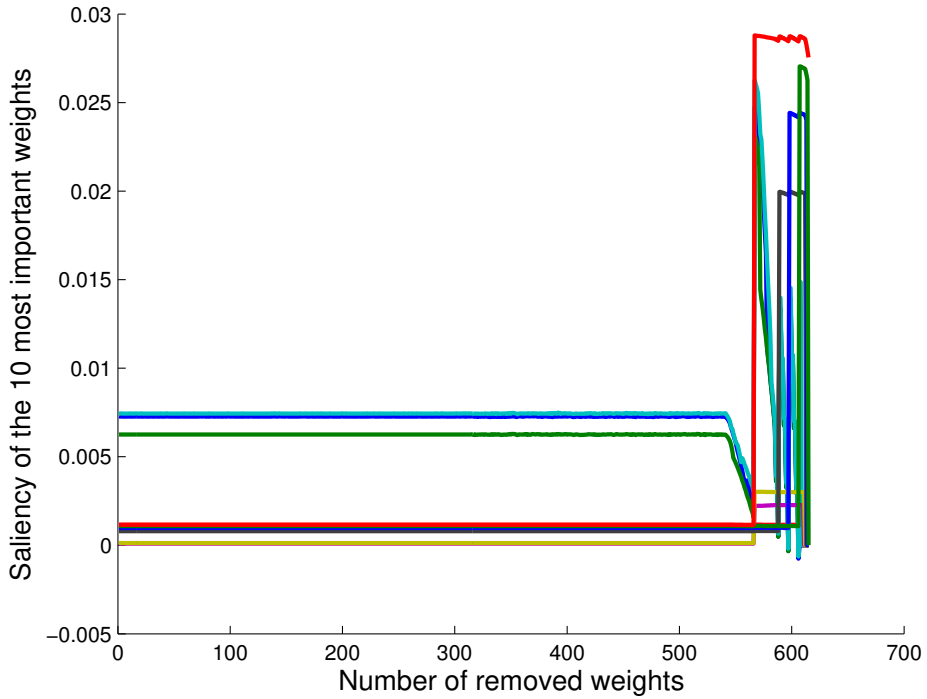


Устойчивость модели

Глубокое обучение предполагает оптимизацию моделей с заведомо избыточной сложностью.

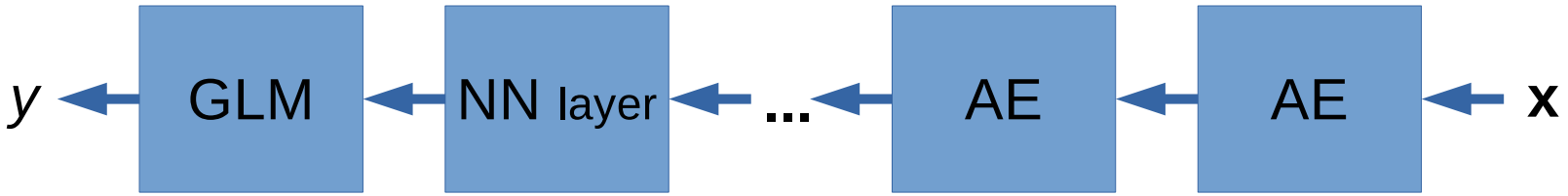
Bakhteev, Strijov. 2019. Comprehensive analysis of gradient-based hyperparameter optimization algorithms // Annals of Operations Research

Neural network optimal brain damage procedure



Salience function $L_j = \frac{w_j^2}{2\mathbf{H}_{jj}^{-1}}$ versus number of removed parameters

Линейная модель, (глубокая) нейросеть и автоэнкодер



$$f = \sigma_k \circ \underset{1\times 1}{\mathbf{w}_k^{\mathsf{T}}} \boldsymbol{\sigma}_{k-1} \circ \mathbf{W}_{k-1}\boldsymbol{\sigma}_{k-2} \circ \cdots \circ \underset{n_2\times 1}{\mathbf{W}_2}\boldsymbol{\sigma}_1 \circ \underset{n_1\times n}{\mathbf{W}_1}\,\underset{n\times 1}{\mathbf{x}}$$

$$S = \sum_{(\mathbf{x}_i, y_i)\in\mathfrak{D}} \big(y_i - f(\mathbf{x}_i)\big)^2 \qquad E_{\mathbf{x}} = \sum_{\mathbf{x}_i\in\mathfrak{D}} \|\mathbf{x}_i - \mathbf{r}(\mathbf{x}_i)\|_2^2$$

Варианты

$E_{\mathbf{x}}$ — ошибка реконструкции автоэнкодера

метод главных компонент: $\mathbf{W}^{\mathsf{T}}\mathbf{W} = \mathbf{I}_n$

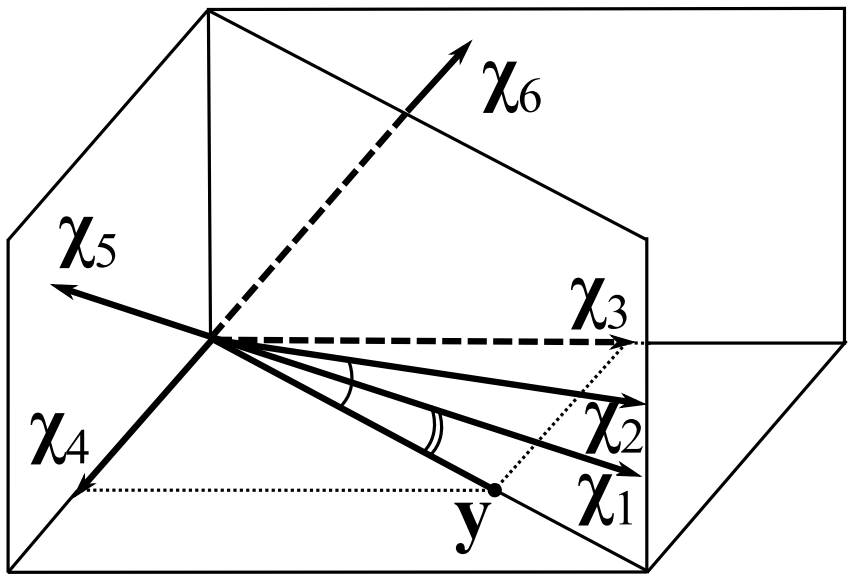
нет блока: $\mathbf{W} = \mathbf{I}_n$, $\sigma = \mathrm{id}$

классификация: $\sigma = \big(1 + \exp(-\cdot)\big)^{-1}$

$+\mathbf{b}$

Выбор устойчивого и точного набора признаков

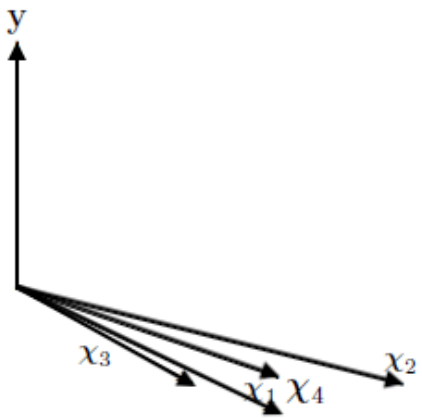
Признаки $\chi_1, \ldots, \chi_6$ — столбцы матрицы плана $\underset{3\times 6}{\mathbf{X}}$.



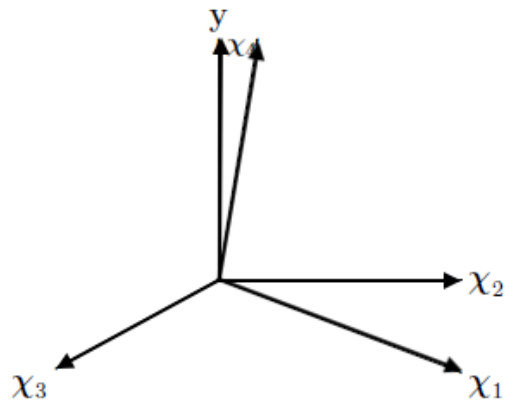
Решение: χ_3, χ_4 ортогональны, их комбинация приближает $\mathbf{y}$, минимизируя ошибку.

Katrutsa, Strijov. 2015. Stress-test procedure for feature selection // Chemometrics

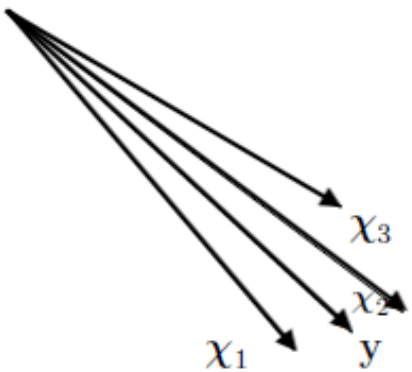
Конфигурации признакового пространства



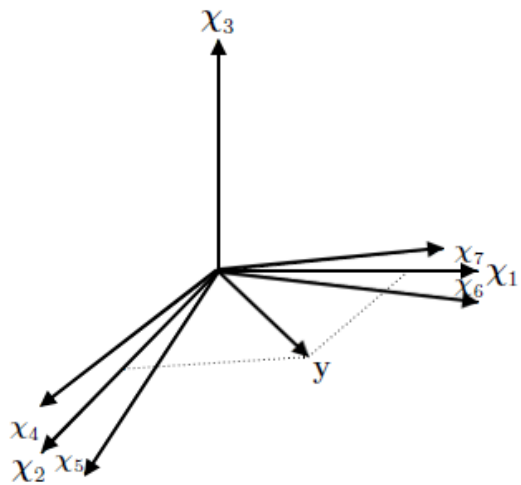
Неадекватный коррелированный



Адекватный случайный



Адекватный избыточный

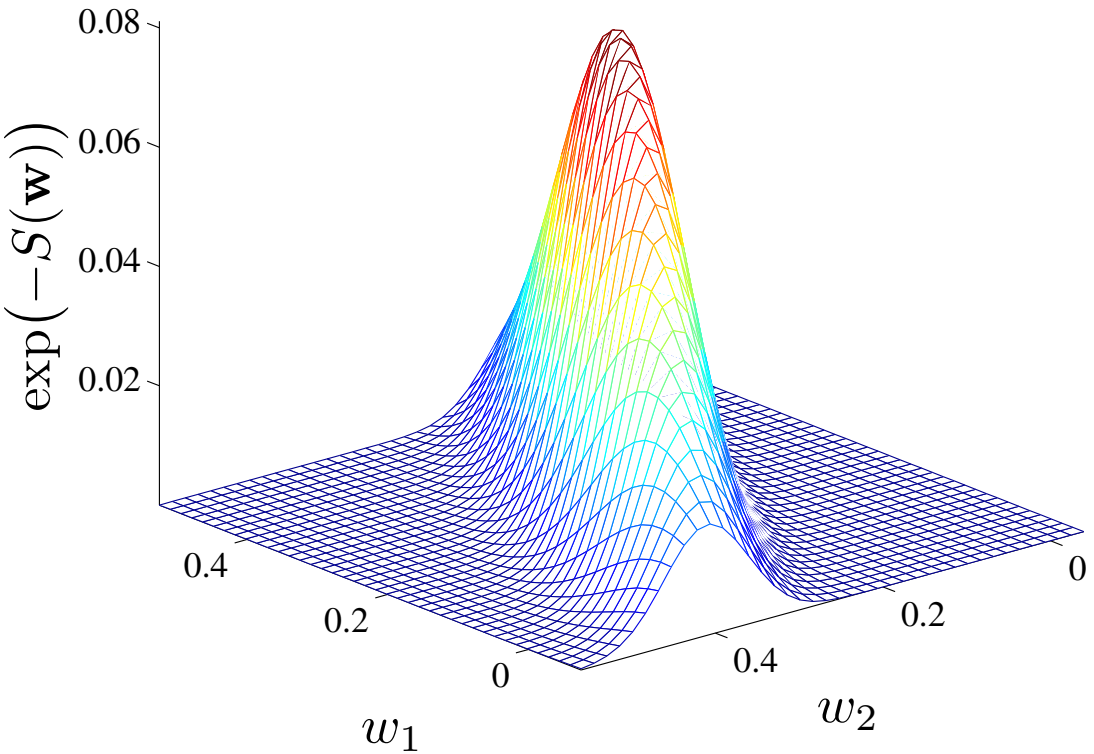


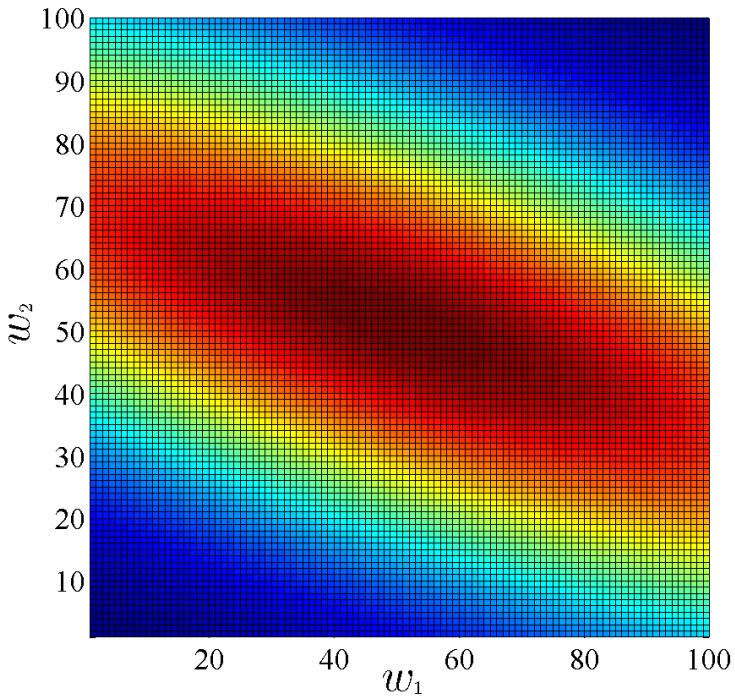
Адекватный коррелированный

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

Эмпирическое распределение параметров модели

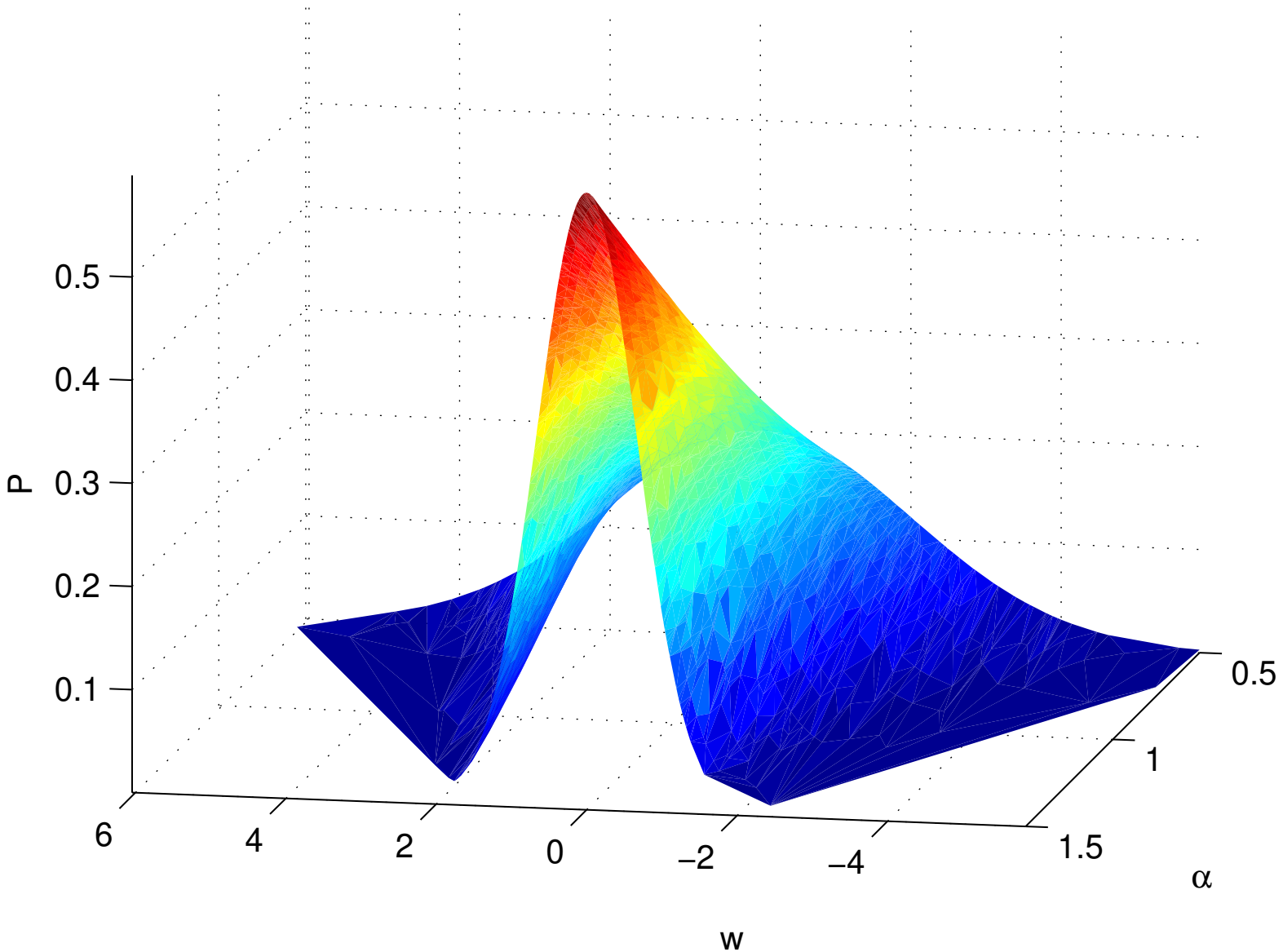
Значение функции ошибки $S(\mathbf{w}|\mathfrak{D}, f)$ зависит от параметров.



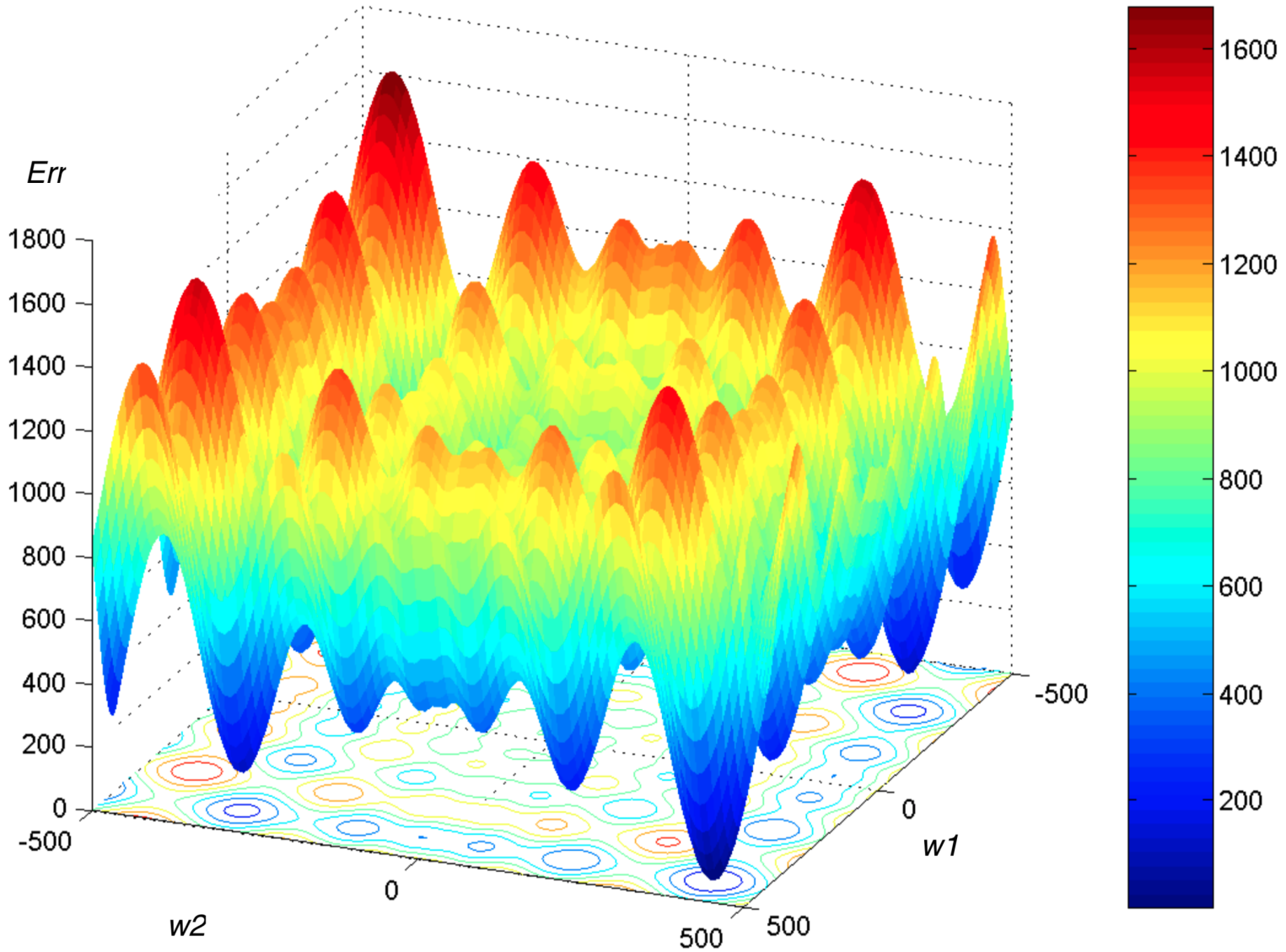


Kuznetsov, Tokmakova, Strijov. 2016. Analytic methods of structure parameter // Informatica

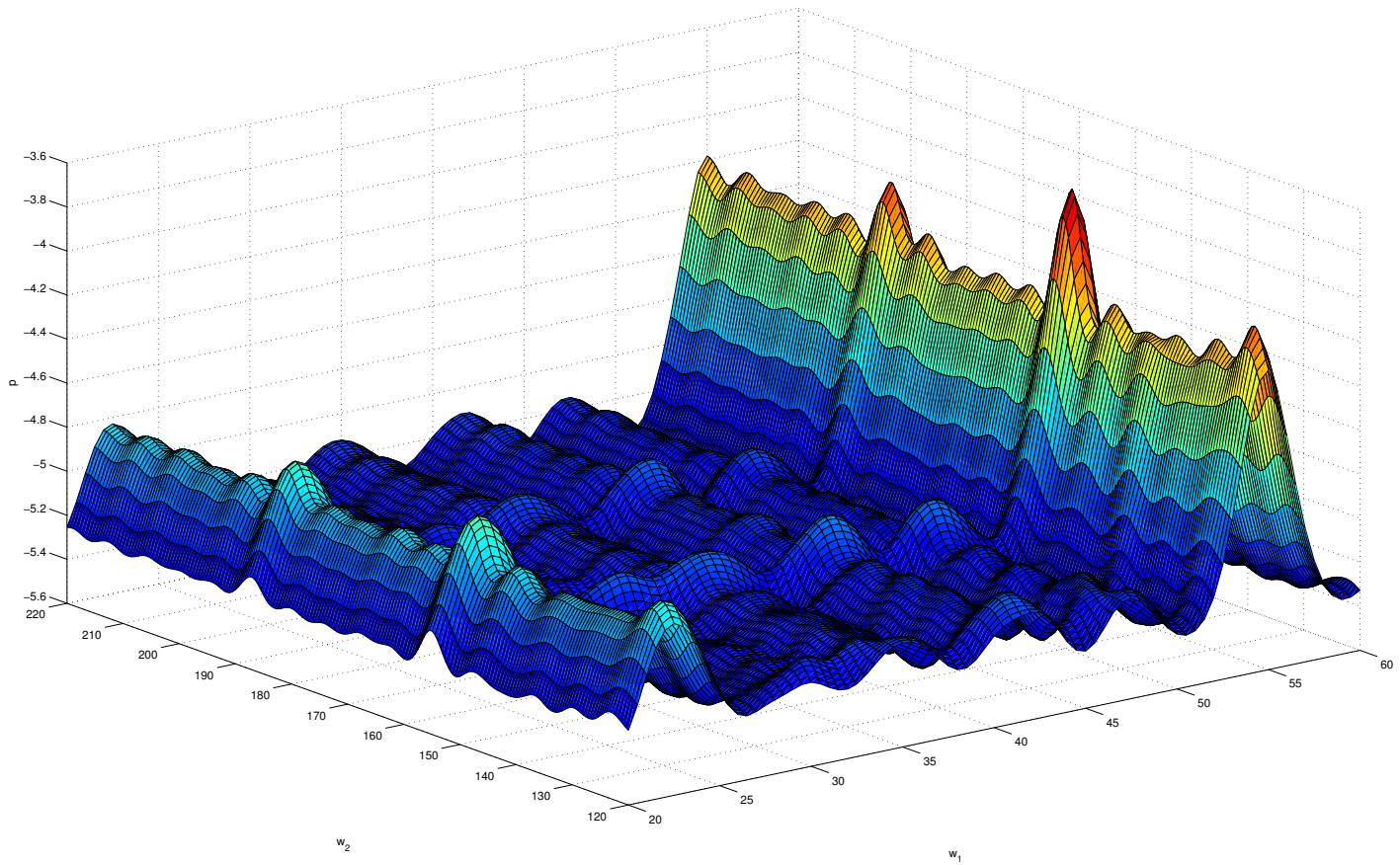
Правдоподобие модели, параметры и их дисперсия



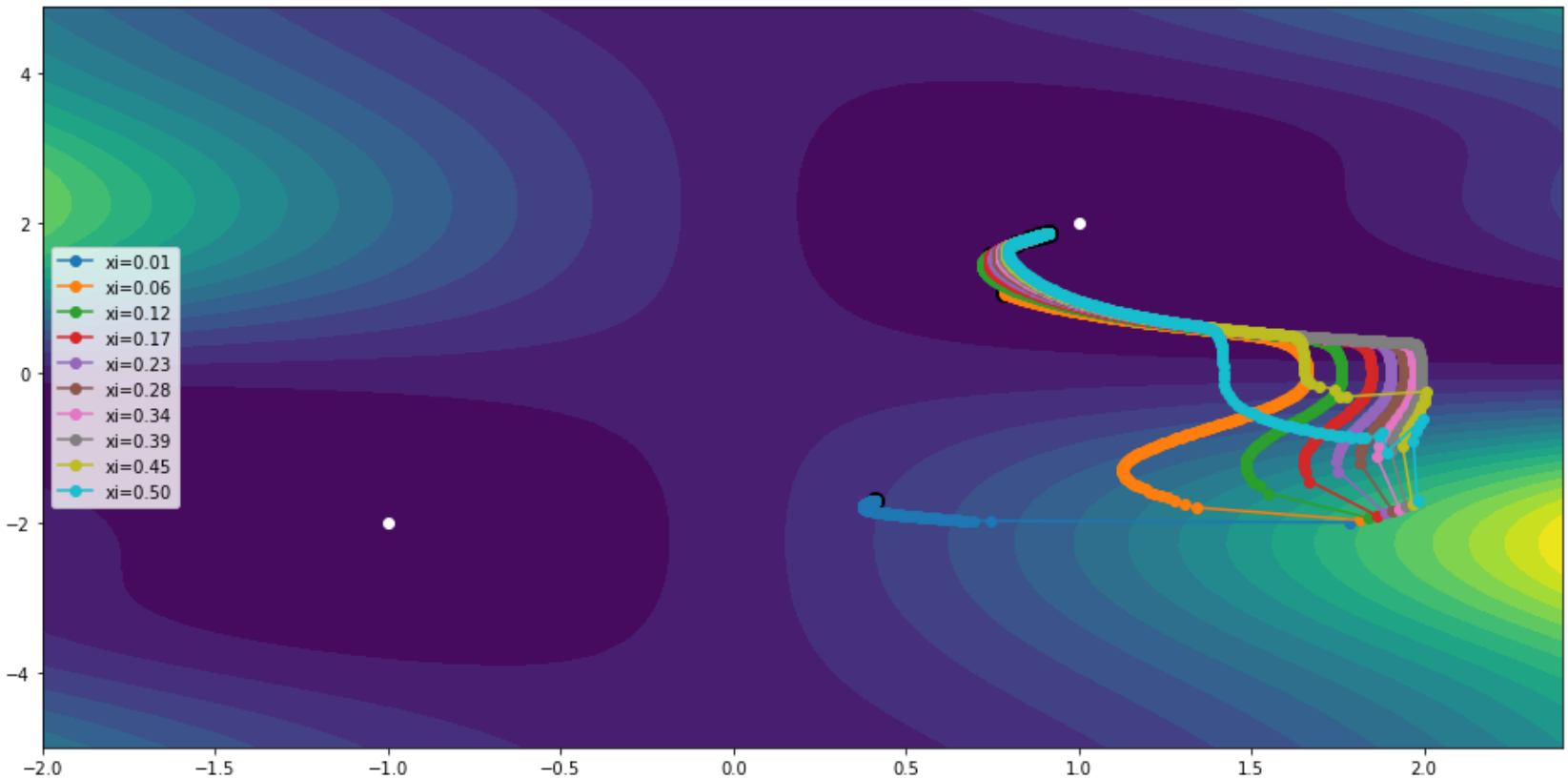
Многоэкстремальность функции ошибки (пример)



Optimal brain surgery



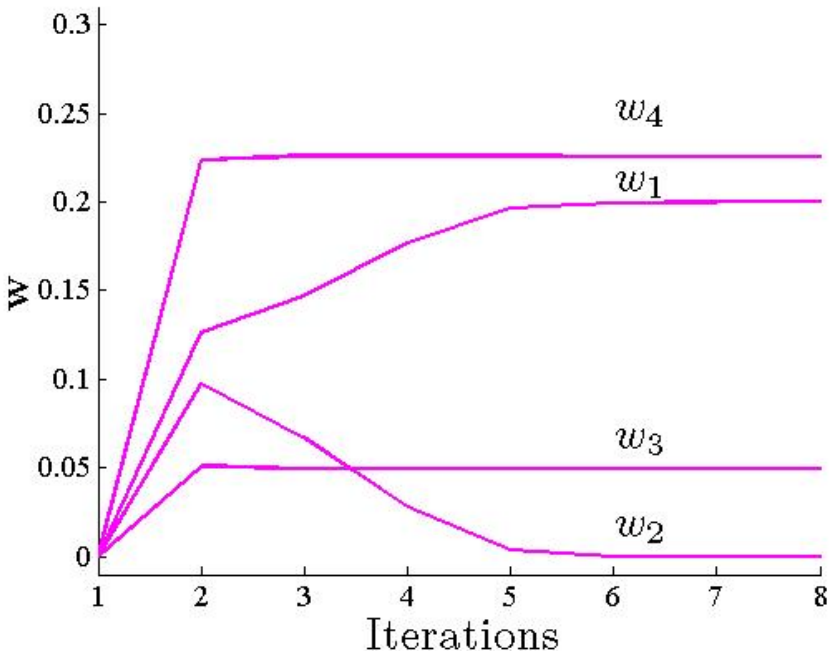
Многоэкстремальность, сходимость и мультистарт



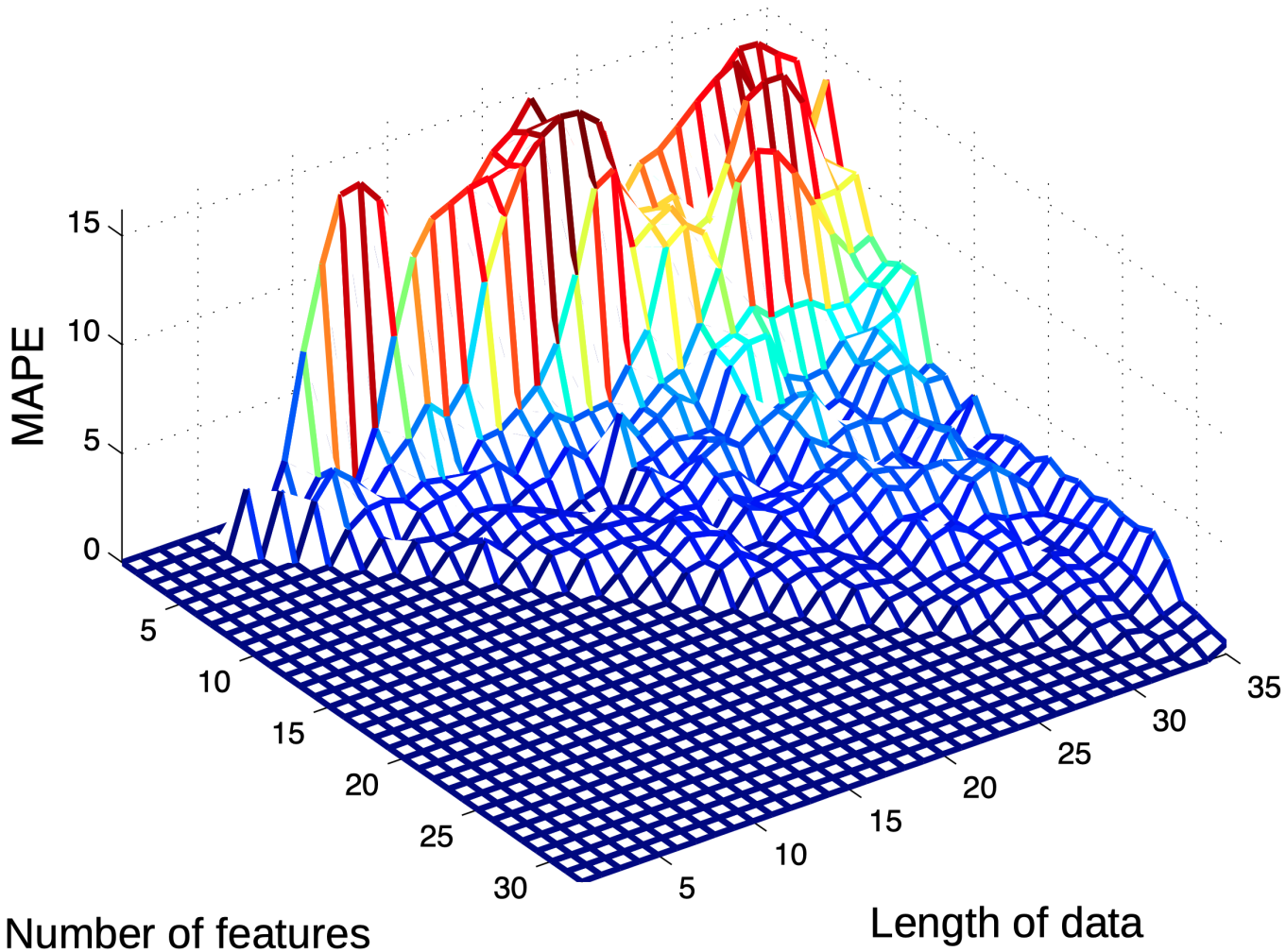
$$\mathbf{w} \in \mathbb{R}^2$$

Стабилизация параметров сети

При нахождении минимума (он может оказаться локальным) параметры стабилизируются.

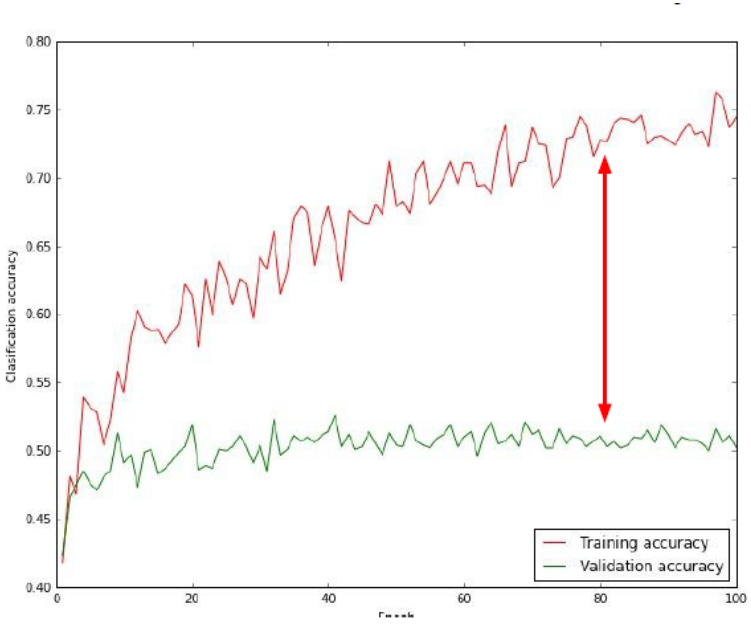


Ошибка (переобученной!) модели

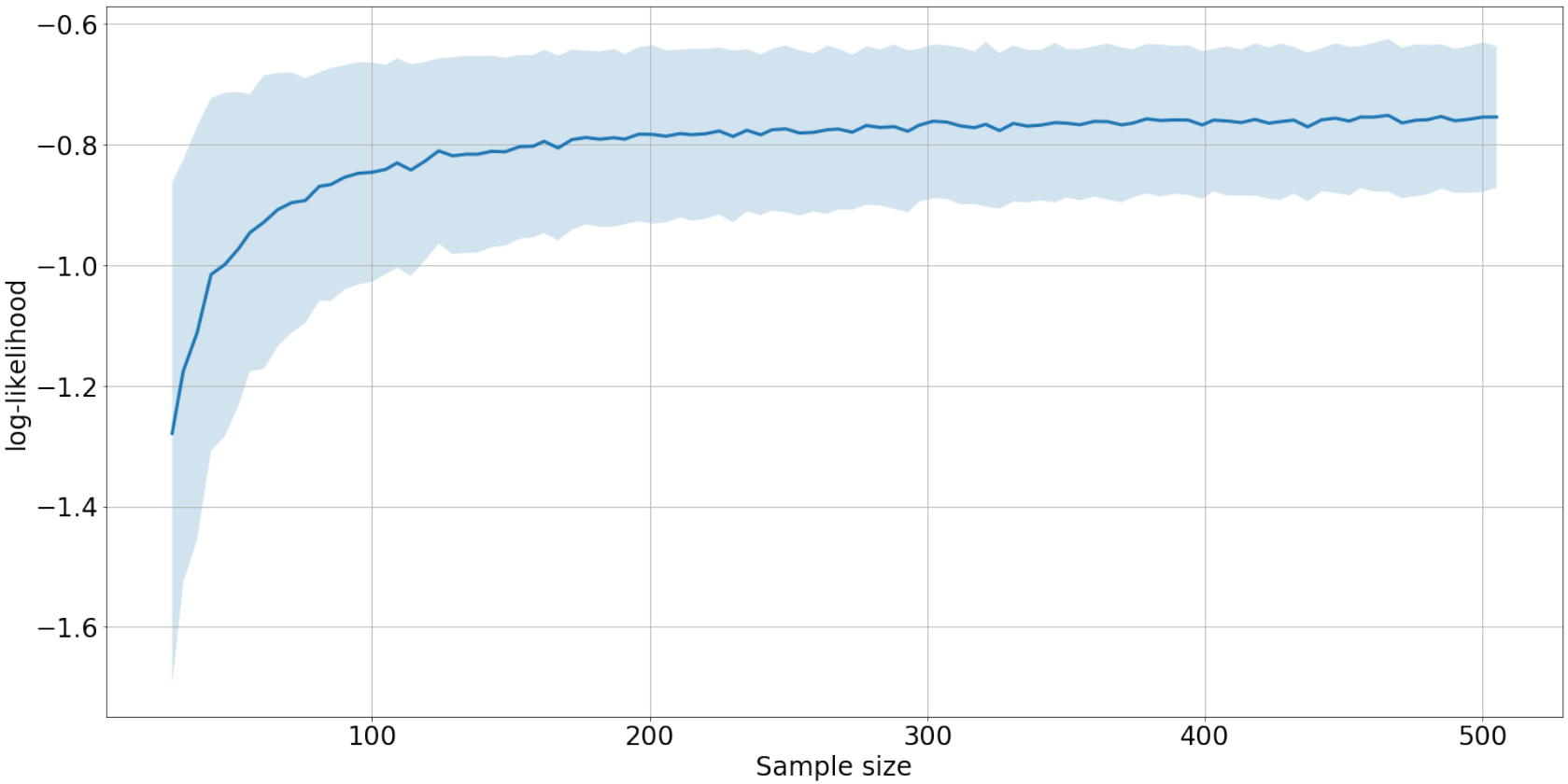


Обучение и контроль

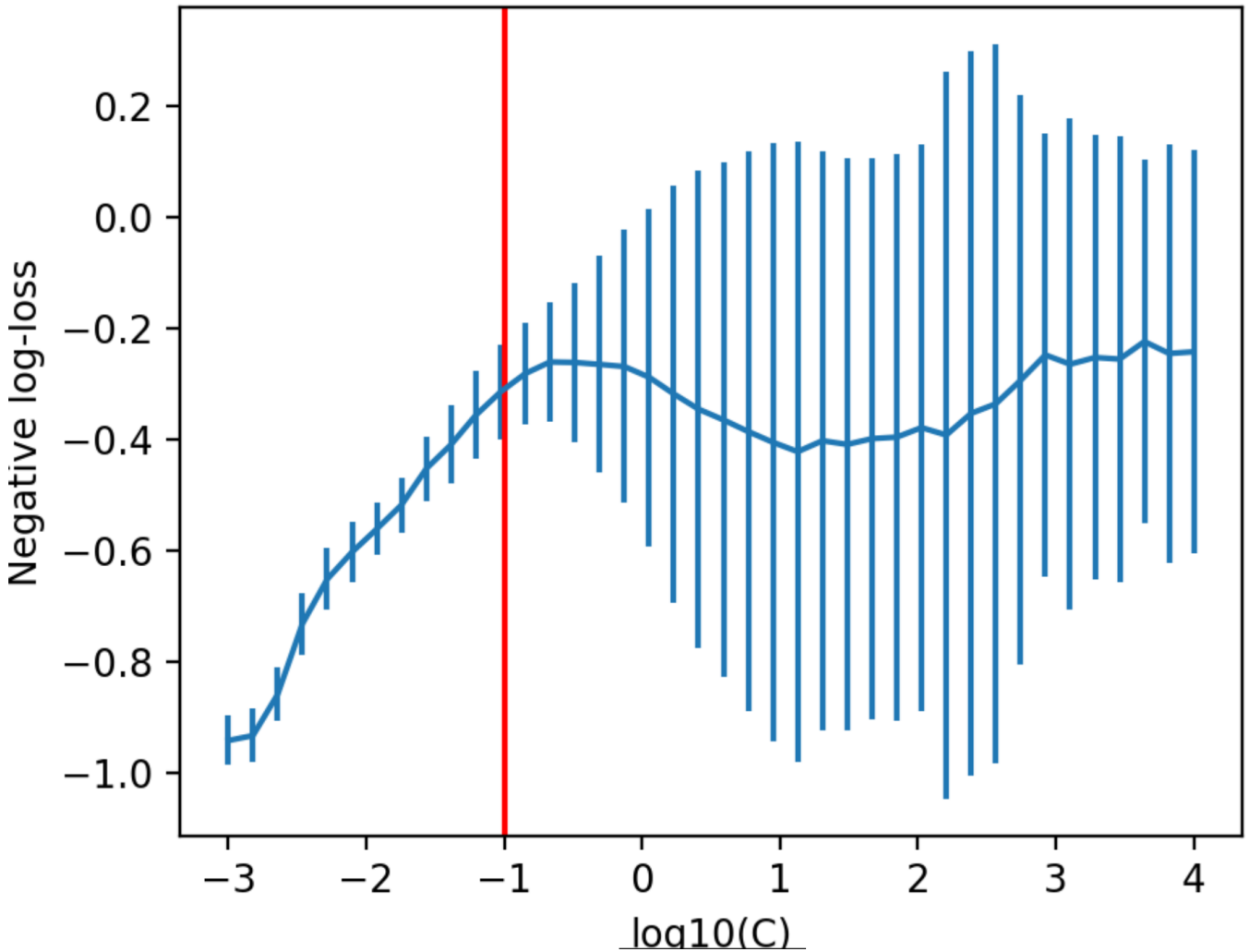
Разница между значениями функции ошибки на обучении и контроле не должна быть существенной.



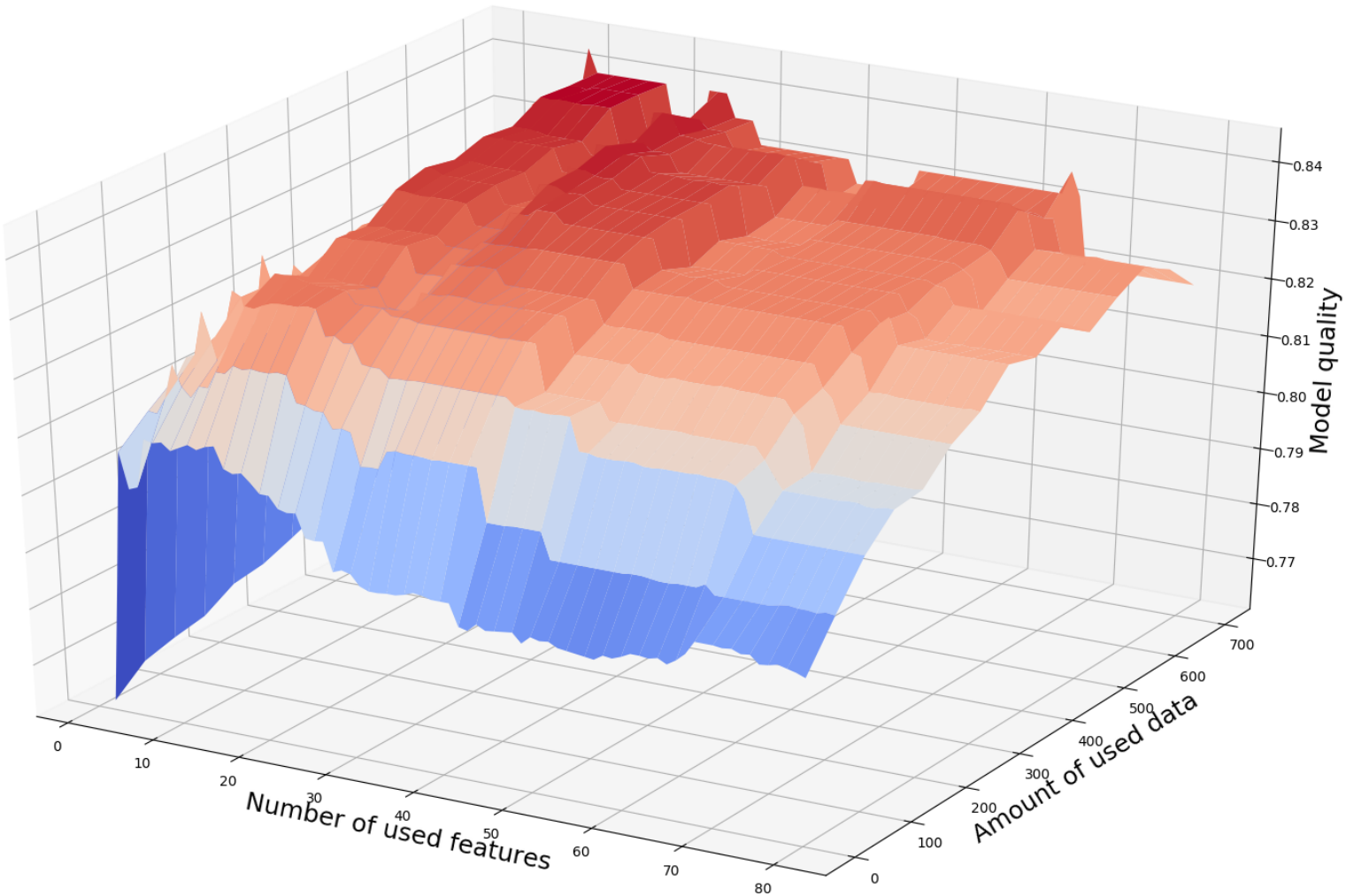
– Ошибка и её дисперсия при пополнении выборки



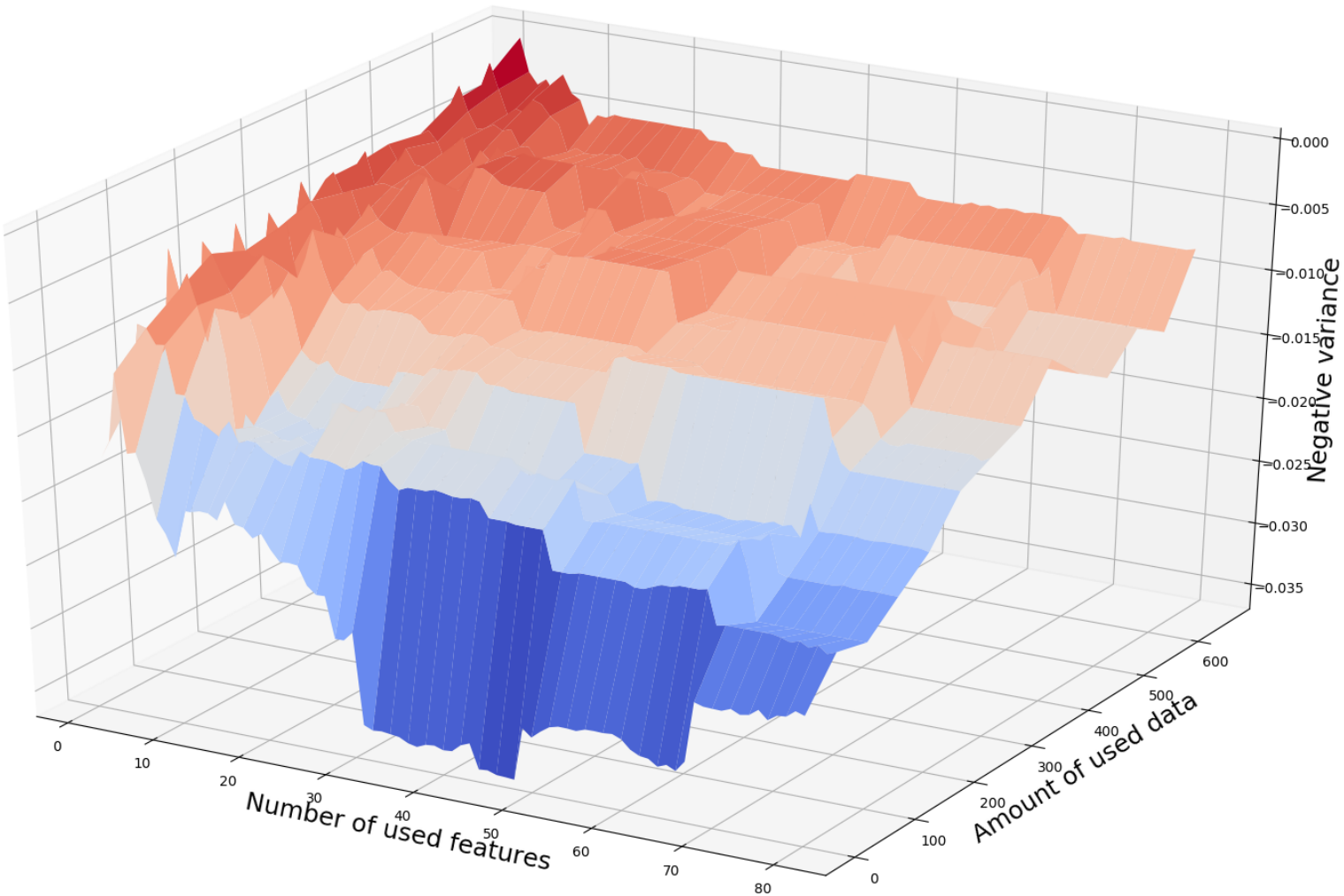
Дисперсия ошибки при повышении сложности модели



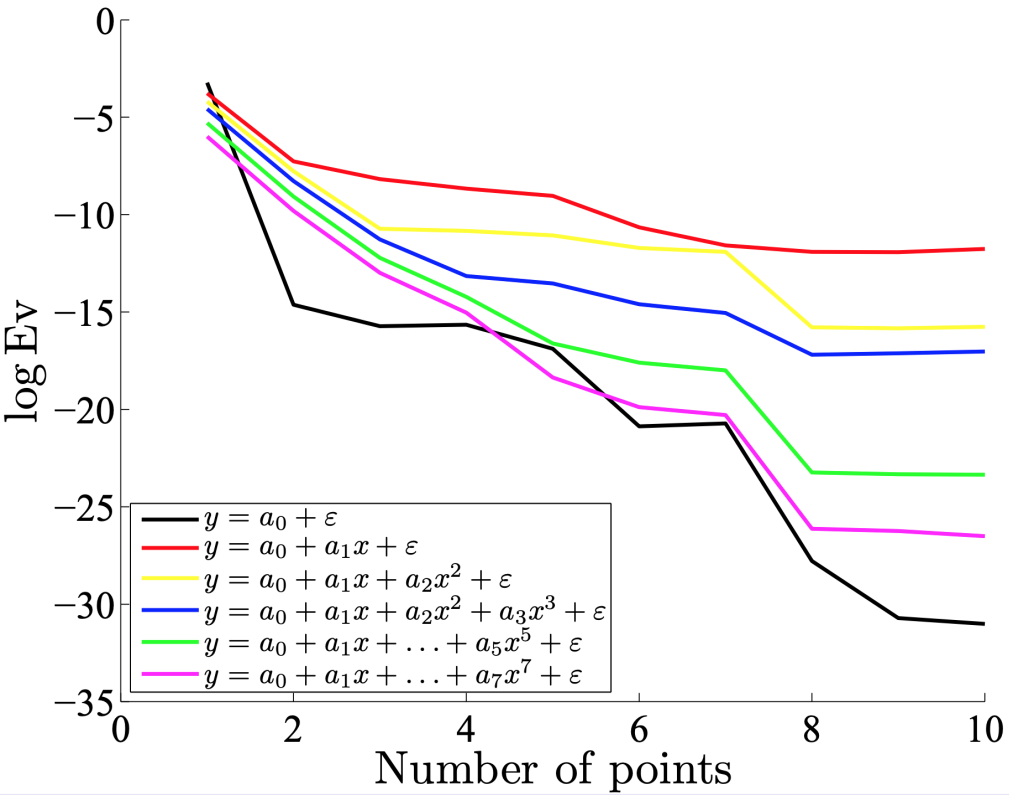
– Ошибка при различных объемах выборки

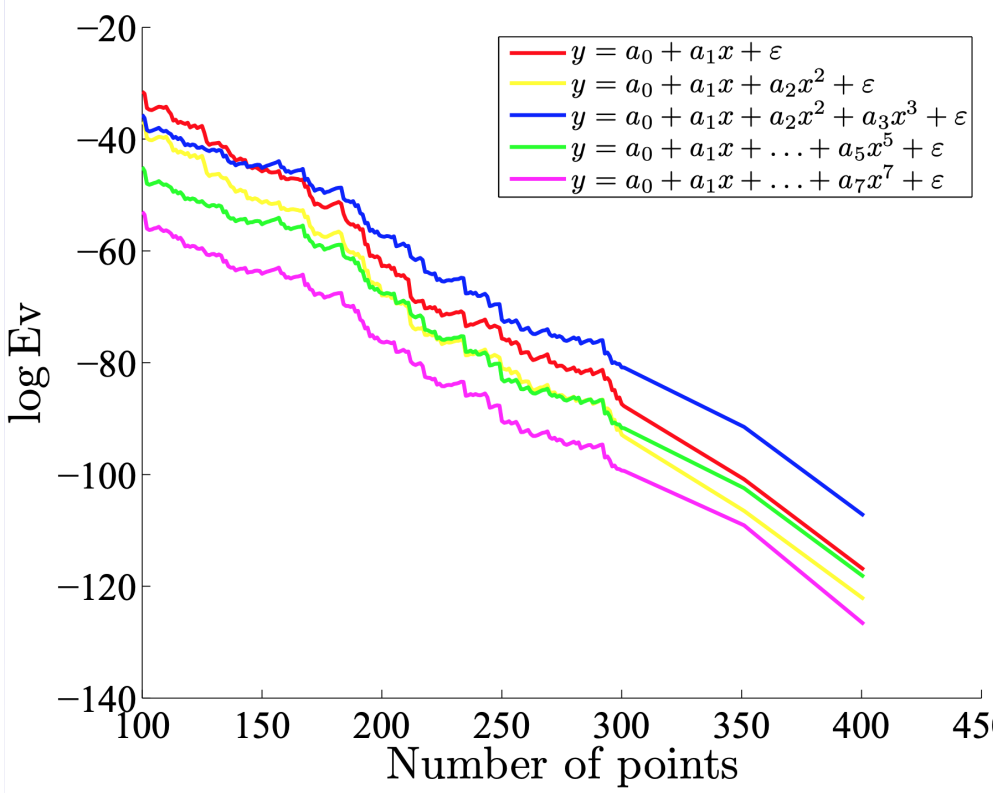


– Дисперсия ошибки при различных объемах выборки

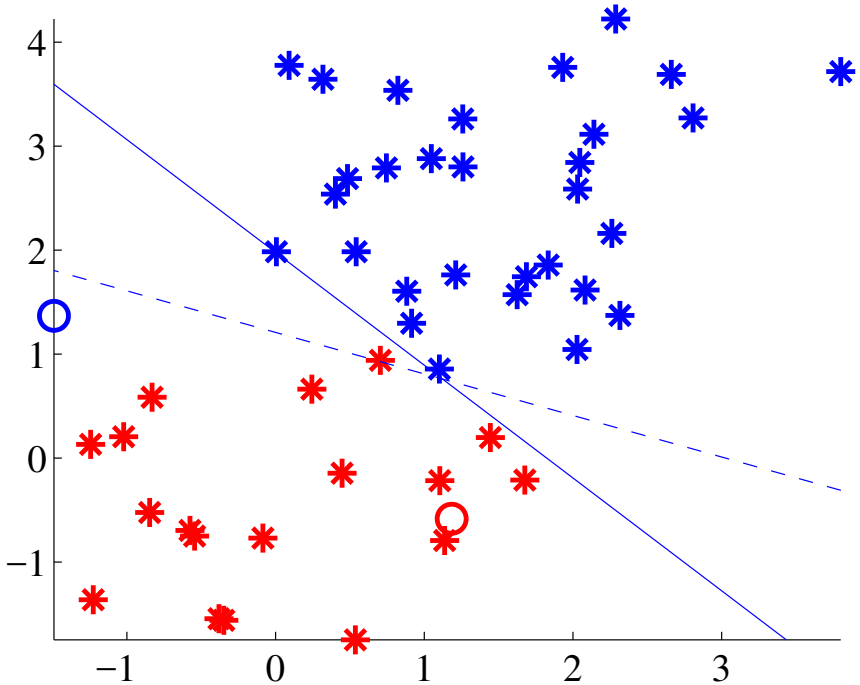


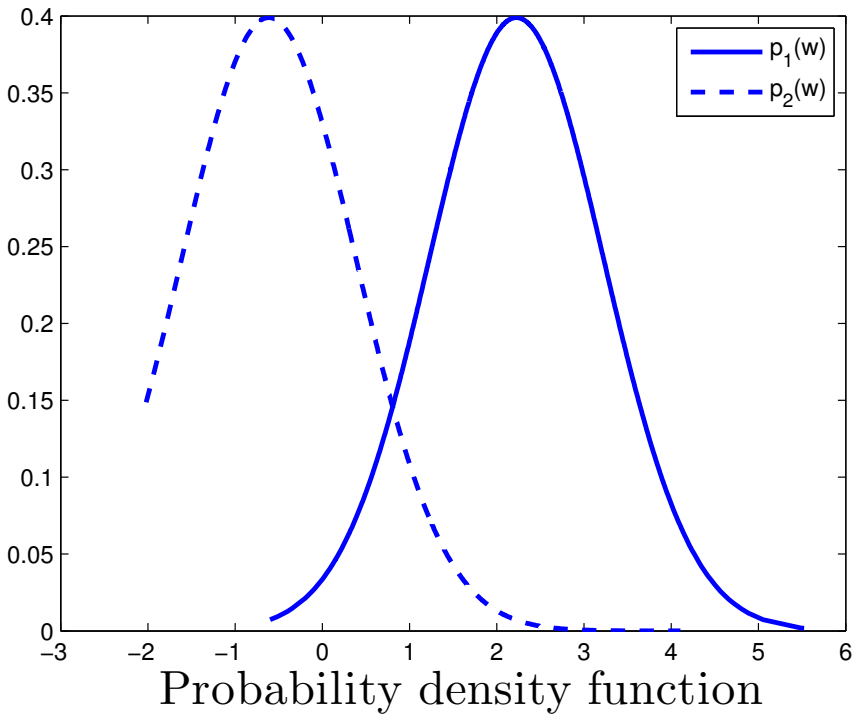
Правдоподобие модели при различных объемах выборки





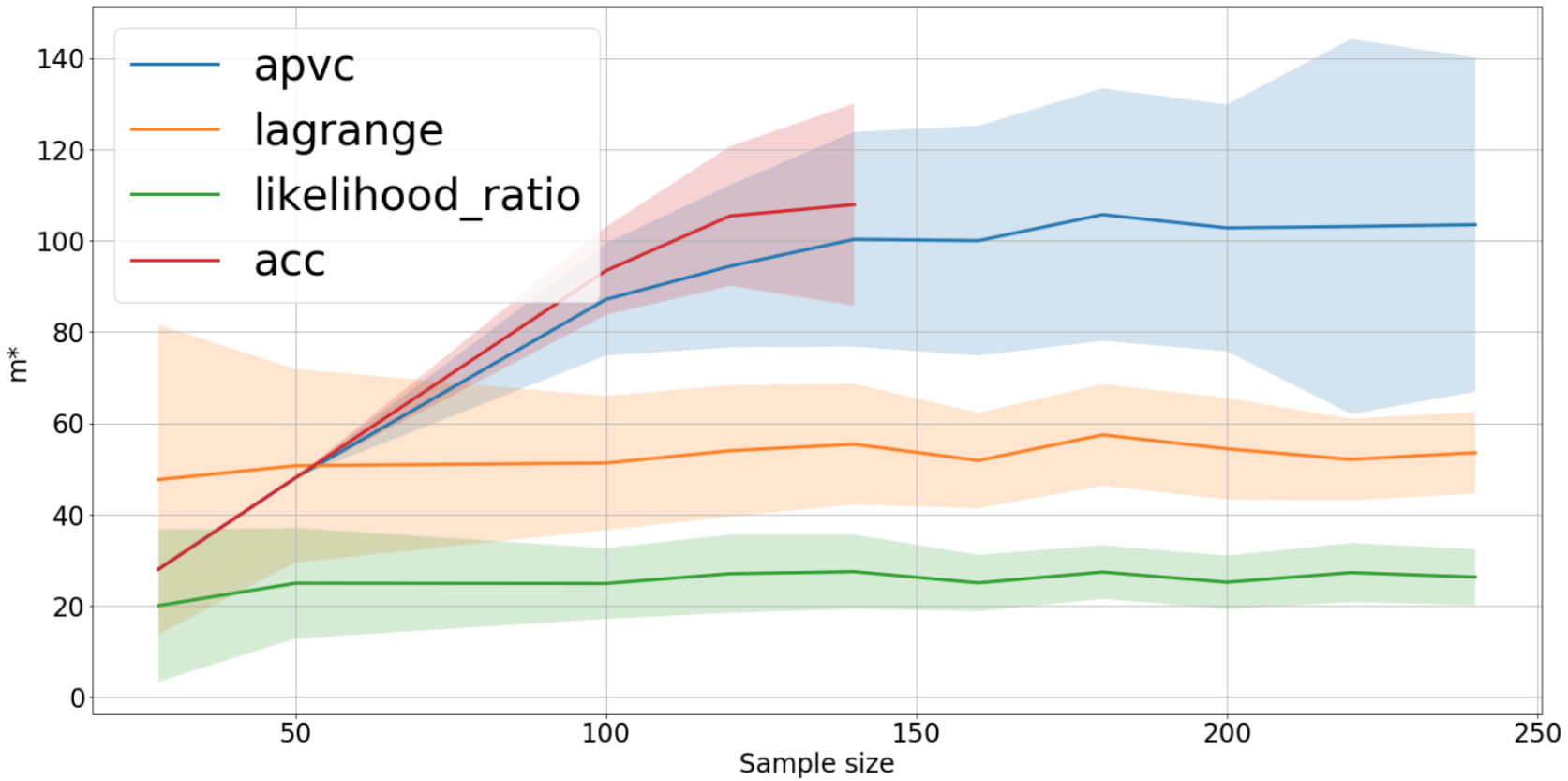
Изменение эмпирического распределения параметров





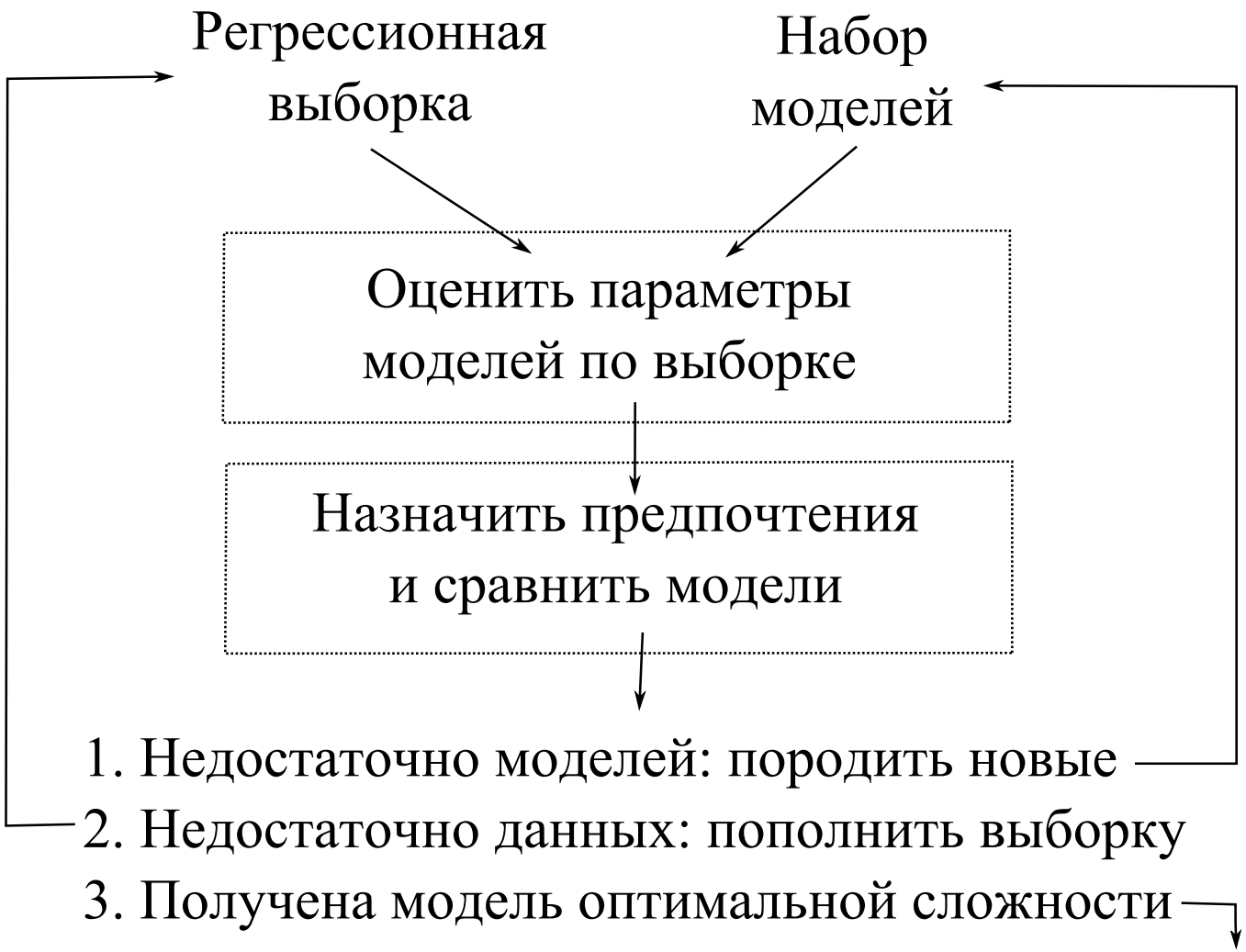
Объем выборки m^* из распределения P достаточен, если выборки X_1, X_2 размера $m > m^*$ из P схожи согласно функции сходства $D(\hat{P}_1, \hat{P}_2)$ между эмпирическими распределениями, полученными на этих выборках.

Объем выборки, спрогнозированной на раннем этапе сбора данных



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

Процедура выбора моделей и пополнения выборки



Последовательный выбор моделей:

точность, сложность, устойчивость