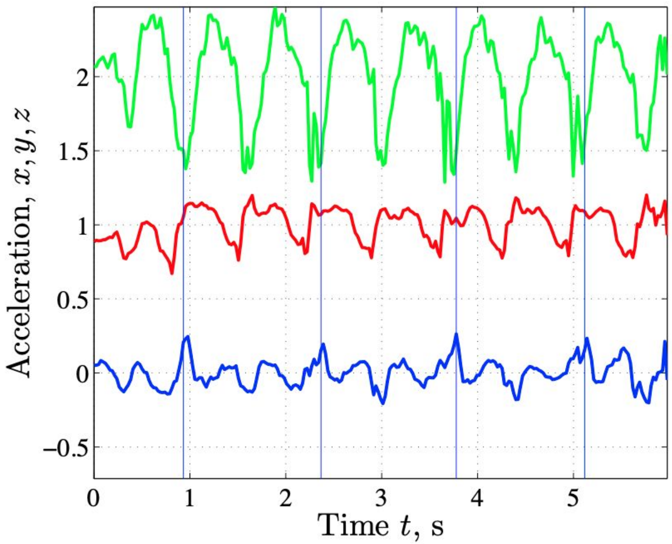
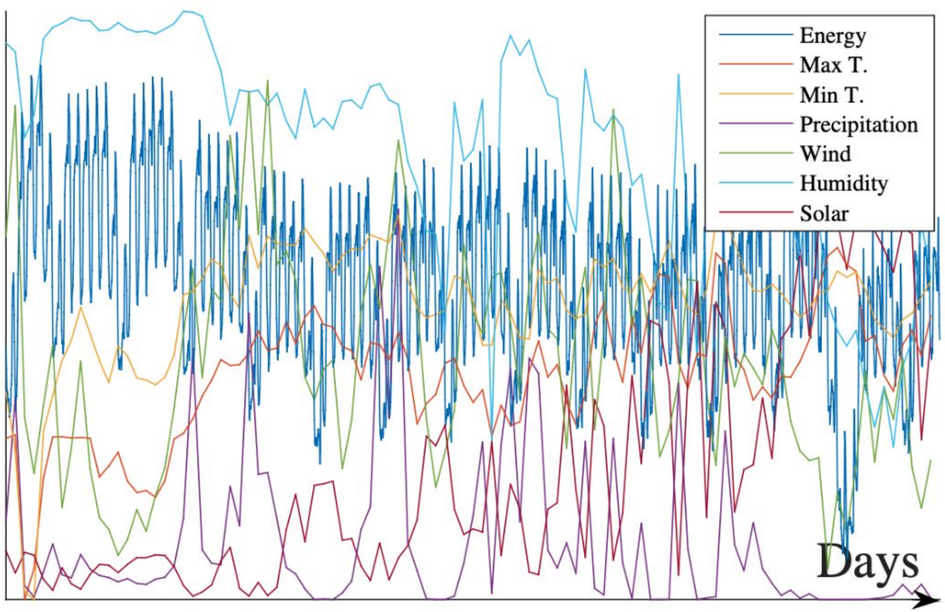
Mathematical methods of forecasting

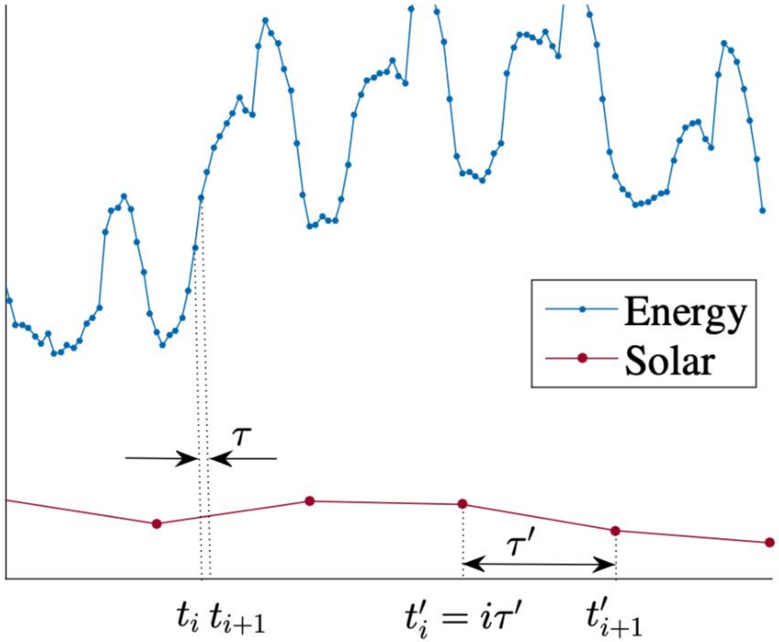
Intelligent systems, Phystech

2022

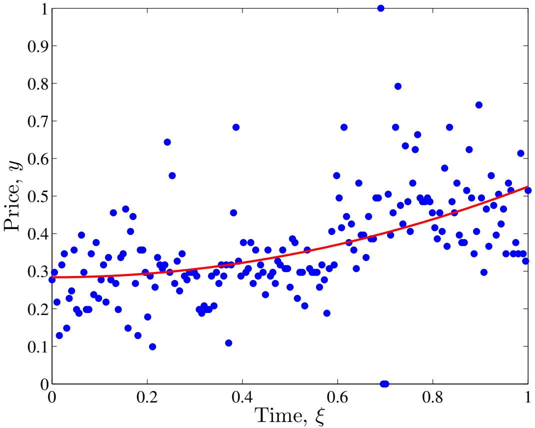


Acceleration, x, y, z
2
1.5
1
0.5
0
−0.5
0
1
2
3
4
5
Time t, s

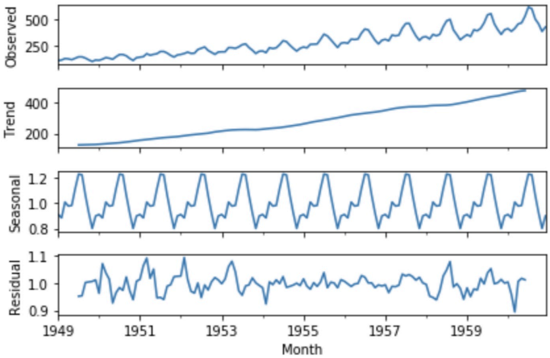

Energy
Max T.
Min T.
Precipitation
Wind
Humidity
Solar
Days


Energy
Solar
τ
τ'
t_i t_{i+1}
$t'_i = i\tau'$
t'_{i+1}

A simple model and its structure $\mathbf{a} \in \mathbb{B}^n$

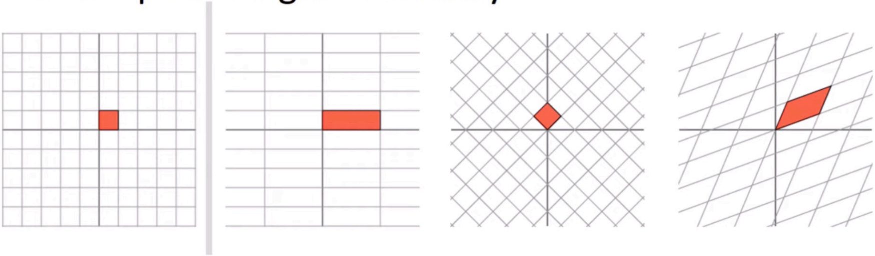


Regression model: $f = w_1 + w_2\xi^1 + w_3\xi^2 + \varepsilon(\xi)$, let $\mathbf{x} = [\xi^0, \xi^1, \xi^2]^\mathsf{T}$,
model to select from: $f = \mathbf{a} \odot \mathbf{w}^\mathsf{T}\mathbf{x}$,
optimal structure: $\hat{\mathbf{a}} = [1, \quad 0, \quad 1]^\mathsf{T}$,
optimal parameters: $\hat{\mathbf{w}} = [0.2839, \text{n/a}, 0.2412]^\mathsf{T}$.


Observed
500
250
Trend
400
200
Seasonal
1.2
1.0
0.8
Residual
1.1
1.0
0.9
1949
1951
1953
1955
1957
1959
Month

Linear Maps (geometrically) are spatial transforms that...

1. Keep gridlines parallel
2. Keep gridlines evenly spaced
3. Keep the origin stationary



(Images taken from the American Math Society.)

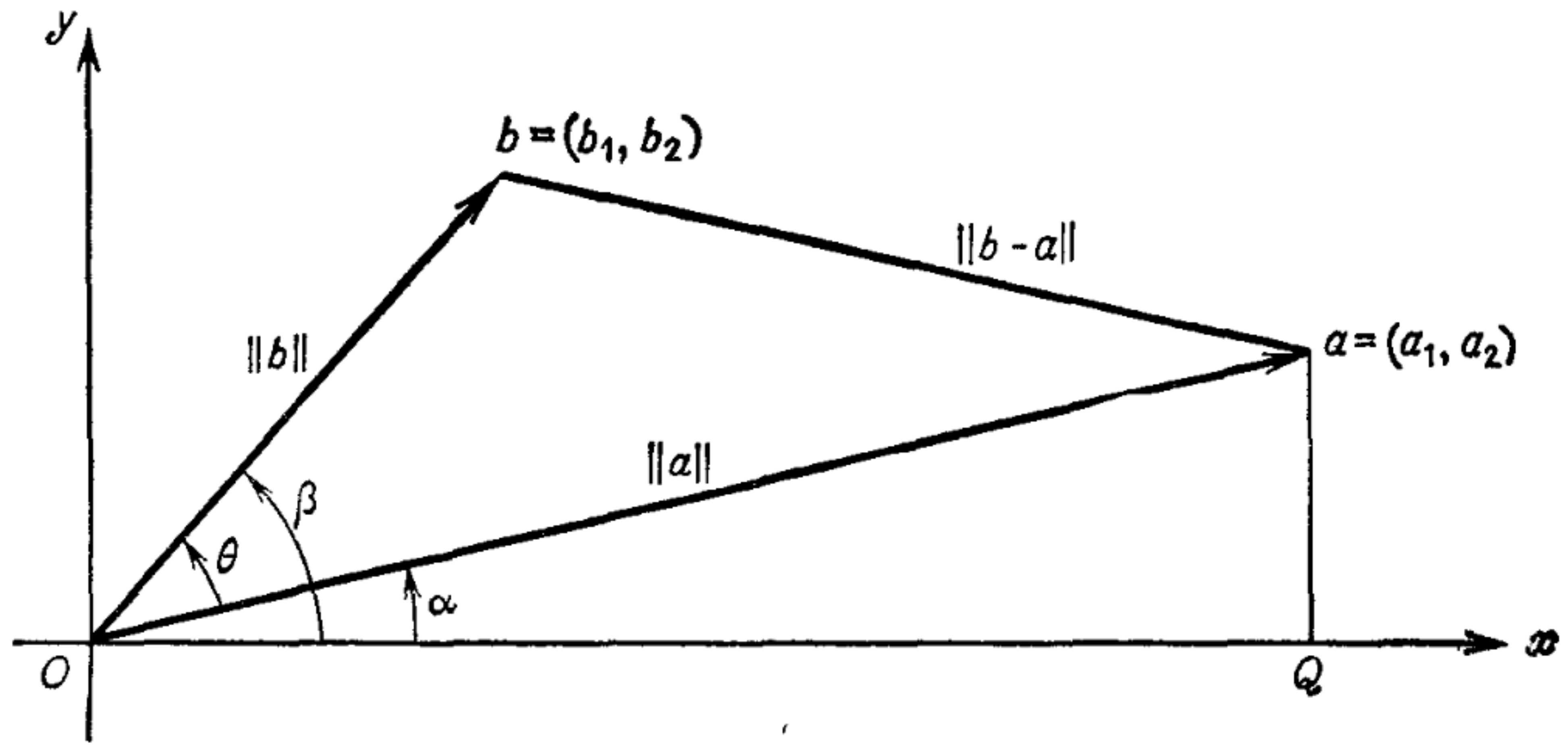


Рис. 3.2. Косинус угла $\theta=\beta-\alpha$.

синуса угла α в виде

$$\sin\alpha=\frac{a_2}{\|a\|},\qquad \cos\alpha=\frac{a_1}{\|a\|}.$$

То же самое справедливо и для вектора b с соответствующим углом β: его синус равен $b_2/\|b\|$, а косинус равен $b_1/\|b\|$. Теперь, поскольку угол θ в точности равен $\beta-\alpha$, его косинус дается хорошо известным тригонометрическим тождеством

$$\cos\theta=\cos\beta\cos\alpha+\sin\beta\sin\alpha=\frac{a_1b_1+a_2b_2}{\|a\|\|b\|}. \tag{1}$$

Числитель этой формулы совпадает со скалярным произведением векторов b и a, откуда мы и получаем требуемое соотношение:

3A. *Косинус угла между двумя векторами равен*

$$\cos\theta=\frac{a^{\mathrm{T}}b}{\|a\|\|b\|}. \tag{2}$$

Отметим, что эта формула правильна и в метрическом отношении: если мы вдвое увеличим вектор b, то как числитель, так и знаменатель увеличатся вдвое и косинус останется неизменным. Изменение знака вектора b на противоположный одновременно изменит знак у $\cos\theta$, что означает изменение угла на 180°.

Замечание. Тот же самый результат может быть получен и из теоремы косинусов, которая связывает длины сторон в произвольном треугольнике и имеет вид

$$\|b-a\|^2=\|b\|^2+\|a\|^2-2\|b\|\|a\|\cos\theta. \tag{3}$$

Если θ является прямым углом, то мы приходим к теореме Пифагора, а при произвольном θ выражение $\|b-a\|^2$ может быть за-

писано в виде $(b-a)^{\mathrm{T}}(b-a)$ и вместо (3) получаем:

$$b^{\mathrm{T}}b-2a^{\mathrm{T}}b+a^{\mathrm{T}}a=b^{\mathrm{T}}b+a^{\mathrm{T}}a-2\|b\|\|a\|\cos\theta.$$

После сокращений мы приходим к формуле (2) для косинуса. Фактически это доказывает и нашу формулу для n-мерного случая, поскольку все происходит лишь в плоском треугольнике Oab.

Теперь мы хотим найти проекцию p. Эта точка должна быть кратна вектору a, т. е. $p=\bar{x}a$, поскольку любая точка этой прямой кратна a, и задача состоит в отыскании коэффициента $\bar{x}$. Для этого нам нужен лишь простой геометрический факт, состоящий в том, что прямая, соединяющая конец вектора b с точкой p, перпендикулярна к вектору a:

$$(b-\bar{x}a)\perp a,\quad\text{или}\quad a^{\mathrm{T}}(b-\bar{x}a)=0,\quad\text{или}\quad \bar{x}=\frac{a^{\mathrm{T}}b}{a^{\mathrm{T}}a}.$$

3B. *Проекция p точки b на прямую, определенную вектором a, задается формулой*

$$p=\frac{a^{\mathrm{T}}b}{a^{\mathrm{T}}a}a. \tag{4}$$

Расстояние (в квадрате) от этой точки до прямой равняется

$$\left\|b-\frac{a^{\mathrm{T}}b}{a^{\mathrm{T}}a}a\right\|^2=b^{\mathrm{T}}b-2\frac{(a^{\mathrm{T}}b)^2}{a^{\mathrm{T}}a}+\left(\frac{a^{\mathrm{T}}b}{a^{\mathrm{T}}a}\right)^2a^{\mathrm{T}}a=$$

$$=\frac{(b^{\mathrm{T}}b)(a^{\mathrm{T}}a)-(a^{\mathrm{T}}b)^2}{(a^{\mathrm{T}}a)}. \tag{5}$$

Это позволяет нам повторить рис. 3.1 уже с указанием формулы для определения точки p (рис. 3.3).

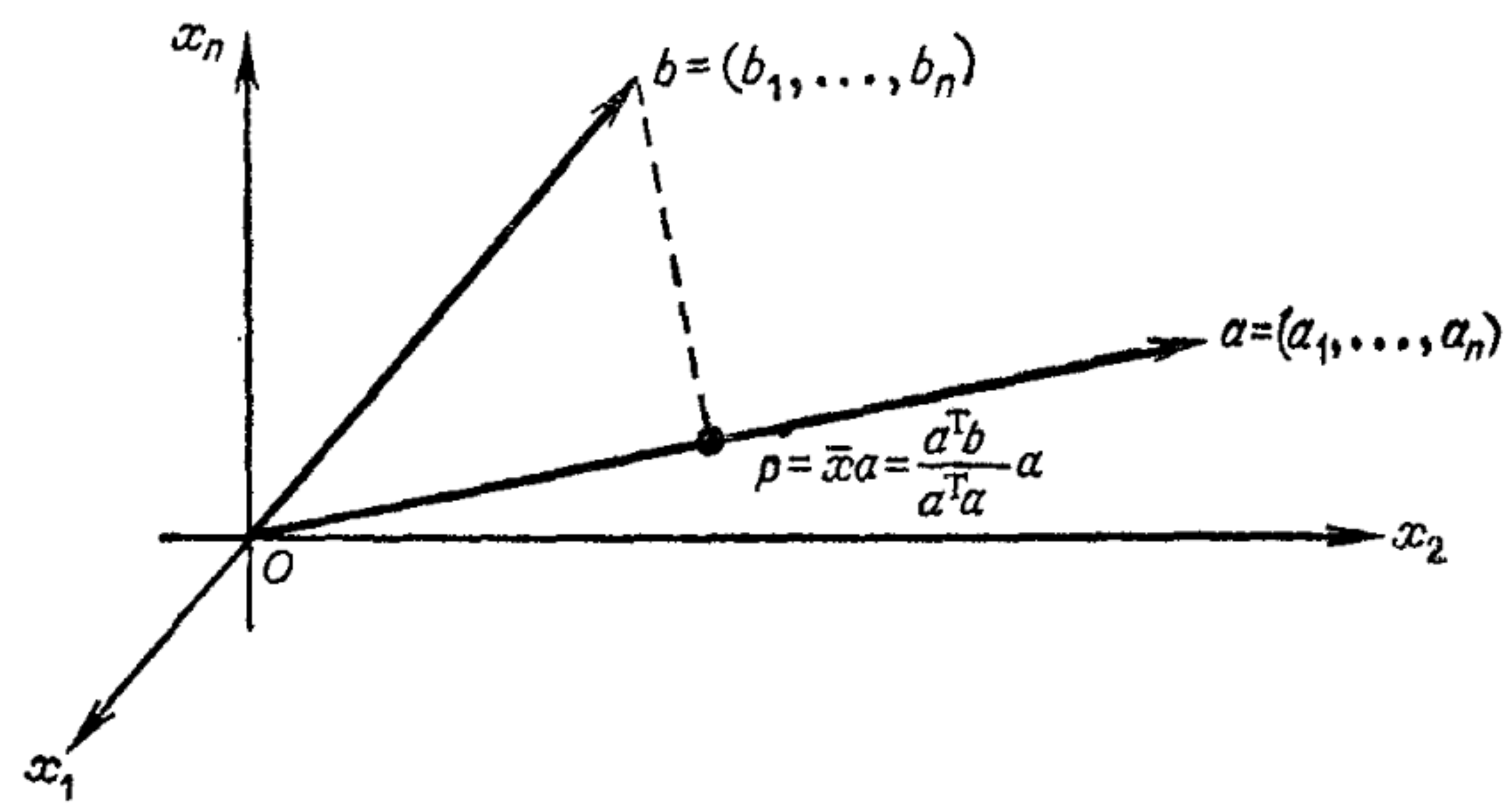


Рис. 3.3. Проекция вектора **b** на вектор **a**.

Требуется минимизировать евклидово расстояние от вектора $\mathbf{y}$ до вектора $\mathbf{Xw}$. Этот вектор лежит в пространстве столбцов матрицы $\mathbf{X}$, так как $\mathbf{Xw}$ — это линейная комбинация столбцов этой матрицы с коэффициентами $w_1, \dots, w_n$. Задача оценки $\mathbf{w}$ эквивалентна задаче нахождения точки $\mathbf{p} = \mathbf{Xw}$, ближайшей к $\mathbf{y}$ и находящейся в пространстве столбцов матрицы $\mathbf{X}$. Следовательно, вектор $\mathbf{p}$ должен быть проекцией $\mathbf{y}$ на пространство столбцов, вектор регрессионных остатков $\mathbf{Xw} - \mathbf{y}$ должен быть ортогонален этому пространству. Рассмотрим произвольный вектор $\mathbf{Xv}$, ортогональный вектору регрессионных остатков $\mathbf{Xw} - \mathbf{y}$:

$$(\mathbf{Xv})^\mathsf{T}(\mathbf{Xw} - \mathbf{y}) = \mathbf{v}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{Xw} - \mathbf{X}^\mathsf{T}\mathbf{y}) = 0.$$

Так как это равенство должно быть справедливо для произвольного вектора $\mathbf{v}$, то $\mathbf{X}^\mathsf{T}\mathbf{Xw} - \mathbf{X}^\mathsf{T}\mathbf{y} = 0$, см. рис. **??**. Если столбцы матрицы $\mathbf{X}$ линейно независимы, то матрица $\mathbf{X}^\mathsf{T}\mathbf{X}$ обратима и уравнение имеет единственное решение относительно параметров

$$\mathbf{w} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}. \tag{38}$$

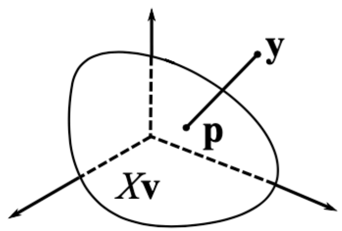


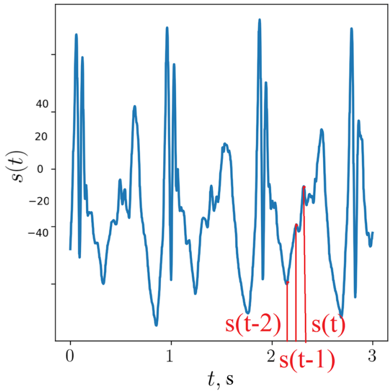
Рис. 6. Проекция вектора зависимой переменной на пространство столбцов матрицы плана.

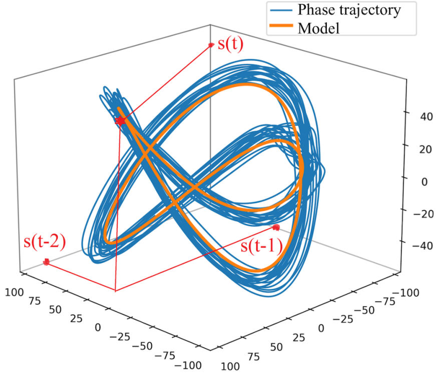
Проекция вектора $\mathbf{y}$ на пространство столбцов матрицы $\mathbf{X}$ имеет вид

$$\mathbf{p} = \mathbf{Xw} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = \mathbf{Py}.$$

Матрица $\mathbf{P} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$ называется матрицей проектирования. Она она идемпотентна, $\mathbf{P}^2 = \mathbf{P}$, и симметрична, $\mathbf{P}^\mathsf{T} = \mathbf{P}$.

Phase trajectory of the accelerometer time series





dim(s) $\approx$ 1000

dim(x) $=$ 4

SSA can be used as a model-free technique so that it can be applied to arbitrary time series including non-stationary time series. The basic aim of SSA is to decompose the time series into the sum of interpretable components such as trend, periodic components and noise with no a-priori assumptions about the parametric form of these components.

Consider a real-valued time series $\mathbb{X} = (x_1, \ldots, x_N)$ of length N. Let $L \ (1 < L < N)$ be some integer called the *window length* and $K = N - L + 1$.

Main algorithm

1st step: Embedding.

Form the *trajectory matrix* of the series $\mathbb{X}$, which is the $L \times K$ matrix

$$\mathbf{X} = [X_1 : \ldots : X_K] = (x_{ij})_{i,j=1}^{L,K} = \begin{bmatrix} x_1 & x_2 & x_3 & \ldots & x_K \\ x_2 & x_3 & x_4 & \ldots & x_{K+1} \\ x_3 & x_4 & x_5 & \ldots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \ldots & x_N \end{bmatrix}$$

where $X_i = (x_i, \ldots, x_{i+L-1})^{\mathrm{T}} \quad (1 \leq i \leq K)$ are *lagged vectors* of size L. The matrix $\mathbf{X}$ is a Hankel matrix which means that $\mathbf{X}$ has equal elements x_{ij} on the anti-diagonals $i + j = \mathrm{const}$.

2nd step: Singular Value Decomposition (SVD).

Perform the singular value decomposition (SVD) of the trajectory matrix $\mathbf{X}$. Set $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ and denote by $\lambda_1, \ldots, \lambda_L$ the *eigenvalues* of $\mathbf{S}$ taken in the decreasing order of magnitude ($\lambda_1 \geq \ldots \geq \lambda_L \geq 0$) and by $U_1, \ldots, U_L$ the orthonormal system of the *eigenvectors* of the matrix $\mathbf{S}$ corresponding to these eigenvalues.

Set $d = \operatorname{rank} \mathbf{X} = \max\{i, \text{ such that } \lambda_i > 0\}$ (note that $d = L$ for a typical real-life series) and $V_i = \mathbf{X}^{\mathrm{T}} U_i / \sqrt{\lambda_i} \ (i = 1, \ldots, d)$. In this notation, the SVD of the trajectory matrix $\mathbf{X}$ can be written as

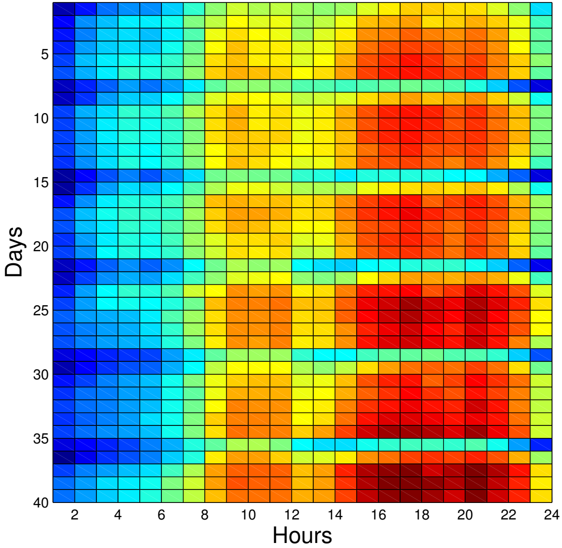
$$\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d,$$

where

$$\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$$

are matrices having rank 1; these are called *elementary matrices*. The collection $(\sqrt{\lambda_i}, U_i, V_i)$ will be called the ith *eigentriple* (abbreviated as ET) of the SVD. Vectors U_i are the left singular vectors of the matrix $\mathbf{X}$, numbers $\sqrt{\lambda_i}$ are the singular values and provide the singular spectrum of $\mathbf{X}$; this gives the name to SSA. Vectors $\sqrt{\lambda_i} V_i = \mathbf{X}^{\mathrm{T}} U_i$ are called vectors of principal components (PCs).

Матрица авторегрессии



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

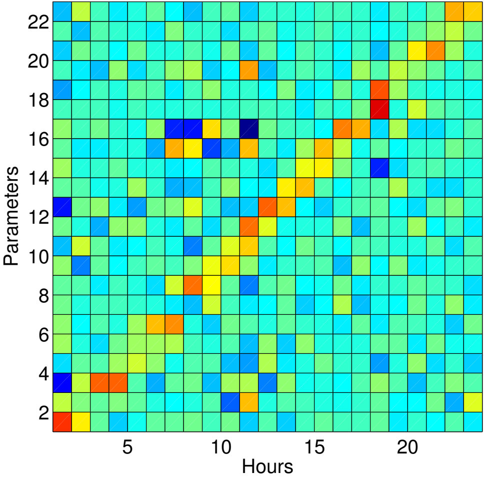
$$\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \mathbf{w}) = \mathbf{X}\mathbf{w},$$

$$y_{m+1} = s_T = \mathbf{w}^\mathsf{T}\mathbf{x}_{m+1}^\mathsf{T}.$$

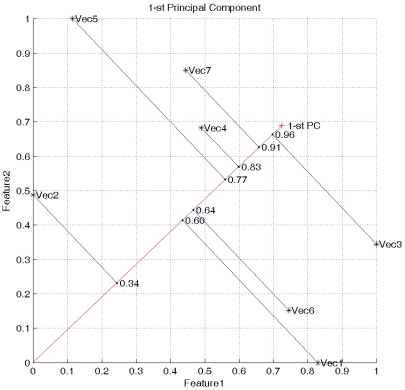
Нейчев Р.Г., Катруца А.М., Стрижов В. Выбор оптимального набора признаков из мультикоррелирующего множества в задаче прогнозирования // Заводская лаборатория. Диагностика материалов, 2016.

How many parameters must be used to forecast?

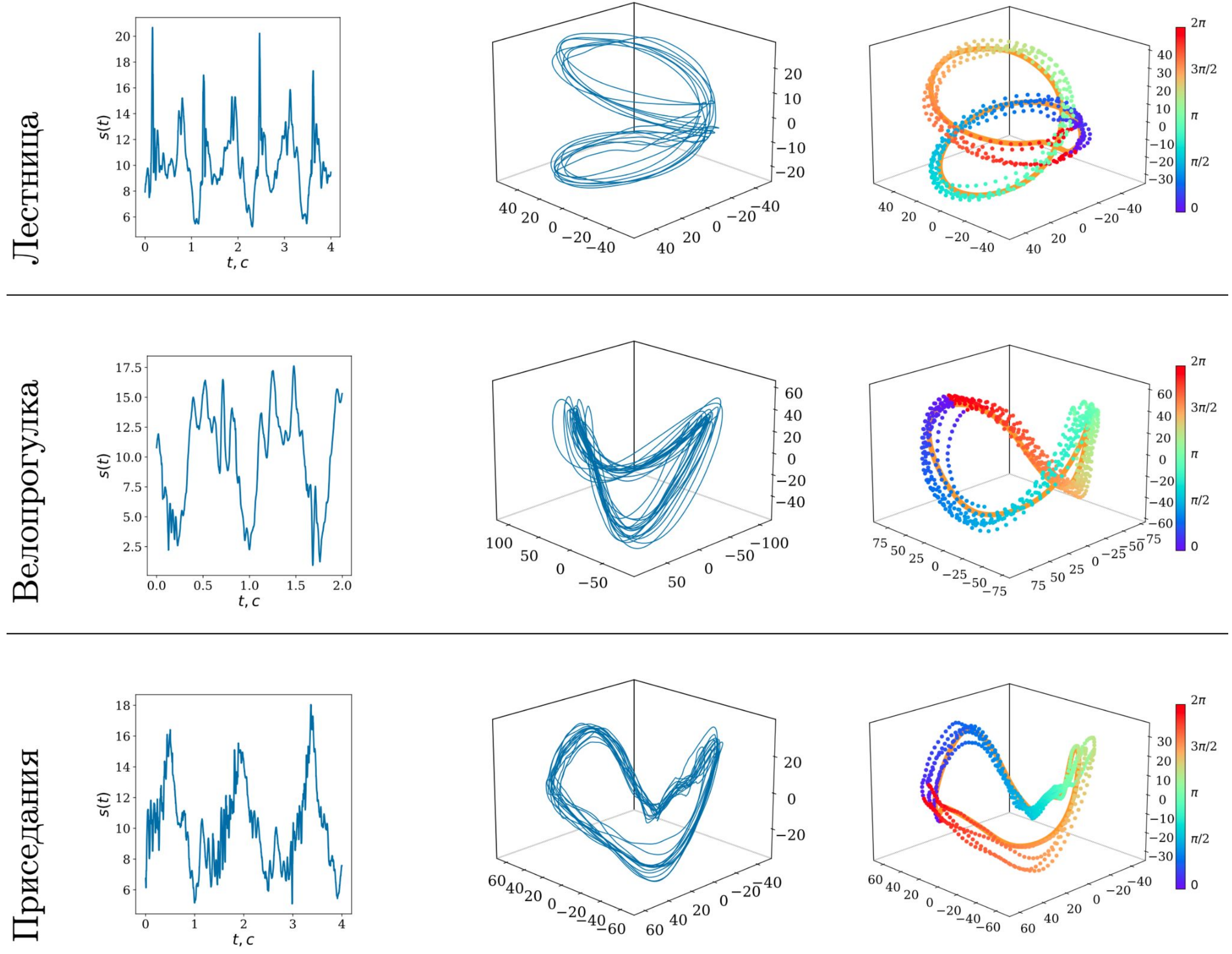
The color shows the value of a parameter for each hour.



Estimate parameters $\mathbf{w}(\tau) = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$, then calculate the sample $s(\tau) = \mathbf{w}^\mathsf{T}(\tau)\mathbf{x}_{m+1}$ for each τ of the next ($m+1$-th) period.



1-st Principal Component
Vec5
Vec7
Vec4
1-st PC
0.96
0.91
0.83
0.77
Vec2
0.64
0.60
Vec3
0.34
Vec6
Vec1
Feature1
Feature2
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1



Лестница
Велопрогулка
Приседания
s(t)
t, c

Рассмотрим квадратичный алгоритм решения этой задачи. Найдем последовательно векторы $\mathbf{u}_k, \mathbf{v}_k$ и сингулярные числа λ_k для $k = 1, \ldots, r$. В качестве этих векторов берутся нормированные значения векторов $\mathbf{a}_k$ и $\mathbf{b}_k$, соответственно

$$\underset{m\times 1}{\mathbf{u}_k} = \frac{\mathbf{a}_k}{\|\mathbf{a}_k\|} \quad \text{и} \quad \underset{1\times n}{\mathbf{v}_k} = \frac{\mathbf{b}_k}{\|\mathbf{b}k\|}.$$

Векторы $\mathbf{a}_k$ и $\mathbf{b}_k$ находятся как пределы последовательностей векторов $\{\mathbf{a}_{k_s}\}$ и $\{\mathbf{b}_{k_s}\}$, соответственно

$$\mathbf{a}_k = \lim_{s\to\infty}(\mathbf{a}_{k_s}) \ \text{ и } \mathbf{b}_k = \lim_{s\to\infty}(\mathbf{b}_{k_s}).$$

Сингулярное число λ_k находится как произведение норм векторов

$$\lambda_k = \|\mathbf{a}_k\| \cdot \|\mathbf{b}_k\|.$$

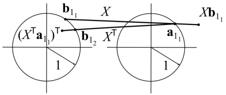


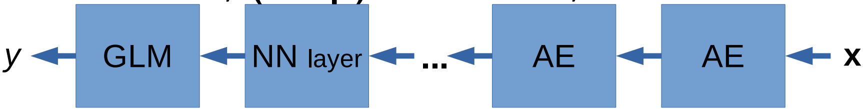
Рис. 13. Итеративная процедура оценивания сингулярных векторов.

Процедура нахождения последовательностей векторов $\mathbf{a}_{k_s}, \mathbf{b}_{k_s}$ $\mathbf{u}_k, \mathbf{v}_k$ начинается с выбора наибольшей по норме строки $\mathbf{b}_{1_1}$ матрицы $\mathbf{X}$. Для $k = 1$ формулы нахождения векторов $\mathbf{a}_{1_s}, \mathbf{b}_{1_s}$ имеют вид:

$$\underset{m\times 1}{\mathbf{a}_{1_s}} = \frac{\mathbf{X}\mathbf{b}_{1_s}^\mathsf{T}}{\mathbf{b}_{1_s}\mathbf{b}_{1_s}^\mathsf{T}}, \quad \underset{1\times n}{\mathbf{b}_{1_{s+1}}} = \frac{\mathbf{a}_{1_s}^\mathsf{T}\mathbf{X}}{\mathbf{a}_{1_s}^\mathsf{T}\mathbf{a}_{1_s}}, \quad s = 1, 2, \cdots.$$

Для вычисления векторов $\mathbf{u}_k, \mathbf{v}_k$ при $k = 2, \ldots, r$ используется вышеприведенная формула, с той разницей, что матрица $\mathbf{X}$ заменяется на скорректированную на k-м шаге матрицу $\mathbf{X}_{k+1} = \mathbf{X}_k - \mathbf{u}_k\lambda_k\mathbf{v}_k$. На рисунке **??** показаны две итерации, $s = 1, 2$, первого шага $k = 1$ упрощенной процедуры нахождения сингулярного разложения.

Linear model, (deep) neural net, and autoencoder



$$\underbrace{f = \sigma_k \circ \underset{1\times 1}{\mathbf{w}_k^{\mathsf{T}}} \boldsymbol{\sigma}_{k-1} \circ \mathbf{W}_{k-1} \boldsymbol{\sigma}_{k-2} \circ \cdots \circ \underbrace{\underset{n_2\times 1}{\mathbf{W}_2} \boldsymbol{\sigma}_1 \circ \underset{n_1\times n}{\mathbf{W}_1} \underset{n\times 1}{\mathbf{x}}}_{E_{\mathbf{x}} = \sum_{\mathbf{x}_i \in \mathfrak{D}} \|\mathbf{x}_i - \mathbf{r}(\mathbf{x}_i)\|_2^2} \in \mathfrak{D}}_{E_D = \sum_{(\mathbf{x}_i, y_i) \in \mathfrak{D}} \left(y_i - f(\mathbf{x}_i)\right)^2}$$

$$S = \lambda_1 E_D + \lambda_2 E_{\mathbf{x}} + \lambda_3 E_{\mathbf{w}} = \boldsymbol{\lambda}^{\mathsf{T}} \mathbf{s}$$

$E_{\mathbf{w}}$ is some regularisation error, for

principal component analysis: $\mathbf{W}^{\mathsf{T}}\mathbf{W} = \mathbf{I}_n$,

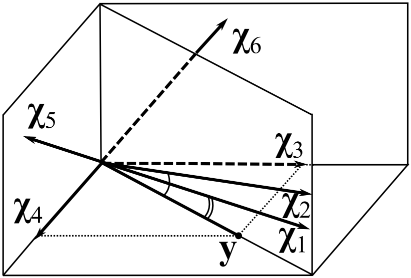
skip block: $\mathbf{W} = \mathbf{I}_n$, $\sigma = \mathrm{id}$,

classification: $\sigma \in \{\text{logistic, softmax, ReLu}, \ldots\}$.

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

Selection of a stable set of features of restricted size

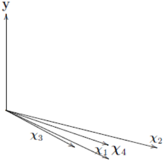
The sample contains multicollinear χ_1, χ_2 and noisy χ_5, χ_6 features, columns of the design matrix $\mathbf{X}$. We want to select two features from six.



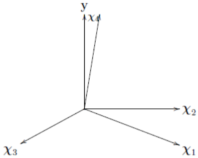
Stability and accuracy for a fixed complexity

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

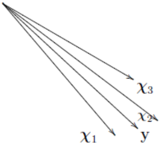
Multicollinear features to forecast: possible configurations



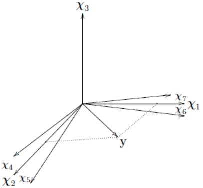
Inadequate and correlated



Adequate and random



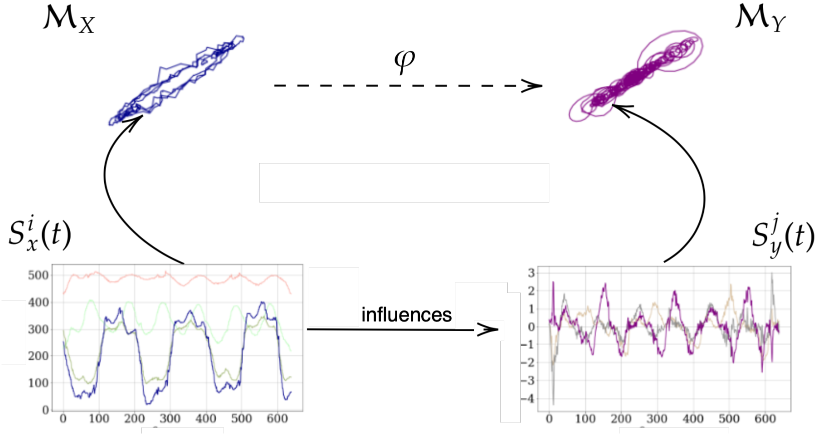
Adequate and redundant



Adequate and correlated

Katrutsa A.M., Strijov V.V. Stresstest procedure for feature selection algorithms // Chemometrics and Intelligent Laboratory Systems, 2015, 142 : 172-183.

Time series and phase space[3]

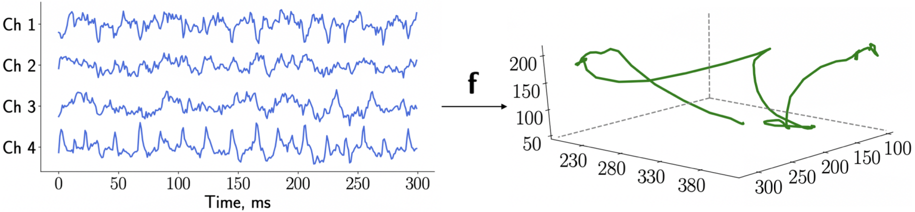


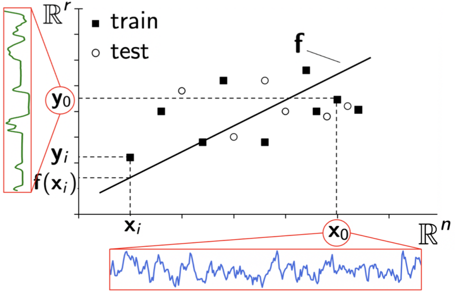
Transform from time domain to frequency domain is essential trick.

[3] Thanks to Ed. Vladimirov

Isachenko R.V., Strijov V.V. Quadratic Programming Optimization with Feature Selection for Non-linear Models // Lobachevskii Journal of Mathematics, 2018, 39(9) : 1179-1187.

Ensemble of models for brain computer interface





$$
\begin{tikzcd}
\mathbf{x} \in \mathbb{R}^n \arrow[rr, "\mathbf{f}"] \arrow[dr, bend right, "\mathbf{W}"'] & & \mathbf{y} \in \mathbb{R}^r \arrow[dl, bend left, "\mathbf{C}"] \\
& \mathbf{t}, \mathbf{u} \in \mathbb{R}^\ell \arrow[ul, bend right, "\mathbf{P}"'] \arrow[ur, bend left, "\mathbf{Q}"] &
\end{tikzcd}
$$

$$\mathbf{x} = \mathbf{P}\mathbf{t} + \mathbf{e}_{\mathbf{x}}$$

$$\mathbf{y} = \mathbf{Q}\mathbf{u} + \mathbf{e}_{\mathbf{y}}$$

$$\text{cov}(\mathbf{t}, \mathbf{u}) \to \max_{\mathbf{P},\mathbf{Q}}$$

Isachenko R.V., Strijov V.V. Quadratic programming feature selection for multicorrelated signal decoding with partial least squares // Expert Systems with Applications. Volume 207, 30 November 2022.

An element of the form $v \otimes w$ is called the **tensor product** of v and w. An element of $V \otimes W$ is a tensor, and the tensor product of two vectors is sometimes called an *elementary tensor* or a *decomposable tensor*. The elementary tensors span $V \otimes W$ in the sense that every element of $V \otimes W$ is a sum of elementary tensors. If bases are given for V and W, a basis of $V \otimes W$ is formed by all tensor products of a basis element of V and a basis element of W.

The tensor product of two vector spaces captures the properties of all bilinear maps in the sense that a bilinear map from $V \times W$ into another vector space Z factors uniquely through a linear map $V \otimes W \to Z$ (see Universal property).

$$\begin{tikzcd} V \times W \arrow[r, "\varphi"] \arrow[dr, "h"'] & V \otimes W \arrow[d, dashed, "\tilde{h}"] \\ & Z \end{tikzcd}$$

Universal property of tensor product: if h is bilinear, there is a unique linear map $\tilde{h}$ that makes the diagram commutative (that is, $h = \tilde{h} \circ \varphi$).

Cartesian product as an object in category theory

Рассмотрим декартово произведение $X \times Y$ двух множеств, состоящее, как обычно, из всех упорядоченных пар $\langle x, y\rangle$ элементов $x \in X$ и $y \in Y$. Проекции произведения $\langle x, y\rangle \mapsto x, \langle x, y\rangle \mapsto y$ на его оси X и Y представляют собой функции $p : X \times Y \to X$, $q : X \times Y \to Y$. Любая функция $h : W \to X \times Y$ из третьего множества W однозначно определяется композициями $p \circ h$ и $q \circ h$. Обратно, если дано множество W и функции f и g, такие, как на последующей диаграмме, то существует единственная функция h, которая делает диаграмму коммутативной; а именно, $hw = \langle fw, gw\rangle$ для каждого $w \in W$:

$$
\begin{tikzcd}
 & W \arrow[dl, "f"'] \arrow[d, "h"] \arrow[dr, "g"] & \\
X & X \times Y \arrow[l, "p"] \arrow[r, "q"'] & Y.
\end{tikzcd}
$$

Таким образом, для данных X и Y функция $\langle p, q\rangle$ универсальна среди всех пар функций, отображающих некоторое множество в X и в Y, поскольку любая другая такая пара $\langle f, g\rangle$ однозначно пропускается (посредством h) через пару $\langle p, q\rangle$. Это свойство определяет декартово произведение единственным образом (с точностью до биекции);

$$
\begin{tikzcd}
 & 3 \arrow[dl, bend right] \arrow[d, dotted] \arrow[dr, bend left] & \\
 & \mathrm{gcd}(36,30) \arrow[dl] \arrow[dr] & \\
36 & & 30
\end{tikzcd}
$$

An arrow $n \to m$ means "n evenly divides m."

In category theory, $\mathrm{gcd}(n,m)$ is the product of n and m

Маклейн

Tae-Danae Bradley