My first scientific paper

Week 7

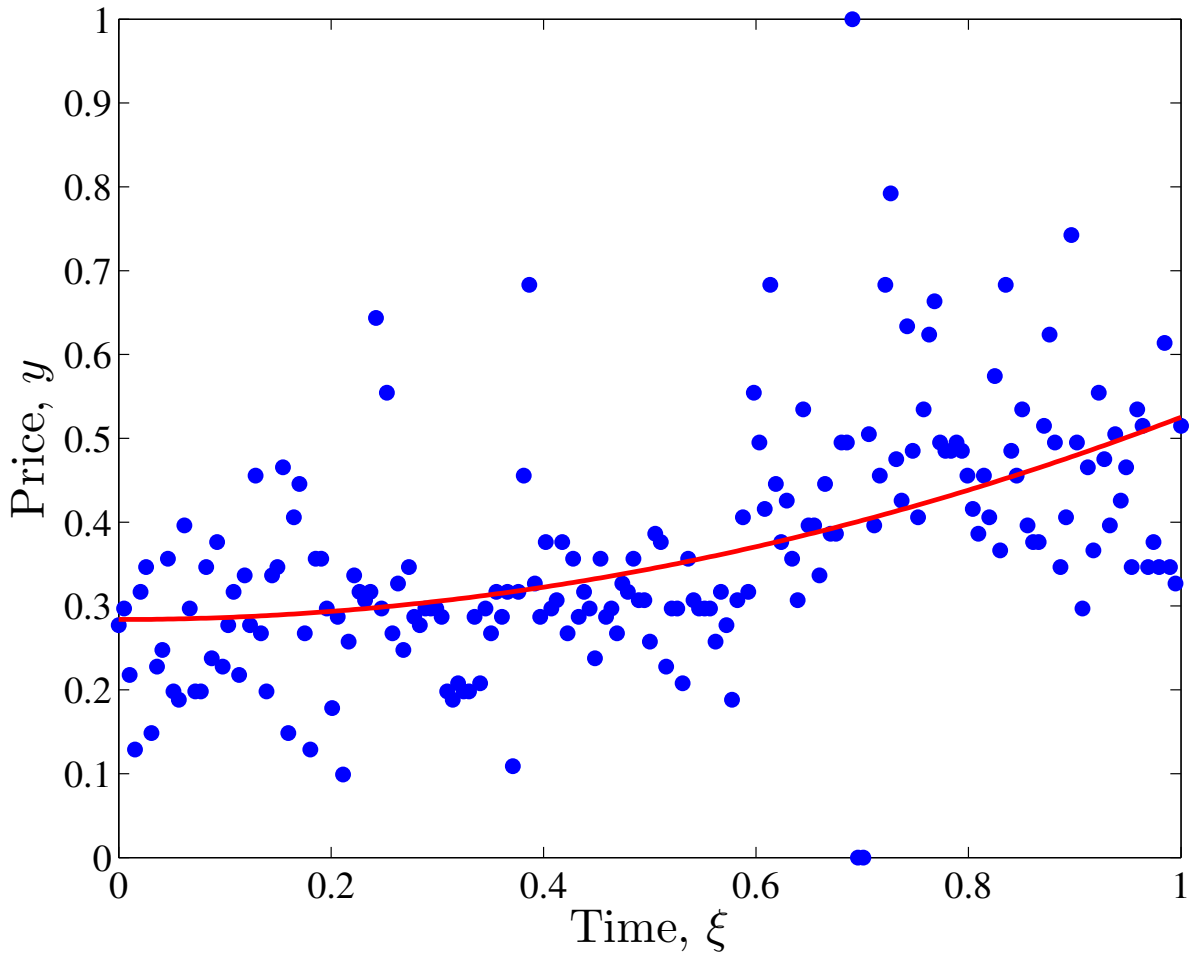
Analyse the error

Vadim Strijov

Moscow Institute of Physics and Technology

2022

Regression model and regression function



The regression model is $f = \mathbf{x}^{\mathsf{T}}\mathbf{w}$, where $x_1 = \xi^0, x_2 = \xi^2$.

The regression function: $y = w_1 + w_2\xi^2 + \varepsilon(\xi)$

with the optimal parameters $\mathbf{w}_0 = [0.2839, 0.2412]^{\mathsf{T}}$.

The neural network: $f = \sigma'\left(\mathbf{w}^{\mathsf{T}}\boldsymbol{\sigma}\left(\mathbf{W}\mathbf{x} + \mathbf{b}\right) + b'\right)$.

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:

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

Probabilistic model

Call the (approximation) model f the parametric family

$$f = f(\mathbf{w}, \mathbf{x}).$$

Call the residue $\varepsilon = f - y$. Assume, for example,

$$\varepsilon \sim \mathcal{N}\left(f, \frac{1}{\beta}\right), \quad \mathbf{w} \sim \mathcal{N}\left(\hat{\mathbf{w}}, \frac{1}{\alpha}\right).$$

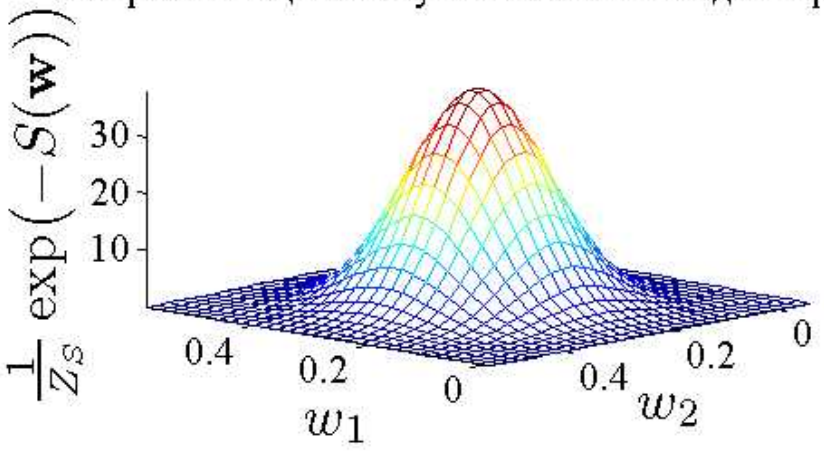
Call the probabilistic model the distribution

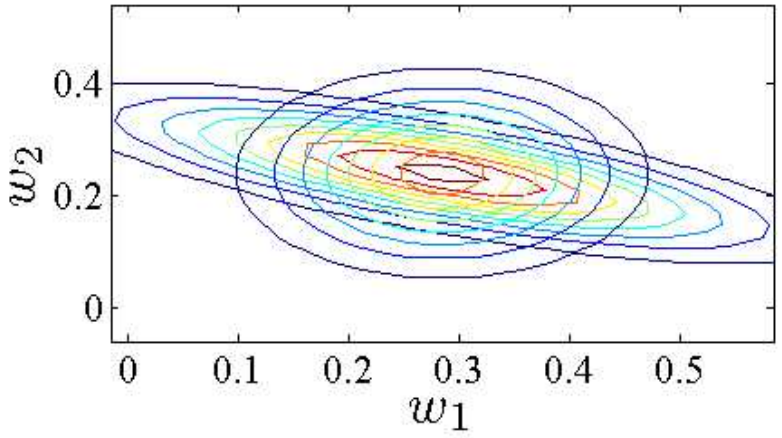
$$p(f|\mathbf{x}, \mathbf{w}).$$

The forecast is the expected value (at some point $\mathbf{x}_0$)

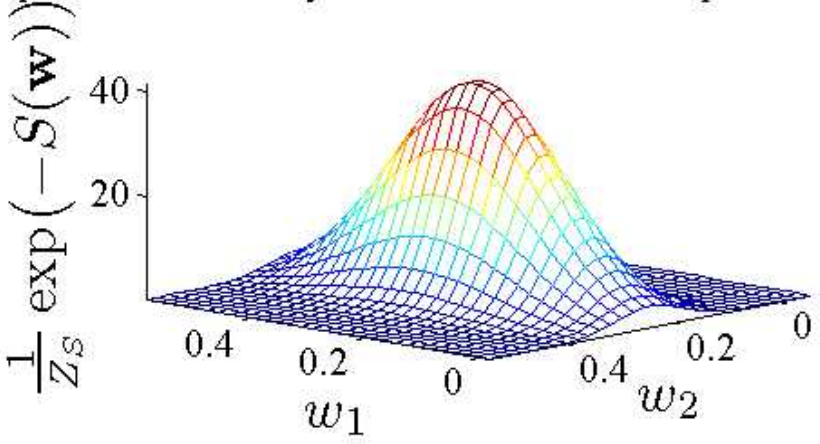
$$\mathrm{E}(f|\mathbf{x}_0, \hat{\mathbf{w}}).$$

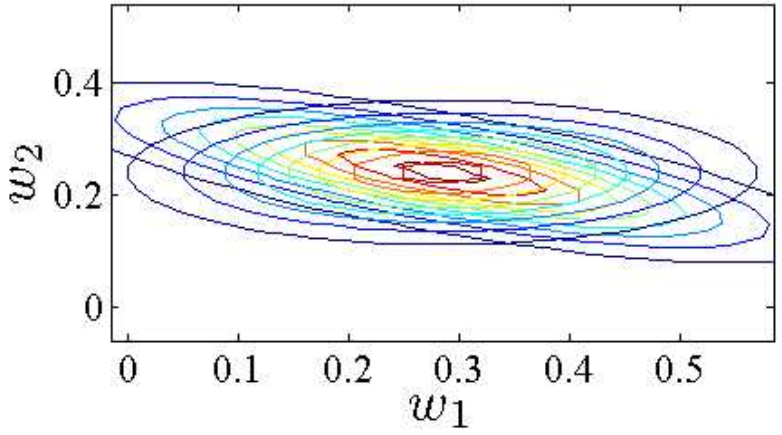
Аппроксимации в случае постоянной дисперсии



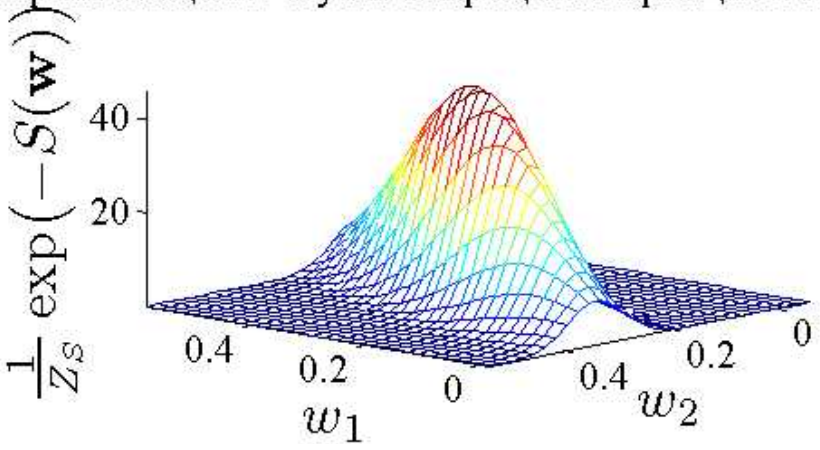


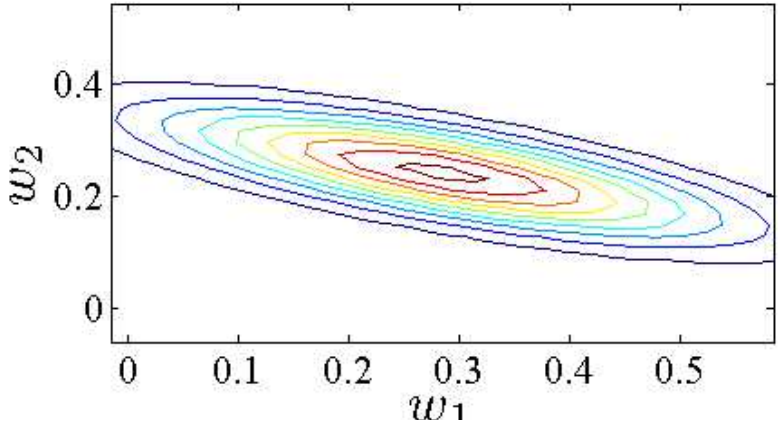
проксимации в случае диагональной матрицы ковариаций



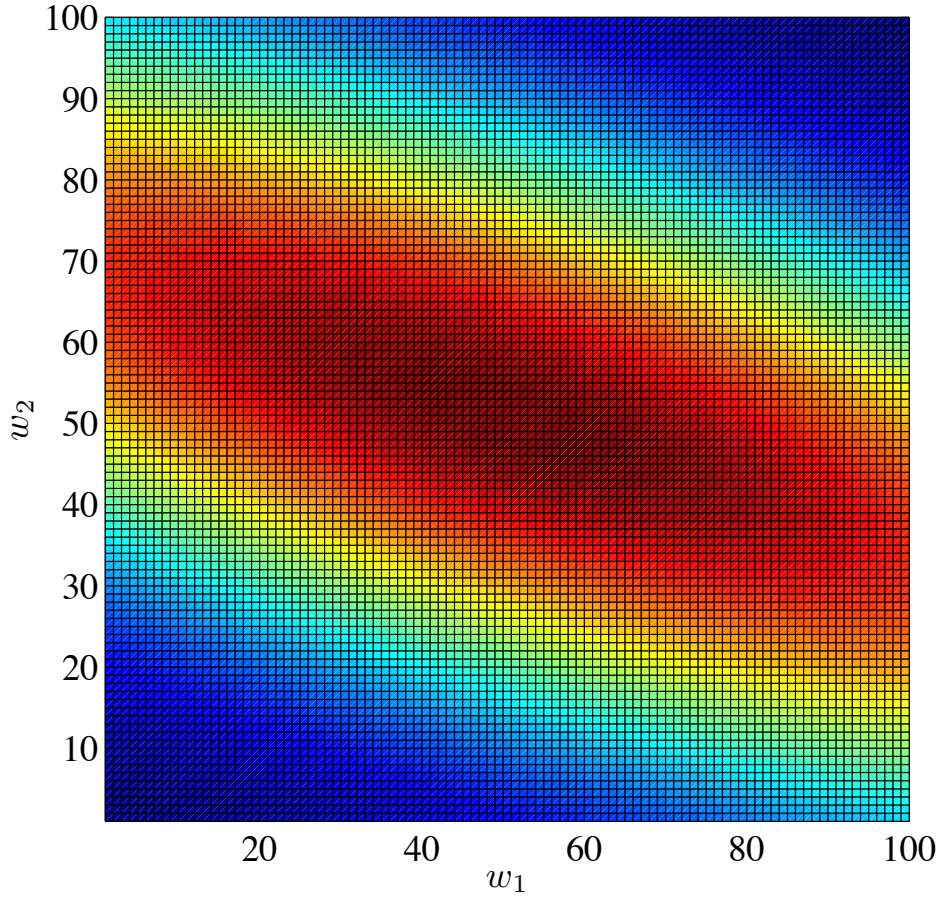


проксимации в случае матрицы ковариаций общего вида





Пространство параметров модели



Сэмплирование параметров модели полным перебором. Цветом обозначено значение функциии $\exp\bigl(-S(\mathbf{w})\bigr)$.

Гипотеза порождения данных для линейной модели

Пусть $\mathbb{E}(\mathbf{y}|X) = \mathbf{f}$ и многомерная случайная величина имеет нормальное распределение

$$p(\mathbf{y}) = (2\pi)^{-\frac{m}{2}} \det{}^{-\frac{1}{2}} \left(B^{-1}\right) \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{f})^T B(\mathbf{y}-\mathbf{f})\right).$$

Рассмотрим три варианта. Элементы вектора $\mathbf{y}$ имеют

1) одинаковую дисперсию и независимы, $\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_l) = 0, i \neq l$,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \beta^{-1} I),$$

2) имеют различную дисперсию и независимы,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \mathrm{diag}(\beta_1, \ldots, \beta_m)^{-1} I)$$

3) описываются ковариационной матрицей общего вида,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, B^{-1});$$

эта матрица симметрична и положительно определена.

Функция правдоподобия данных

Функция вероятности появления зависимой переменной имеет вид

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, B, f) \stackrel{\text{def}}{=} p(D|\mathbf{w}, \beta, f) = \frac{\exp(-E_D)}{Z_D(B)}.$$

Функция ошибки, соответствующая математическому ожиданию регрессионной модели при данной гипотезе, определена как

$$E_D = \frac{1}{2}(\mathbf{y} - \mathbf{f})^T B(\mathbf{y} - \mathbf{f}).$$

Коэффициент Z_D определен выражением, нормирующим функцию плотности нормального распределения

$$Z_D(B) = (2\pi)^{\frac{m}{2}} \det{}^{\frac{1}{2}} \left(B^{-1}\right).$$

Функция правдоподобия данных при $B = \beta I$

Для гомоскедаксичного случая функция ошибки равна

$$E_D = \frac{1}{2}\beta \sum_{i \in \mathcal{I}} (y_i - f(\mathbf{w}, \mathbf{x}_i))^2 ,$$

а нормирующий множитель

$$Z_D(\beta) = \left(\frac{2\pi}{\beta}\right)^{\frac{m}{2}} .$$

Априорное (sic!) распределение параметров модели

Из принятой гипотезы порождения данных следует нормальность распределения параметров, $\mathbf{w} \sim \mathcal{N}(\mathbf{w_0}, A^{-1})$:

$$p(\mathbf{w}|A, f) = \frac{\exp(-E_\mathbf{w})}{Z_\mathbf{w}(A)}.$$

Функция-штраф за большое значение параметров модели для принятого распределения определена как

$$E_\mathbf{w} = \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T A(\mathbf{w} - \mathbf{w}_0).$$

Нормирующая константа $Z_\mathbf{w}$ равна

$$Z_\mathbf{w}(A) = (2\pi)^{\frac{n}{2}} \det{}^{\frac{1}{2}}(A^{-1}).$$

При равенстве дисперсий элементов вектора параметров

$$Z_\mathbf{w}(\alpha) = \left(\frac{2\pi}{\alpha}\right)^{\frac{m}{2}} \quad \text{и} \quad E_\mathbf{w} = \frac{1}{2}\alpha\|\mathbf{w}\|^2.$$

Байесовский вывод, первый уровень

Апостериорное распределение параметров модели для заданных матриц A, B имеет вид

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

Элементы этого выражения и соответствующие им параметры:

- $p(\mathbf{w}|D, A, B, f)$ — апостериорное распределение параметров,
- $\mathbf{w}_{\text{MP}} = \arg\max p(\mathbf{w}|D, A, B, f)$ — наиболее вероятные параметры,
- $p(D|\mathbf{w}, B, f)$ — функция правдоподобия данных,
- $\mathbf{w}_{\text{ML}} = \arg\max p(D|\mathbf{w}, B, f)$ — наиболее правдоподобные параметры,
- $p(\mathbf{w}|A, f)$ — априорное распределение параметров,
- $p(D|A, B, f)$ — функция правдоподобия модели.

Апостериорное распределение параметров, частный случай

Апостериорное распределение параметров модели для заданных матриц A, B

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

Записывая функцию ошибки $S = E_{\mathbf{w}} + E_D$ в виде

$$S(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T A(\mathbf{w} - \mathbf{w}_0) + \frac{1}{2}(\mathbf{y} - \mathbf{f})^T B(\mathbf{y} - \mathbf{f}),$$

получаем вместо вышестоящего выражение

$$p(\mathbf{w}|D, A, B, f) \propto \frac{\exp(-S(\mathbf{w}))}{Z_S},$$

где Z_S — нормирующий множитель.

Апостериорное распределение параметров, частный случай

При рассмотрении частных случаев ковариационных матриц $B = \beta I_m$ и $A = \alpha I_n$ и при $\mathbf{w}_0 = \mathbf{0}$ апостериорное распределение параметров принимает вид

$$p(\mathbf{w}|D, \alpha, \beta, f) = \frac{p(D|\mathbf{w}, \beta, f)p(\mathbf{w}|\alpha, f)}{p(D|\alpha, \beta, f)}.$$

а функция ошибки —

$$S(\mathbf{w}) = \frac{1}{2}\alpha\|\mathbf{w}\|^2 + \frac{1}{2}\beta\|\mathbf{y} - \mathbf{f}\|^2.$$

Параметры α и β в последнем выражении играют роль регуляризирующих множителей.

Функция ошибки включает две матрицы ковариации

Согласно первому уровню Байесовского вывода

$$S(\mathbf{w}|D, f) = \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MP}})^{\mathsf{T}} A(\mathbf{w} - \mathbf{w}_{\text{MP}}) + \frac{1}{2}(\mathbf{f} - \mathbf{y})^{\mathsf{T}} B(\mathbf{f} - \mathbf{y}).$$

Имеется девять возможных вариантов гипотезы порождения данных.

Обратная ковариационная матрица параметров	Обратная ковариационная матрица зависимой переменной
$A = \alpha I_n$	$B = \beta I_m$
$A = \text{diag}(\alpha_1, \ldots, \alpha_n)$	$B = \text{diag}(\beta_1, \ldots, \beta_m)$
A	B

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:

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

Зависимость среднего значения ошибки от объема выборки

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

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

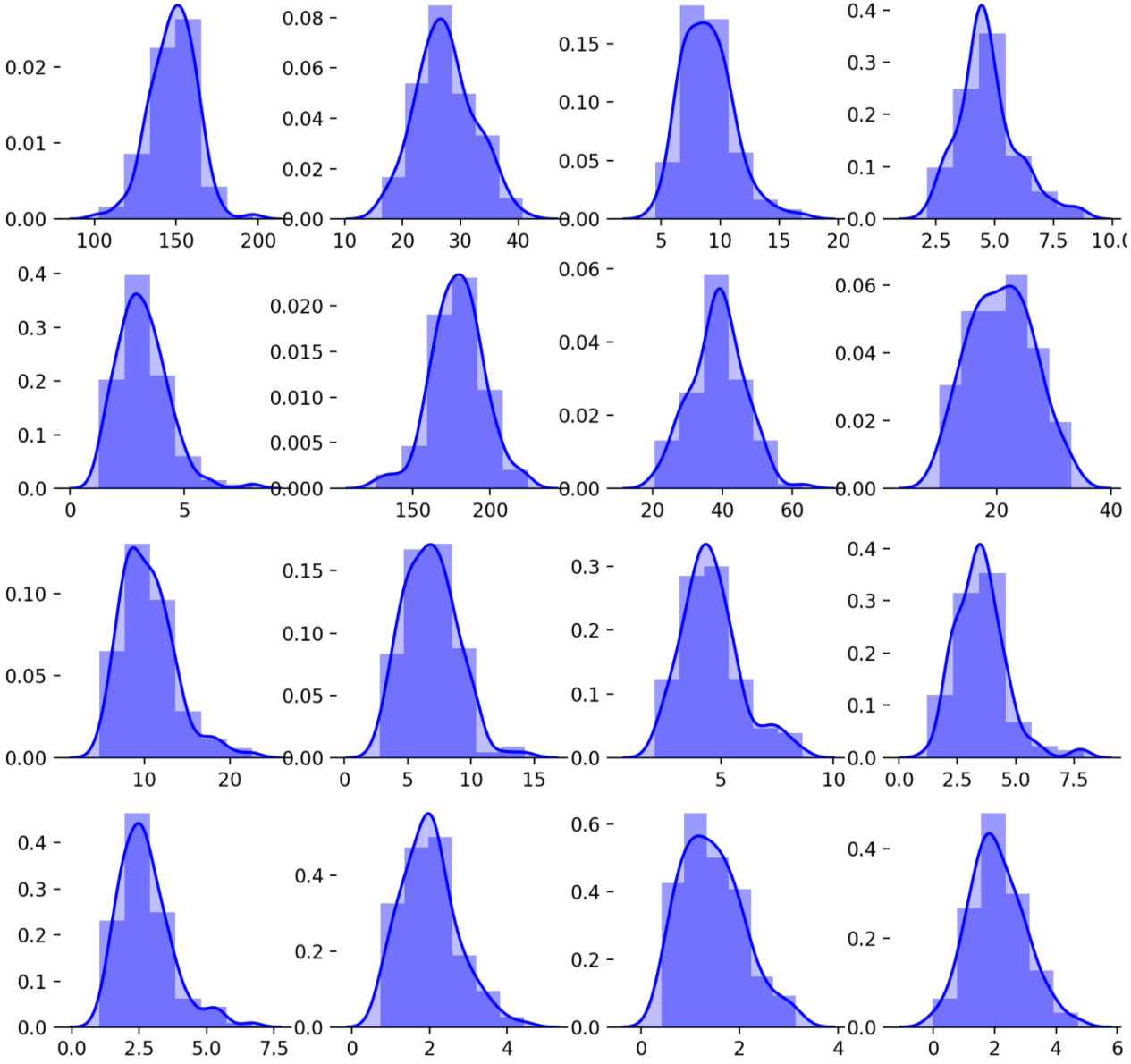
Повторим процедуру на ограниченном объеме выборки, например:

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

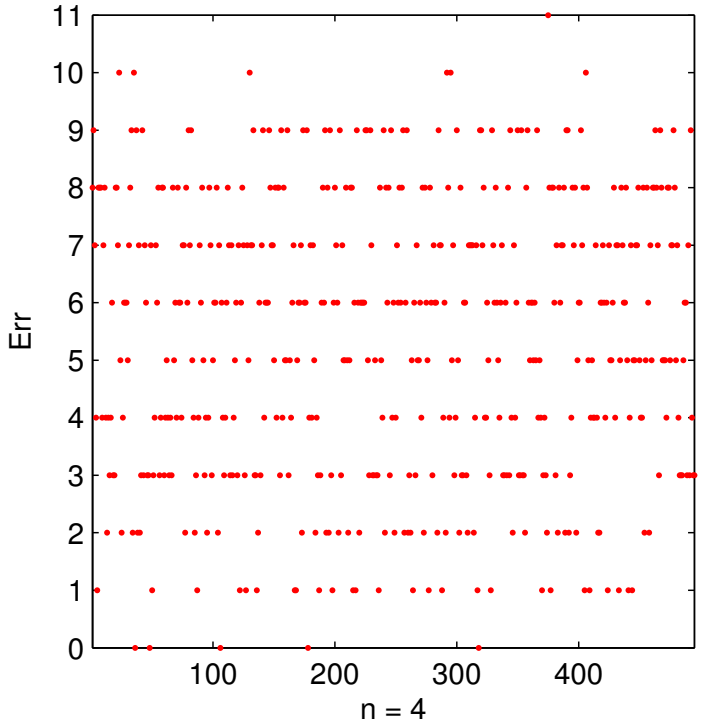
где M — наибольший объем доступной выборки.
Построим график зависимости ошибки и стандартного отклонения от объема выборки.

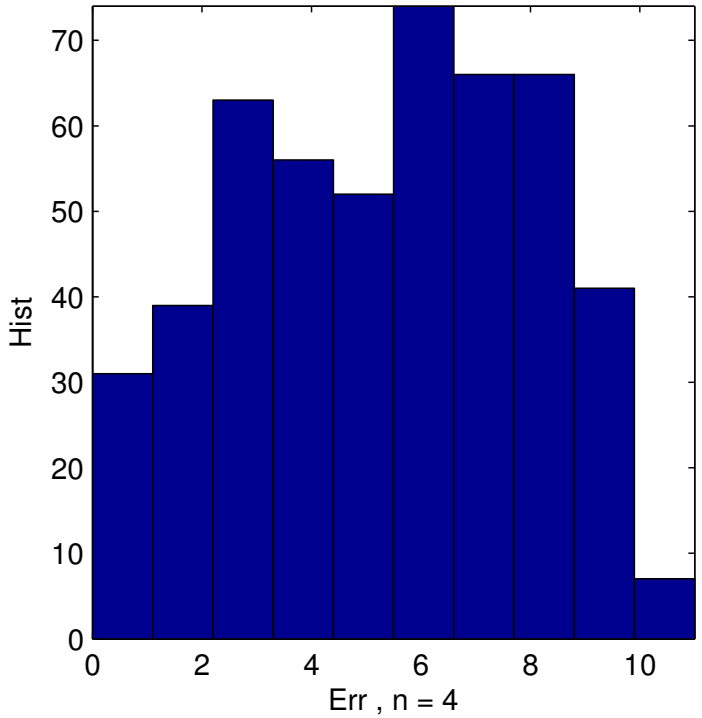
How to check the i.i.d hypothesis

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



Robustness of the classification





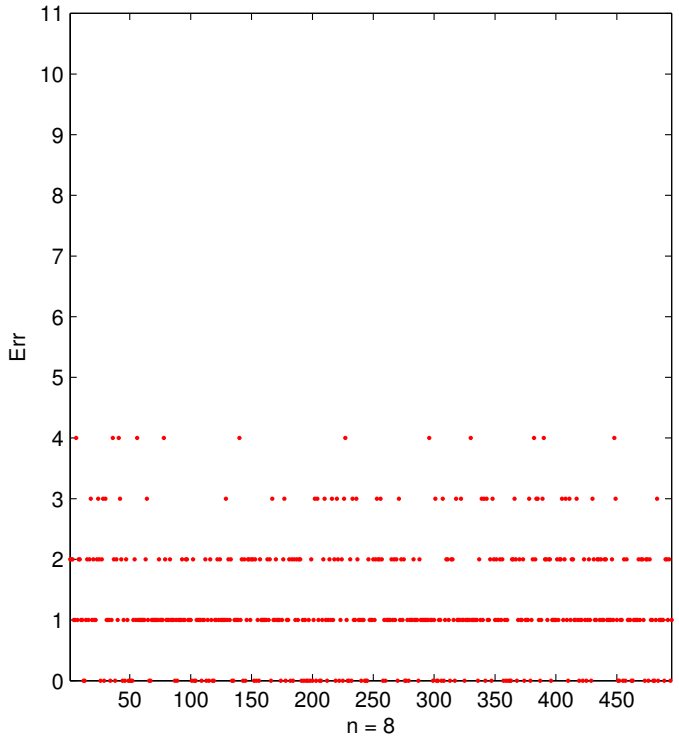
X-axis: experiment number

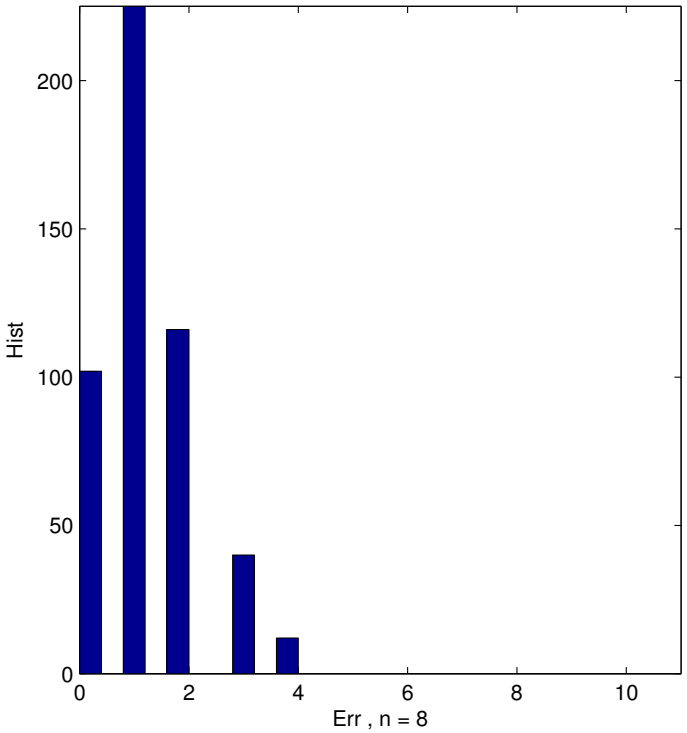
Y-axis: error, misclassified patients

X-axis: error

Y-axis: histogram of given error

Robustness of the classification

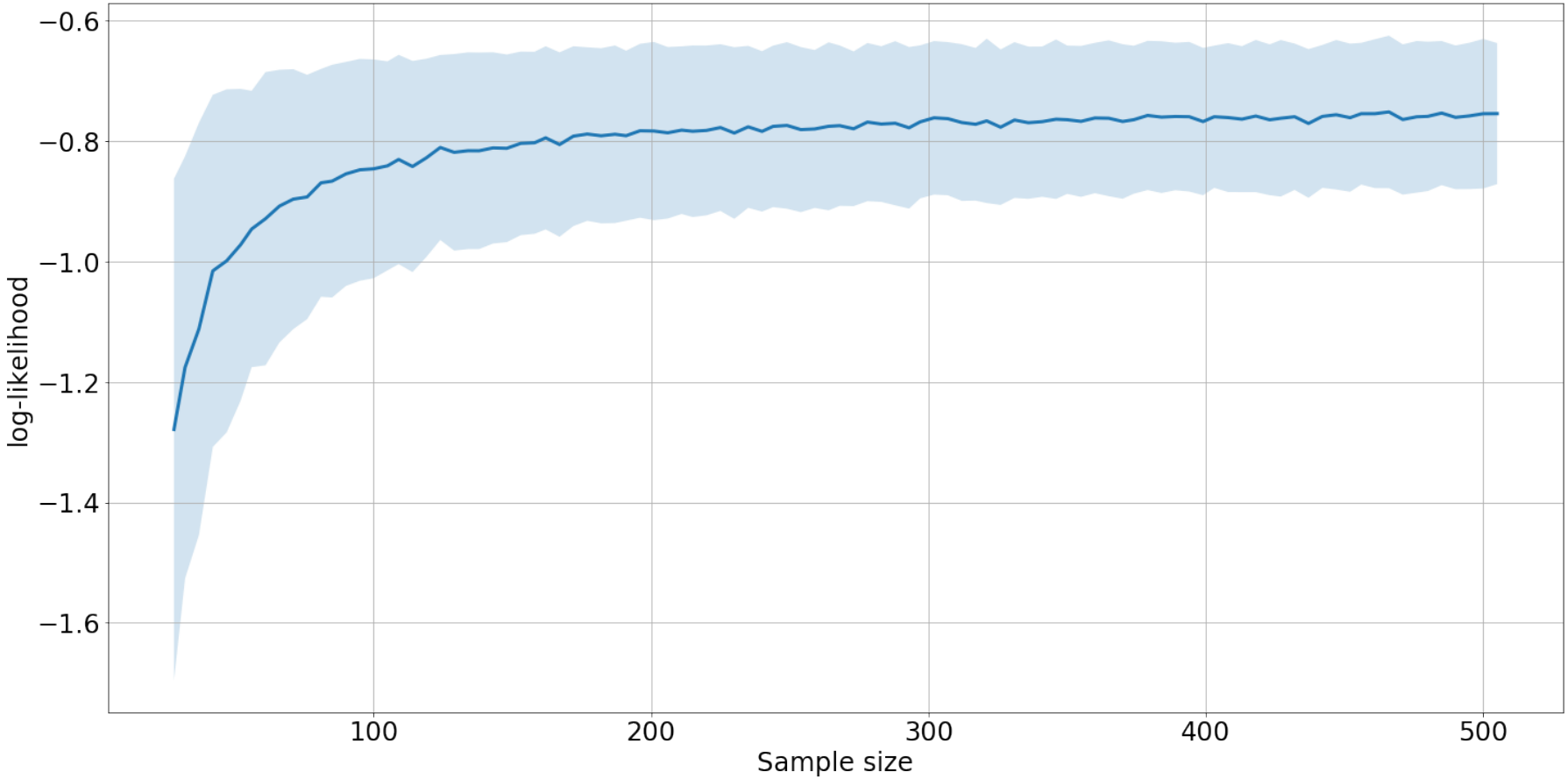




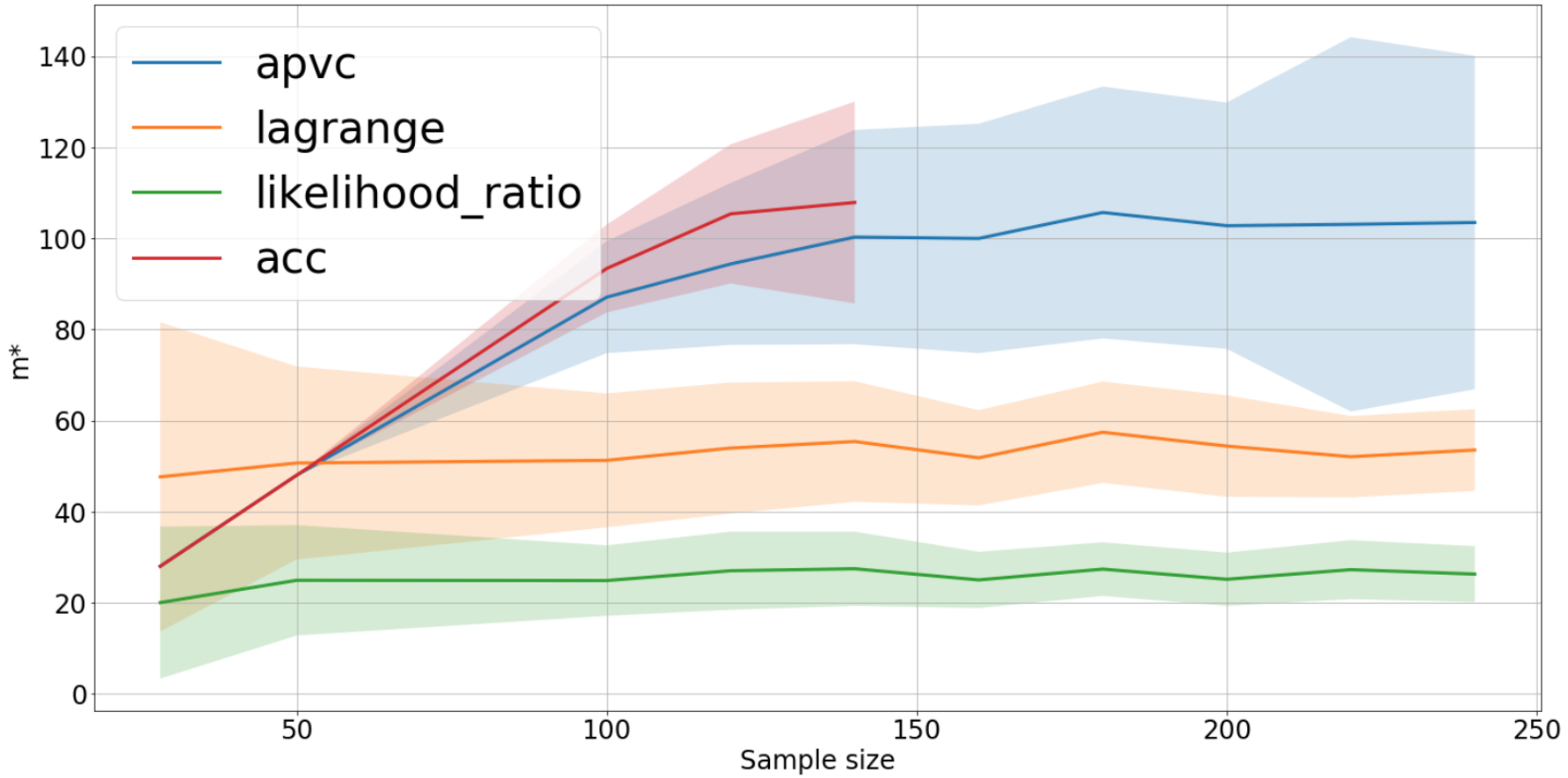
X-axis: experiment number
Y-axis: error, misclassified patients

X-axis: error
Y-axis: histogram of given error

– Error and its variance for a reinforced sample set

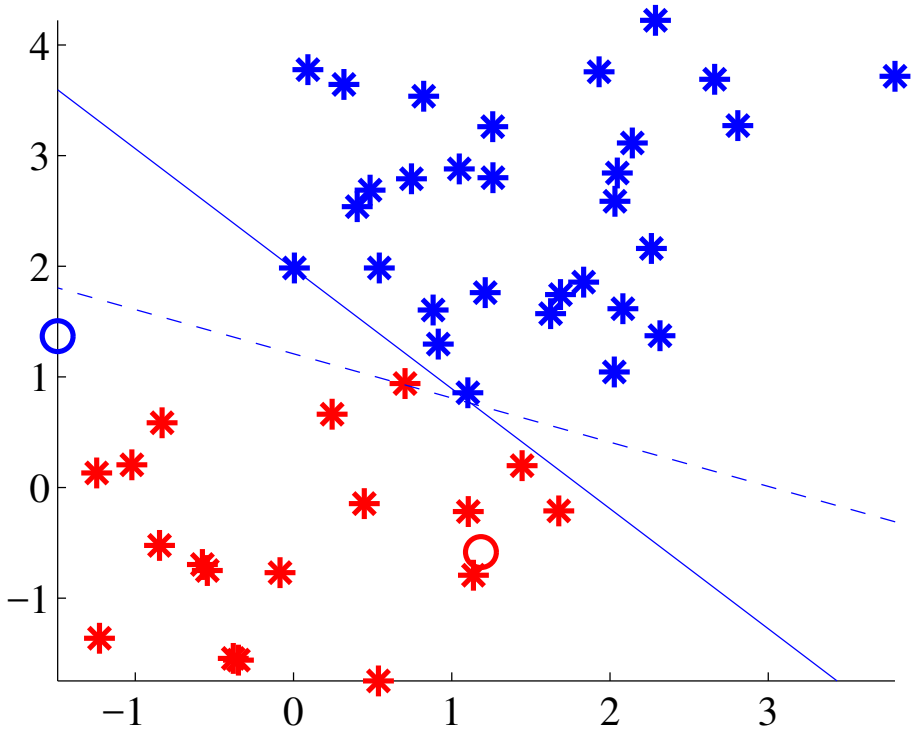


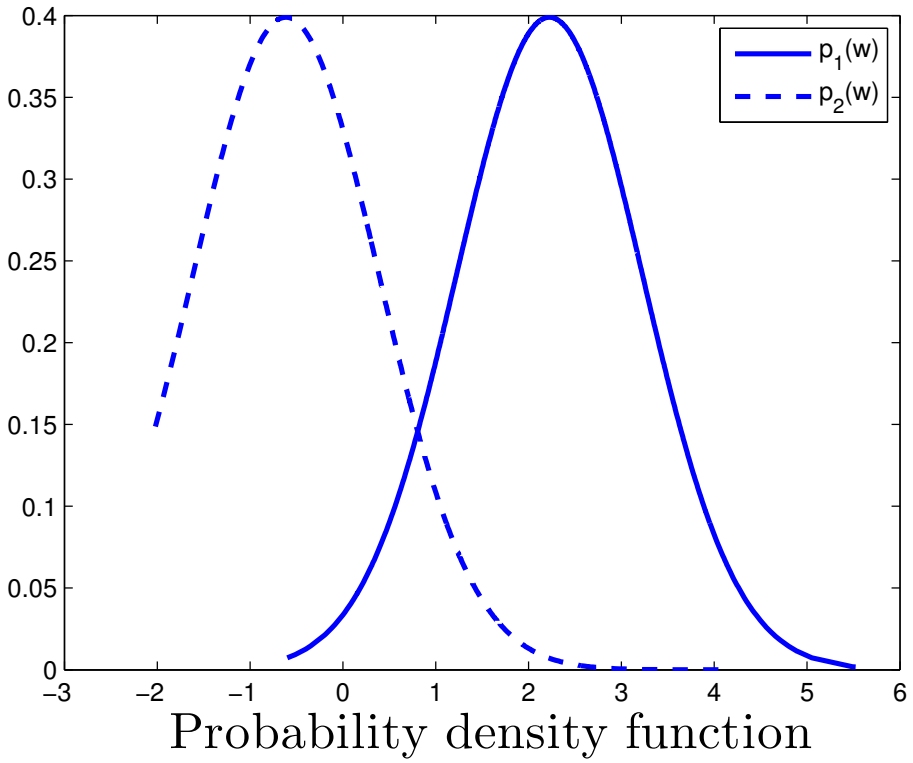
Объем выборки, спрогнозированной на раннем этапе сбора данных



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

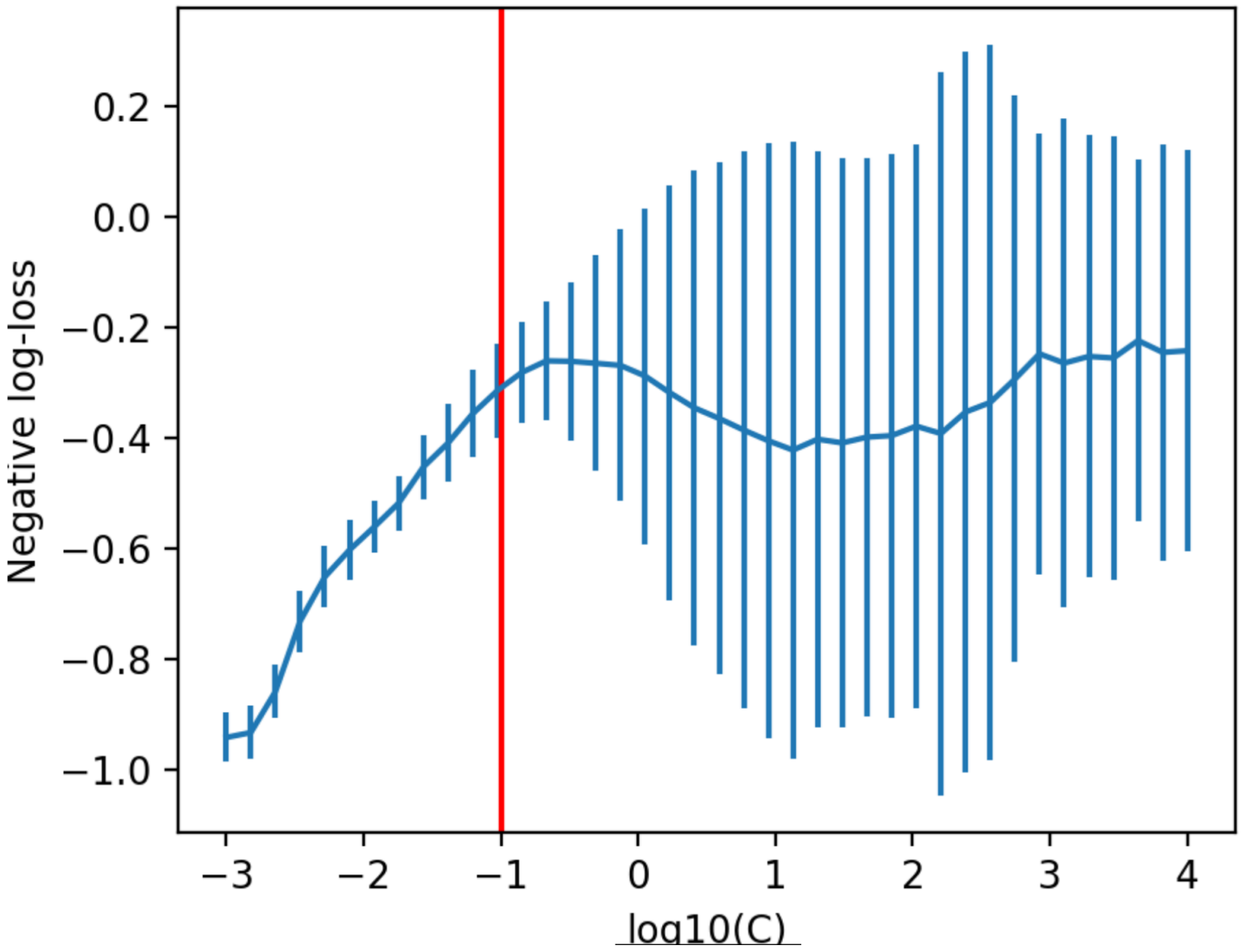
Изменение эмпирического распределения параметров



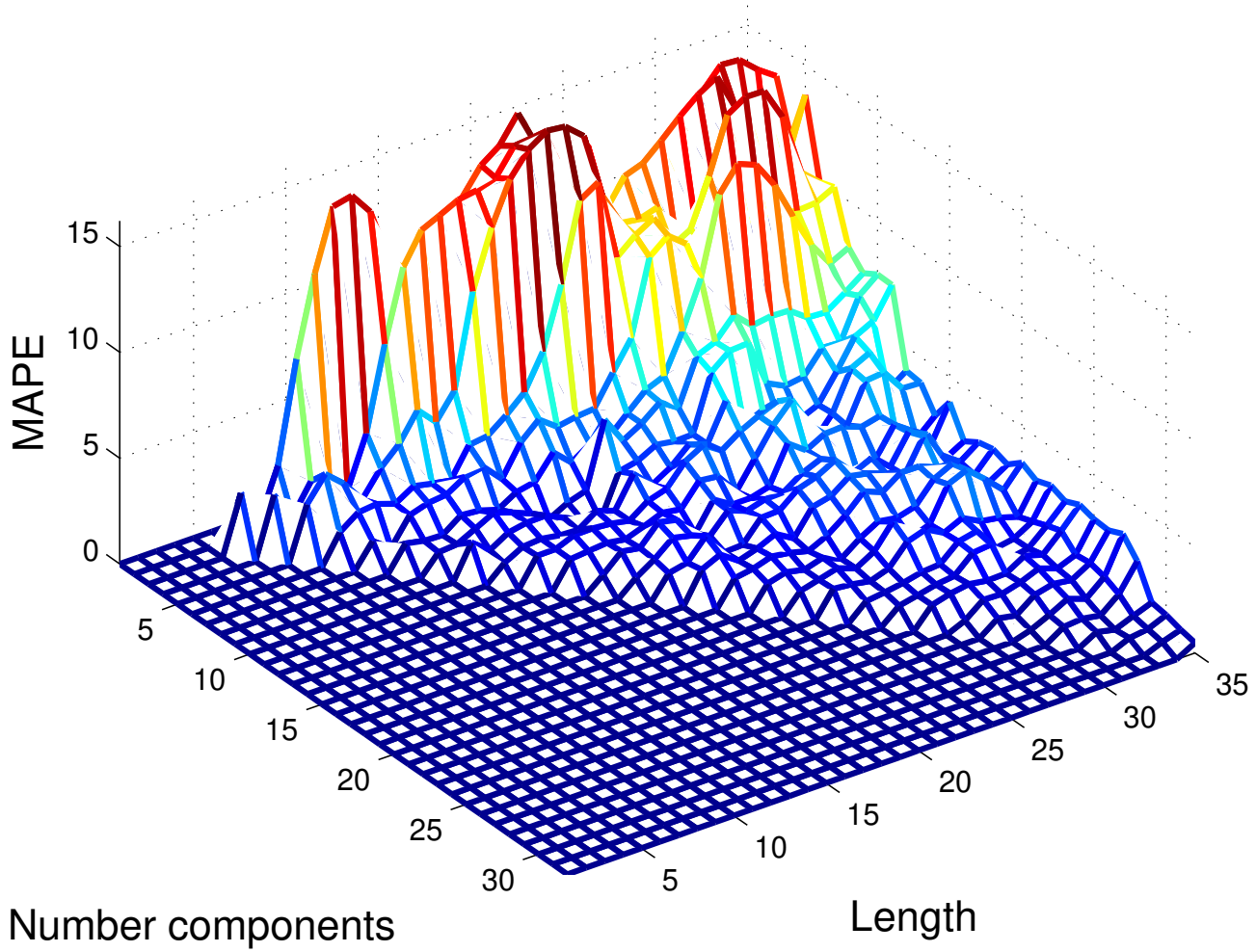


Объем выборки m^* из распределения P достаточен, если выборки X_1, X_2 размера $m > m^*$ из P схожи согласно функции сходства $D(\hat{P}_1, \hat{P}_2)$ между эмпирическими распределениями, полученными на этих выборках.

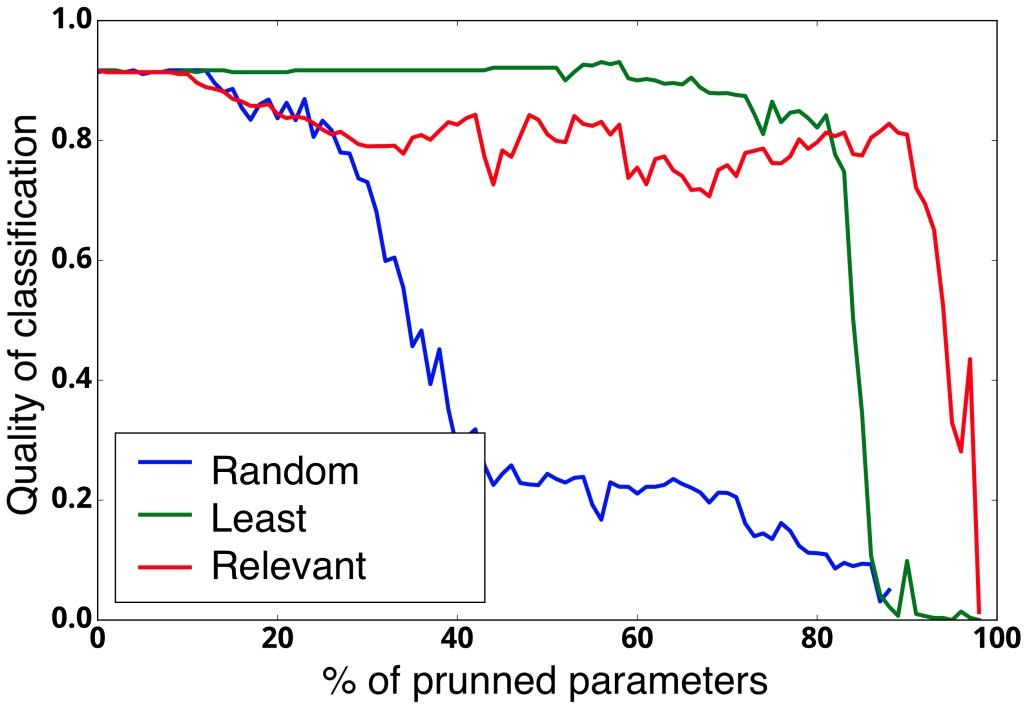
Error variance and increasing of model complexity



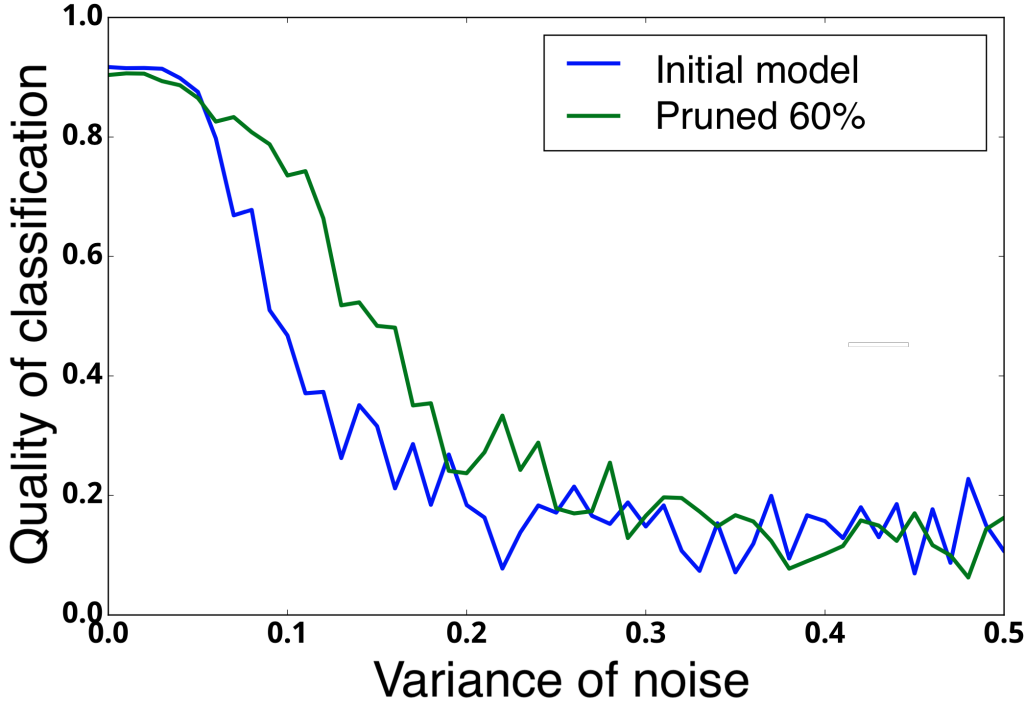
Процедура выбора моделей и пополнения выборки



The evidence of models with an excessive number of parameters **does not change significantly** when the parameters are removed



Redundancy of parameters



Stability of model

Deep learning suggests to optimise models with obviously excessive complexity.

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