My first scientific paper

Week 5

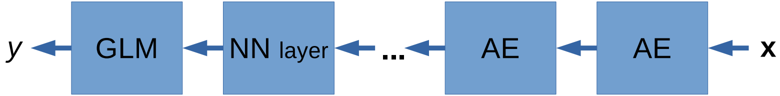
Highlight the principles

Vadim Strijov

Moscow Institute of Physics and Technology

2022

Linear model, (deep) neural net, and autoencoder



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

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

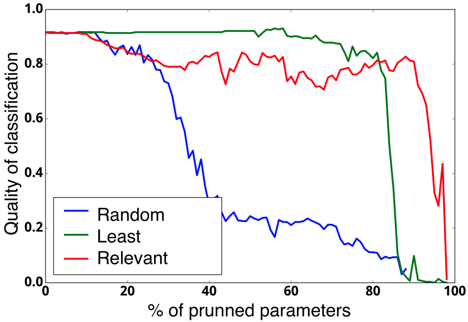
Variants

- $E_{\mathbf{x}}$ is reconstruction error
- principal component analysis: $\mathbf{W}^{\top}\mathbf{W} = \mathbf{I}_n$
- skip block: $\mathbf{W} = \mathbf{I}_n$, $\sigma = \mathrm{id}$
- classification: $\sigma = \big(1 + \exp(-\cdot)\big)^{-1}$

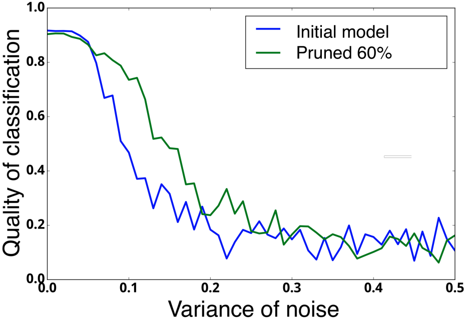
$+\mathbf{b}$

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

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

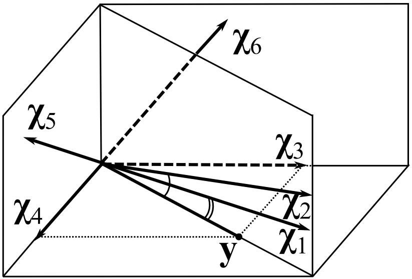
Deep learning suggests to optimise models with obviously excessive complexity.

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

Select an accurate and stable set of features

Features $\chi_1, \ldots, \chi_6$ are columns of the design matrix $\underset{3\times 6}{\mathbf{X}}$.

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



Solution: χ_3, χ_4 are orthogonal; their linear combination fits $\mathbf{y}$.

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

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

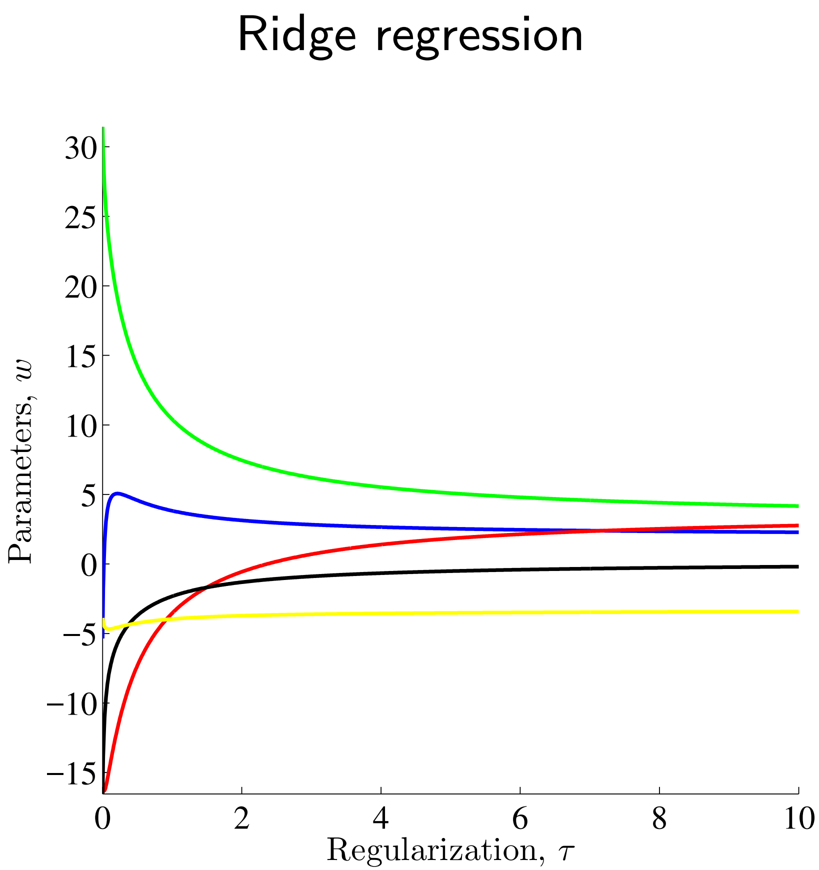
$$[a_{p,1}, \dots, a_{p,\nu}, a_{q,\nu+1}, \dots, a_{q,n}] \to \mathbf{a}'_p,$$

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

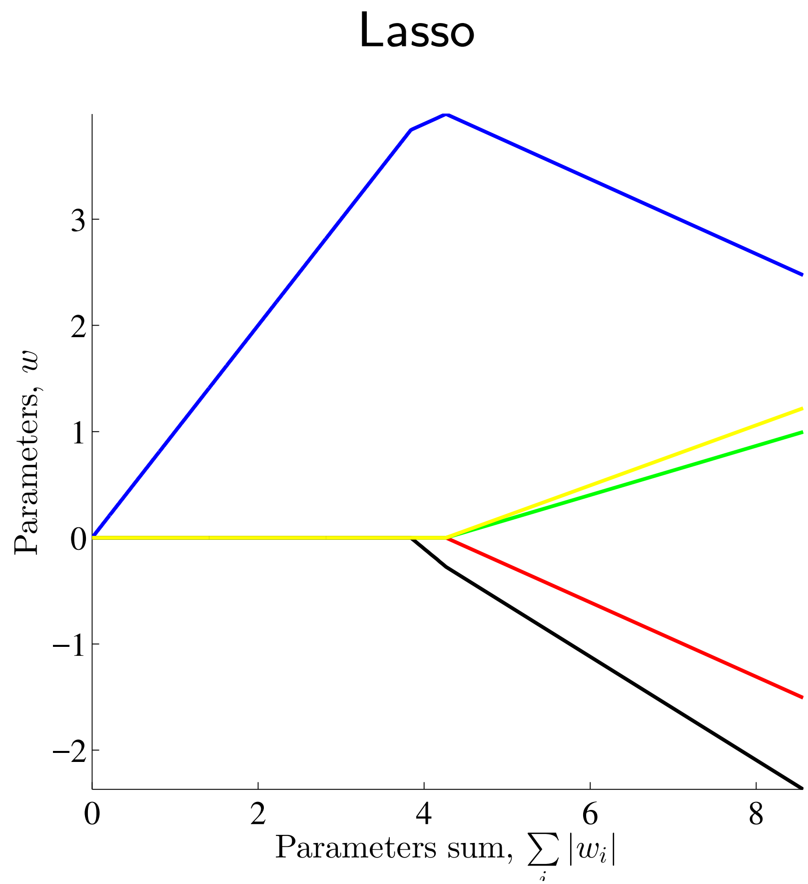
5. choose random numbers $\eta_1, \dots, \eta_Q \in \{1, \dots, n\}$;
6. replace values in positions $\eta_1, \dots, \eta_Q$ of the vectors $\mathbf{a}'_p, \mathbf{a}'_q$ for random values from $\{1, \dots, k\}$;
7. repeat items 2-6 $P/2$ times;
8. evaluate the obtained models.

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

Model parameters with regularization



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



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

Probabilistic model selection

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

Posterior — Likelihood — Prior

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

Evidence
(to select a model)

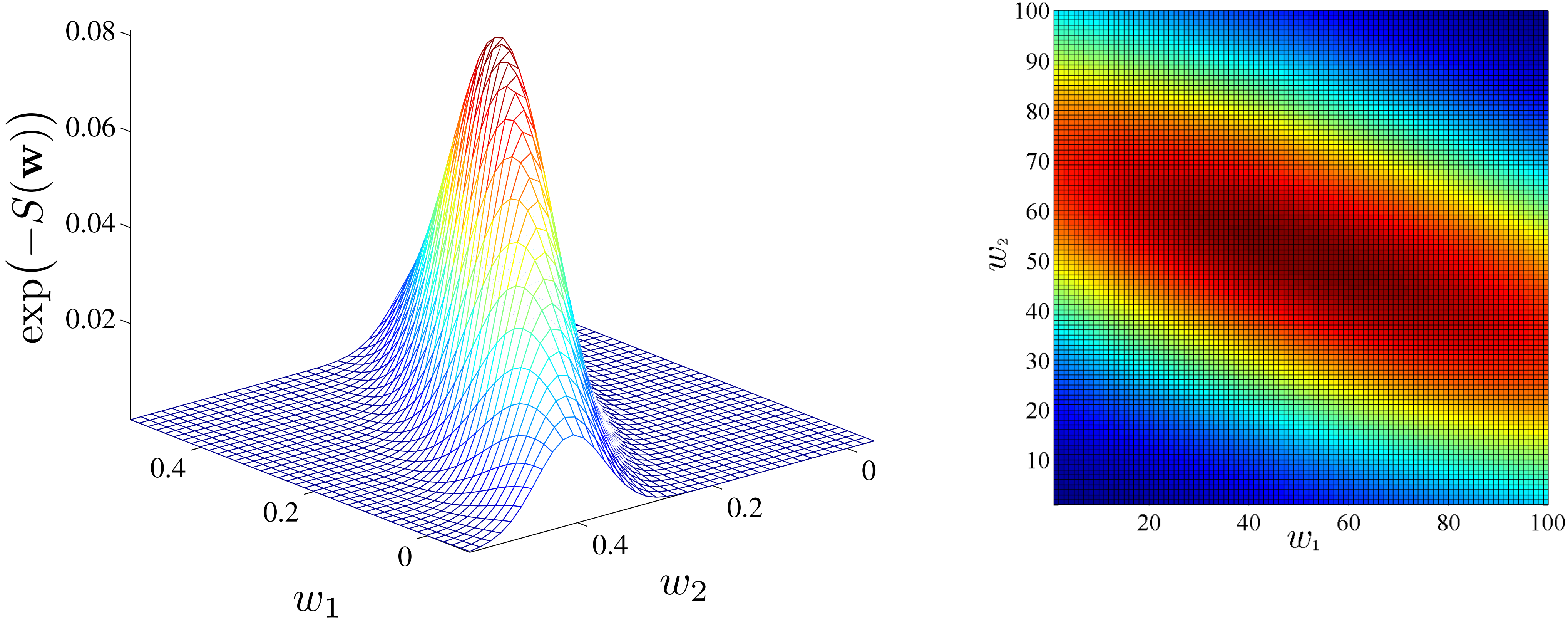
Write the error function given hyperparameters $\mathbf{A}, \mathbf{B}$

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

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

Empirical distribution of model parameters

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

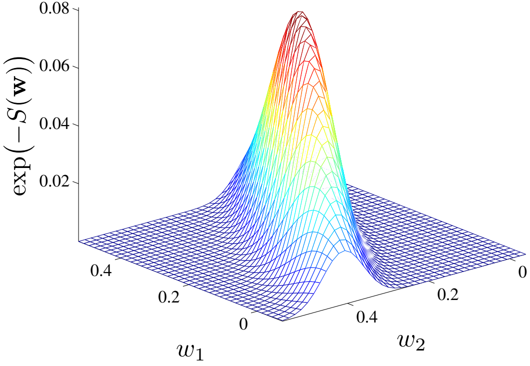


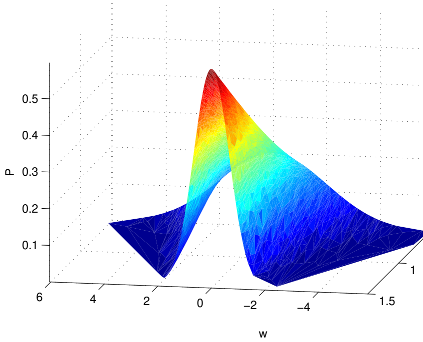
x-axis and y-axis: parameters $\mathbf{w}$, z-axis: $\exp(-S(\mathbf{w}))$

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

Empirical distribution of model parameters

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



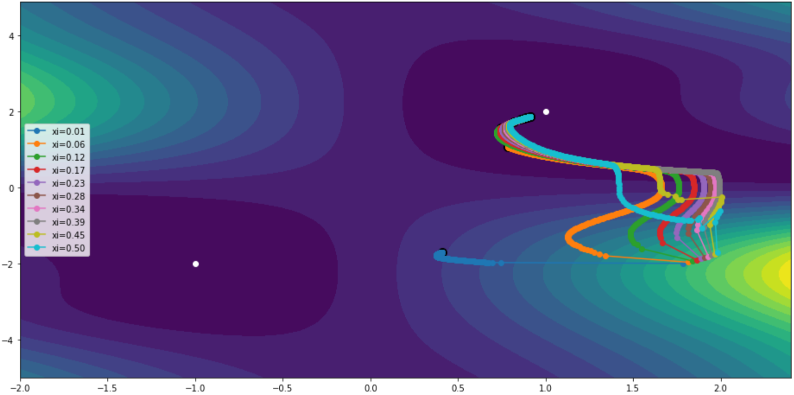


x- and y-axis: parameters $\mathbf{w}$, z-axis: $\exp(-S(\mathbf{w}))$

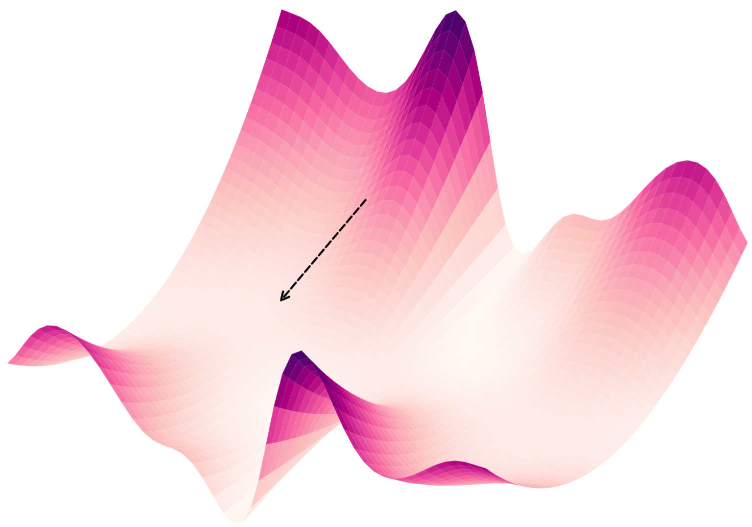
x-axis: parameters $\mathbf{w}$, y-axis: variance α, z-axis: $p(\mathbf{w}|\mathfrak{D}, \alpha)$

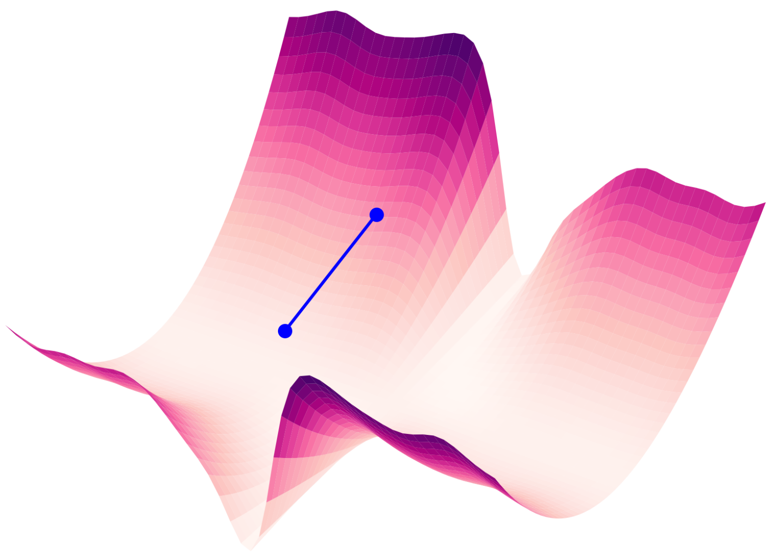
Kuznetsov, Tokmakova, Strijov. 2016. Analytic methods of structure parameter // Informatica

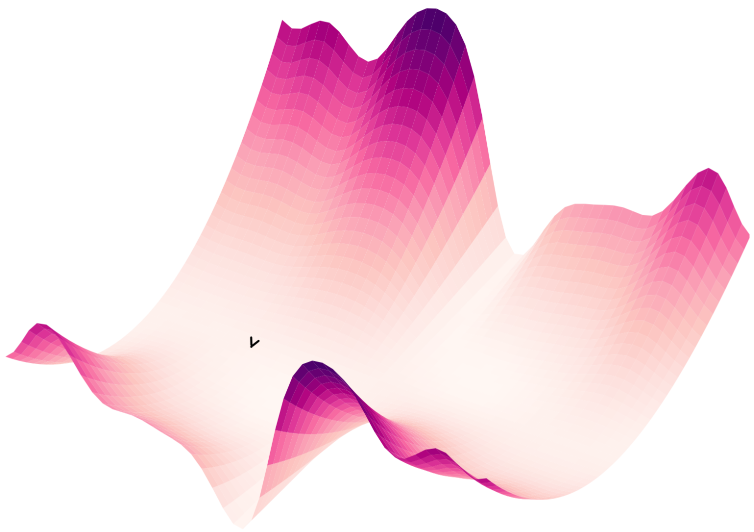
Multiple extremes, convergence and multi-start

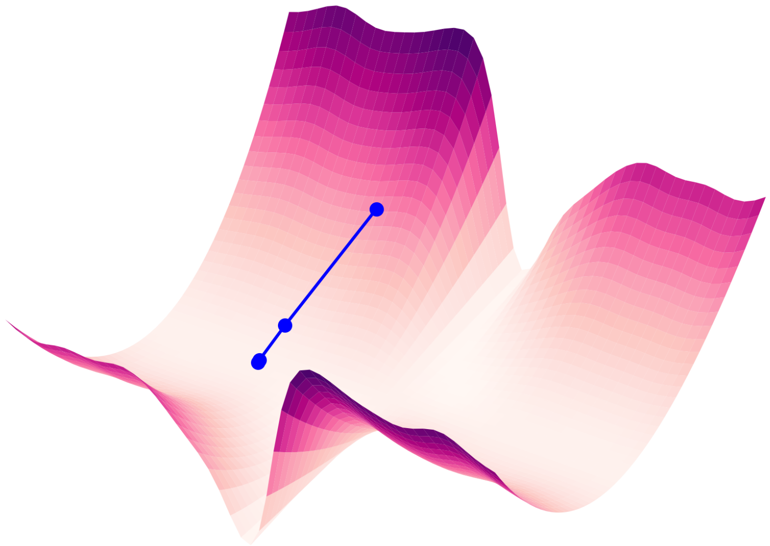


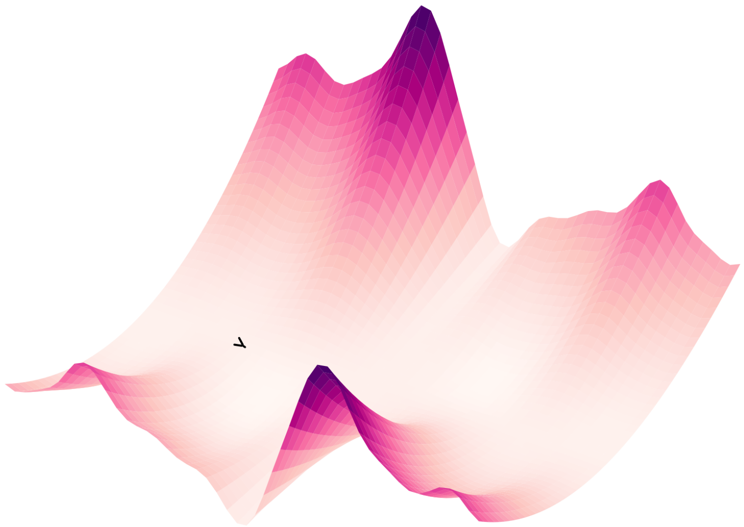
$$\mathbf{w} \in \mathbb{R}^2$$

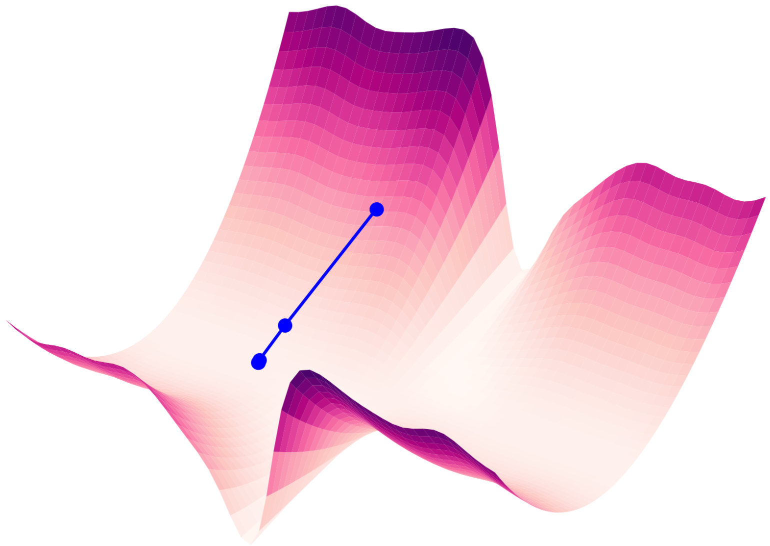


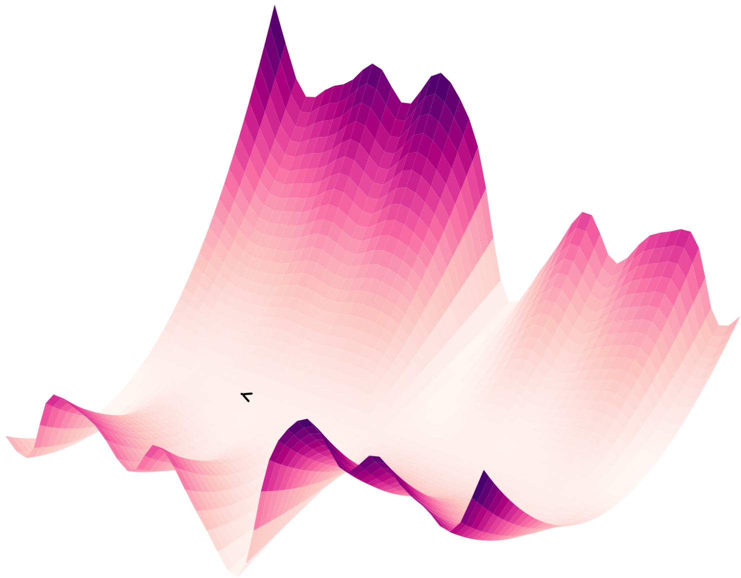


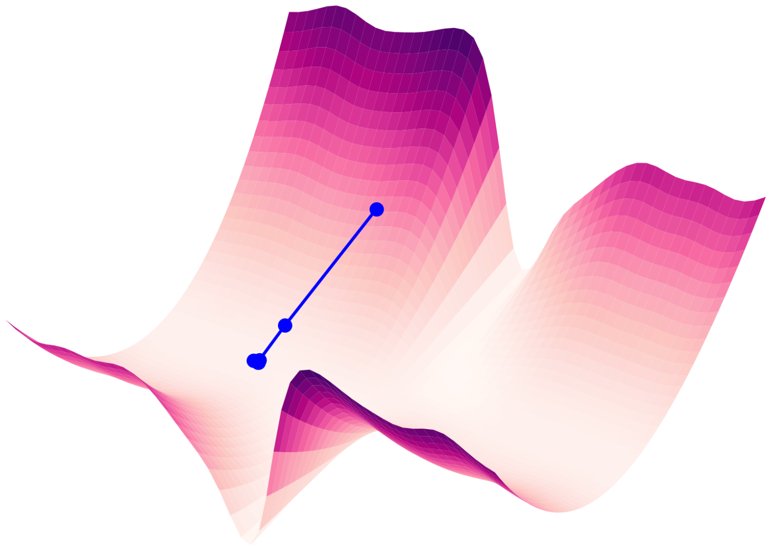


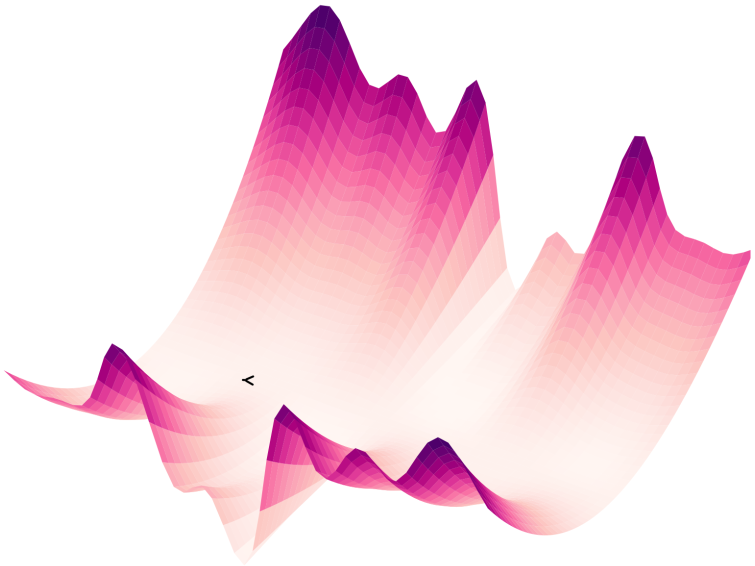
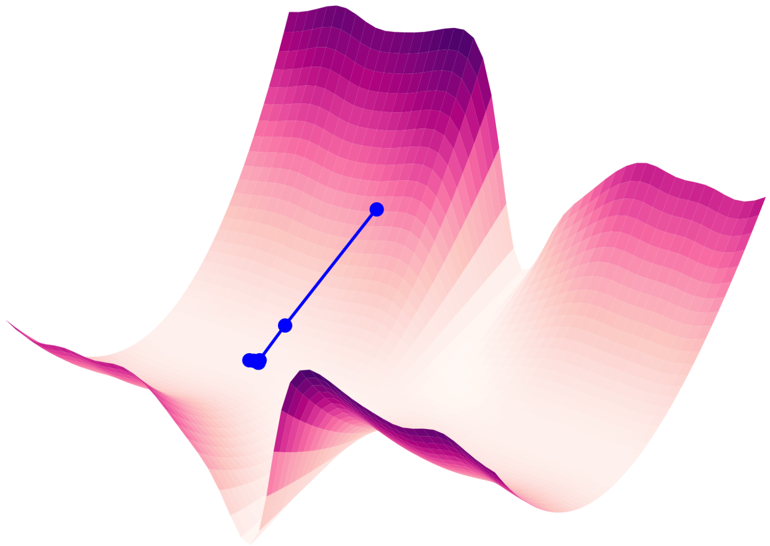


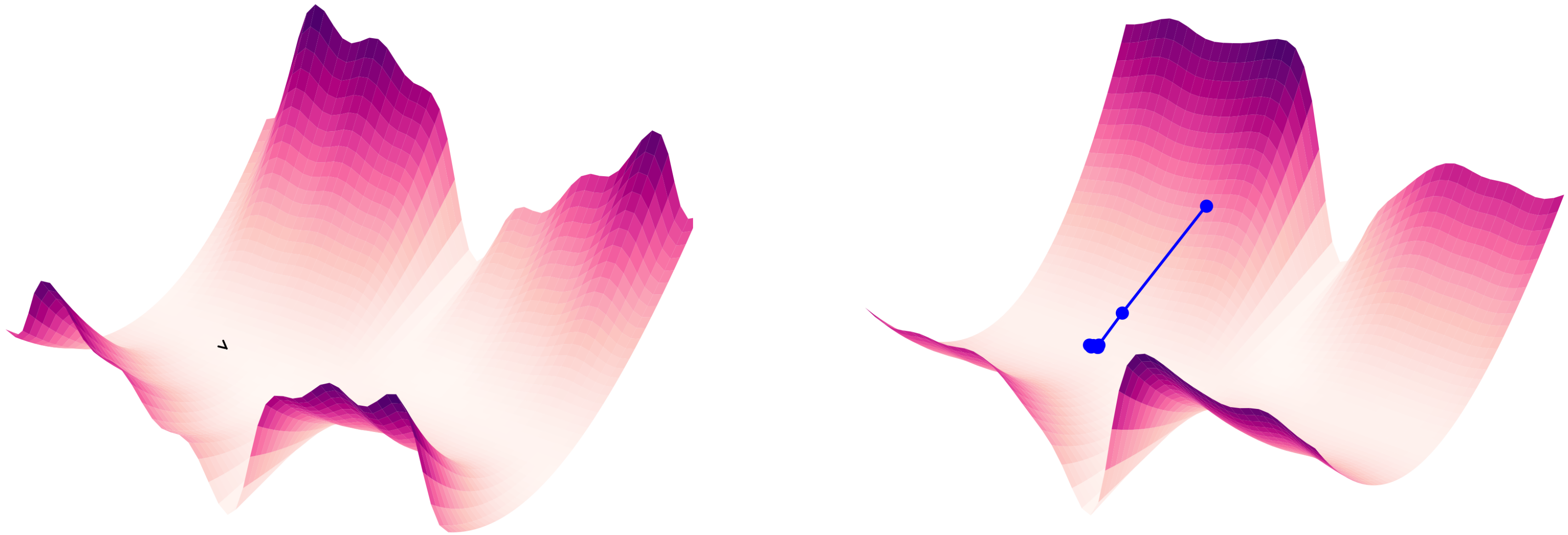




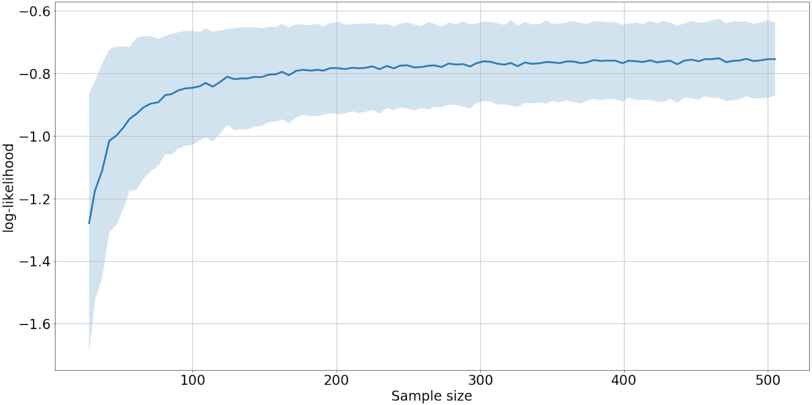




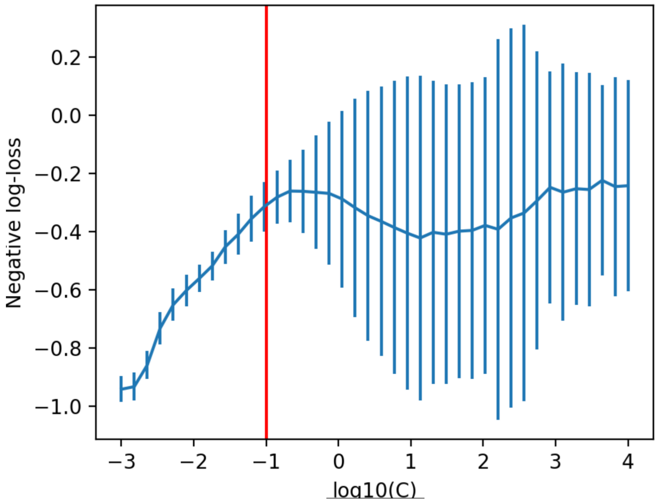





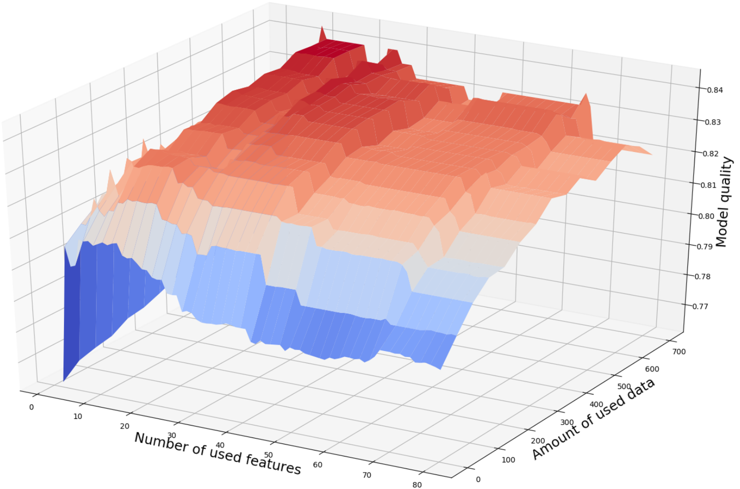
– Error and its variance for a reinforced sample set



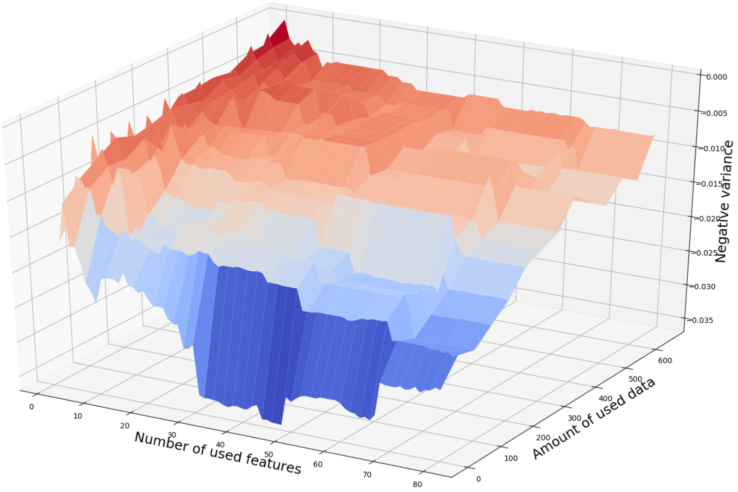
Error variance and increasing of model complexity



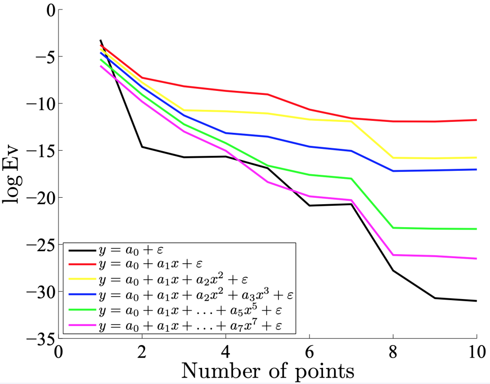
– Error for various sample sizes

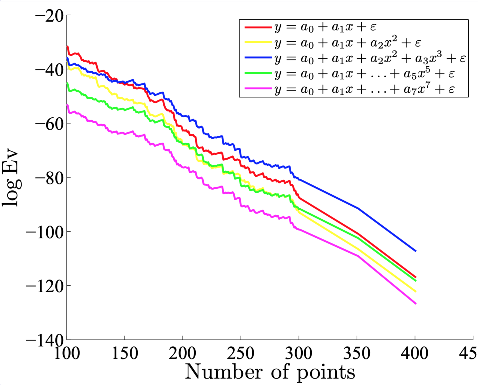


– Error variance for various sample sizes

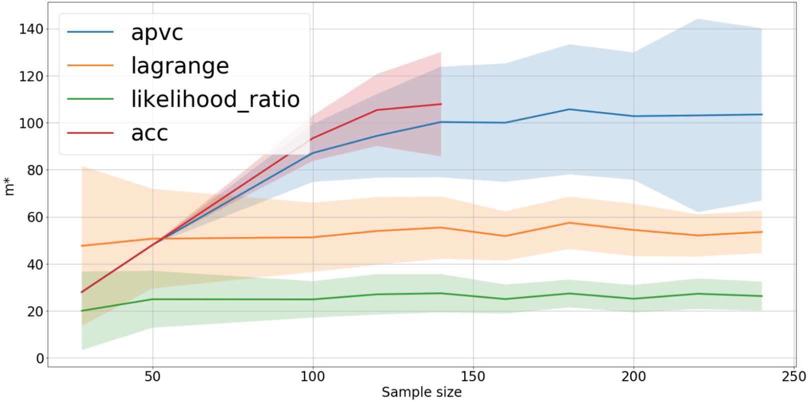


The most plausible model when data arrive



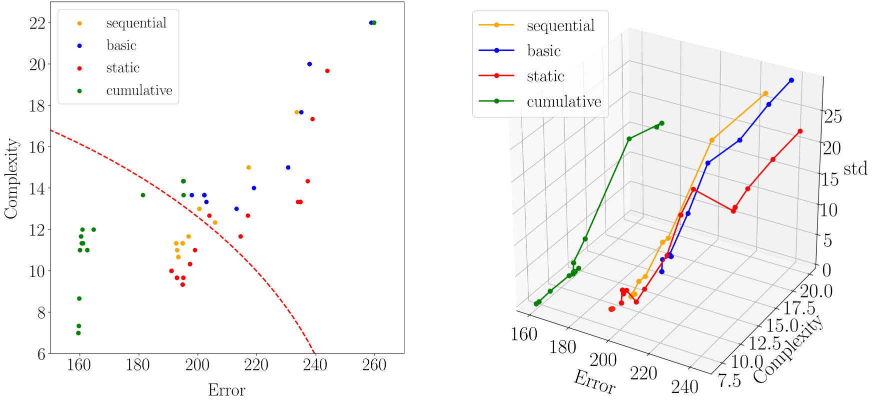


Sample size predicted at an early stage in data collection



Given a sample size of m, one has to forecast the minimum sufficient sample size

Процедура оптимизации структуры



Изменение точности, сложности и устойчивости моделей при итерациях генетического алгоритма

Sequential model selection:
accuracy, complexity, stability