

Школа прикладной математики и информатики МФТИ: перспективы развития

А.М. Райгородский,
К.В. Воронцов,
В.В. Стрижов

Определение и свойства больших данных

Большими данными называются такие выборки, для которых скорость роста объема выборки превышает скорость роста вычислительных мощностей.

Свойства больших данных

Объем. Увеличение мощностей хранения данных, появление новых типов данных, разработка новых способов извлечения знаний из данных.

Разнообразие. Новые источники данных: социальные медиа, видео и аудио коллекции, мобильные приложения, данные поисковых систем, мониторинг интернета вещей.

Скорость. Новые способы передачи данных: непрерывное вещание 4K и передача данных с помощью сетей LiFi со скоростью до 1 Gb/s.

Гипотеза порождения данных

Простая выборка — это случайная, однородная, независимая выборка (i.i.d.):

- ▶ выборка из генеральной совокупности выполнена случайным образом,
- ▶ прецеденты выбраны из одного распределения,
- ▶ вероятностная мера на пространстве представима в виде произведения вероятностных мер на компонентах.

Появление больших данных привело к разрушению гипотезы о простоте выборки.

Классификация пациентов в иммунологии

Ожидание:

для аналитической обработки больших данных требуются алгоритмы субквадратичной сложности относительно объема выборки.

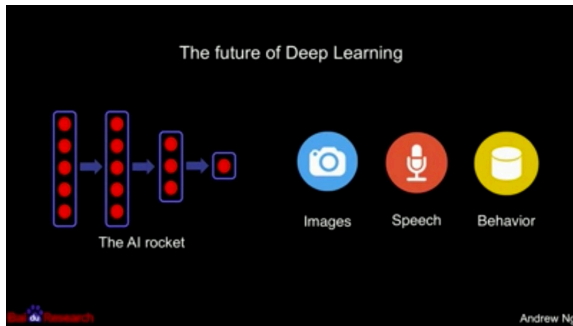
Реальность:

исследователи предлагают алгоритмы линейной и сублинейной сложности.

Model selection for time series forecasting

The Internet of things is the world of networking devices (portables, vehicles, buildings) embedded with sensors and software.

- ▶ Environment and energy monitoring
- ▶ Medical and health monitoring
- ▶ Consumer support, sales monitoring
- ▶ Urban management and manufacturing



Классификация пациентов в иммунологии

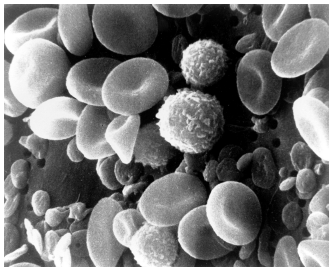
Два типа пациентов с нарушениями работы сердечно-сосудистой системы:

$y \in \{A1, A3\}$; каждому пациенту сопоставлен набор биомаркеров x .

Классы	→ Группы пациентов	Пациентам присвоены метки классов “A1 (оперирован)” и “A3 (группа риска)”.
--------	--------------------	--

Объекты	→ Пациенты	Обследованы 14 пациентов в группе “A1” и 17 пациентов “A3”.
---------	------------	---

Признаки	→ Маркеры	20 биомаркеров: K, L, K/M, L/M, K/N, K/O,...
----------	-----------	--



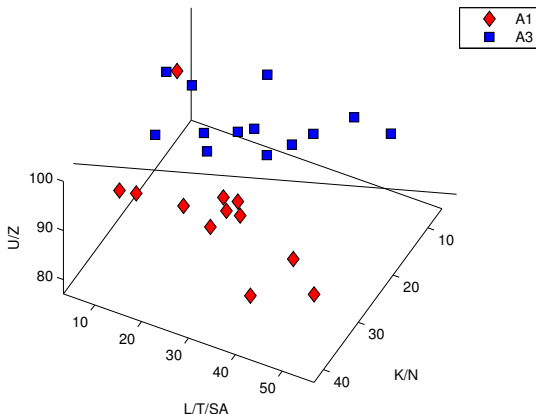
https://en.wikipedia.org/wiki/Blood_cell

Таблица объект–признак (пациент–маркер)

Класс	ID пациента	K	L	K/M	L/M	
A1	C001	58.3	16.7	0.52	0.00	etc.
A1	C004	40.2	6.0	NaN	NaN	
A1	C005	54.3	13.1	NaN	NaN	
A1	C008	48.7	9.8	0.05	0.02	
A3	023	46.6	21.2	0.40	0.08	
A3	026	50.7	26.2	0.12	0.00	
A3	027	45.3	24.5	0.05	0.02	
A3	D037	46.3	13.1	1.23	0.13	
				etc.		

Как показать, что группы различимы?

Линейная классификация



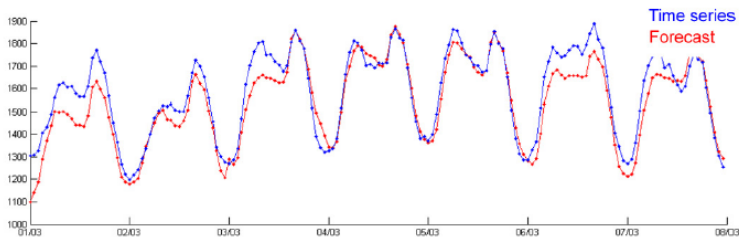
Результатами классификации (A1 против A3) являются
параметры математической модели

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

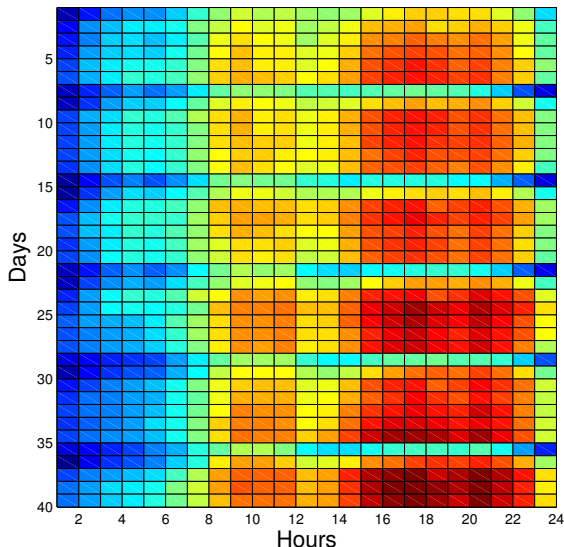
Многомерные данные в разных временных шкалах

Прогнозирование объема потребления и цен на электроэнергию

- ▶ Заданы временные ряды цен, потребления, температуру, влажность, информацию о времени измерения и праздничных днях.
- ▶ Временные ряды измеряются в различных шкалах и имеют различную периодичность.



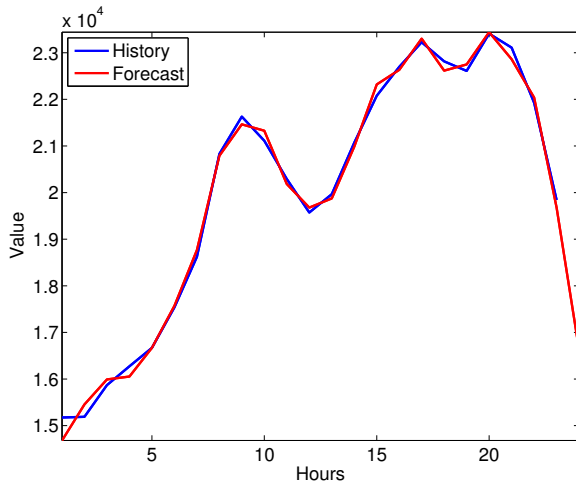
Матрица авторегрессии



$$\mathbf{X}^* = \left[\begin{array}{c|c} S_T & \mathbf{x}_{m+1} \\ \hline \mathbf{y} & \mathbf{X} \end{array} \right],$$
$$\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \mathbf{w}) = \mathbf{X}\mathbf{w},$$
$$y_{m+1} = S_T = \mathbf{w}^T \mathbf{x}_{m+1}.$$

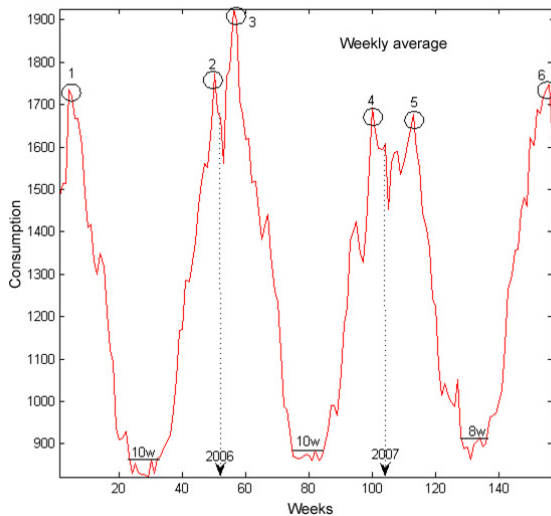
Нейчев Р.Г., Катруца А.М., Стрижов В. Выбор оптимального набора признаков из мультикоррелирующего множества в задаче прогнозирования // Заводская лаборатория. Диагностика материалов, 2016.

The one-day forecast: expected error is 3.1% working day, 3.7% week-end

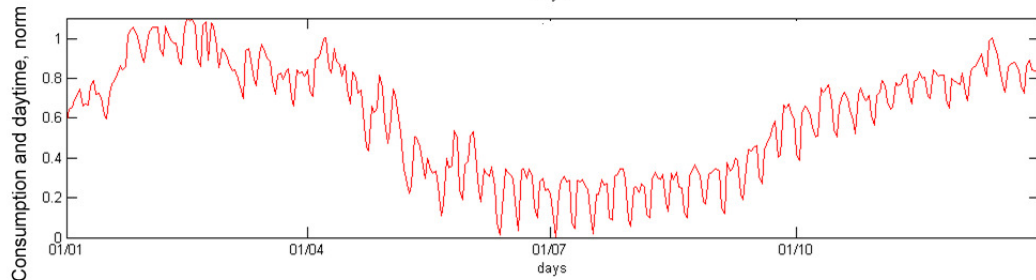
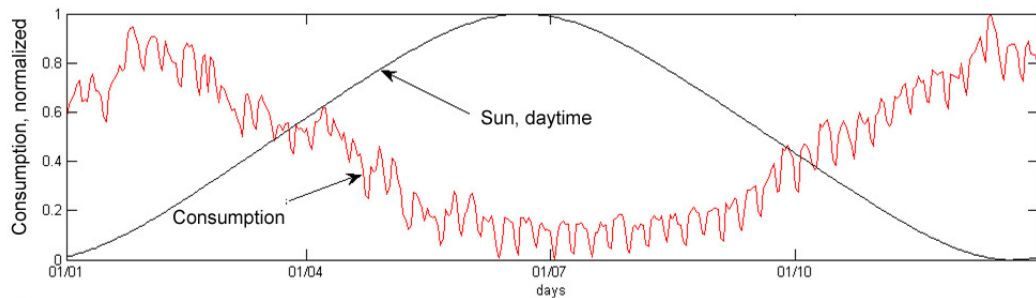


The model $\hat{y} = f(\mathbf{X}, \mathbf{w})$ could be a linear model, neural network, deep NN, SVN, ...

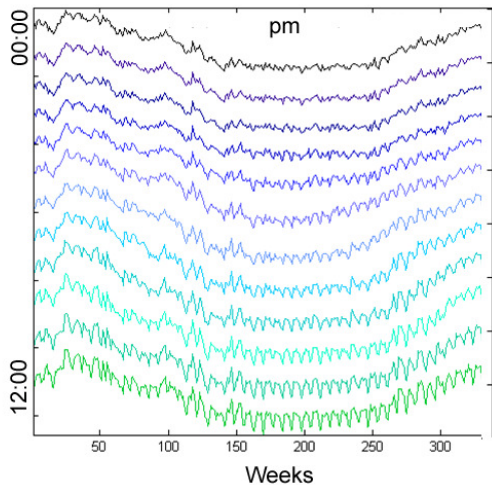
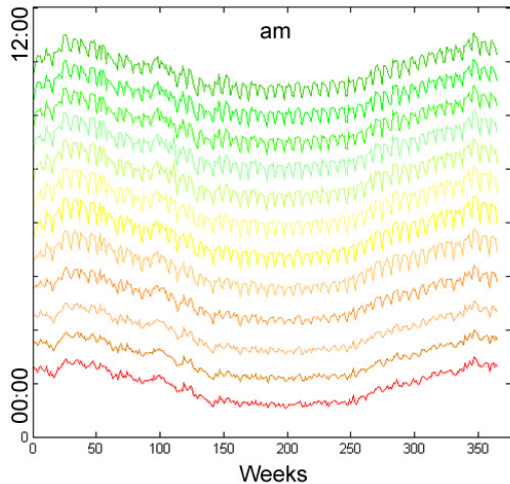
Structure of energy consumption



Sunrise bias: one-year daytime and consumption

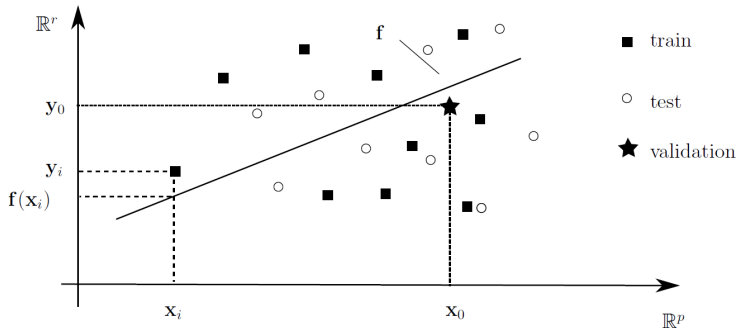


One-hour line, day-by-day during a year: autoregressive analysis

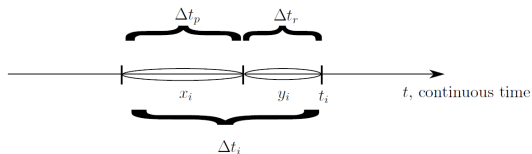


Design matrix and online rolling validation procedure

Forecast is a mapping from p -dimensional objects space to r -dimensional answers



$$\mathbf{X}^* = \left[\begin{array}{c|c} \mathbf{x} & \mathbf{y} \\ \hline \mathbf{X} & \mathbf{Y} \\ \hline 1 \times n & 1 \times r \\ m \times n & m \times r \end{array} \right]$$



The model performance criteria and forecast errors

Stability:

- ▶ the error does not change significantly following small changes in time series,
- ▶ the distribution of the model parameters does not change.

Complexity:

- ▶ the number of parameters (elements in superposition) is minimal,
- ▶ the minimum description length principle holds the William Ockham's rule.

Error: the residue $\varepsilon_j = \hat{y}_j - y_j$ for

- ▶ mean absolute error and (symmetric) mean absolute percent error

$$sMAPE = \frac{1}{r} \sum_{j=1}^r \frac{2|\varepsilon_j|}{|\hat{y}_j + y_j|}.$$

Case 2. Sales planning: to forecast the goods consumption

Retailers' daily routines

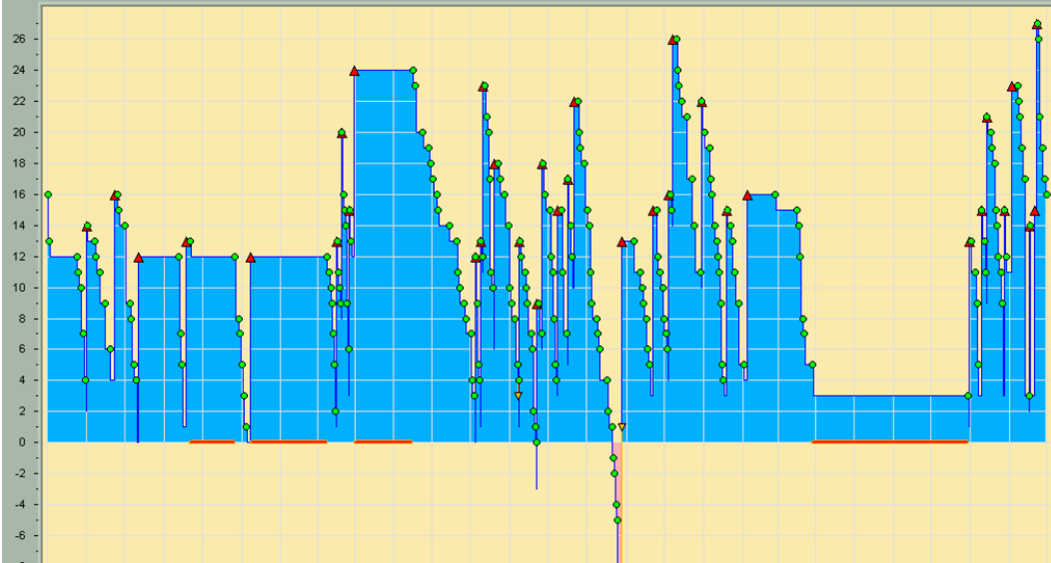
- ▶ custom inventory
- ▶ consumer demand forecasting



- ▶ There given historical time series of the volume off-takes: foodstuff
over 60000 items in 5000 groups
- ▶ Let the time series be homoscedastic: its variance is time-constant
- ▶ Minimize the loss function to forecast the next sample

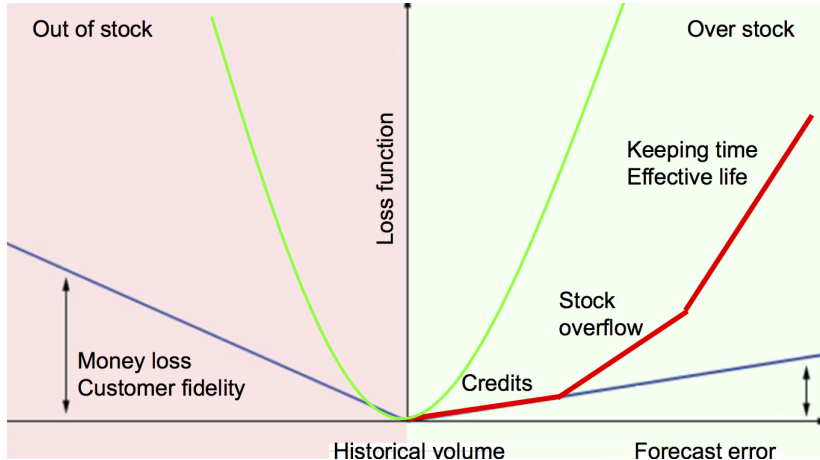
Custom inventory

НОУАН Напиток ШИПОВНИКА 1л

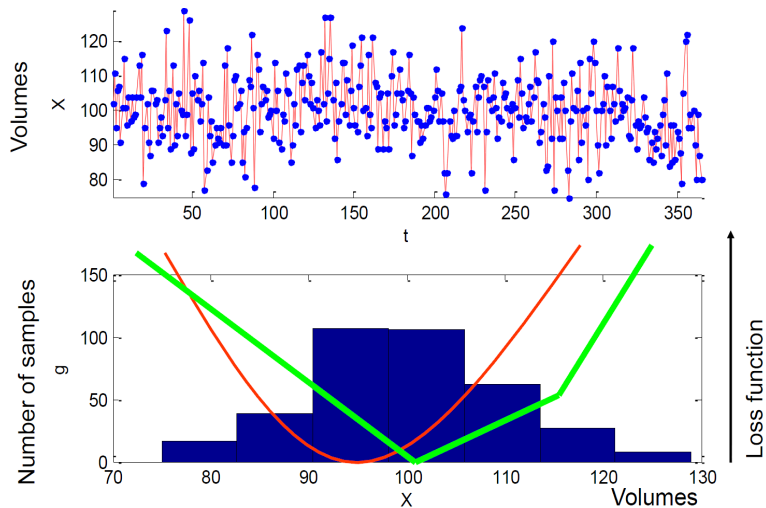


The performance criterion is **minimum loss of money**

Error functions: **quadratic**, **linear**, **asymmetric**.



The time series of residues and its histogram



There given historgam

$$H = \{X_i, g_i\}_{i=1}^m$$

and loss function

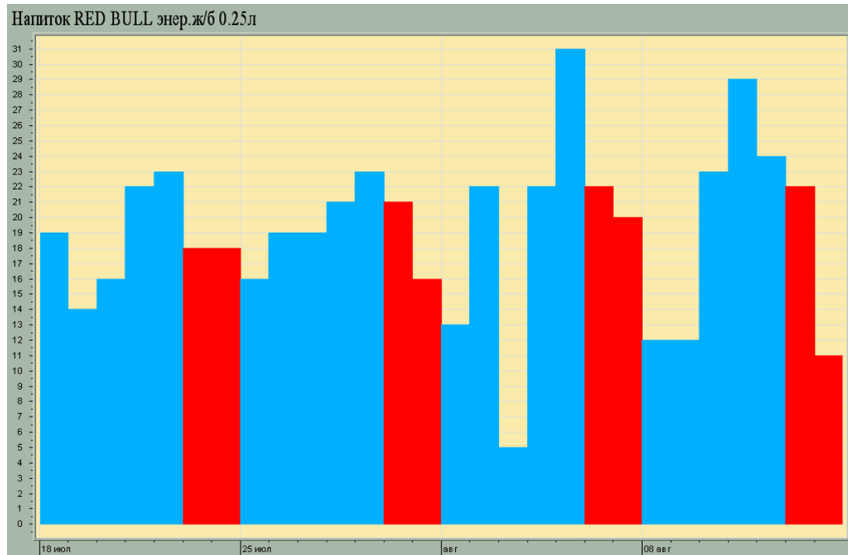
$$L = L(Z, X).$$

The optimal forecast is

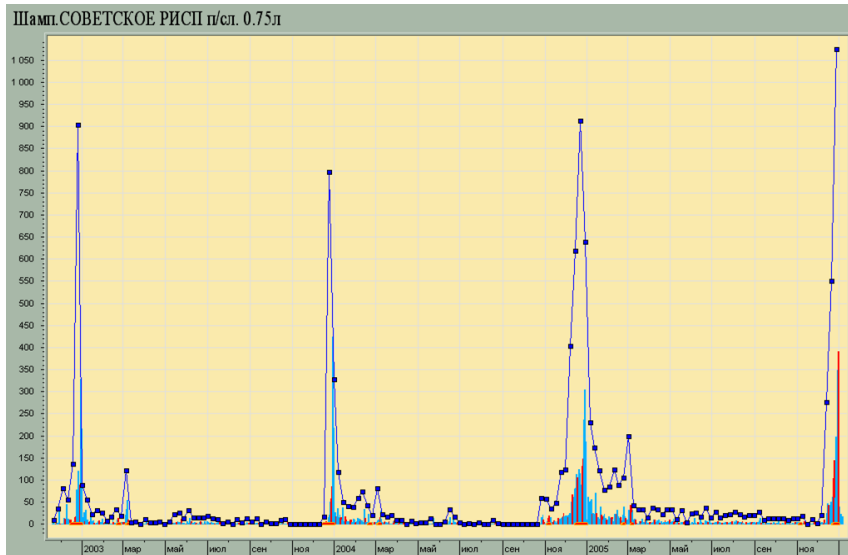
$$\tilde{X} = \arg \min_{Z \in \{X_1 \dots X_m\}} \sum_{i=1}^m g_i L(Z, X_i)$$

of this convolution.

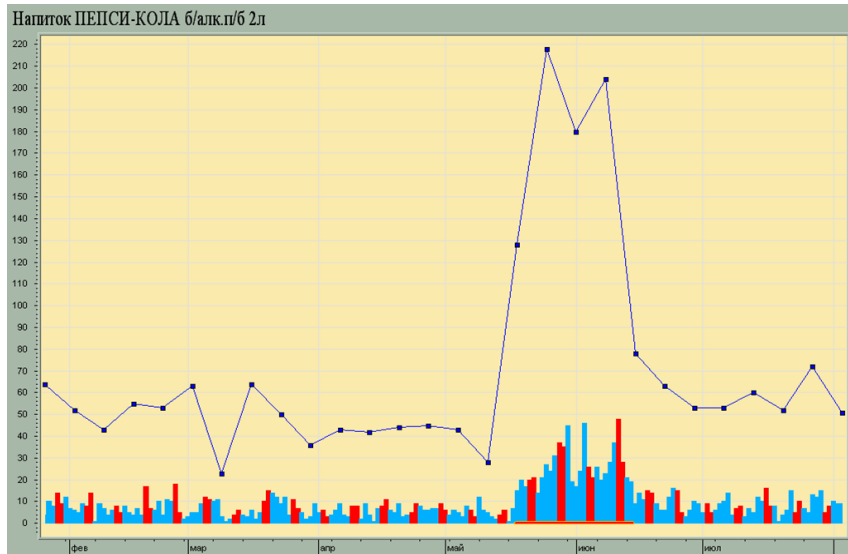
Beverage: the week periodicity daily over four weeks



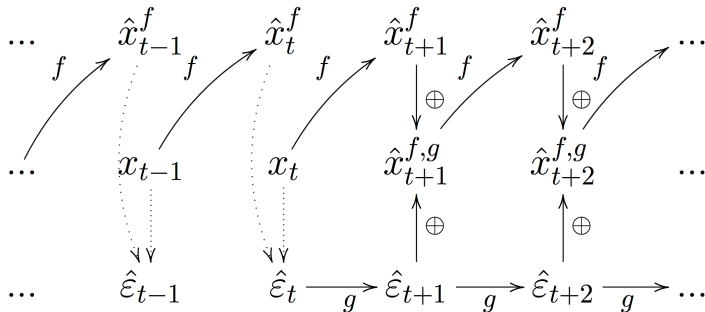
Sparkling wine: holidays weekly over three years



Promotional actions



Forecast the residues to boost performance



1. Model f forecasts $n(g)$ history ends $\hat{x}_t^f, \dots, \hat{x}_{t-n(g)+1}^f$ for one sample.
2. Compute $n(g)$ residues $\hat{\epsilon}_t, \dots, \hat{\epsilon}_{t-n(g)+1}$ as $\hat{\epsilon}_{t-k} = x_{t-k} - \hat{x}_{t-k}^f$.
3. Function g forecasts residues $\hat{\epsilon}_{t+i}$ ahead $\max(i)$ time-ticks.
4. Combine forecasts $\hat{x}_{t+i}^{f,g} = \hat{x}_{t+i}^f + \hat{\epsilon}_{t+i}$ computing f for each sample $\hat{x}_{t+i}^f$.

Case 3. Forecasting volumes of Russian railways freight transportation

Keep a hierarchical structure of time series without losing performance

Forecast with hierarchical aggregation of

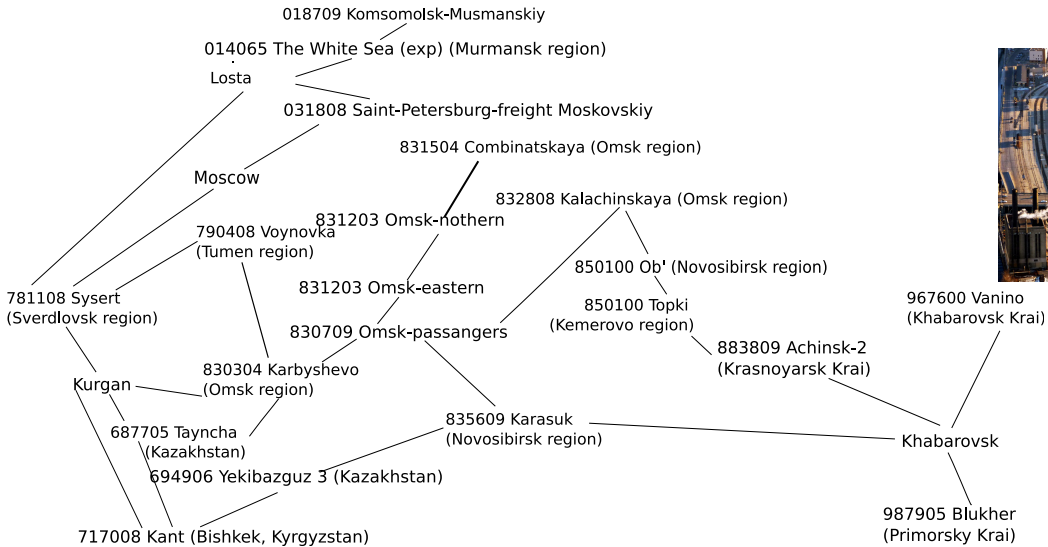
- ▶ types of freight in
- ▶ stations, regions, and roads,
- ▶ for a day, week, month, and quarter,
- ▶ counting all combinations above.

Satisfy the conditions:

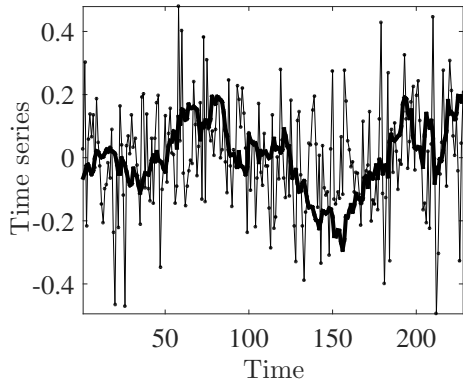
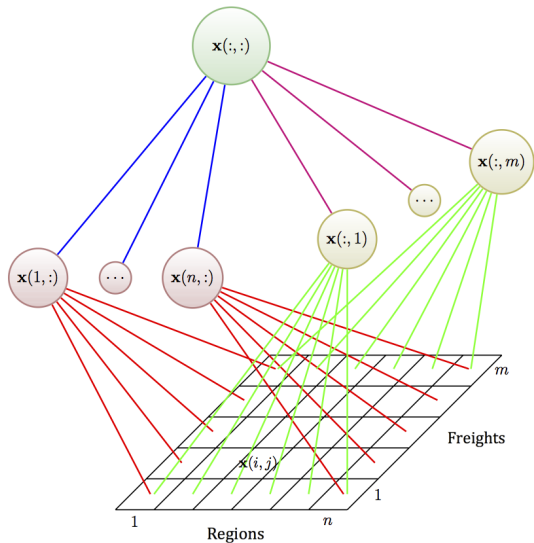
- ▶ minimize error,
- ▶ incorporate important external factors,
- ▶ respect hierarchical structure,
- ▶ do not exceed physical bounds of forecast values.



The railroad map counts ~ 78 regions, ~ 4000 stations, and ~ 100 rail-yards



Independent forecasts might be inconsistent with aggregated ones



Hierarchical data χ , independent forecasts $\hat{\chi}$ and reconciled forecasts $\hat{\varphi}$

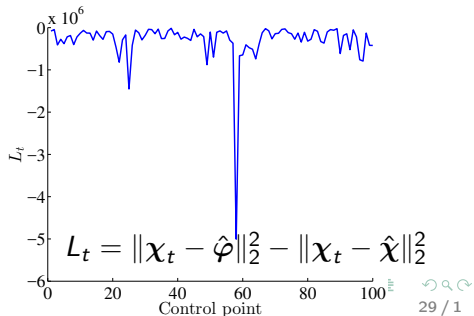
$$\mathbf{S} = \begin{bmatrix} \begin{array}{c|c|c|c|c} -1 & 1 & \dots & 1 & 0 & \dots & 0 \\ -1 & 0 & \dots & 0 & 1 & \dots & 1 \\ \hline 0 & -1 & \dots & 0 & 0 & \dots & 0 \\ \dots & & \dots & & & \dots & \\ 0 & 0 & \dots & -1 & 0 & \dots & 0 \\ \hline 0 & 0 & \dots & 0 & -1 & \dots & 0 \\ \dots & & \dots & & & \dots & \\ 0 & 0 & \dots & 0 & 0 & \dots & -1 \end{array} & \begin{array}{c|c|c|c|c} 0 & 0 & \dots & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \hline 1 & 1 & \dots & 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & & \dots & & & \dots & & & \dots & \\ 0 & 0 & \dots & 0 & \dots & 1 & 1 & \dots & 1 \\ \hline 1 & 0 & \dots & 0 & \dots & 1 & 0 & \dots & 0 \\ \dots & & \dots & & & & & \dots & \\ 0 & 0 & \dots & 1 & \dots & 0 & 0 & 0 & \dots & 1 \end{array} \end{bmatrix}, \chi_t = \begin{bmatrix} x_t(:, :) & \text{top} \\ x_t(n, :) & \text{regions} \\ \dots & \\ x_t(:, m) & \text{freights} \\ \dots & \\ x_t(1, m) & \text{given regions} \\ \dots & \\ x_t(n, 1) & \text{given freights} \\ \dots & \end{bmatrix}$$

Consistency condition $\mathbf{S}\chi_t = \mathbf{0}$, where the link matrix is $(2 + n + m) \times (1 + n + m + nm)$

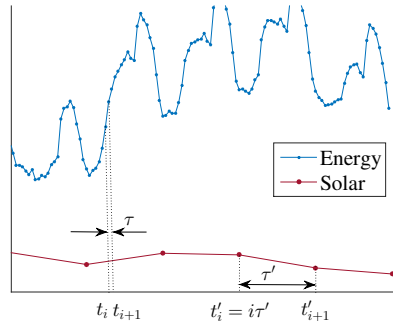
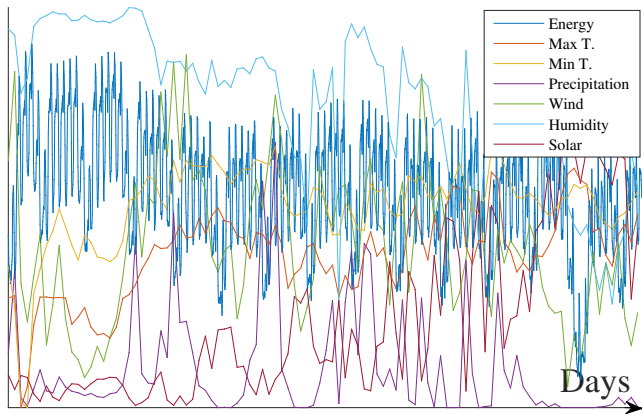
There given independent forecasts $\hat{\chi} \notin \mathbf{A}$, $\hat{\chi} \in \mathbf{B}$ to make reconciled $\hat{\varphi}$ s.t.

- **consistency** $\hat{\varphi} \in \mathbf{A} = \{\chi \in \mathbb{R}^d \mid \mathbf{S}\chi = \mathbf{0}\}$,
- **physical limitations** $\hat{\varphi} \in \mathbf{B}$,
- **precision** $l_h(\chi_{T+1}, \hat{\varphi}) \leq l_h(\chi_{T+1}, \hat{\chi})$ for any hierarchical data $\chi_{T+1} \in \mathbf{A} \cap \mathbf{B}$.

$$\hat{\varphi} = \chi_{proj} = \arg \min_{\chi \in \mathbf{A} \cap \mathbf{B}} l_h(\chi, \hat{\chi})$$

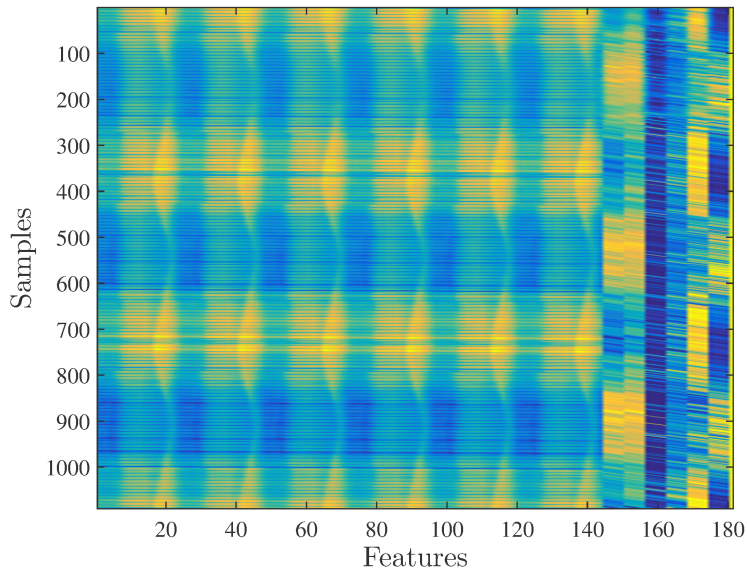


Case 4. Internet of things, multiscale dataset for vector forecasting



Each real-valued time series $\mathbf{s} = [s_1, \dots, s_i, \dots, s_T]$, $s_i = s(t_i)$, $0 \leq t_i \leq t_{\max}$ is a sequence of observations of some real-valued signal $s(t)$ with its own sampling rate τ .

To boost the forecast quality include external variables



Models and features

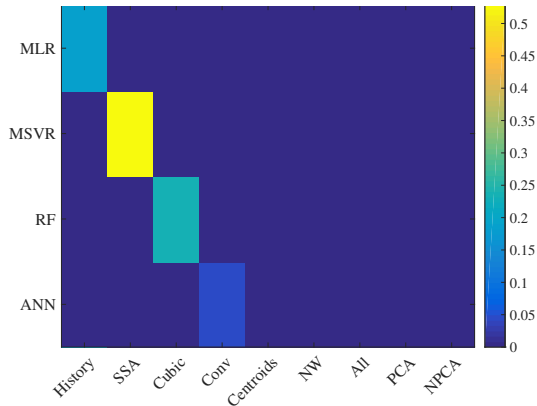
Models:

- ▶ Baseline method: $\hat{s}_i = s_{i-1}$.
- ▶ Multivariate linear regression (MLR) with l_2 -regularization. Regularization coefficient: 2
- ▶ SVR with multiple output. Kernel type: RBF, p_1 : 2, p_2 : 0, γ : 0.5, λ : 4.
- ▶ Feed-forward ANN with single hidden layer, size: 25
- ▶ Random forest (RF). Number of trees: 25 , number of variables for each decision split: 48.

Feature combinations:

- ▶ History: the standard regression-based forecast with no additional features.
- ▶ SSA, Cubic, Conv, Centroids, NW: history + a particular feature.
- ▶ All: all of the above, with no feature selection.
- ▶ PCA and NPCA: all generation strategies with feature selection.

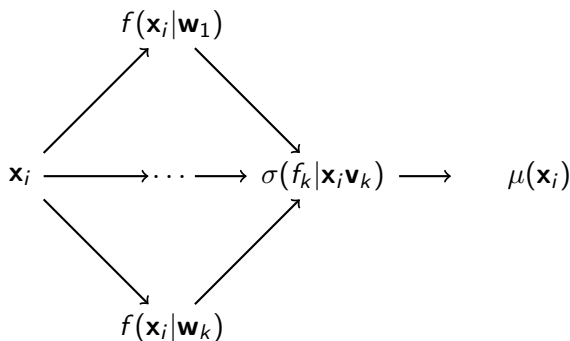
Feature analysis



Test subset MAPE

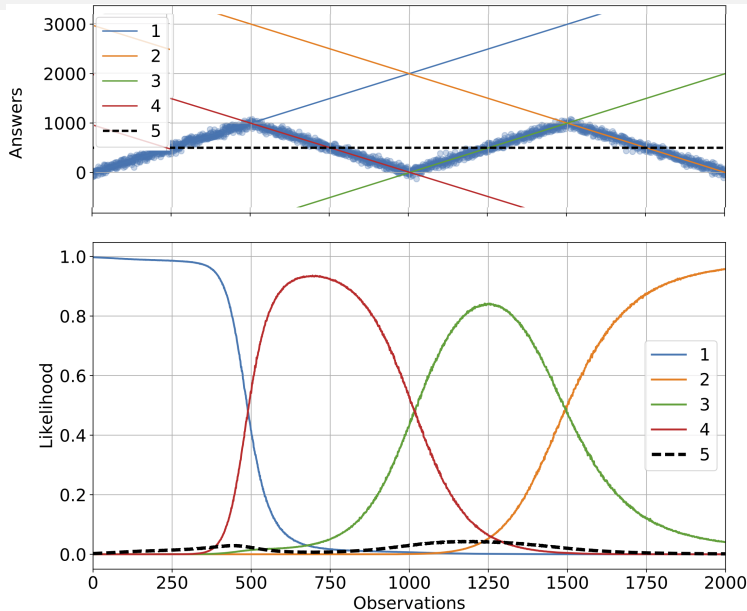
Ratio of times each combination of model and feature performed best for at least one of the time series (7) or error functions (6), all (6) data sets ($6 \times 7 \times 6 = 252$ cases).

Моделирование разнородных данных: смесь экспертов



$$\hat{y}_i = \mu(\mathbf{x}_i) = \sum_{k=1}^K \left(\frac{\exp(\mathbf{v}_k^T \mathbf{x}_i)}{\sum_{\kappa=1}^K \exp(\mathbf{v}_\kappa^T \mathbf{x}_i)} f(\mathbf{x}_i | \mathbf{w}_k) \right).$$

Mixture of experts as a model selection procedure



The likelihood of every expert given a data sample, the model

$$\mathbf{f} = \mathbf{a}(\mathbf{h}_1(\dots \mathbf{h}_K(\mathbf{x}))) (\mathbf{w}),$$

where $\mathbf{h}_k$ are autoencoders and $\mathbf{a}$ is a softmax classifier:

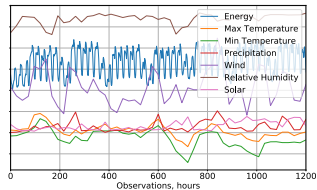
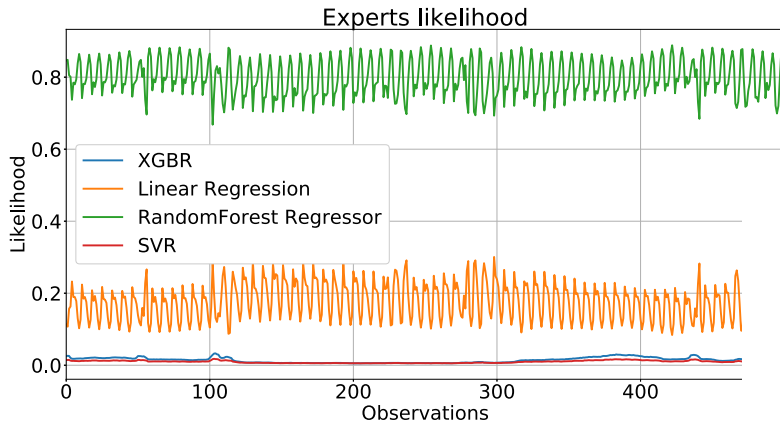
$$\mathbf{f}(\mathbf{w}, \mathbf{x}) = \frac{\exp(\mathbf{a}(\mathbf{x}))}{\sum_j \exp(a_j(\mathbf{x}))},$$

$$\mathbf{a}(\mathbf{x}) = \mathbf{W}_2^\top \tanh(\mathbf{W}_1^\top \mathbf{h}_1),$$

$$\mathbf{h}_k(\mathbf{x}) = \sigma(\mathbf{W}_k \mathbf{x} + \mathbf{b}_k),$$

where $\mathbf{w}$ minimizes the error function.

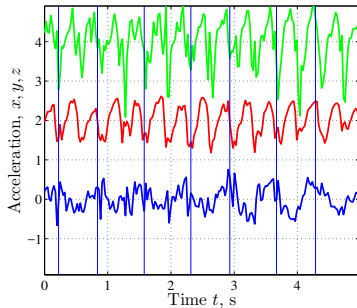
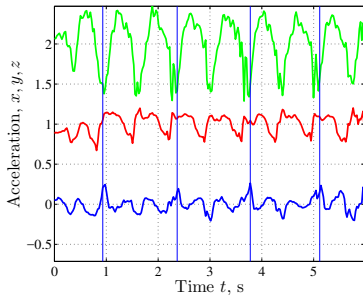
Heterogeneous model mixture for time series forecasting



Построение сетей глубокого обучения



Временные ряды акселерометра
(шаг и бег):

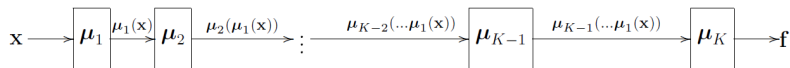


Motrenko A., Strijov V. Extracting fundamental periods to segment biomedical signals // Journal Of Biomedical And Health Informatics, 2016.

Сеть глубокого обучения как суперпозиция моделей

Модель классификации — суперпозиция функций

$$\mathbf{f}(\mathbf{x}, \mathbf{w}) = \mu_1(\mu_2(\dots \mu_K(\mathbf{x}))) : \mathbb{R}^n \rightarrow [0, 1]^M,$$



где μ_k , $k \in \{1, \dots, K\}$ — модели из класса нейронных сетей с параметрами $\mathbf{w}_k$, $k \in \{1, \dots, K\}$, $\mathbf{w} = [\mathbf{w}_1, \dots, \mathbf{w}_K]^T$,

$$\mathbf{f}(\mathbf{x}; \mathbf{w}) = [p(y_1 = 1 | \mathbf{x}; \mathbf{w}), \dots, p(y_M = 1 | \mathbf{x}; \mathbf{w})]^T.$$

Функция ошибки

$$S(\mathbf{w} | \mathcal{D}) = - \sum_{\mathbf{x} \in \mathcal{D}} \sum_{\xi=1}^M [y_t = 1] \ln(f_{\xi}(\mathbf{x}, \mathbf{w})).$$

Сеть глубокого обучения как суперпозиция моделей

Автокодировщик — суперпозиция блоков

$$\mu_k = \phi(\mathbf{g}(\mathbf{x})), \quad k = 1, \dots, K$$

где кодирующий блок

$$\mathbf{g}(\mathbf{x}) = \sigma(\mathbf{W}_g \mathbf{x} + \mathbf{b}_g).$$

Двухслойная нейронная сеть

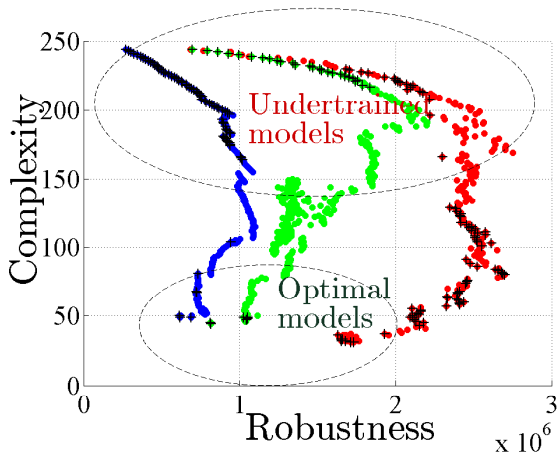
$$\mathbf{a}(\mathbf{x}, \mathbf{w}) = \mathbf{W}_2^T \tanh(\mathbf{W}_1^T \mathbf{x}),$$

$$\mu(\mathbf{x}, \mathbf{w}) = \frac{\exp(\mathbf{a}(\mathbf{x}, \mathbf{w}))}{\sum_{j=1}^n \exp(a_j(\mathbf{x}, \mathbf{w}))},$$

где $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$ — матрицы весов первого и второго слоев нейронной сети.

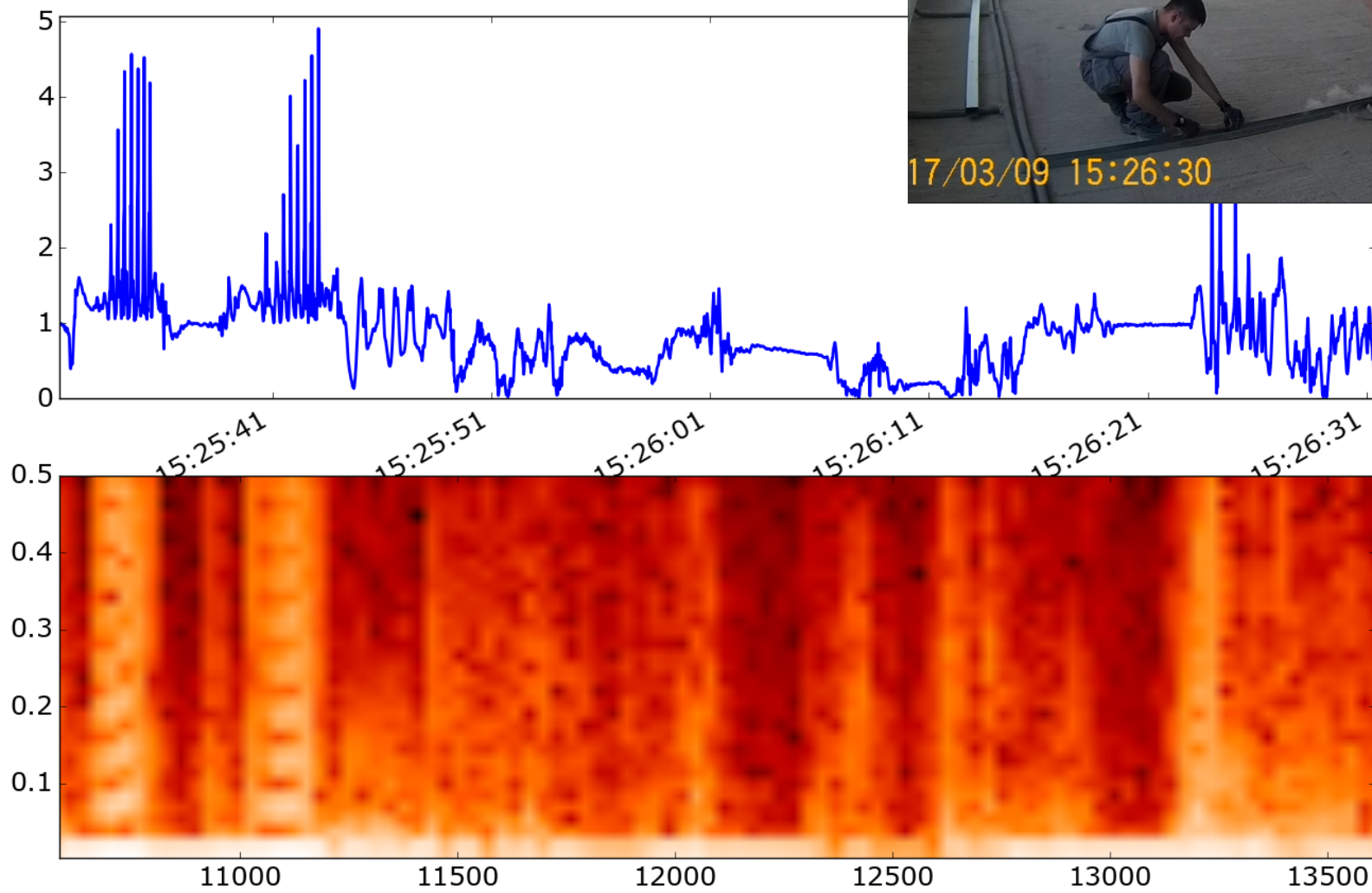
Автоматическое порождение моделей

Анализ моделей с точки зрения устойчивости и сложности



Попова М. С., Стрижов В. В. Выбор оптимальной модели классификации физической активности по измерениям акселерометра // Информатика и ее применения, 2015.

Анализ поведения на производстве



Local transformations in superposition with the deep neural network

Feature generation by local transformations

- ▶ parameters of Singular Structure Analysis approximation of time series $\mathbf{s}$,
- ▶ Fast Fourier Transformation of $\mathbf{s}$,
- ▶ parameters of polynomial/spline approximation of $\mathbf{s}$

could reduce complexity this model down to complexity of logistic regression.

To generate the features $\mathbf{x}$

1. Optimize transformation parameters $\hat{\mathbf{b}}$ given model parameters $\hat{\mathbf{w}}$

$$\hat{\mathbf{b}} = \arg \min Q(g(\mathbf{b}, \mathbf{x})).$$

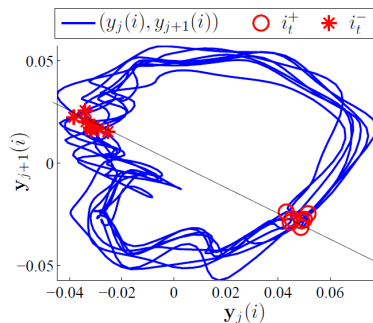
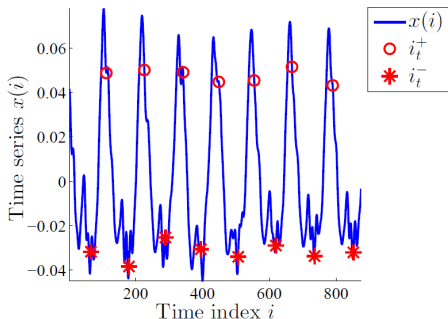
2. Include the vector $\hat{\mathbf{b}}$ the feature space to solve the main problem

$$\hat{\mathbf{w}} = \arg \min S(\mathbf{w} | \mathbf{f}(\mathbf{w}, \phi), \mathbf{y}), \quad \text{where} \quad \phi(\hat{\mathbf{b}}, \mathbf{s}) \subseteq \mathbf{x}.$$

Human gate detection with time series segmentation

Find dissection of the trajectory of principal components $\mathbf{y}_j = \mathbf{H}\mathbf{v}_j$, where $\mathbf{H}$ is the Hankel matrix and $\mathbf{v}_j$ are its eigenvectors:

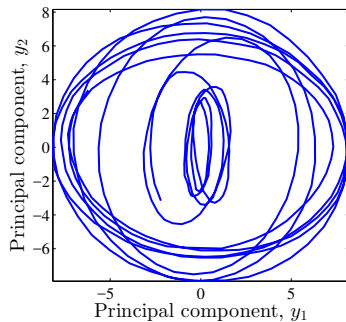
$$\frac{1}{N}\mathbf{H}^\top\mathbf{H} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top, \quad \mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_N).$$



Parameters of the local models: SSA

For time series $\mathbf{s}$ construct the Hankel matrix with a period k and shift p , so that for $\mathbf{s} = [s_1, \dots, s_T]$ the matrix

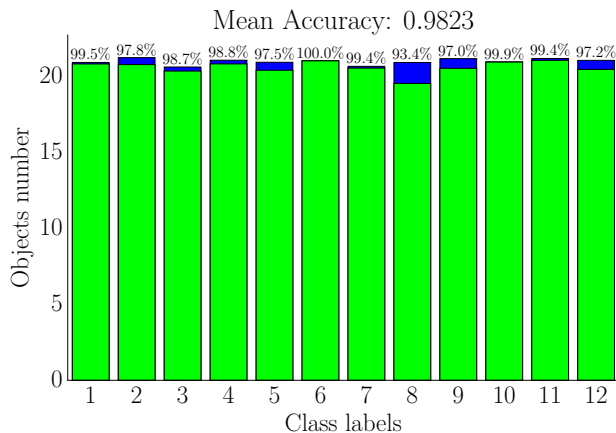
$$\mathbf{H}^* = \left[\begin{array}{c|cccc} s_T & s_{T-1} & \dots & s_{T-k+1} \\ \vdots & \ddots & \ddots & \vdots \\ s_{k+2} & s_{k+1} & \dots & s_2 \\ s_k & s_{k-1} & \dots & s_1 \end{array} \right].$$



Reconstruct the regression to the first column of the matrix $\mathbf{H}^* = [\mathbf{h}, \mathbf{H}]$ and denote its least square parameters as the feature vector

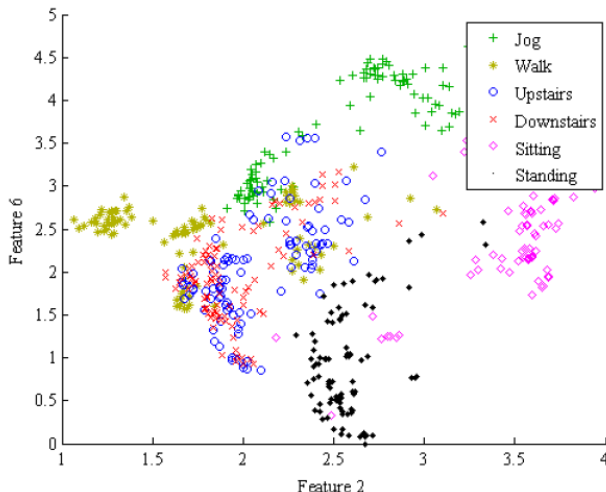
$$\phi(\mathbf{s}) = \arg \min \|\mathbf{h} - \mathbf{H}\phi\|_2^2.$$

Performance of the human physical activities classification model



- 1) walk forward
- 2) walk left
- 3) walk right
- 4) go upstairs
- 5) go downstairs
- 6) run forward
- 7) jump up and down
- 8) sit and fidget
- 9) stand
- 10) sleep
- 11) elevator up
- 12) elevator down

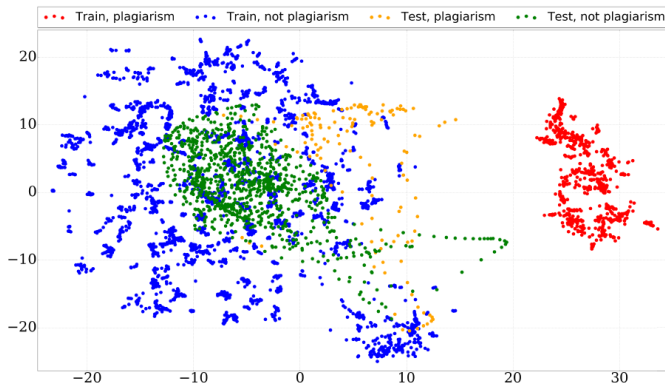
Порождение и выбор метрического пространства



Обучение многообразий

Обнаружение заимствований в тексте

$$\mathbf{F} : \mathbf{X} \in \mathbb{R}^D \rightarrow \mathbf{Z} \in \mathbb{R}^d, \quad d \ll D.$$



Тематическое моделирование и матричное разложение



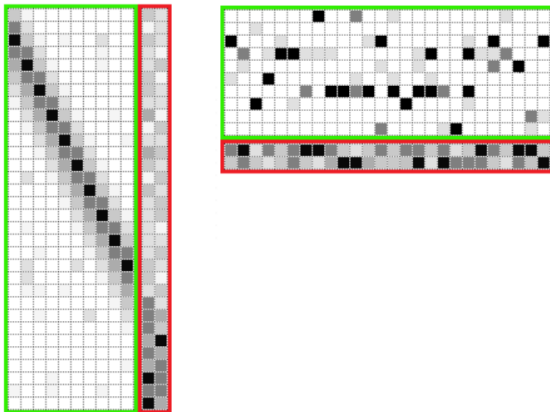
$w_1, \dots, w_{n_d}$:

Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найлены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Разложение разреженных матриц

$$\underset{d \times w}{\mathbf{X}} = \underset{d \times t}{\mathbf{\Phi}} \underset{t \times w}{\mathbf{\Theta}}$$

documents, words, topics

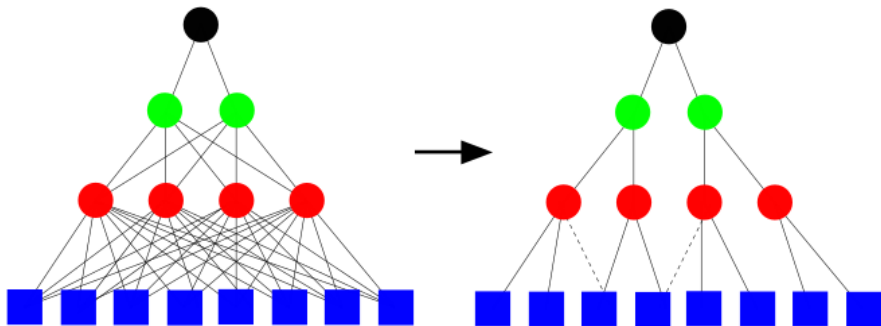


Hierarchical tree building

- Nodes: documents, topics, root.
- Extracting tree from matrix:

$$Topic(d) = \arg \max_t \Theta_{t,d}$$

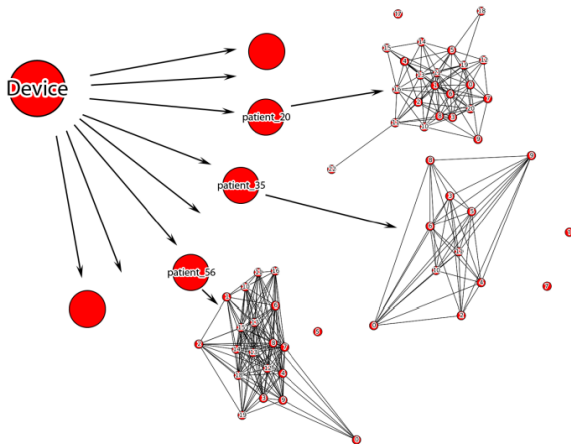
$$MultiTopic(d) = \arg \max_t \Theta_{t,d} \cup \{t | \Theta_{t,d} > \varepsilon\}$$



Grid visualization

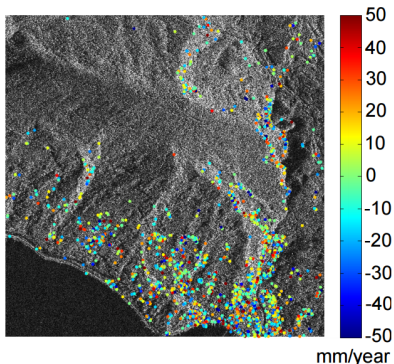


Кластерные структуры больших данных

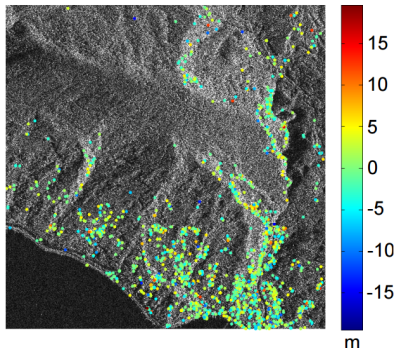


Жариков И. Н., Ишкина Ш. Х., Воронцов К.В. Кластеризация множества состояний человека в информационном анализе электрокардиосигналов // Машинное обучение и анализ данных, 2016.

Анализ спутниковых снимков



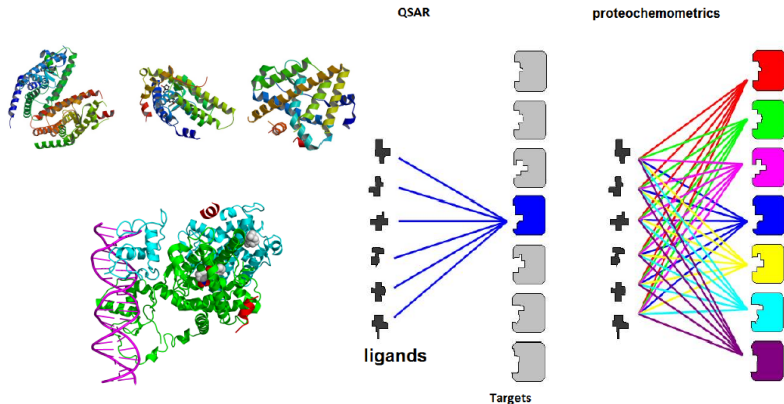
Скорость движения вдоль направления наблюдения



Погрешность в высоте

Рудаков К.В., Адуенко А.А., Рейер И.А., Василейский А.С., Карелов А.И., Стрижов В.В.
Алгоритмы выделения и совмещения устойчивых отражателей на спутниковых снимках // Компьютерная оптика, 2015.

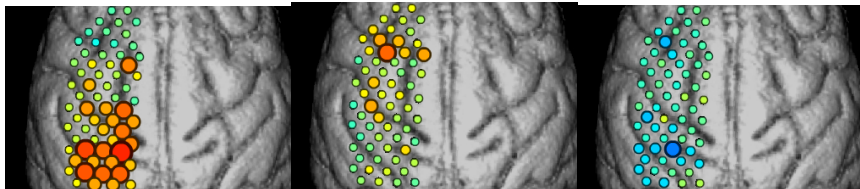
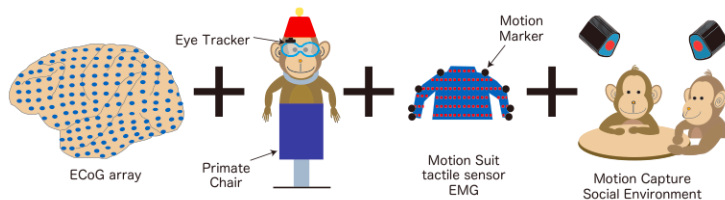
Согласованное построение моделей биологической активности ядерных рецепторов



Popova M. S., Isaev A.O., Strijov V.V. Feature selection and multi-task prediction of biological activity for nuclear receptors // Machine Learning and Data Analysis, 2016.

Анализ пространственно-временных данных

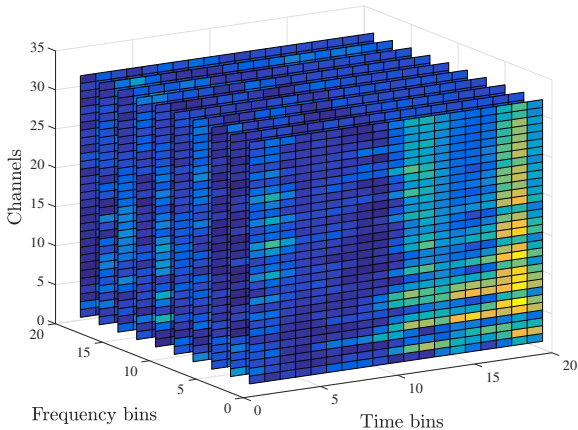
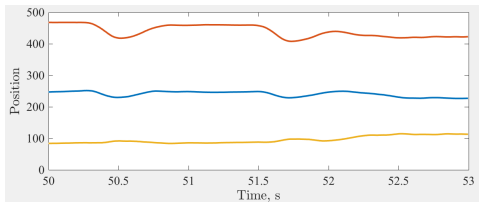
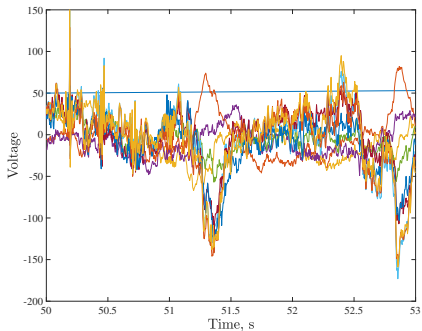
Прогнозирование движений конечностей по электрокортикограммам (ECoG) с учетом движений глаз



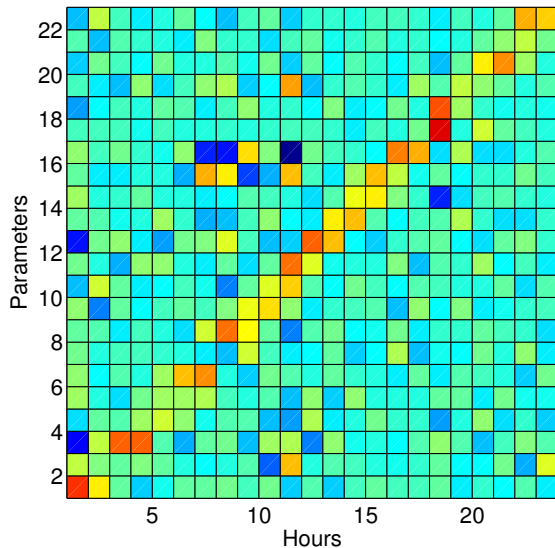
<http://neurotycho.org/>

Макарчук Г. И., Задаянчук А. И., Стрижов В. В. Преобразования временных рядов для декодирования движения руки с помощью ECoG сигналов (electrocorticographic signals) у обезьян // Машинное обучение и анализ данных, 2016.

Source signals and scalogram of ECoG

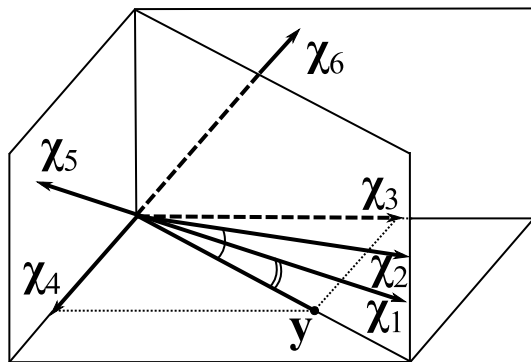


Case 6. How many parameters must be used? Relations of 24 hourly models



Selection of a stable set of features of restricted size

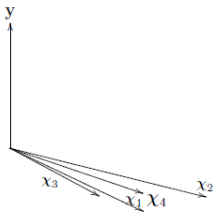
The sample contains multicollinear χ_1, χ_2 and noisy χ_5, χ_6 features, columns of the design matrix $\mathbf{X}$. We want to select two features from six.



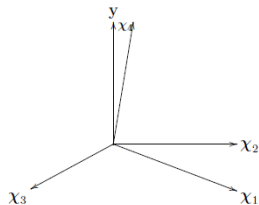
Stability and accuracy for a fixed complexity

The solution: χ_3, χ_4 is an orthogonal set of features minimizing the error function.

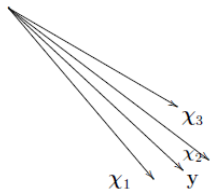
Multicollinear features and the forecast: possible configurations



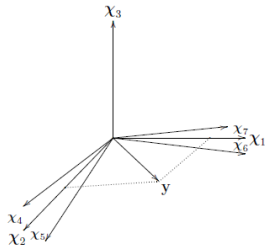
Inadequate and correlated



Adequate and random



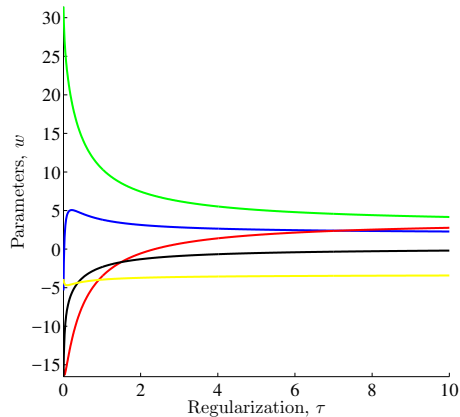
Adequate and redundant



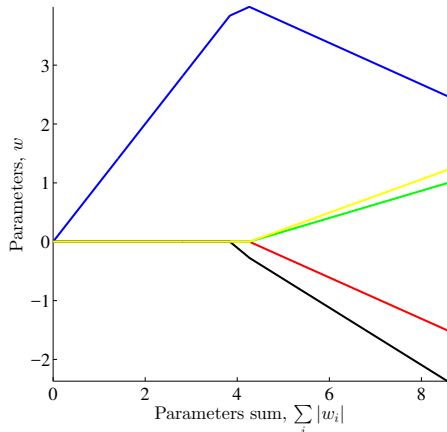
Adequate and correlated

Model parameter values with regularization

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

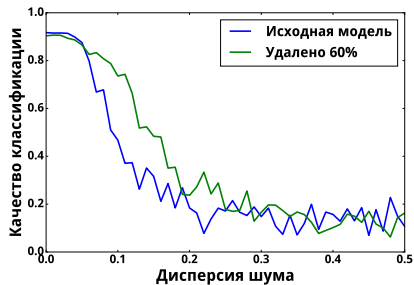


$$S(\mathbf{w}) = \|\mathbf{f}(\mathbf{w}, \mathbf{X}) - \mathbf{y}\|^2 + \gamma^2 \|\mathbf{w}\|^2$$

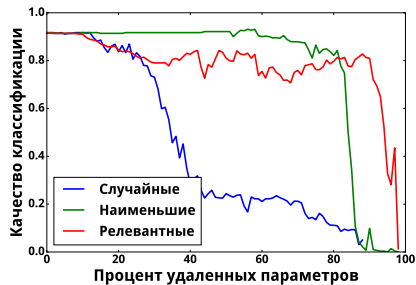


$$S(\mathbf{w}) = \|\mathbf{f}(\mathbf{w}, \mathbf{X}) - \mathbf{y}\|^2, \quad T(\mathbf{w}) \leq \tau$$

Сложность модели: зачем?



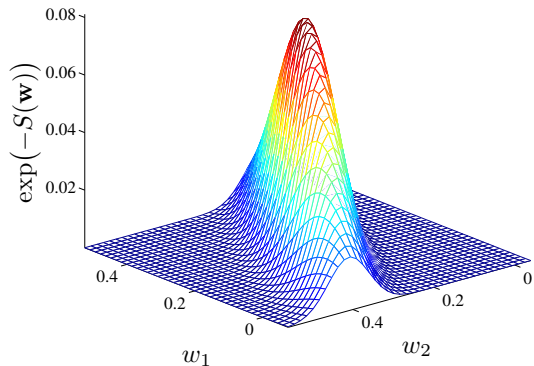
Устойчивость моделей при возмущении выборки



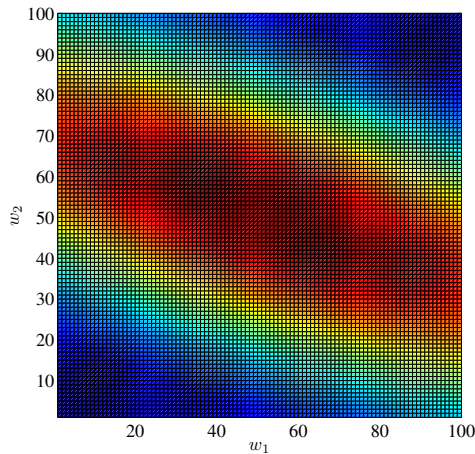
Качество классификации при удалении параметров

Empirical distribution of model parameters

There given a sample $\{\mathbf{w}_1, \dots, \mathbf{w}_K\}$ of realizations of the m.r.v. $\mathbf{w}$ and an error function $S(\mathbf{w}|\mathcal{D}, \mathbf{f})$. Consider the set of points $\{s_k = \exp(-S(\mathbf{w}_k|\mathcal{D}, \mathbf{f})) | k = 1, \dots, K\}$.



$$p(\mathbf{w}|\mathcal{D}, \mathbf{A}, \mathbf{B}, \mathbf{f}) = \frac{1}{Z_{\mathcal{D}}} p(\mathcal{D}|\mathbf{w}, \mathbf{B}, \mathbf{f}) p(\mathbf{w}|\mathbf{A}, \mathbf{f})$$



Minimize number of similar and maximize number of relevant features

Introduce a feature selection method $QP(\text{Sim}, \text{Rel})$ to solve the optimization problem

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

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

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

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

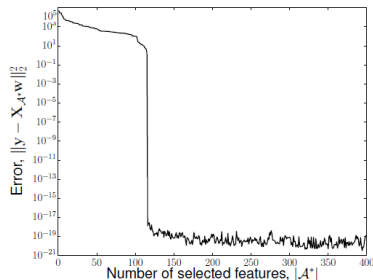
$$\mathbf{b} = [b_i] = \text{Rel}(\chi_i),$$

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 $\text{Sim} \rightarrow \min$, number of correlated to the target $\text{Rel} \rightarrow \max$.

Evaluation criteria for the diesel NIR spectra data set

Method	C_p	RSS	$\ln \frac{\lambda_1}{\lambda_n}$ SVD	VIF	BIC
QP (ρ, ρ) ($\tau = 10^{-9}$)	-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. and Strijov V. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria // Expert Systems with Applications, 2017.

Инструменты анализа больших данных

Языки программирования

- ▶ Matlab, Python

Библиотеки глубокого обучения

- ▶ Caffe, Theano, Lasagne

Пользовательские системы анализа больших данных

- ▶ IBM Watson, Google TensorFlow

Ускорители глубокого обучения

- ▶ Titan DIGITS DevBox, CUDA

Сервера распределенных облачных вычислений

- ▶ Amazon Web Services, Microsoft Azure

Разрушение гипотезы о простоте выборки привело к появлению методов

- ▶ мультимоделирования,
- ▶ онлайн-обучения,
- ▶ мультизадачного обучения,
- ▶ глубокого обучения,
- ▶ структурного обучения.

Проблемы и прогнозы

Проблемы развития методов анализа больших данных:

- ▶ недостаточно высокая скорость возникновения открытых коллекций,
- ▶ относительно высокая стоимость разметки данных (assessment, mecturk, captcha),
- ▶ сложности с организацией обмена информацией, ...

постепенно решаются.

Ожидается появление методов

- ▶ прогнозирования сложных и уникальных структур,
- ▶ прогнозирования прогностических моделей,
- ▶ автоматического переноса опыта моделирования из одной предметной области в другую,
- ▶ автоматического создания оптимальных сетей глубокого обучения.