Business intelligence III: Feature generation and model selection for multiscale time series forecasting

Vadim Strijov

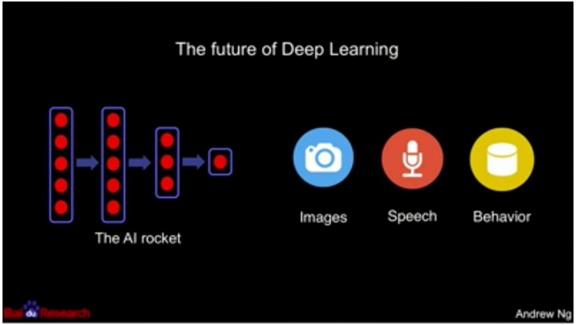
Moscow Institute of Physics and Technology

AINL FRUCT
2016, 10 - 12 of November

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



Case 1. Energy consumption and price forecasting, 1-day ahead hourly

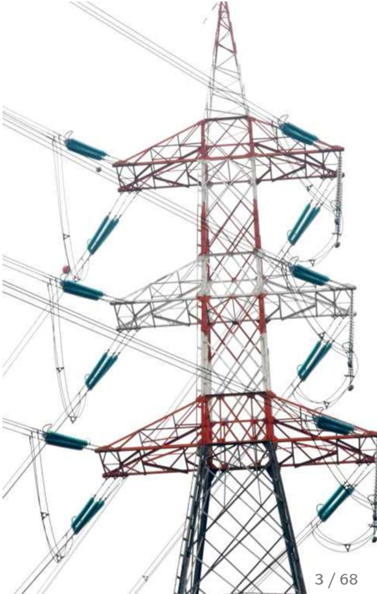
The components of multivariate time series with periodicity

Time series:

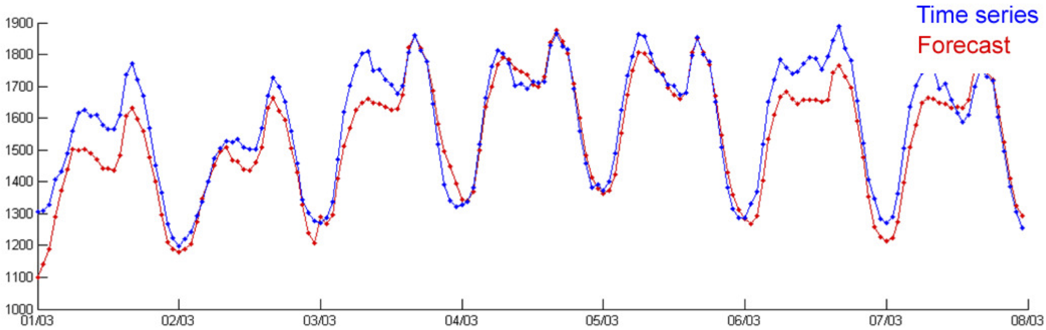
- energy price,
- consumption,
- daytime,
- temperature,
- humidity,
- wind force,
- holiday schedule.

Periodicity:

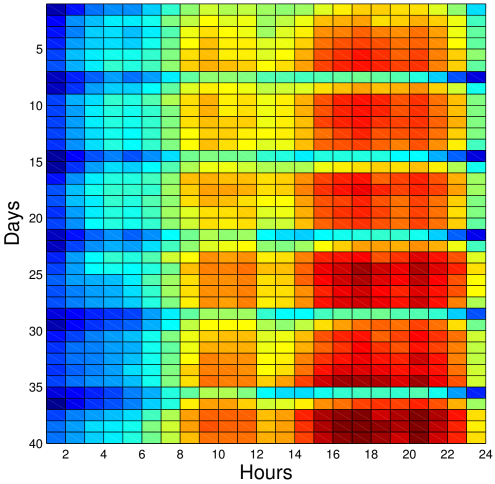
- one year seasons (temperature, daytime),
- one week,
- one day (working day, week-end),
- a holiday,
- aperiodic events.



Energy consumption one-week forecast for each hour



The autoregressive matrix, five weeks



The autoregressive matrix and the linear model

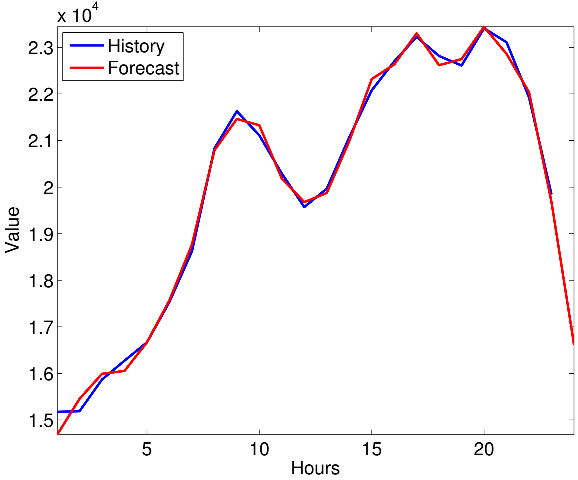
$$\underset{(m+1)\times(n+1)}{\mathbf{X}^*} = \left[\begin{array}{c|ccc} \hat{s}_T & s_{T-1} & \dots & s_{T-\kappa+1} \\ \hline s_{(m-1)\kappa} & s_{(m-1)\kappa-1} & \dots & s_{(m-2)\kappa+1} \\ \dots & \dots & \dots & \dots \\ s_{n\kappa} & s_{n\kappa-1} & \dots & s_{n(\kappa-1)+1} \\ \dots & \dots & \dots & \dots \\ s_{\kappa} & s_{\kappa-1} & \dots & s_1 \end{array}\right] = \left[\begin{array}{c|c} \underset{1\times 1}{\hat{s}_T} & \underset{1\times n}{\mathbf{x}_{m+1}} \\ \hline \underset{m\times 1}{\mathbf{y}} & \underset{m\times n}{\mathbf{X}} \end{array}\right].$$

In terms of linear regression:

$$\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \mathbf{w}) = \mathbf{X}\mathbf{w},$$

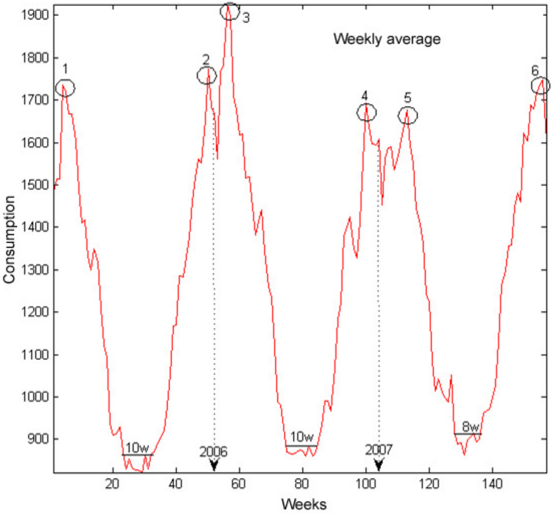
$$\hat{y}_{m+1} = \hat{s}_T = \langle \mathbf{x}_{m+1}, \hat{\mathbf{w}} \rangle.$$

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

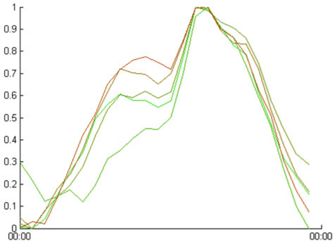


The model $\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \mathbf{w})$ could be a linear model, neural network, deep NN, SVN, . . .

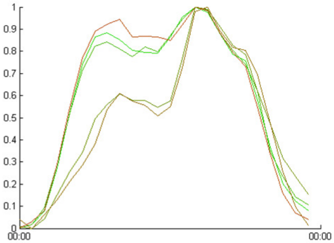
Structure of energy consumption



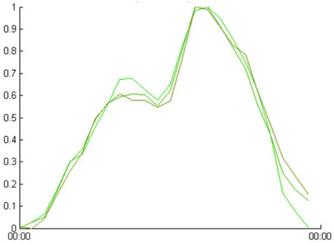
Similarity of daily consumption



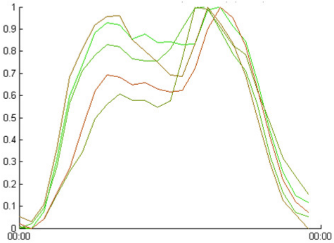
Five days of one week



One day of five consequent weeks

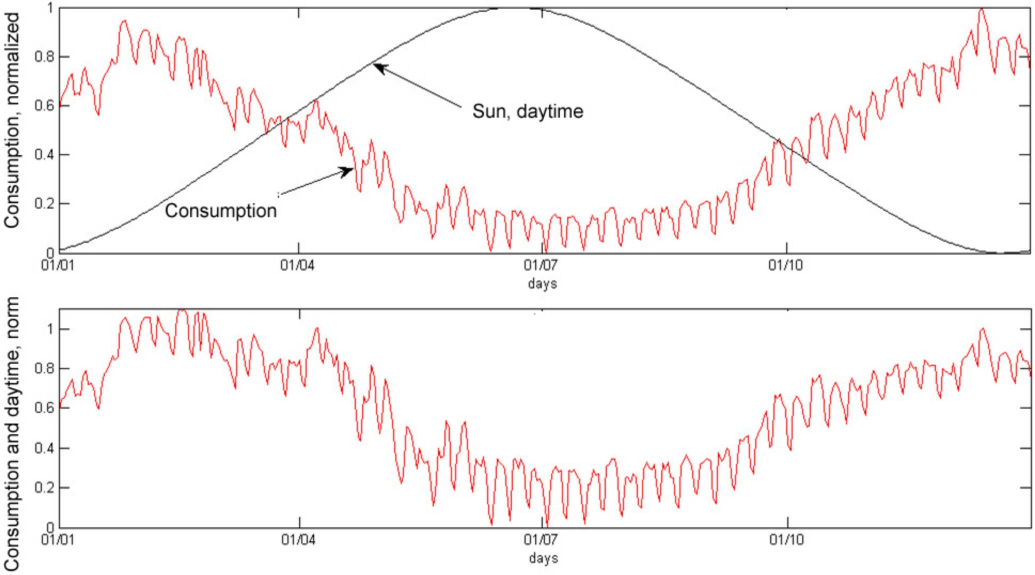


The same weekday and month of three years

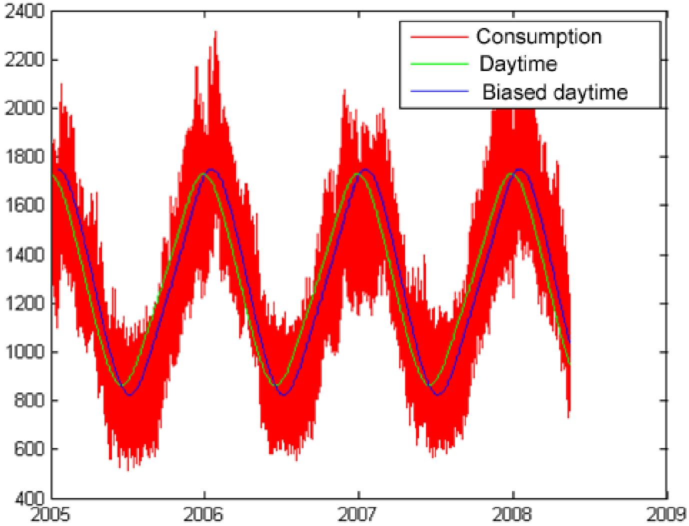


The same weekday of five months

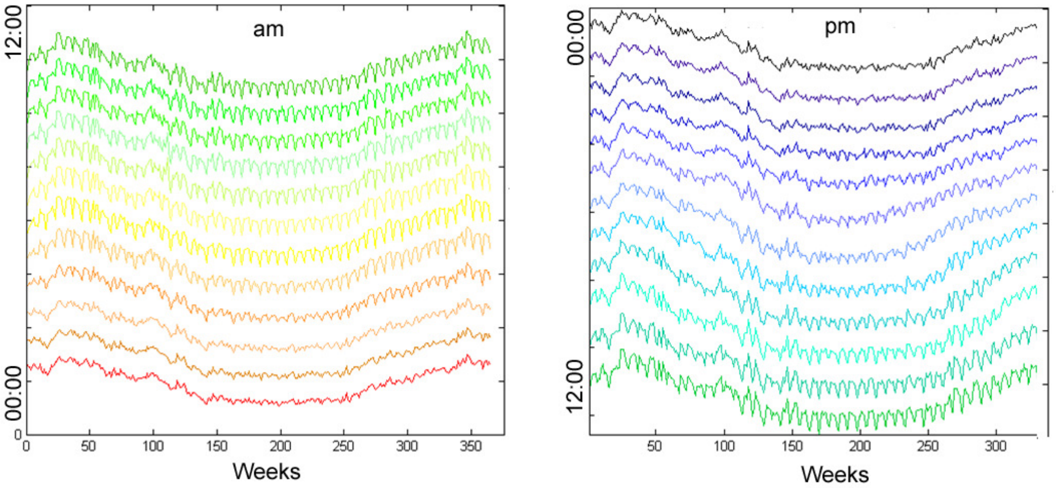
Sunrise bias: one-year daytime and consumption



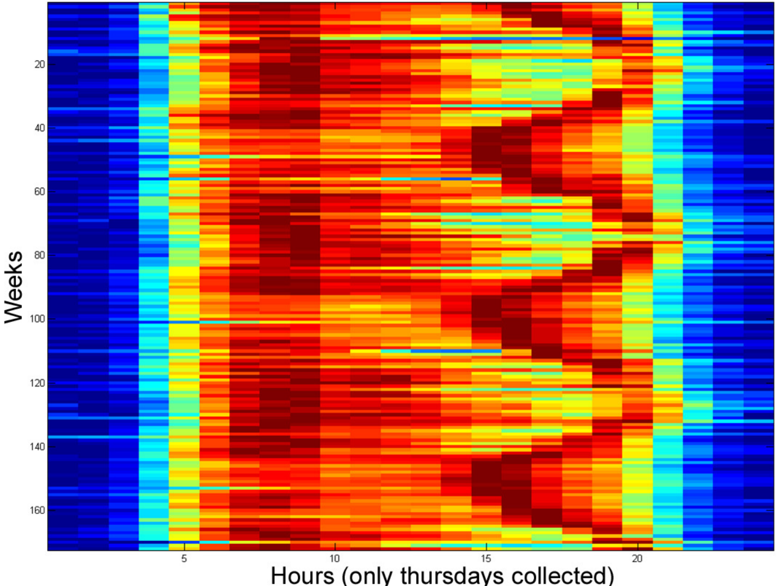
Biased and original daytime to fit consumption over years



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



Daily similarity



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.

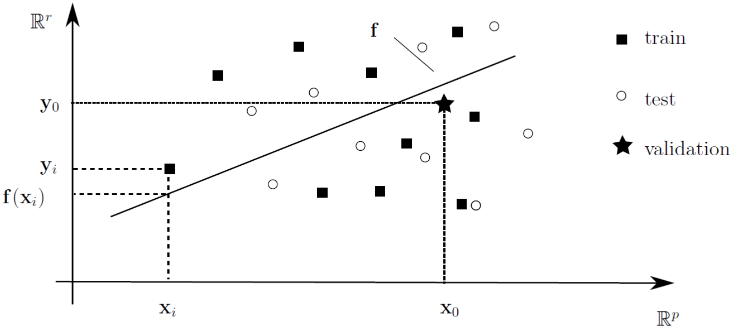
Error: the residue $\varepsilon_j = \hat{y}_j - y_j$ for

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

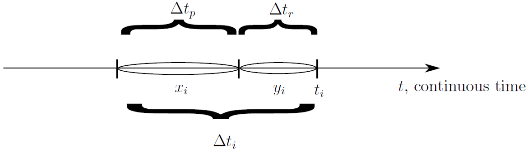
$$RSS = \sum_{j=1}^{r} \varepsilon_j^2, \quad MAPE = \frac{1}{r}\sum_{j=1}^{r} \frac{|\varepsilon_j|}{|y_j|}, \quad sMAPE = \frac{1}{r}\sum_{j=1}^{r} \frac{2|\varepsilon_j|}{|\hat{y}_j + y_j|}.$$

Design matrix

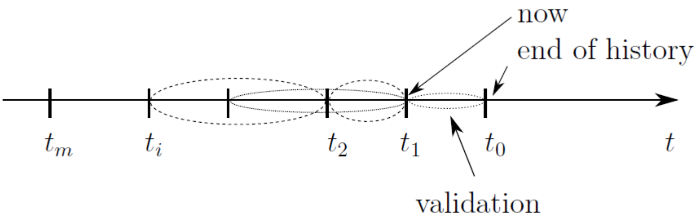
Forecast is a mapping from p-dimensional objects space to r-dimensional answers space.



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



Rolling validation



$$\mathbf{Z} = \left[\begin{array}{c|c} \dots & \dots \\ \hline \underset{1\times n}{\mathbf{x}_{\text{val},k}} & \underset{1\times r}{\mathbf{y}_{\text{val},k}} \\ \underset{m_{\min}\times n}{\mathbf{X}_{\text{train},k}} & \underset{m_{\min}\times r}{\mathbf{Y}_{\text{train},k}} \\ \hline \dots & \dots \end{array}\right] \uparrow k$$

The rolling validation procedure

1) construct the validation vector $\mathbf{x}^*_{\text{val},k}$ for time series of the length Δt_r as the first row of the design matrix $\mathbf{Z}$,
2) construct the rest rows of the design matrix $\mathbf{Z}$ for the time after t_k and present it as
3) optimize model parameters $\mathbf{w}$ using $\mathbf{X}_{\text{train},k}, \mathbf{Y}_{\text{train},k}$ and compute residues $\varepsilon_k = \mathbf{y}_{\text{val},k} - \mathbf{f}(\mathbf{x}_{\text{val}_k}, \mathbf{w})$ and MAPE,
4) increase k and repeat.

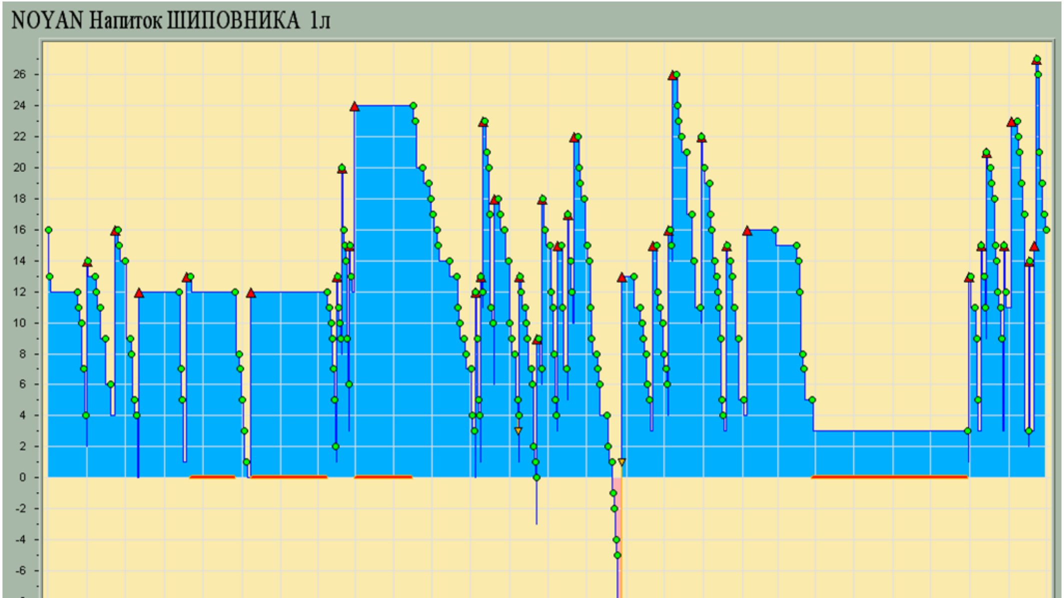
Case 2. Sales planning: to forecast the goods consumption

Retailers' daily routines:

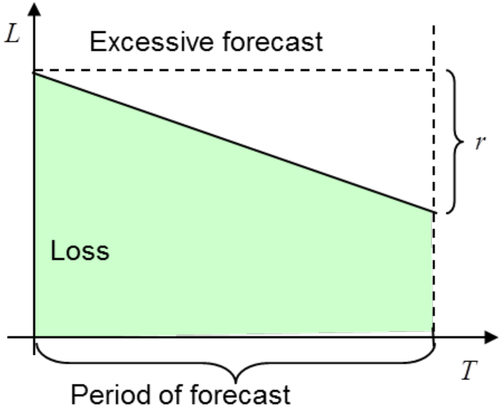
- custom inventory,
- calculation of optimal insurance stocks,
- consumer demand forecasting.

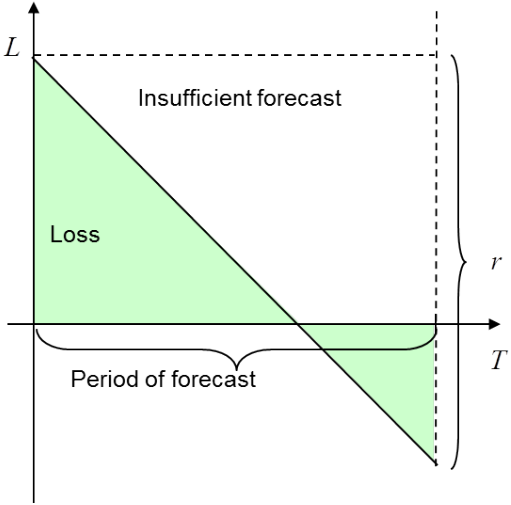
- There given historical time series of the volume off-takes: foodstuff.
- Let the time series be homoscedastic: its variance is time-constant.
- Minimizing the loss function one must forecast the next sample.

Custom inventory



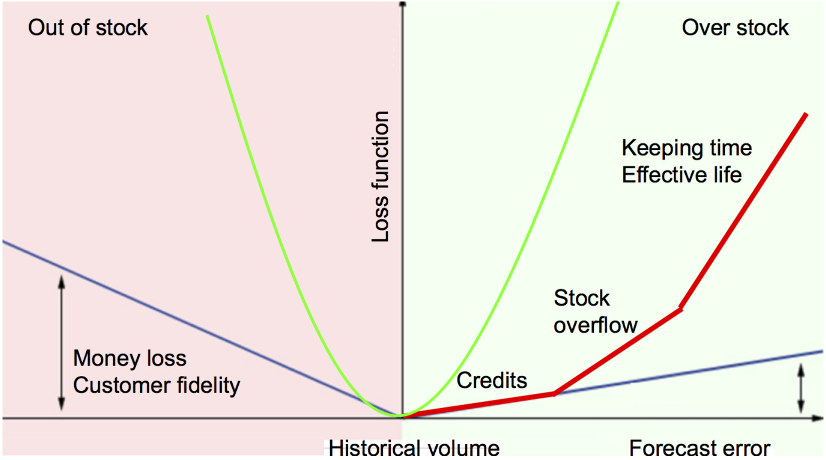
Excessive forecast and insufficient forecast lead to loss



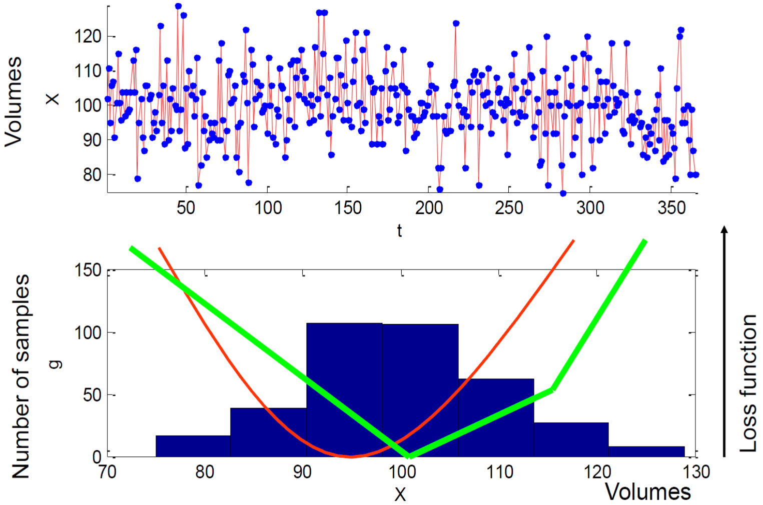


The performance criterion is **minimum loss of money**

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

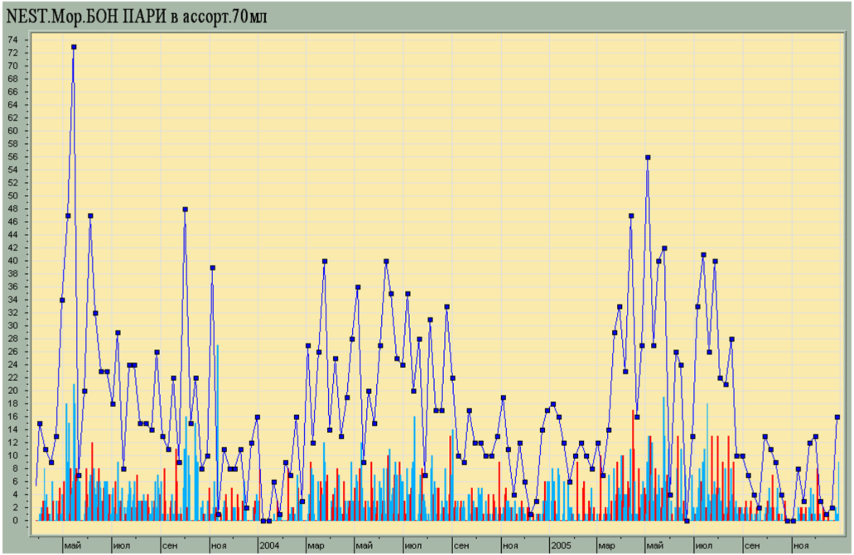
$$L = L(Z, X).$$

The optimal forecast is

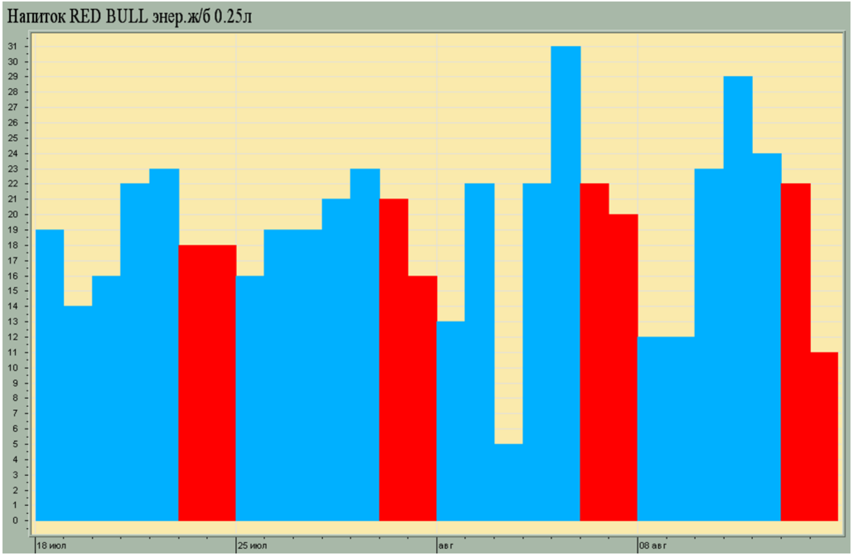
$$\tilde{X} = \mathop{\arg\min}_{Z \in \{X_1 \ldots X_m\}} \sum_{i=1}^{m} g_i L(Z, X_i)$$

of this convolution.

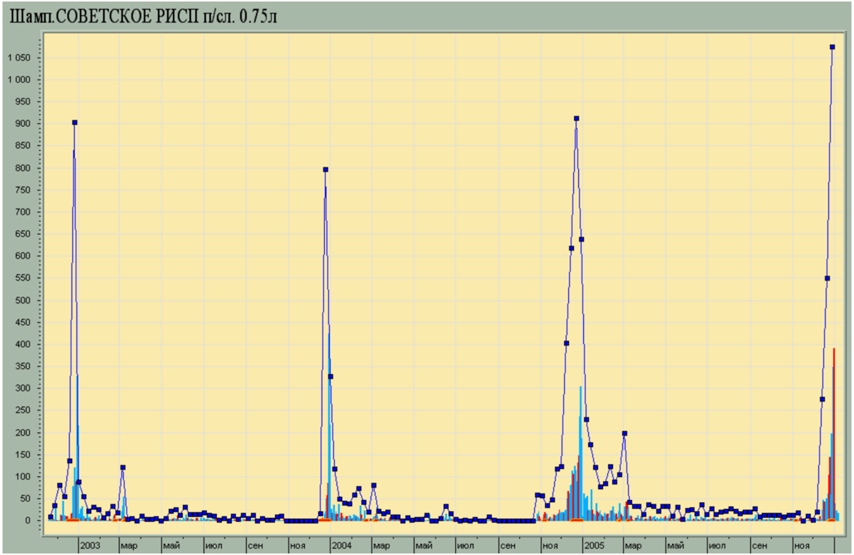
Candies: the seasonality and trend weekly over three years



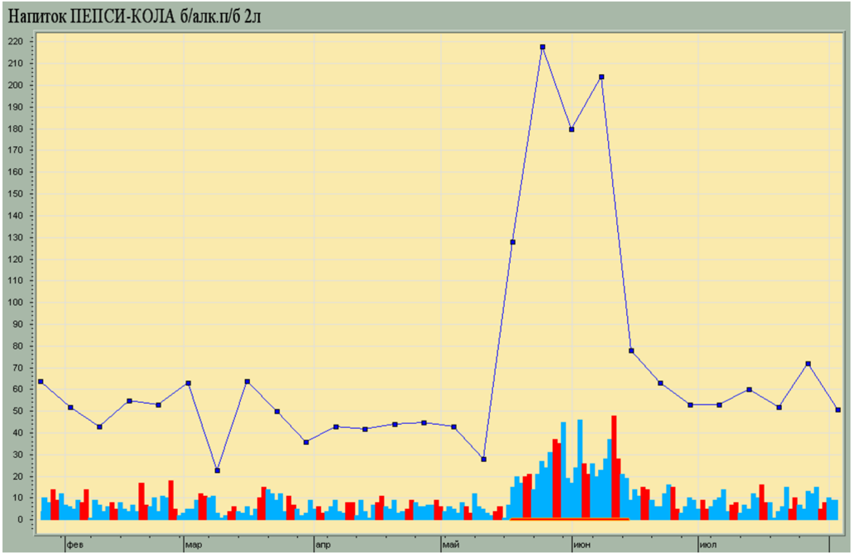
Beverage: the week periodicity daily over four weeks



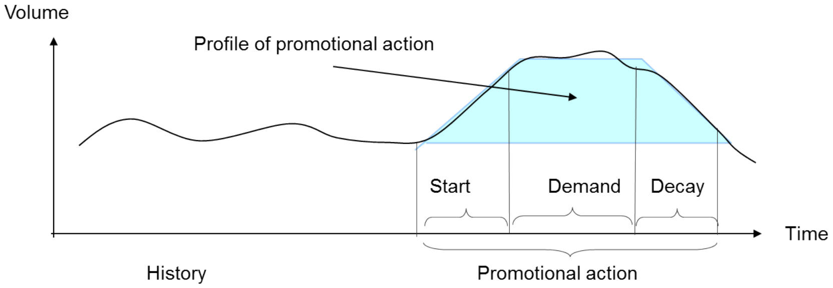
Sparkling wine: holidays weekly over three years



Promotional actions



Promotional profile extraction



- Hypothesis: the shape of the profile (excluding the profile parameters) does not depend of duration of the action.
- Problem: to forecast the customer demand during the promotional action.

Forecast the residues to boost performance

$$\begin{tikzcd}
\cdots & \hat{x}^f_{t-1} \arrow[dd, dotted, bend right] & \hat{x}^f_t \arrow[dd, dotted, bend right] & \hat{x}^f_{t+1} \arrow[d, "\oplus"] & \hat{x}^f_{t+2} \arrow[d, "\oplus"] & \cdots \\
\cdots \arrow[ur, bend left, "f"] & x_{t-1} \arrow[ur, bend left, "f"] \arrow[d, dotted] & x_t \arrow[ur, bend left, "f"] \arrow[d, dotted] & \hat{x}^{f,g}_{t+1} \arrow[ur, bend left, "f"] & \hat{x}^{f,g}_{t+2} \arrow[ur, bend left, "f"] & \cdots \\
\cdots & \hat{\varepsilon}_{t-1} & \hat{\varepsilon}_t \arrow[r, "g"'] & \hat{\varepsilon}_{t+1} \arrow[r, "g"'] \arrow[u, "\oplus"'] & \hat{\varepsilon}_{t+2} \arrow[r, "g"'] \arrow[u, "\oplus"'] & \cdots
\end{tikzcd}$$

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

Case 3. Forecasting volumes of Russian railways freight transportation

Keep a hierarchical structure of time series without loosing performance

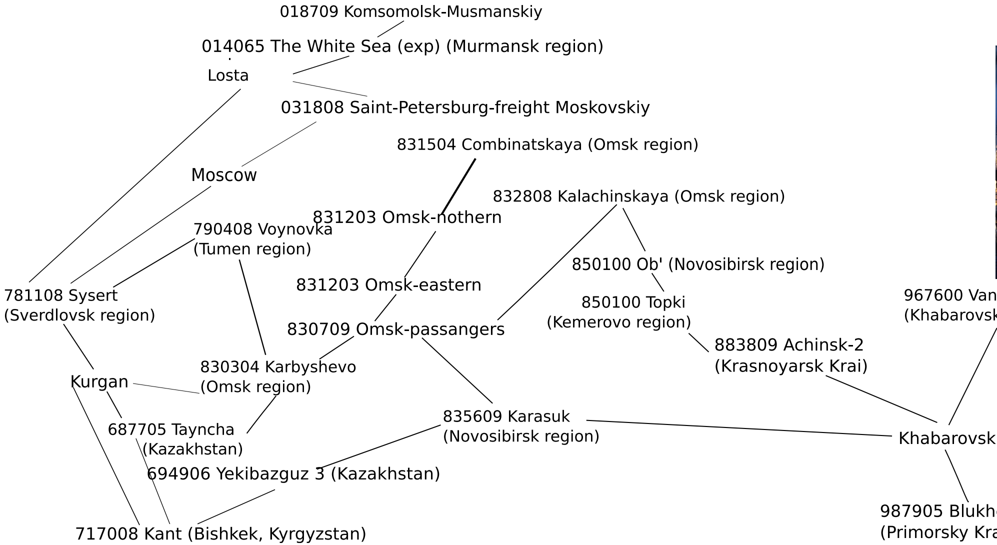
Forecast with hierarchical aggregation of

- 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



018709 Komsomolsk-Musmanskiy

014065 The White Sea (exp) (Murmansk region)

Losta

031808 Saint-Petersburg-freight Moskovskiy

831504 Combinatskaya (Omsk region)

Moscow

832808 Kalachinskaya (Omsk region)

831203 Omsk-nothern

790408 Voynovka
(Tumen region)

850100 Ob' (Novosibirsk region)

831203 Omsk-eastern

781108 Sysert
(Sverdlovsk region)

850100 Topki
(Kemerovo region)

967600 Vanino
(Khabarovsk Krai)

830709 Omsk-passangers

883809 Achinsk-2
(Krasnoyarsk Krai)

830304 Karbyshevo
(Omsk region)

Kurgan

835609 Karasuk
(Novosibirsk region)

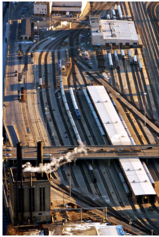
Khabarovsk

687705 Tayncha
(Kazakhstan)

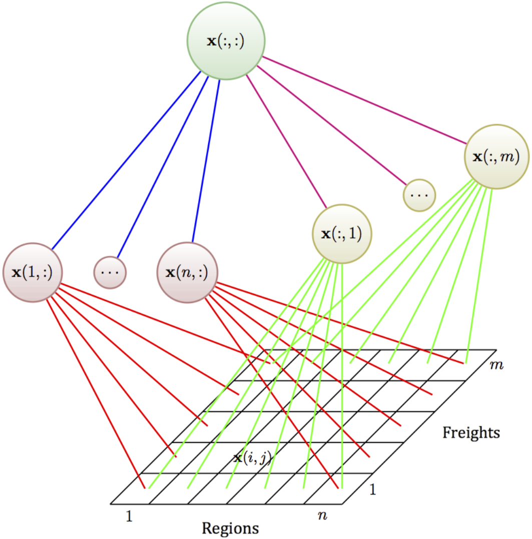
694906 Yekibazguz 3 (Kazakhstan)

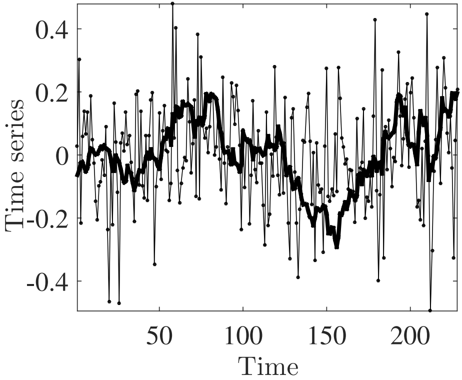
987905 Blukher
(Primorsky Krai)

717008 Kant (Bishkek, Kyrgyzstan)



Independent forecasts might be inconsistent with aggregated ones





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

$$\chi_t = \begin{pmatrix} x_t(:,:) \\ \dots \\ x_t(n,1) \\ \dots \\ x_t(n,m) \end{pmatrix}, \quad \hat{\chi} = \begin{pmatrix} \hat{x}(:,:) \\ \dots \\ \hat{x}(n,1) \\ \dots \\ \hat{x}(n,m) \end{pmatrix}, \quad \hat{\varphi} = \begin{pmatrix} \hat{y}(:,:) \\ \dots \\ \hat{y}(n,1) \\ \dots \\ \hat{y}(n,m) \end{pmatrix}.$$

Consistency condition $\mathbf{S}\chi_t = \mathbf{0}$, $t = 1, \dots, T$, where the link matrix $\mathbf{S}$ of size $(2+n+m) \times (1+n+m+nm)$ has the form

$$\mathbf{S} = \left(\begin{array}{c|ccc|ccc|cccc|c|cccc} -1 & 1 & \dots & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 \\ -1 & 0 & \dots & 0 & 1 & \dots & 1 & 0 & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 \\ \hline 0 & -1 & \dots & 0 & 0 & \dots & 0 & 1 & 1 & \dots & 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & & \dots & & & \dots & & & & \dots & & & & & \dots & \\ 0 & 0 & \dots & -1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & \dots & 1 & 1 & \dots & 1 \\ \hline 0 & 0 & \dots & 0 & -1 & \dots & 0 & 1 & 0 & \dots & 0 & \dots & 1 & 0 & \dots & 0 \\ \dots & & \dots & & & \dots & & & & \dots & & & & & \dots & \\ 0 & 0 & \dots & 0 & 0 & \dots & -1 & 0 & 0 & \dots & 1 & \dots & 0 & 0 & \dots & 1 \end{array}\right).$$

The regions

$$x_t(:,:) = \sum_{i=1}^{n} x_t(i,:);$$

The freights

$$x_t(:,:) = \sum_{j=1}^{m} x_t(:,j);$$

Freights given a region

$$x_t(i,:) = \sum_{j=1}^{m} x_t(i,j),$$

$i = 1, \dots n;$

Regions given a freight

$$x_t(:,j) = \sum_{i=1}^{n} x_t(i,j),$$

$j = 1, \dots m;$

$t = 1, \dots, T.$

Performance of independent and reconciliated forecasts

There given link matrix $\mathbf{S}$, admissible sets $\mathbf{A}$, $\mathbf{B}$ and independent forecasts $\hat{\chi} \notin \mathbf{A}, \quad \hat{\chi} \in \mathbf{B}$ **to make** reconciliated forecast $\hat{\varphi}$ subject to

- consistency $\hat{\varphi} \in \mathbf{A}$, $\mathbf{A} = \{\chi \in \mathbb{R}^d \mid \mathbf{S}\chi = \mathbf{0}\}$,
- 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 $\chi_{T+1} \in \mathbf{A} \cap \mathbf{B}$.

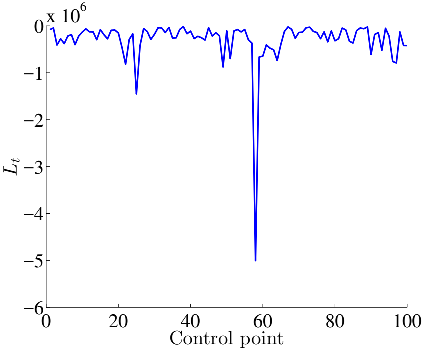
Theorem [Maria Stenina, 2014]

Given listed conditions, the projection vector

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

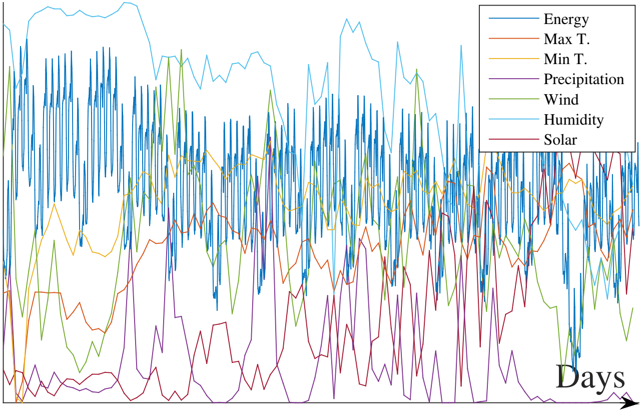
is guaranteed to satisfy the requirements of consistency, physical limitations and precision.

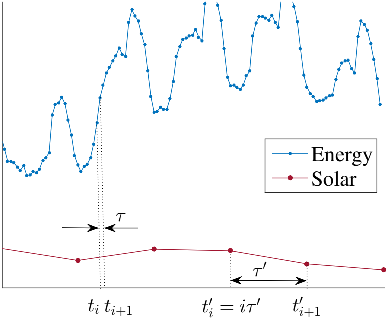
The solution of the optimization problem $\hat{\varphi} = \underset{\chi \in \mathbf{A} \cap \mathbf{B}}{\arg\min} \|\chi - \hat{\chi}\|_2^2$ demonstrates decrease in loss for all control samples.



$$L_t = \|\chi_t - \hat{\varphi}\|_2^2 - \|\chi_t - \hat{\chi}\|_2^2$$

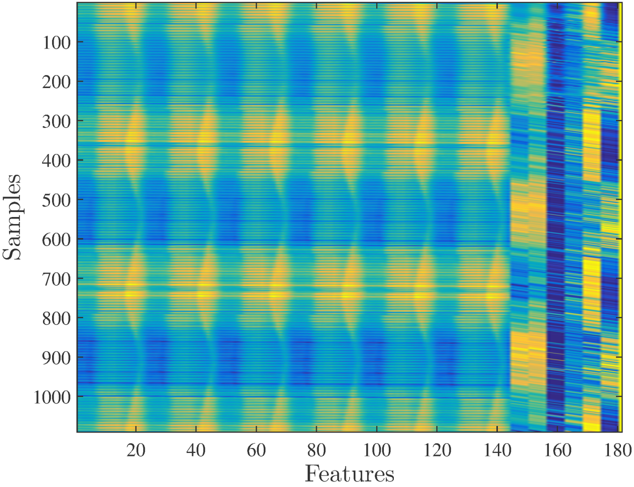
Case 4. Internet of things, multiscale dataset for vector forecasting

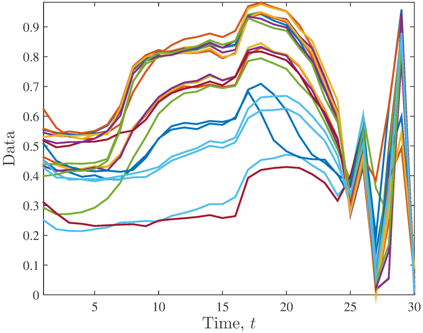




Each real-valued time series $\mathbf{s} = [s_1, \dots, s_i, \dots, s_T]$, $s_i = s(t_i)$, $0 \le t_i \le t_{\max}$ is a sequence of observations of some real-valued signal $s(t)$ with its own sampling rate τ.

To boost the forecast quality include external variables





Generate features with nonparametric transformation functions

Univariate

Formula	Output dimension
$\sqrt{x}$	1
$x\sqrt{x}$	1
$\arctan x$	1
$\ln x$	1
$x \ln x$	1

Bivariate

Plus	$x_1 + x_2$
Minus	$x_1 - x_2$
Product	$x_1 \cdot x_2$
Division	$\frac{x_1}{x_2}$
	$x_1\sqrt{x_2}$
	$x_1 \ln x_2$

Nonparametric aggregation: sample statistics

Nonparametric transformations include basic data statistics:

- Sum or average value of each row $\mathbf{x}_i$, $i = 1, \ldots, m$:

$$\phi_i = \sum_{j=1}^{n} x_{ij}, \text{ or } \phi'_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}.$$

- Min and max values: $\phi_i = \min_j x_{ij}$, $\phi'_i = \max_j x_{ij}$.
- Standard deviation:

$$\phi_i = \frac{1}{n-1}\sqrt{\sum_{j=1}^{n}(x_{ij} - \text{mean}(\mathbf{x}_i))^2}.$$

- Data quantiles: $\phi_i = [X_1, \ldots, X_K]$, where

$$\sum_{j=1}^{n}[X_{k-1} < x_{ij} \leq X_k] = \frac{1}{K}, \text{ for } k = 1, \ldots, K.$$

Nonparametric transformations: Haar's transform

Applying Haar's transform produces multiscale representations of the same data.

Assume that $n = 2^K$ and init $\phi_{i,j}^{(0)} = \phi_{i,j}'^{(0)} = x_{ij}$ for $j = 1, \ldots, n$.

To obtain coarse-graining and fine-graining of the input feature vector $\mathbf{x}_i$, for $k = 1, \ldots, K$ repeat:

- data averaging step

$$\phi_{i,j}^{(k)} = \frac{\phi_{i,2j-1}^{(k-1)} + \phi_{i,2j}^{(k-1)}}{2}, \quad j = 1, \ldots, \frac{n}{2^k},$$

- and data differencing step

$$\phi_{i,j}'^{(k)} = \frac{\phi_{i,2j}'^{(k-1)} - \phi_{i,2j-1}'^{(k-1)}}{2}, \quad j = 1, \ldots, \frac{n}{2^k}.$$

The resulting multiscale feature vectors are $\phi_i = [\phi_i^{(1)}, \ldots, \phi_i^{(K)}]$ and $\phi_i' = [\phi_i'^{(1)}, \ldots, \phi_i'^{(K)}]$.

Monotone functions

- **By grow rate**

Function name	Formula	Constraints
Linear	$w_1 x + w_0$	
Exponential rate	$\exp(w_1 x + w_0)$	$w_1 > 0$
Polynomial rate	$\exp(w_1 \ln x + w_0)$	$w_1 > 1$
Sublinear polynomial rate	$\exp(w_1 \ln x + w_0)$	$0 < w_1 < 1$
Logarithmic rate	$w_1 \ln x + w_0$	$w_1 > 0$
Slow convergence	$w_0 + w_1 / x$	$w_1 \neq 0$
Fast convergence	$w_0 + w_1 \cdot \exp(-x)$	$w_1 \neq 0$

- **Other**

Soft ReLu	$\ln(1 + e^x)$	
Sigmoid	$1/(w_0 + \exp(-w_1 x))$	$w_1 > 0$
Softmax	$1/(1 + \exp(-x))$	
Hiberbolic tangent	$\tanh(x)$	
softsign	$\frac{\lvert x \rvert}{1+\lvert x \rvert}$	

Collection of parametric transformation functions

Function name	Formula	Output dim.	Num. of args	Num. of pars
Add constant	$x + w$	1	1	1
Quadratic	$w_2 x^2 + w_1 x + w_0$	1	1	3
Cubic	$w_3 x^3 + w_2 x^2 + w_1 x + w_0$	1	1	4
Logarithmic sigmoid	$1/(w_0 + \exp(-w_1 x))$	1	1	2
Exponent	$\exp x$	1	1	0
Normal	$\frac{1}{w_1\sqrt{2\pi}} \exp\left(\frac{(x-w_2)^2}{2w_1^2}\right)$	1	1	2
Multiply by constant	$x \cdot w$	1	1	1
Monomial	$w_1 x^{w_2}$	1	1	2
Weibull-2	$w_1 w_2 x^{w_2 - 1} \exp -w_1 x^{w_2}$	1	1	2
Weibull-3	$w_1 w_2 x^{w_2 - 1} \exp -w_1 (x - w_3)^{w_2}$	1	1	3
...	...	...	...	...

Parametric transformations

Optimization of the transformation function parameters **b** is iterative:

1. Fix the vector $\hat{\mathbf{b}}$, collected over all the primitive functions $\{g\}$, which generate features ϕ:

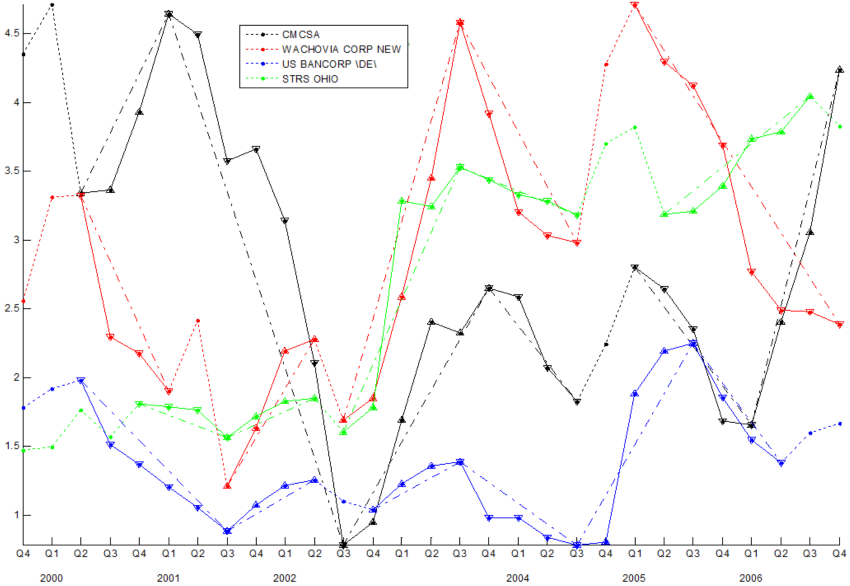
$$\hat{\mathbf{w}} = \arg\min S\big(\mathbf{w}|\mathbf{f}(\mathbf{w}, \mathbf{x}), \mathbf{y}\big), \quad \text{where} \quad \phi(\hat{\mathbf{b}}, \mathbf{s}) \subseteq \mathbf{x}.$$

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

$$\hat{\mathbf{b}} = \arg\min S\big(\mathbf{b}|\mathbf{f}(\hat{\mathbf{w}}, \mathbf{x}), \mathbf{y}\big).$$

Repeat these steps until vectors $\hat{\mathbf{w}}, \hat{\mathbf{b}}$ converge.

Markup the time series, two types of marks: Up and Down



Parameters of the local models

More feature generation options:

- Parameters of SSA approximation of the time series $\mathbf{x}^{(q)}$.
- Parameters of the FFT of each $\mathbf{x}^{(q)}$.
- Parameters of polynomial/spline approximation of each $\mathbf{x}^{(q)}$.

Parameters of the local models: SSA

For the 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

$$\mathbf{H}^* = \left[\begin{array}{c|ccc} s_T & \dots & s_{T-k+1} \\ \vdots & \ddots & \vdots \\ s_{k+p} & \dots & s_{1+p} \\ s_k & \dots & s_1 \end{array}\right], \quad \text{where } 1 \geqslant p \geqslant k.$$

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

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

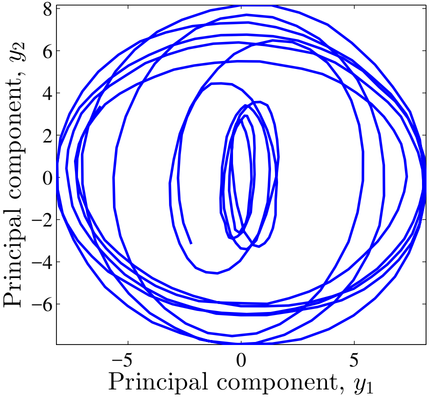
For the orignal feature vector $\mathbf{x} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(Q)}]$ use the parameters $\phi(\mathbf{x}^{(q)})$, $q = 1, \dots, Q$ as the features.

SSA: principal components

Compute the SVD of covariance matrix of $\mathbf{H}$

$$\frac{1}{N}\mathbf{H}^{\mathsf{T}}\mathbf{H} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathsf{T}}, \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$$

and find the principal components $\mathbf{y}_j = \mathbf{H}\mathbf{v}_j$.



Models and features

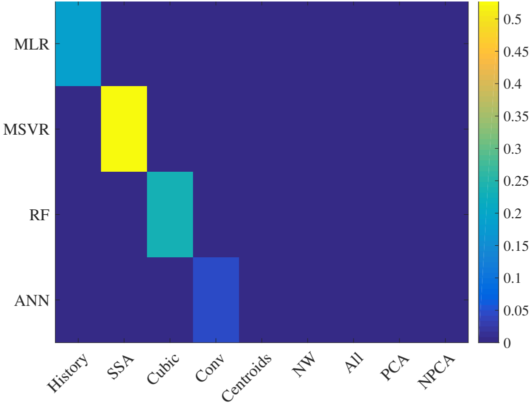
Models:

- 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.

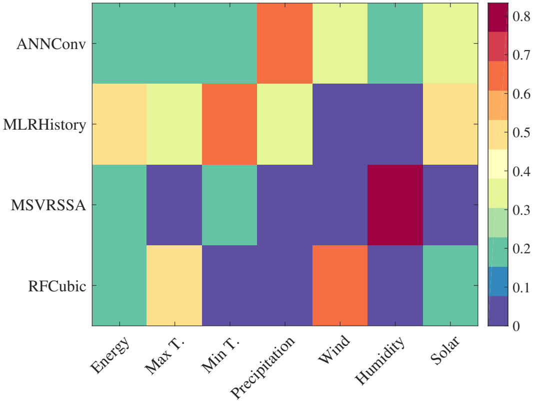
Feature analysis



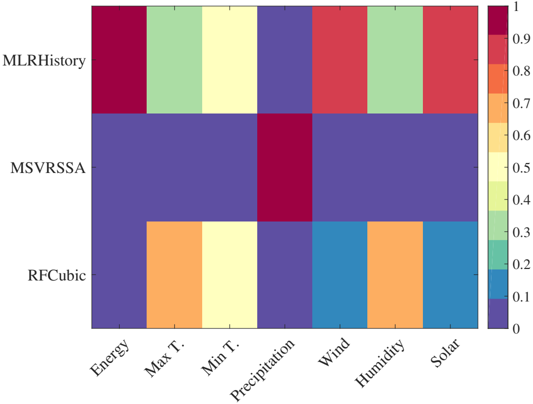
Test subset MAPE

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).

Best models

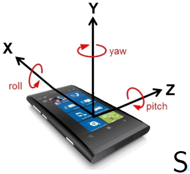


Test subset residues



Standard deviations of test subset residues

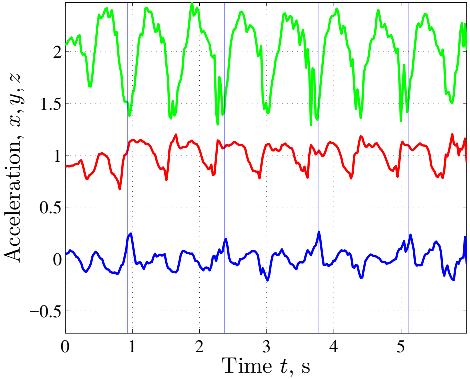
Case 5. Classification of human physical activity with mobile devices



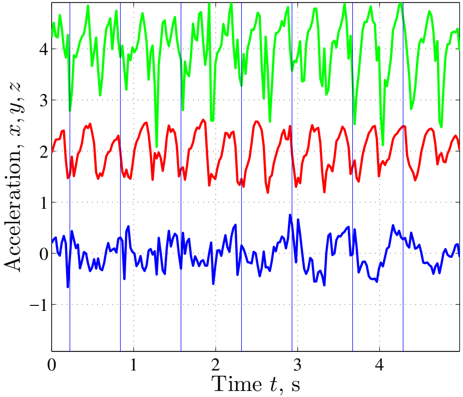
3D-projection of acceleration time series to spatial axis

$$\mathbf{x} = \{acc_x(t); acc_y(t); acc_z(t)\}_{t=1}^{n} \mapsto \mathbf{y} \in \mathbb{R}^S.$$

Slow walking



Jogging



Local transformations for deep learning neural network

Model $\mathbf{f} = \mathbf{a}(\mathbf{h}_N(\ldots\mathbf{h}_1(\mathbf{x})))(\mathbf{w})$ contains autoencoders $\mathbf{h}_k$ and softmax classifier $\mathbf{a}$:

$$\mathbf{f}(\mathbf{w},\mathbf{x}) = \frac{\exp(\mathbf{a}(\mathbf{x}))}{\sum_j \exp(a_j(\mathbf{x}))}, \qquad \mathbf{a}(\mathbf{x}) = \mathbf{W}_2^{\mathsf{T}}\mathbf{tanh}(\mathbf{W}_1^{\mathsf{T}}\mathbf{x}), \qquad \mathbf{h}_k(\mathbf{x}) = \boldsymbol{\sigma}(\mathbf{W}_k\mathbf{x} + \mathbf{b}_k),$$

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

Feature generation by local transformations:

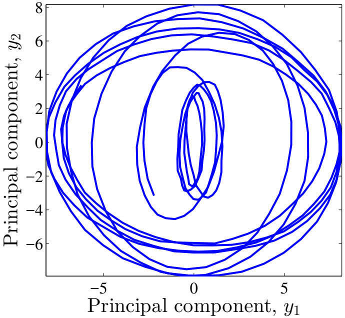
- parameters of SSA approximation of the time series $\mathbf{x}$,
- FFT of $\mathbf{x}$,
- parameters of polynomial/spline approximation,

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

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

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



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

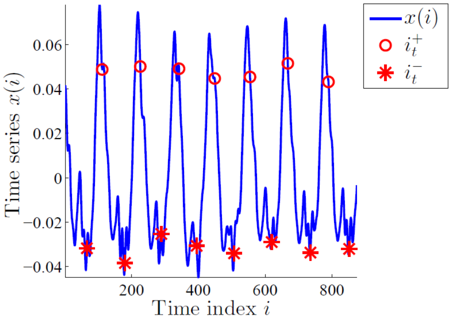
$$\phi(\mathbf{s}) = \arg\min \|\mathbf{h} - \mathbf{H}\phi\|_2^2.$$

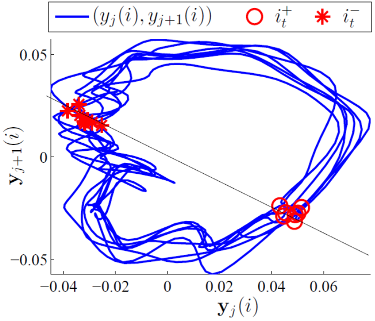
For the original feature vector $\mathbf{x} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(Q)}]$ use the parameters $\phi(\mathbf{x})$ as an item.

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:

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





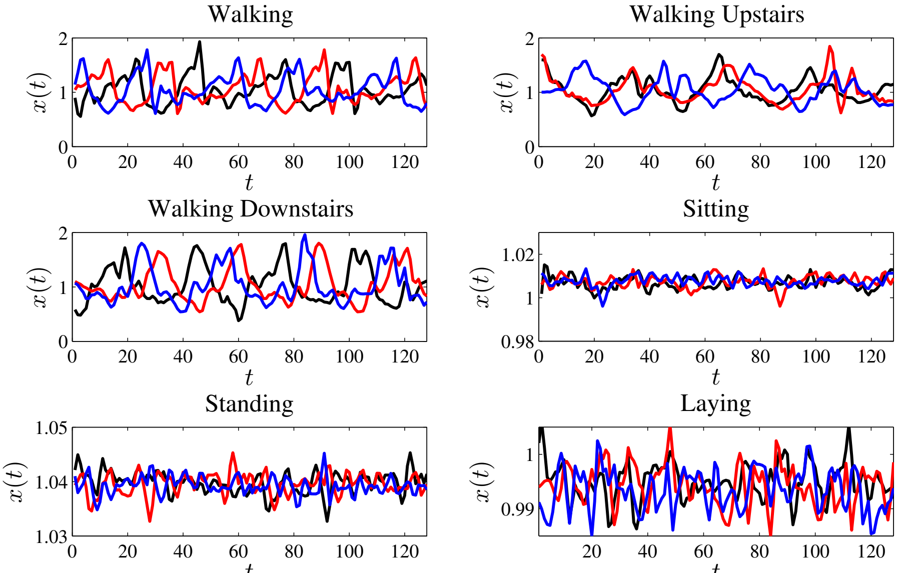
Metric features: distances to the centroids of local clusters

Apply kernel trick to the time series.

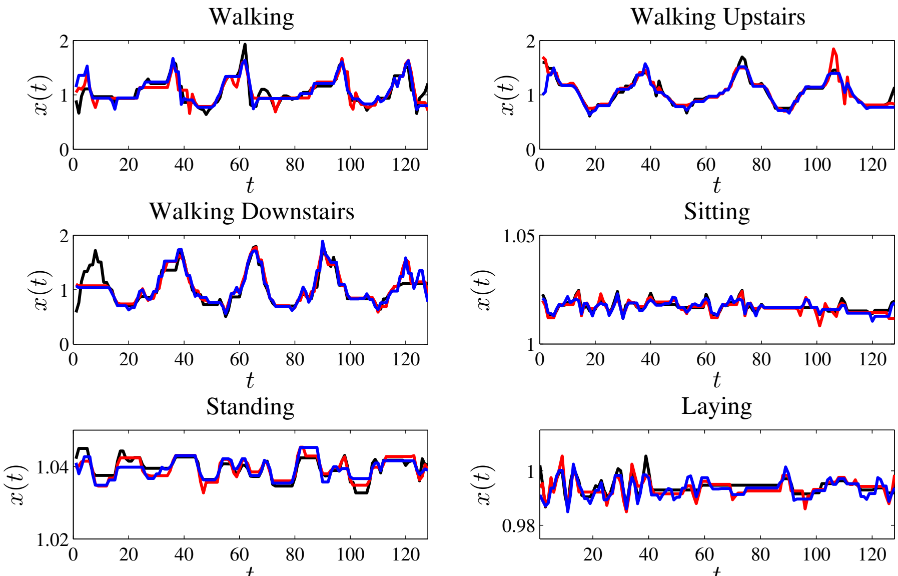
1. For objects $\mathbf{x}_i$ from $\mathbf{X}$ compute k-mean centroids $\mathbf{c}$.
2. Use distance function ρ to combine feature vector

$$\phi_i = [\rho(\mathbf{c}_1, \mathbf{x}_i), \dots, \rho(\mathbf{c}_p, \mathbf{x}_i)] \in \mathbb{R}^p_+.$$

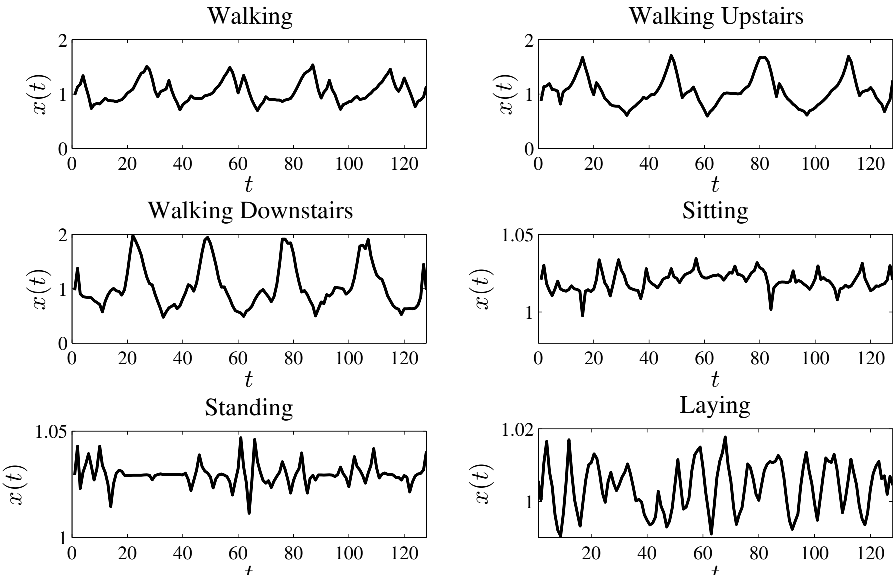
Computing distances to centroids of the time series: **sources**



Computing distances to centroids of the time series: **aligned series**



Computing distances to centroids of the time series: **centroids**



Walking

Walking Upstairs

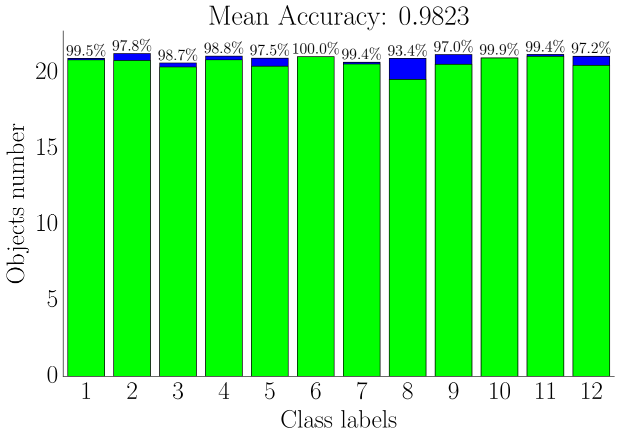
Walking Downstairs

Sitting

Standing

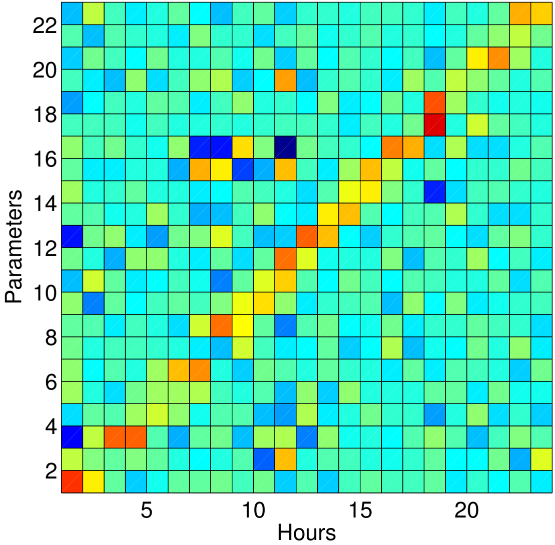
Laying

Performance of the human physical activities classification model



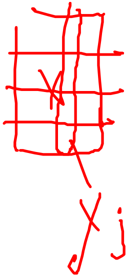
1) walk forward
2) walk left
3) walk right
4) go upstairs
5) go downstairs
6) run forward
7) jump up and down
8) sit and fidget
9) stand
10) sleep
11) elevator up
12) elevator down

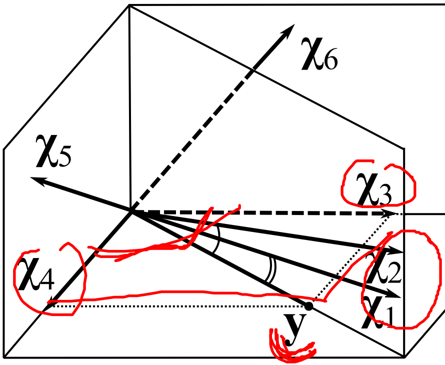
Case 6. How many parameters must be used? Relations of 24 hourly models



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.



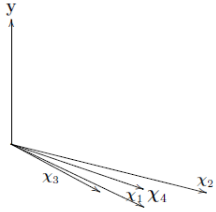


→ 3 items in D

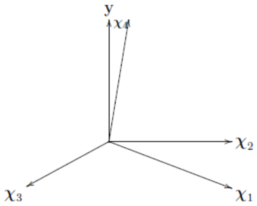
Stability and accuracy for a fixed complexity

The solution: $\chi_3, \chi_4$is an orthogonal set of features minimizing the error function.

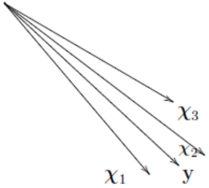
Multicollinear features and the forecast: possible configurations



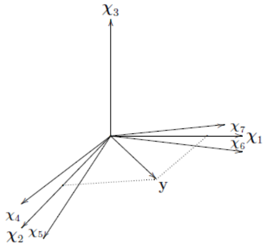
Inadequate and correlated



Adequate and random



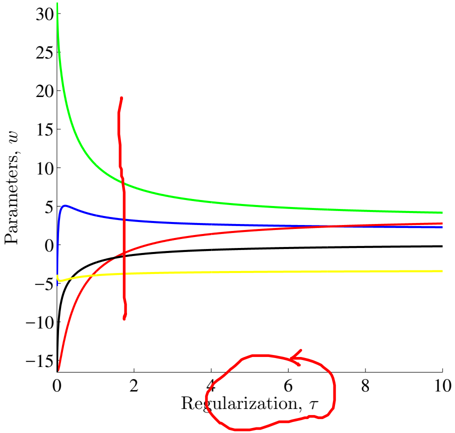
Adequate and redundant



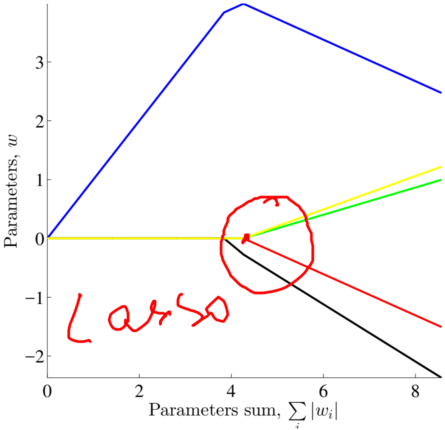
Adequate and correlated

Model parameter values with regularization

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



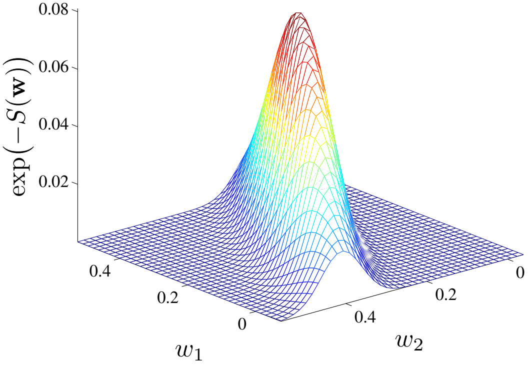
$$S(\mathbf{w}) = \|\mathbf{f}(\mathbf{w}, \mathbf{X}) - \mathbf{y}\|^2 + \gamma^2\|\mathbf{w}\|^2$$

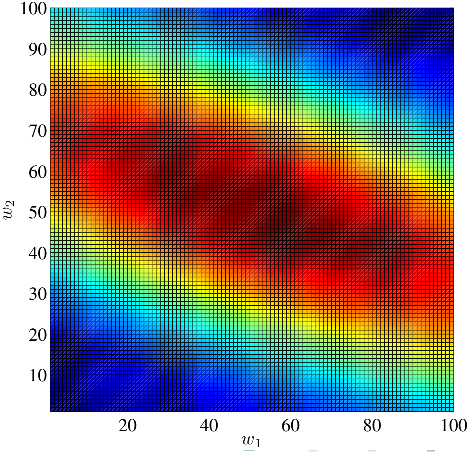


$$S(\mathbf{w}) = \|\mathbf{f}(\mathbf{w}, \mathbf{X}) - \mathbf{y}\|^2, \; T(\mathbf{w}) \leqslant \tau$$

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 $S(\mathbf{w}|\mathfrak{D}, \mathbf{f})$. Consider the set of points $\{s_k = \exp\bigl(-S(\mathbf{w}_k|\mathfrak{D}, \mathbf{f})\bigr)|k = 1, \dots, K\}$.





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

Minimize number of similar and maximize number of relevant features

Dec 5th, 7⁰⁰ pm MSK + 3 weeks

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

Ladder networks

$$\mathbf{a}^* = \arg\min_{\mathbf{a}\in\mathbb{B}^n} \underbrace{\mathbf{a}^\mathsf{T}\mathbf{Q}\mathbf{a}}_{\text{min}} - \underbrace{\mathbf{b}^\mathsf{T}\mathbf{a}}, \longrightarrow a = [0, 1, \dots, 1]$$

$a \in \mathbb{B}^n$

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

PSD

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

convex problem with relaxation

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

vs to send papers

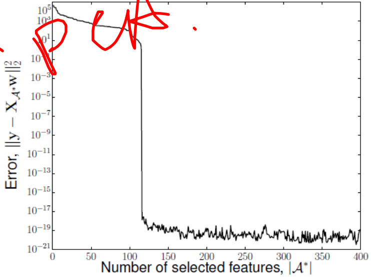
$$\mathbf{b} = [b_i] = \text{Rel}(\chi_i),$$

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 Sim → min, number of correlated to the target Rel → max.

Evaluation criteria for the diesel NIR spectra data set

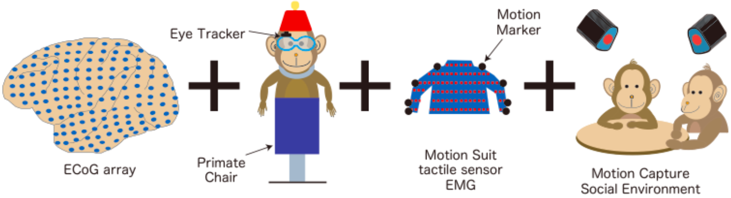
Method	C_p	RSS	$\ln \frac{\lambda_1}{\lambda_n}$ SVD	VIF	BIC
QP (ρ, ρ) $(\tau = 10^{-9})$	-110	$1.37 \cdot 10^{-18}$	-25.7	$6.43 \cdot 10^6$	548.38
Genetic	-110.88	$7.68 \cdot 10^{-30}$	-24	$8.13 \cdot 10^5$	534.19
LARS	$3.22 \cdot 10^{21}$	$2.07 \cdot 10^{-7}$	-28.3	$7.94 \cdot 10^7$	529.47
Lasso	$2.5 \cdot 10^{28}$	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	$1.07 \cdot 10^{22}$	$1.79 \cdot 10^3$

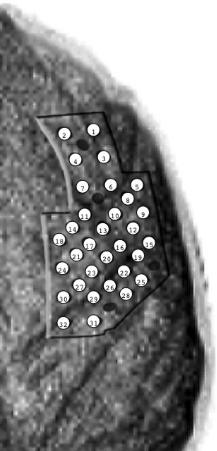


p uA.

Dependence of residual norm on the number of selected features QP(Sim, Rel).

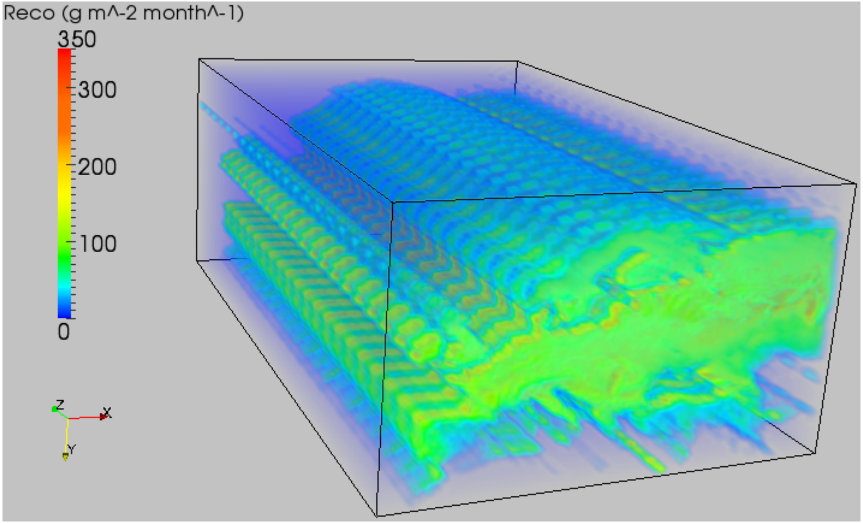
ECoG brain signals to forecast upper limbs' movements


ECoG array
Eye Tracker
Primate Chair
Motion Marker
Motion Suit tactile sensor EMG
Motion Capture Social Environment



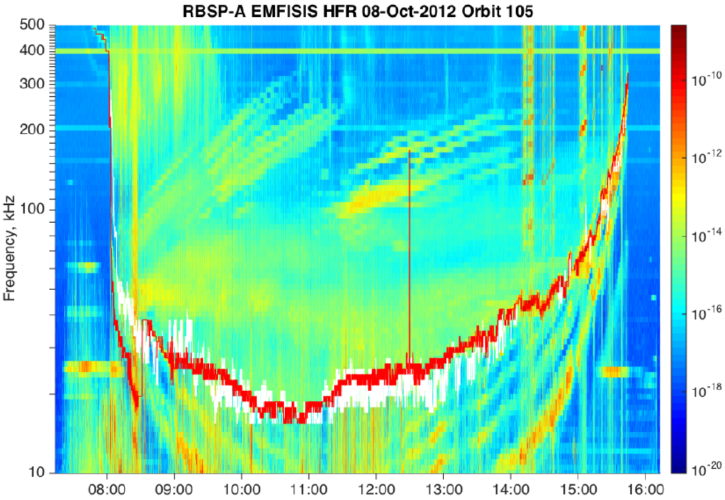
- Neurotycho.org food-tracking dataset: 32 epidural electrodes captures brain signals of the monkey (ECoG),
- 11 sensors track movements of the hand, contralateral to the implant side,
- experiment duration is 15 minutes,
- the experimenter feeds the monkey ≈ 4.5 per minute,
- ECoG and motion data were sampled at 1KHz and 120Hz, respectively.

Weather forecasting and geo (spatio) temporal dataset



Monthly Ecosystem Respiration by M. Reichstein, GHG-Europe

Spatio temporal dataset, frequency versus time, given spatial position



Electric field measurement, the Van Allen probes by I. Zhelavskaya, Skoltech

Подробнее о вышеизложенных задачах

Авторегрессионное прогнозирование и выбор признаков

- Катруца А., Стрижов В. Проблема мультиколлинеарности при выборе признаков в регрессионных задачах // Информационные технологии, 2015, 1 : 8-18.
- Нейчев Р.Г., Катруца А.М., Стрижов В. Выбор оптимального набора признаков из мультикоррелирующего множества в задаче прогнозирования // Заводская лаборатория. Диагностика материалов, 2016, 3.
- Мотренко А.П., Рудаков К.В., Стрижов В.В. Учет влияния экзогенных факторов при непараметрическом прогнозировании временных рядов // Вестник Московского университета. Серия 15. Вычислительная математика и кибернетика, 2016. А

Классификация временных рядов акселерометра

- Ignatov A., Strijov V. Human activity recognition using quasiperiodic time series collected from a single triaxial accelerometer // Multimedia Tools and Applications, 2015, 17.05.2015 : 1-14.
- Motrenko A.P., Strijov V.V. Extracting fundamental periods to segment human motion time series // Journal of Biomedical and Health Informatics, 2016.
- Попова М.С., Стрижов В.В. Выбор оптимальной модели классификации физической активности по измерениям акселерометра // Информатика и ее применения, 2015, 9(1) : 79-89.
- Попова М. С., Стрижов В.В. Построение сетей глубокого обучения для классификации временных рядов // Системы и средства информатики, 2015, 25(3) : 60-77.
- Гончаров А.В., Стрижов В.В. Метрическая классификация временных рядов со взвешенным выравниванием относительно центроидов классов // Информатика и ее применения, 2016, 2.

Согласование прогнозов иерархических временных рядов

- Стенина М.М., Стрижов В.В. Согласование прогнозов при решении задач прогнозирования иерархических временных рядов // Информатика и ее применения, 2015, 9(2) : 77-89.
- Стенина М., Стрижов В. Согласование агрегированных и детализированных прогнозов при решении задач непараметрического прогнозирования // Системы и средства информатики, 2014, 24(2) : 21-34.

Model selection for time series forecasting

Thanks to the Chair of Intelligent Systems, MIPT

- Anastasia Motrenko
- Mikhail Kuznetsov
- Alexandr Aduenko
- Arsentiy Kuzmin
- Maria Stenina
- Alexandr Katrutsa
- Oleg Bakhteev
- Maria Popova
- Andrey Kulunchakov
- Mikhail Karasikov
- Radoslav Neychev
- Alexey Goncharov
- Roman Isachenko
- Maria Vladimirova

`http://machinelearning.ru`
`strijov@phystech.edu`