Model generation and selection using coherent Bayesian inference

Vadim Strijov
Visiting Professor at Laboratoire d'Informatique de Grenoble,
Apprentissage : modeles et algorithmes

April 22^{nd}, 2015

Problem significance

To get an accurate and stable forecast we develop the methods of model selection from the set of admissible basic models.

Our approach

Optimization of parameters for an arbitrary model is a non-trivial optimization problem. Our approach is to simplify the problem by considering sets of the successively generated stable models of given complexity.

Regression analysis: problem statement

We solve a regression problem:

estimate the conditional expectation $\mathsf{E}(Y|\mathbf{x}) = f(\mathbf{w}_0, \mathbf{x})$.

The sample: $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I} = \{1, \dots, m\}$. The set $\mathfrak{G}$ is a set of parametric basic functions $g(\mathbf{b}, \mathbf{x}')$.

Regression model

$$f = f(\mathbf{w}, \mathbf{x}) = g_1(\mathbf{b}_1, \mathbf{x}'_1) \circ \dots \circ g_r(\mathbf{b}_r, \mathbf{x}'_r)(\mathbf{x}),$$

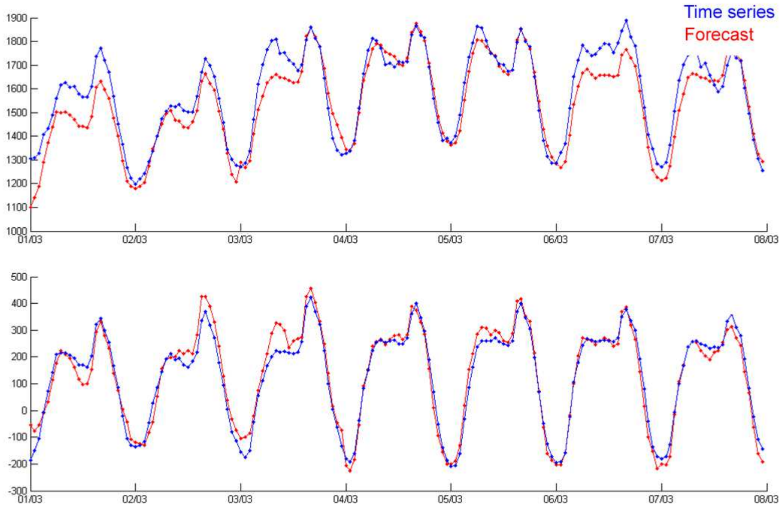
$$f : \mathbb{W} \times \mathbb{X} \to \mathbb{Y}, \quad \text{or elementwise:} \quad f : (\mathbf{w}, \mathbf{x}) \mapsto y,$$

is chosen from the successively generated set $\mathfrak{F}$.

We find the regression function, the restriction of the model over the set of parameters

$$\hat{f}|_{\mathbb{W} \ni \mathbf{w} = \mathbf{w}_0} : \mathbb{X} \to \mathbb{Y}.$$

Energy consumption one-week forecast, an example



The periodic components of the multivariate time series

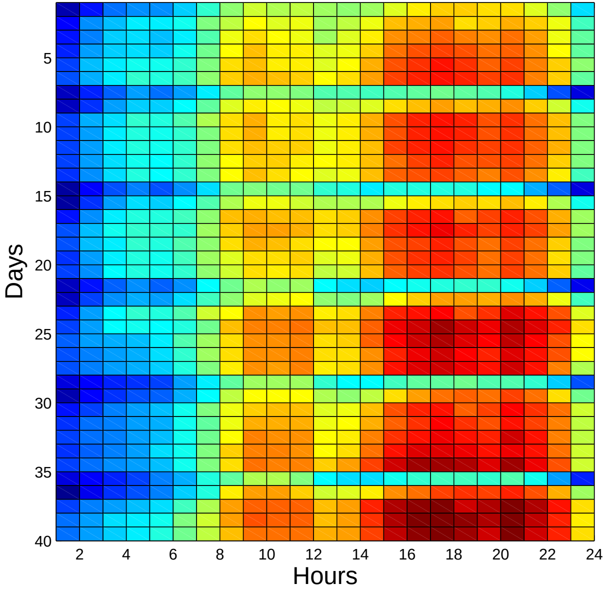
The time series:

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

Periods:

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

The autoregressive matrix, five week-ends



$$\underset{(m+1)\times(n+1)}{\mathbf{X}^*} = \left(\begin{array}{c|cccc} s_T & s_{T-1} & \dots & s_{T-\kappa+1} \\ \hline \\ s_{(m-1)\kappa} & s_{(m-1)\kappa-1} & \dots & s_{(m-2)\kappa+1} \\ \dots & \dots & \dots & \dots \\ s_{n\kappa} & s_{n\kappa-1} & \dots & s_{n(\kappa-1)+1} \\ \dots & \dots & \dots & \dots \\ s_\kappa & s_{\kappa-1} & \dots & s_1 \end{array}\right).$$

In a nutshell,

$$\mathbf{X}^* = \left[\begin{array}{c|c} \underset{1\times1}{s_T} & \underset{1\times n}{\mathbf{x}_{m+1}} \\ \hline \underset{m\times1}{\mathbf{y}} & \underset{m\times n}{\mathbf{X}} \end{array}\right].$$

In terms of linear regression:

$$\mathbf{y} = \mathbf{X}\mathbf{w},$$

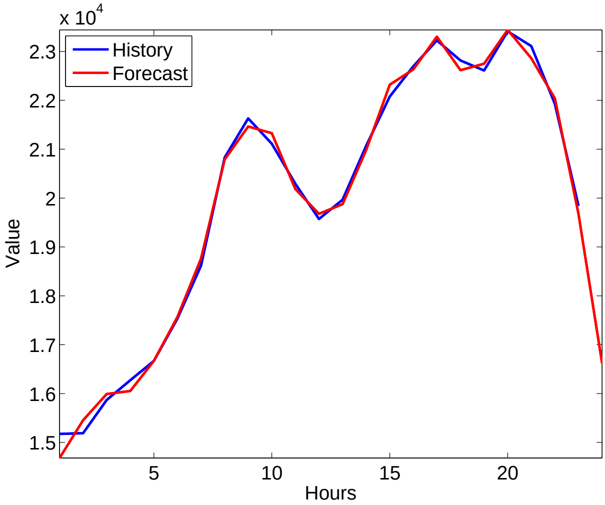
$$y_{m+1} = s_T = \mathbf{w}^{\mathsf{T}}\mathbf{x}_{m+1}^{\mathsf{T}}.$$

Introduce a set of the primitive functions $G = \{g_1, \dots, g_r\}$,
for example $g_1 = 1$, $g_2 = \sqrt{x}$, $g_3 = x$, $g_4 = x\sqrt{x}$, etc.

The generated set of features $\mathbf{X} =$

$$\left(\begin{array}{ccc|c|ccc}
g_1 \circ s_{T-1} & \dots & g_r \circ s_{T-1} & \dots & g_1 \circ s_{T-\kappa+1} & \dots & g_r \circ s_{T-\kappa+1} \\
\hline
g_1 \circ s_{(m-1)\kappa-1} & \dots & g_r \circ s_{(m-1)\kappa-1} & \dots & g_1 \circ s_{(m-2)\kappa+1} & \dots & g_r \circ s_{(m-2)\kappa+1} \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
g_1 \circ s_{n\kappa-1} & \dots & g_r \circ s_{n\kappa-1} & \dots & g_1 \circ s_{n(\kappa-1)+1} & \dots & g_r \circ s_{n(\kappa-1)+1} \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
g_1 \circ s_{\kappa-1} & \dots & g_r \circ s_{\kappa-1} & \dots & g_1 \circ s_1 & \dots & g_r \circ s_1
\end{array}\right).$$

The one-day forecast (an example)



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

Ill-conditioned matrix, or curse of dimensionality

Assume we have hourly data on price/consumption for three years. Then the matrix $\underset{(m+1)\times(n+1)}{\mathbf{X}^*}$ is

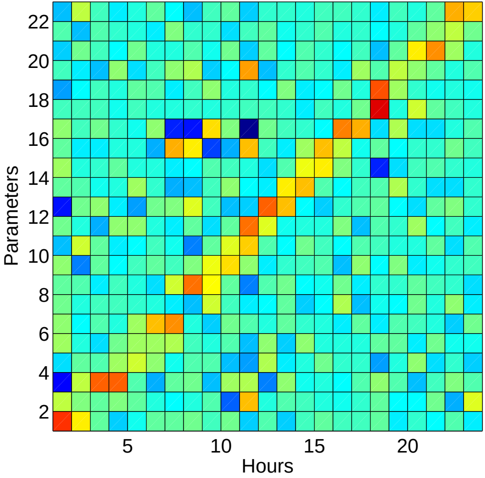
$$156 \times 168, \text{ in details: } 52\text{w} \cdot 3\text{y} \times 24\text{h} \cdot 7\text{d};$$

- for 6 time series the matrix $\mathbf{X}$ is 156×1008,
- for 4 primitive functions it is 156×4032,

$$m << n.$$

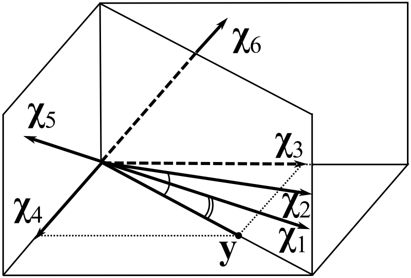
The autoregressive matrix could be considered as *ill-conditioned* and *multi-correlated*. The model selection procedure is required.

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.

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

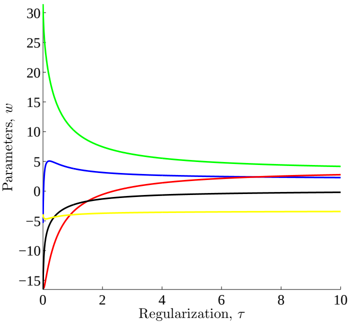


Stability and accuracy for a fixed complexity

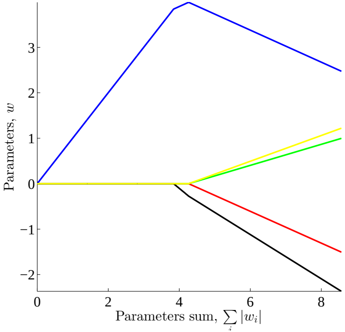
The solution: $\chi_3, \chi_4$is an orthogonal set of features minimizing the error function.

Algorithms: GMDH, Stepwise, Ridge, Lasso, Stagewise, FOS, LARS, Genetics, ...

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

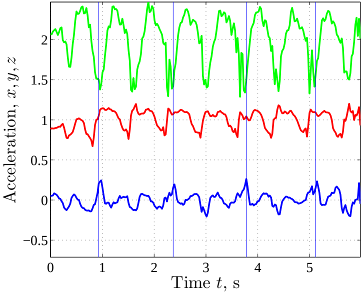


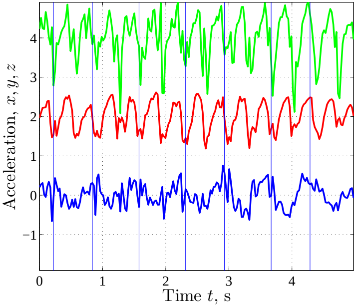
$$S(\mathbf{w}) = \|\mathbf{f}(\mathbf{w}, \mathbf{X}) - \mathbf{y}\|^2 + \gamma^2\|\mathbf{w}\|^2$$



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

Examples of accelerometric time series for slow walking and jogging:





3-dimensional time series of acceleration projections to spatial axis

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

Class labels y_i correspond to one of $S = 6$ types of activity: Jogging, Walking, Upstairs, Downstairs, Sitting, Standing.

Construct a classifier

$$\mathbf{f} = \mathbf{a}(\mathbf{h}_N(\ldots \mathbf{h}_1(\mathbf{x}))),$$

where $\mathbf{h}_k$ are autoencoding blocks of the form

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

and $\mathbf{a}$ is multinomial logistic regression classifier

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

Vectorize matrices $\mathbf{W}_1 \in \mathbb{R}^{n\times N_h}, \mathbf{W}_2 \in \mathbb{R}^{N_h\times S}$ of parameters of each layer to obtain vector of model parameters

$$\mathbf{w} = \text{vec}(\mathbf{W}_1^\mathsf{T}|\mathbf{W}_2^\mathsf{T}) \in \mathbb{R}^k.$$

Here number N_h of neurons in the hidden layer — the *structure parameter* of the model — is fixed.

Model structure

Parameter w_j of model $\mathbf{f}$ is called *active*, if $w_j \neq 0$.
The set of active indices $\mathcal{A} = \{j : w_j \neq 0\} \subseteq \mathcal{J}$ is called *structure* $\mathcal{A}$ of model $\mathbf{f}$.

Each structure $\mathcal{A} \subseteq \mathcal{J}$ defines a model $\mathbf{f}_\mathcal{A}$

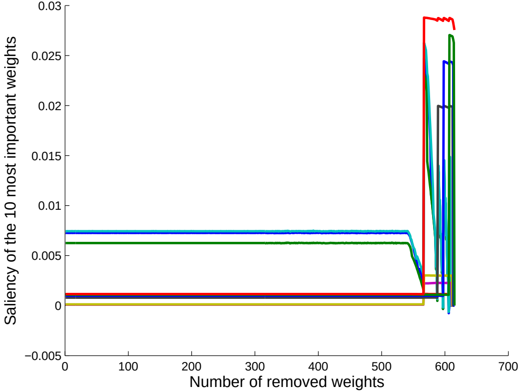
$$\mathbf{f}_\mathcal{A} : \hat{\mathbf{w}}_\mathcal{A} \in \mathbb{R}^k,$$

where $\hat{\mathbf{w}}_\mathcal{A} \in \mathbb{R}^k$ is an optimal parameter vector of $\mathbf{f}_\mathcal{A}$ which minimizes error function

$$S(\mathbf{w}|\mathcal{L}) = -\sum_{i \in \mathcal{K}} \sum_{\xi=1}^{S} t_{i\xi} \ln\bigl(p_\xi(\mathbf{x}_i, \mathbf{w})\bigr), \quad \mathbf{p}(\mathbf{x}) = \frac{\exp(\mathbf{a}(\mathbf{x}))}{\sum_j \exp(a_j(\mathbf{x}))},$$

computed at learning subset of $\mathfrak{D}$, defined by set of indices $\mathcal{L}$.
We chose optimal model $\hat{\mathbf{f}}_\mathcal{A}$ from a set $\mathfrak{F}$ of *admissive models*:

$$\mathfrak{F} = \bigcup_{\mathcal{A} \subseteq \mathcal{J}} \{\mathbf{f}_\mathcal{A}\}.$$



Dependency of a salency $L_j = \frac{w_j^2}{2\mathbf{H}_{jj}^{-1}}$ from a number of removed parameters.

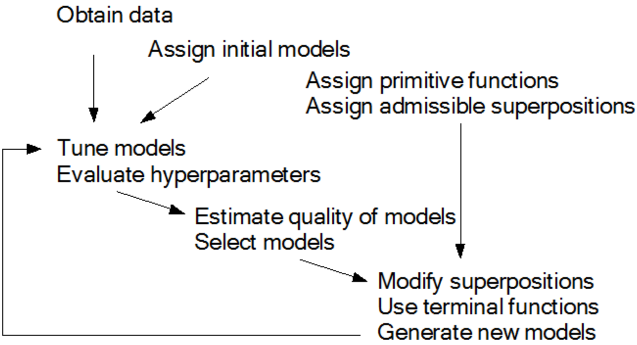
The basic goal of research

To develop a methodology for selection of successively generated models for regression and classification problems.

The approach

a) we successively generate a set of regression models,
b) we investigate space of model parameters,
c) we compare model elements by estimating a covariance matrix and its parameters,
d) we choose the model according to the MDL principle.

Consequent model generation



History of the problem

1	Stepwise method of model selection	M. A. Efroimson, 1960.
2	Regularization for the inverse problem	A. N. Tikhonov, 1963.
3	Group method of data handling	A. G. Ivakhnenko, 1971.
4	Optimal brain damage	Y. LeCun, 1999.
5	Model hyperparameters estimation	Y. Nabney, 2004.
6	Symbol regression	I. Zelinka, D. Koza, 2004.
7	Least angle regression	B. Efron, T. Hastie, 2002.
8	Entropy methods for MDL	P. Gruenwald, 2006.
9	MDL principle in regression	J. Rissanen, 2009.
10	Learning of Bayesian network structure	T. Jaakkola, 2012.

Distribution of the dependent random variable $\mathbf{y} = \boldsymbol{\mu}^{-1}(\mathbf{X}, \mathbf{w})$ belongs to the *exponential family*

$$p(\mathbf{y}|\boldsymbol{\eta}) = h(\mathbf{y})g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{y})\right) \qquad \text{(ED)}$$

with a vector $\boldsymbol{\eta}$ of parameters. The secial cases: normal (ND) and binomial (BD) distributions:

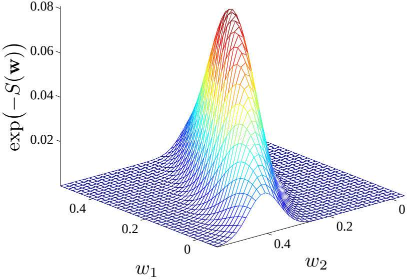
$$p(\mathfrak{D}|\mathbf{B}, \mathbf{w}, \mathbf{f}) = (2\pi)^{-\frac{m}{2}} |\mathbf{B}^{-1}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{f})^{\mathsf{T}}\mathbf{B}(\mathbf{y} - \mathbf{f})\right), \qquad \text{(ND)}$$

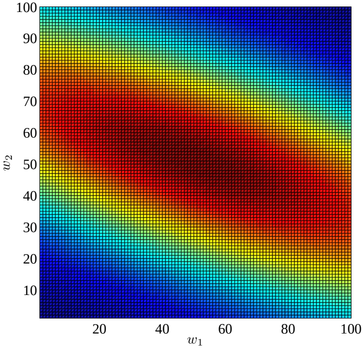
$$p'(\mathfrak{D}|\mathbf{w}, \mathbf{f}) = \prod_{i \in \mathcal{I}} f_i^{y_i}(1 - f_i)^{1-y_i}. \qquad \text{(BD)}$$

Distributions $p(\mathfrak{D}|\mathbf{B}, \mathbf{w}, \mathbf{f})$ and $p(\mathbf{w}|\mathbf{A}, \mathbf{f})$: different cases

Dependent variable $\mathbf{y}$	Model parameters $\mathbf{w}$
$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \sigma_{\mathbf{y}}^2\mathbf{I}) \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{f}, \beta^{-1}\mathbf{I})$	$\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, \sigma_{\mathbf{w}}^2\mathbf{I}) \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$
$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \operatorname{diag}^{-1}(\beta_1, \dots, \beta_m)\mathbf{I})$	$\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, \operatorname{diag}^{-1}(\alpha_1, \dots, \alpha_n)\mathbf{I})$
$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \mathbf{B}^{-1})$	$\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, \mathbf{A}^{-1})$

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

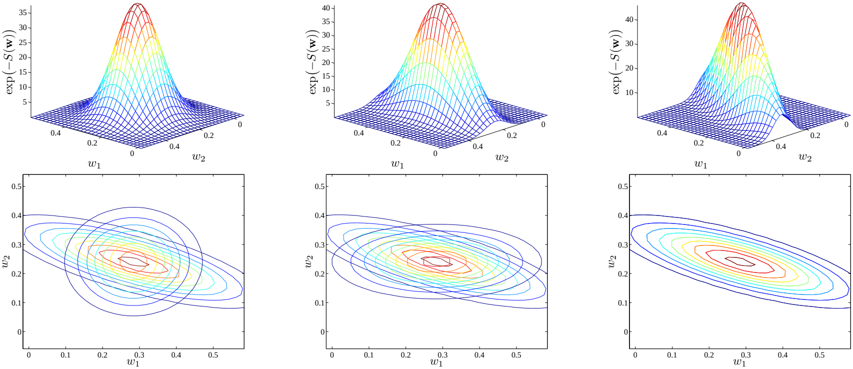




x- and y-axis: parameters $\mathbf{w}$, z-axis: $\exp\bigl(-S(\mathbf{w})\bigr)$.

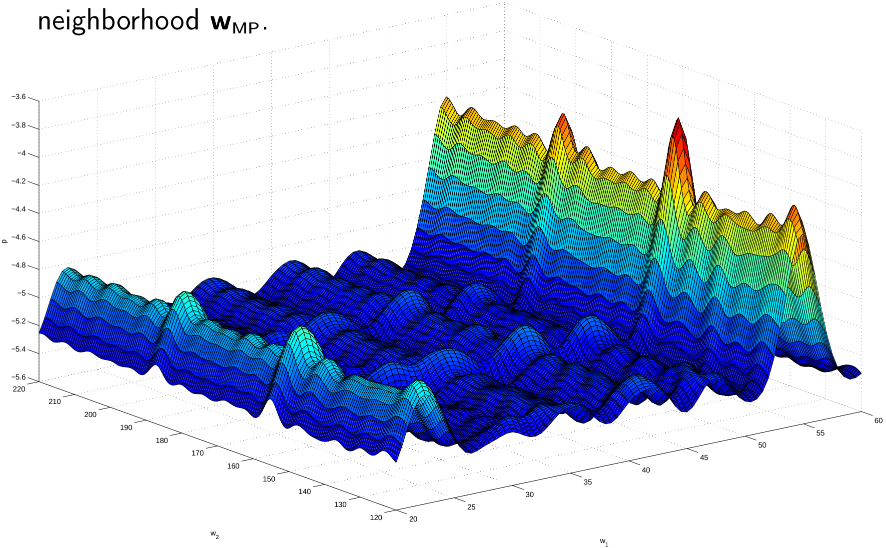
Approximate the set of points $\{s_k\}$ by a function $p(\mathbf{w}|\mathbf{A})$ (ND), considering assumptions about the covariance matrix $\mathbf{A}^{-1}$ type:

$\mathbf{A} = \alpha\mathbf{I}, \quad \alpha \geqslant 0;$ $\qquad \mathbf{A} = \mathbf{diag}(\alpha_1, \dots, \alpha_n);$ $\quad \mathbf{A}, \quad \mathbf{w}^\mathsf{T}\mathbf{A}\mathbf{w} \geqslant 0.$



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

Distribution of parameters $\mathbf{w}$ beyond the most probable neighborhood $\mathbf{w}_{\mathrm{MP}}$.



Posterior parameter distribution

for the given sample $\mathfrak{D}$, model $\mathbf{f} = \mathbf{f}(\mathbf{w}, \mathbf{X})$ and matrices $\mathbf{A}, \mathbf{B}$:

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

The elements of this expression and the corresponding parameters:

$p(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}, \mathbf{f})$ — posterior parameter distribution,

$\mathbf{w}_{\text{MP}} = \arg\max p(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}, \mathbf{f})$ — most probable parameters,

$p(\mathfrak{D}|\mathbf{w}, \mathbf{B}, \mathbf{f})$ — data likelihood,

$\mathbf{w}_{\text{ML}} = \arg\max p(\mathfrak{D}|\mathbf{w}, \mathbf{B}, \mathbf{f})$ — most plausible parameters,

$p(\mathbf{w}|\mathbf{A}, \mathbf{f})$ — prior distribution,

$p(\mathfrak{D}|\mathbf{A}, \mathbf{B}, \mathbf{f})$ — model likelihood.

Coherent Bayesian inference: model selection

For a set of models $\mathfrak{F} = \{f_1, \dots, f_K\}$ to approximate $\mathfrak{D}$

$$p(f_k|D) = \frac{p(D|f_k)p(f_k)}{\sum_{q=1}^{K} p(D|f_k)p(f_k)}.$$

$p(f_k)$ — prior probability,
$p(D|f_k)$ — model evidence,
$p(f_k|D)$ — posterior probability.

Select the most evident model by comparison

$$\frac{p(f_k|D)}{p(f_q|D)} = \frac{p(D|f_k)p(f_k)}{p(D|f_q)p(f_q)}$$

since the denominator does not depend on the model.

Assuming equal prior probability of the models from the set $\mathfrak{F}$,

$$p(f_k) = p(f_q)$$

maximize the model evidence.

Writing the error function $S(\mathbf{w})$ in the following form,

$$S(\mathbf{w}) = -\ln p(\mathfrak{D}|\mathbf{w},\mathbf{B},\mathbf{f})p(\mathbf{w}|\mathbf{A},\mathbf{f}) = E_{\mathbf{w}} + E_{\mathfrak{D}},$$

we obtain the following posterior distribution:

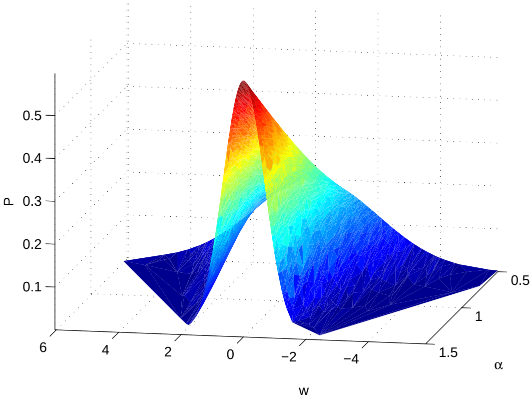
$$p(\mathbf{w}|\mathfrak{D},A,B,f) \propto \frac{\exp\bigl(-S(\mathbf{w})\bigr)}{Z_S}.$$

The case of normal distribution for the dependent variable (ND)

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

The case of binomial distribution for the dependent variable (BD)

$$S(\mathbf{w}) = E_{\mathbf{w}} + \sum_{i\in\mathcal{I}}\bigl(y_i\ln f_i + (1-y_i)\ln(1-f_i)\bigr).$$



x-axis: w is a model parameter.

y-axis: α is an inverted covariance,

z-axis: $p(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}, \mathbf{f})$ is a distribution of parameters.

There is given a sample $\mathfrak{D}$, a set of models $\mathfrak{F} = \{f_k\}$, $k \in \mathcal{K}$ and prior probabilities $p(f_k)$.

The problem is to find the most plausible model f_k:

$$\hat{k} = \arg\max_{k \in \mathcal{K}} p(f_k|\mathfrak{D}) =$$
$$\arg\max_{k \in \mathcal{K}} \int\limits_{\mathbf{w} \in \mathbb{W}_k} p(\mathfrak{D}|\mathbf{w}, \mathbf{B}_k, \mathbf{f}_k) p(\mathbf{w}|\mathbf{A}_k, \mathbf{f}_k) d\mathbf{w}.$$

Posterior model probability

$$p(f_k|\mathfrak{D}) = \frac{1}{p(\mathfrak{D})} p(\mathfrak{D}|f_k) p(f_k),$$

where the function $p(\mathfrak{D}|f_k)$ of the sample $\mathfrak{D}$, with a fixed model f_k is a model likelihood. The normalized coefficient doesn't depend on the model.

Finding the most probable parameters

There is given a sample $\mathfrak{D}$, a model $\mathbf{f} = \mathbf{f}(\mathbf{w}, \mathbf{x})$, a data generation assumption, and an error function

$$S(\mathbf{w}|\mathfrak{D}, \hat{\mathbf{A}}, \hat{\mathbf{B}}, \mathbf{f}) = -\ln\big(p(\mathfrak{D}|\mathbf{w}, \mathbf{B}, \mathbf{f})p(\mathbf{w}|\mathbf{A}, \mathbf{f})\big).$$

The goal is to find parameters $\mathbf{w}_{\text{MP}}$ of the model f

$$\mathbf{w}_{\text{MP}} = \arg\min_{\mathbf{w}\in\mathbb{W}} S(\mathbf{w}|\mathfrak{D}, \hat{\mathbf{A}}, \hat{\mathbf{B}}, \mathbf{f}).$$

The covariance matrix estimation

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg\max_{\mathbf{A}\in\mathbb{R}^{n^2}, \mathbf{B}\in\mathbb{R}^{m^2}} \int_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w}, \mathbf{B}, \mathbf{f})p(\mathbf{w}|\mathbf{A}, \mathbf{f})d\mathbf{w}.$$

Theorem (2014)

The linear model likelihood for the data generation assumption (ND) has the form

$$p(\mathfrak{D}|\mathbf{A},\mathbf{B}) = \frac{|\mathbf{B}|^{\frac{1}{2}}|\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{m}{2}}|\mathbf{K}|^{\frac{1}{2}}} \exp\left(\frac{1}{2}\mathbf{y}^{\mathsf{T}}(\mathbf{C}^{\mathsf{T}}\mathbf{K}\mathbf{C}-\mathbf{B})\mathbf{y}\right),$$

and its logarithm has the form $\ln p(\mathfrak{D}|\mathbf{A},\mathbf{B}) =$

$$= -\frac{1}{2}\big(\ln|\mathbf{K}| + m\ln 2\pi - \ln|\mathbf{B}| - \ln|\mathbf{A}| - \mathbf{y}^{\mathsf{T}}(\mathbf{C}^{\mathsf{T}}\mathbf{K}\mathbf{C}-\mathbf{B})\mathbf{y}\big).$$

Here

$$\mathbf{K} = \mathbf{X}^{\mathsf{T}}\mathbf{B}\mathbf{X} + \mathbf{A}, \quad \mathbf{C} = \mathbf{K}^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{B}.$$

Theorem (2013)

For the data generation assumption(ND) with the fixed covariance matrices $\mathbf{A}^{-1}, \mathbf{B}^{-1}$ the iterative algorithm of parameters estimation,

$$\Delta \mathbf{w}_{k+1} = (\mathbf{J}^\mathsf{T}\mathbf{J})^{-1}\left(\mathbf{J}^\mathsf{T}\big(\mathbf{y} - \mathbf{f}(\mathbf{w}, \mathbf{X})\big) - \frac{1}{\beta}\mathbf{A}^{-1}\mathbf{w}_k\right),$$

finds a minimum of the error function of general form $S(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}, \mathbf{f})$ with the convergence of vectors sequence $\mathbf{w}_k$.

Remark

The iterative algorithm $\mathbf{w}_{k+1} = \Delta\mathbf{w}_{k+1} + \mathbf{w}_k$ requires the initial value $\mathbf{w}_0$. The sequence $\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2$ monotonically decreases due to increase of the step k.

Theorem (2013)

For the data generation assumption (BD) with the fixed covariance matrices $\mathbf{A}^{-1}, \mathbf{B}^{-1}$ the iterative algorithm of parameters estimation for the generalized linear model,

$$\Delta\mathbf{w}_{k+1} = (\mathbf{X}^\mathsf{T}\mathbf{B}\mathbf{X} + \mathbf{A})^{-1}\mathbf{X}^\mathsf{T}\mathbf{B}^\mathsf{T}\mathbf{y} - \mathbf{w}_k, \quad \text{variant:}$$

$$\Delta\mathbf{w}_{k+1} = (\mathbf{X}^\mathsf{T}\mathbf{B}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{B}\big(\mathbf{X}\mathbf{w}_k - \mathbf{B}^{-1}(\mathbf{f} - \mathbf{y})\big) + \frac{1}{2}\mathbf{w}_k^\mathsf{T}\mathbf{A}\mathbf{w}_k,$$

finds a local minimum of the error function of general form with the convergence of vectors sequence $\mathbf{w}_k$.

Let the vector of parameters $\mathbf{w}_0 = [w_{1(0)}, \dots, w_{n(0)}]^\mathsf{T}$ be fixed.

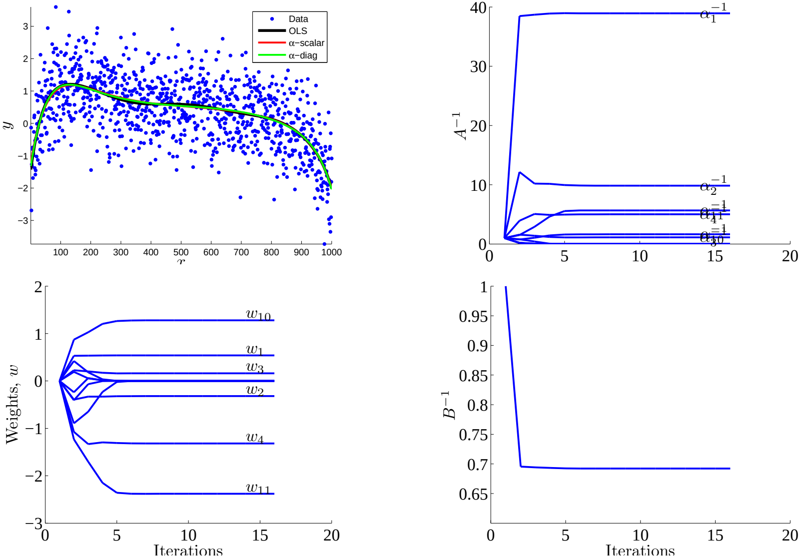
Theorem (2013)

In a neighborhood of the parameters $\mathbf{w}_0$ the covariance matrix estimations $\mathbf{A}^{-1}, \mathbf{B}^{-1}$ for the data generation assumption (ND) has the form

$$\alpha_i = \frac{1}{2}\lambda_i \left(\sqrt{1 + \frac{4}{(w_i - w_{i(0)})^2 \lambda_i}} - 1 \right), \text{ where } \lambda_i = \beta \mathbf{diag}(h_i),$$

$$\beta = \frac{m - \gamma}{2(\mathbf{f} - \mathbf{y})^\mathsf{T} \mathbf{B}' (\mathbf{f} - \mathbf{y})}, \text{ где } \gamma = \sum_{j=1}^{W} \frac{\lambda_j}{\lambda_j + \alpha_j}.$$

The sequences $\|\mathbf{A}_{k+1} - \mathbf{A}_k\|^2$ and $\|\beta_{k+1} - \beta_i\|^2$ monotonically decrease due to increase of the step k.


Data
OLS
α−scalar
α−diag
y
x
A^{-1}
α_1^{-1}
α_2^{-1}
Weights, w
w_{10}
w_1
w_3
w_2
w_4
w_{11}
Iterations
B^{-1}
Iterations

The set of basic functions $\mathfrak{G}$

There is given a set $\mathfrak{G} = \{\text{id}, g_1, \dots, g_l | g = g(\mathbf{b}, \mathbf{x}')\}$, that is, there are given

1) the function $g : (\mathbf{b}, \mathbf{x}') \mapsto \mathbf{x}''$,
2) its parameters $\mathbf{b}$,
3) arity $v(g)$ of the function g and an order of arguments,
4) a domain $\text{dom}(g)$ and a codomain $\text{cod}(g)$.

Consider a model $f(\mathbf{w}, \mathbf{x})$ given by a superposition

$$f(\mathbf{w}, \mathbf{x}) = (g_{i(1)} \circ \dots \circ g_{i(K)})(\mathbf{x}), \text{ где } \mathbf{w} = [\mathbf{b}_{i(1)}^{\mathsf{T}}, \dots, \mathbf{b}_{i(K)}^{\mathsf{T}}]^{\mathsf{T}}.$$

An admissible superposition f

is a superposition such that

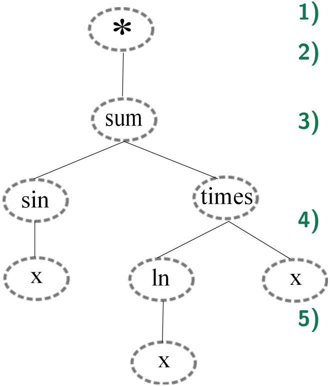
$$\text{cod}(g_{i(k+1)}) \subseteq \text{dom}(g_{i(k)}), \text{ для всех } k = 1, \dots, K-1.$$

To generate the models we use

1) the set dom($\mathbf{x}$),
2) the set of basic functions $\mathfrak{G} = \{\text{id}, g\}$, $g : \mathbf{x} \mapsto \mathbf{x}'$,
3) the set Gen of rules for superposition generation,
4) the set Rem of rules for isomorphic superpositions simplification and estimation.

We propose the following basic methods for the superpositions generation:

- inductive generation,
- structure learning,
- direct search.

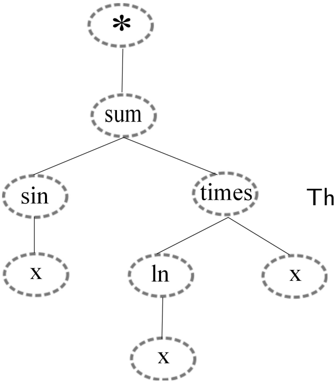


$$f = \sin(x) + (\ln x)x$$

1) the root $*$ of the tree Γ_f has the single vertex,
2) other vertices V_i correspond to the functions $g_r \in \mathfrak{G}$: $V_i \mapsto g_r$,
3) the number of children V_j of the vertex V_i equals to an arity of the corresponding function g_r: $\text{val}(V_j) = v(g_{r(i)})$,
4) the domain of the function $g_{r(i)}$ of a child V_j contains the codomain of the function $g_{r(j)}$ of the parent V_i: $\text{dom}(g_{r(i)}) \supseteq \text{cod}(g_{r(j)})$,
5) an order of vertices traversal with a parent vertex V_i corresponds to the order of arguments of the corresponding function $g_{r(i)}$,
6) the leaves Γ_f correspond to the independent variables, elements of the vector $\mathbf{x}$.

The link matrix $\mathbf{Z}_f$ for the tree Γ_f

	sum	times	ln	sin	x
$*$	1	0	0	0	0
sum	0	1	1	0	0
times	0	0	0	1	1
ln	0	0	0	0	1
sin	0	0	0	0	1



$f = \sin(x) + (\ln x)x$

The link probability matrix $\mathbf{P}_f$ for the tree Γ_f

	sum	times	ln	sin	x
$*$	0.7	0.1	0.1	0.1	0.2
sum	0.2	0.7	0.8	0.1	0.2
times	0.1	0.3	0	0.8	0.8
ln	0.2	0.1	0.3	0.1	0.9
sin	0.1	0.2	0.1	0	0.8

$\mathfrak{Z}$ is a set of matrices corresponding to the superpositions from $\mathfrak{F}$.

There is given a sample $\mathfrak{D} = \{(\mathbf{D}_k, f_k)\}$ where the element $\mathbf{D}_k = (\underset{m\times n}{\mathbf{X}}, \underset{m\times 1}{\mathbf{y}})$, there given $\mathfrak{G}$ and $\mathfrak{F} = \{f_s \mid \mathbf{f}_s : (\hat{\mathbf{w}}_k, \mathbf{X}) \mapsto \mathbf{y}, s \in \mathbb{N}\}$.

The goal

to find an algorithm $a : \mathbf{D}_k \mapsto f_s$ following the condition

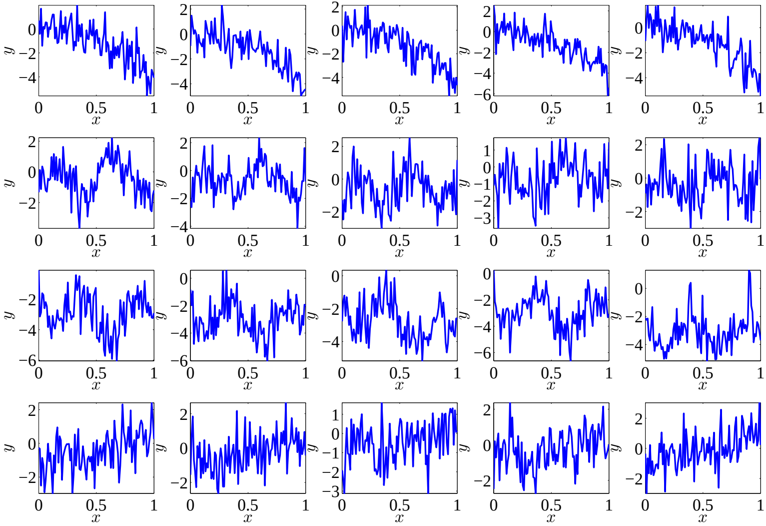
$$\mathbf{Z}_{f_s} = \arg\max_{\mathbf{Z}\in\mathfrak{Z}} \sum_{i,j} P_{ij} \times Z_{i,j}.$$

The index $\hat{s}$, что $f_{\hat{s}}$ provides a minimum for the error function S:

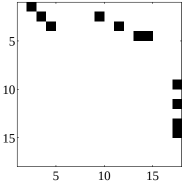
$$\hat{s} = \arg \min_{s\in\{1,\dots,|\mathfrak{F}|\}} S(f_s \mid \hat{\mathbf{w}}_k, \mathbf{D}_k),$$

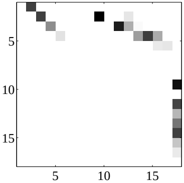
where $\hat{\mathbf{w}}_k$ is an optimal vector of parameters f_s for each $f_s \in \mathfrak{F}$ with the fixed $\mathbf{D}_k$:

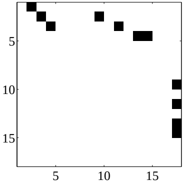
$$\hat{\mathbf{w}}_k = \arg \min_{\mathbf{w}\in\mathbb{W}_s} S(\mathbf{w} \mid f_s, \mathbf{D}_k).$$


y
x
0
0.5
1

Initial and forecasted superposition







$$f = w_1 \cos(w_2 x + w_3) + w_4 x + w_5 \ln(w_6 x + w_7) + w_8,$$

$$f = \cos(x) + x + \ln(x), \qquad \mathbf{w} = [1, 1, 0, 1, 1, 1, 0, 0]^\mathsf{T}.$$

The set $\mathcal{A}$ uniquely defines a model $f_{\mathcal{A}} \in \mathfrak{F}$.

The successive modification procedure

Add: to add an index j to the set $\mathcal{A}_k = \mathcal{A}_{k-1} \cup \{j\}$, that corresponds to the maximum value of the model likelihood

$$\hat{j} = \arg\max_{j \in \mathcal{J} \setminus \mathcal{A}_k} p(f_{\mathcal{A}_k} | \mathbf{w}_{\text{MP}}, \mathbf{A}, \mathbf{B}, \mathfrak{D}).$$

Del: to remove an index j from the set $\mathcal{A}_k = \mathcal{A}_{k-1} \setminus \{j\}$ to maximum increase the stability, $\hat{j} = \arg\max_{j \in \mathcal{A}_k} Q(f_{\mathcal{A}_k} | \mathbf{w}_{\text{MP}}, \mathbf{A}, \mathbf{B}, \mathfrak{D})$:

$$\hat{j} = \arg\max_{j \in \mathcal{A}_{k-1}} \sum_{g=t-\hat{i}+1}^{t} q_g^j, \quad \text{где} \qquad \hat{i} = \sum_{g=1}^{t} \left[\eta_g^2 > \eta_t \right].$$

The stages Add and Del repeated independently such that the inequality holds on each stage: $\max_{\text{Add-Del} k \in \mathbb{N}} \big(\mathcal{E}(f_{\mathcal{A}'_k}) \big) - \mathcal{E}(f_{\mathcal{A}_k}) \leqslant \Delta \mathcal{E}$.

The algorithm is repeated while the expectation of the likelihood function $\mathsf{E}\mathcal{E}(f_{\mathcal{A}_k})$ remains constant.

We approximate error function S by

$$\Delta S = S(\mathbf{w}_0 + \Delta\mathbf{w}) - S(\mathbf{w}_0) = \frac{1}{2}\Delta\mathbf{w}^{\mathsf{T}}\mathbf{H}\Delta\mathbf{w}$$

near its local optimum $\mathbf{w}_0$. Here $\Delta w = \mathbf{w} - \mathbf{w}_0$ and $\mathbf{H}$ stands for Hessian matrix of S.

Since

$$w_j = 0 \equiv \mathbf{e}_j^{\mathsf{T}}\Delta\mathbf{w} + w_j = 0,$$

we specify Lagrange function

$$L = \frac{1}{2}\Delta\mathbf{w}^{\mathsf{T}}\mathbf{H}\Delta\mathbf{w} - \lambda_i(\mathbf{e}_j^{\mathsf{T}}\Delta\mathbf{w} + w_j)$$

for conditional optimization $\Delta S \to \min$, $\mathbf{e}_j^{\mathsf{T}}\Delta\mathbf{w} + w_j = 0$.
The optimal pruning criterion is then given by

$$\hat{j} = \operatorname*{argmin}_{j\in\mathcal{A}} L_j, \text{ where } L_j = \frac{w_j^2}{2[\mathbf{H}^{-1}]_{j,j}}.$$

Decomposition of the covariance matrix $\mathbf{A}^{-1}$

Consider the condition numbers $\eta_j = \frac{\lambda_{\max}}{\lambda_j}$ in the singular decomposition of the covariance matrix $\mathbf{A}^{-1}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}^2$. Find covariance of the parameters $\mathbf{w}$

$$\mathbf{Var}(\mathbf{w}) = \frac{1}{\beta}(\mathbf{V}^\mathsf{T})^{-1}\mathbf{\Lambda}^{-2}\mathbf{V}^{-1} = \frac{1}{\beta}\mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V}^\mathsf{T},$$

where β is an inverse covariance of the residuals, and the covariance of the parameter w_j is a j-th diagonal element $\mathbf{Var}(\mathbf{w})$.

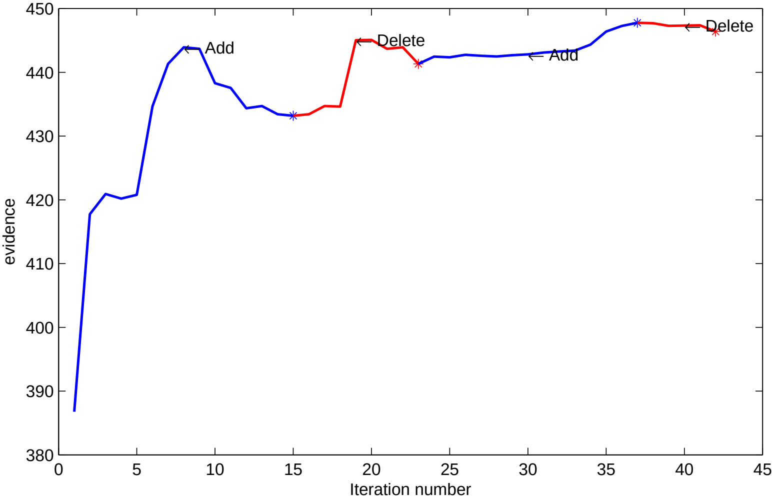
Removal of the index $\hat{j}$ from the set $\mathcal{A}_k = \mathcal{A}_{k-1} \setminus \{\hat{j}\}$

$$\hat{j} = \arg\max_{j \in \mathcal{A}_{k-1}} \sum_{g=t-\hat{i}+1}^{t} q_g^j, \quad \text{where} \qquad \hat{i} = \sum_{g=1}^{t} \left[\eta_g^2 > \eta_t\right], \textit{where}$$

$$\beta\mathbf{var}(w_i) = \sum_{j=1}^{n} \frac{v_{ij}^2}{\lambda_j^2} = (q_{i1} + q_{i2} + \ldots + q_{in})$$

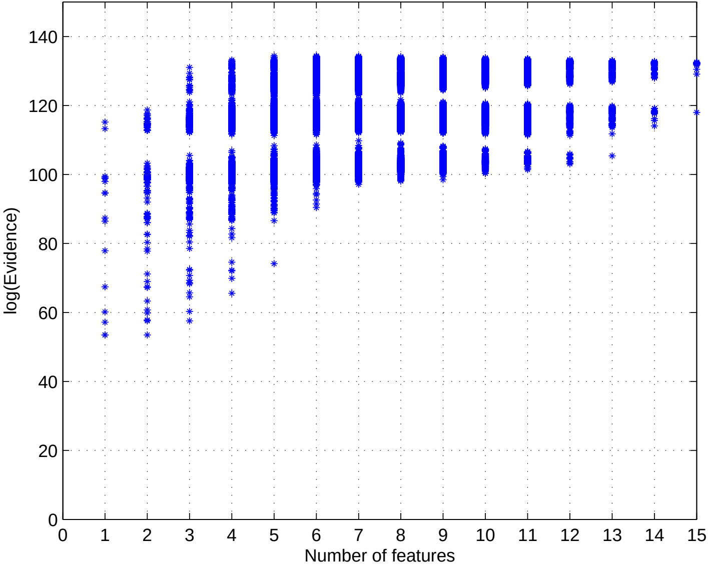
makes maximum increase the model stability $f_{\mathcal{A}_k}$ on the pair of steps $k, k-1$.

Likelihood maximization during the successive model modification

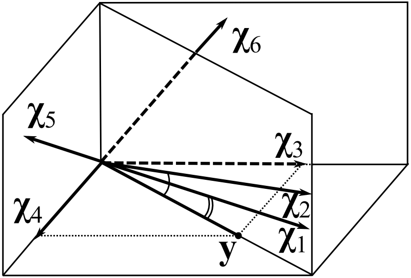


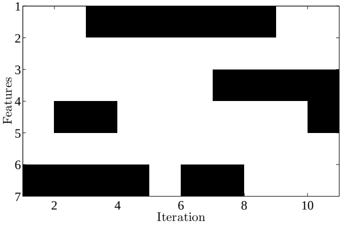
x-axis: iterations k, y-axis: likelihood $p(f_{\mathcal{A}_k}|\mathbf{w}_{\text{MP}}, \mathbf{A}, \mathbf{B}, \mathfrak{D})$.

Change of likelihood at the arbitrary modification

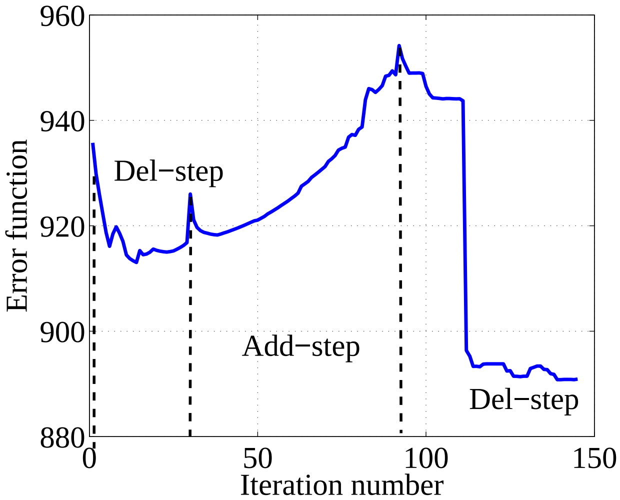


Choice of the most plausible and stable model

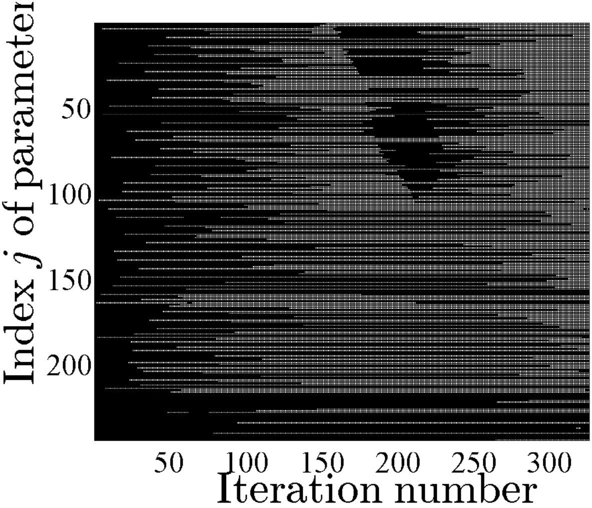




x-axis: the iterations k, y-axis: the indices of the elements j, the black rectangle: the index j added to the set $\mathcal{A}_k$.



Modification procedure runs until the process stabilizes. The termination criterion



Black cells denote active parameters w_j $j \in \mathcal{A}$.

To compare models we use three quality criteria for model **f** with parameter vector **w**: complexity, robustness and precision.

Complexity C is the size of the set $\mathcal{A}$ of active parameters:

$$C(\mathbf{w}) = \sum_{i=1}^{k} [w_i \neq 0].$$

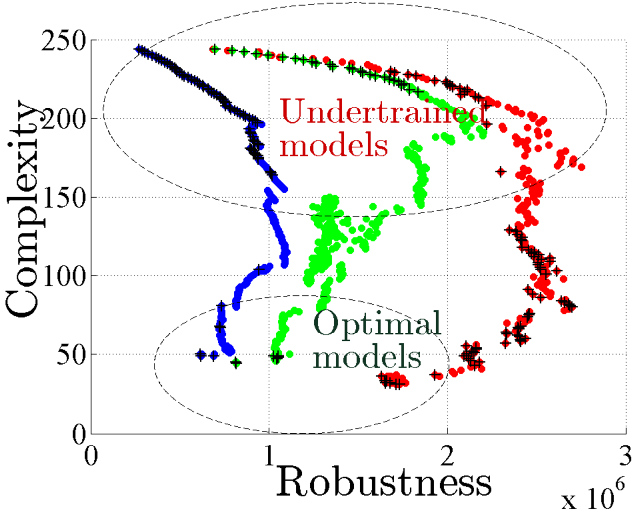
Robustness $\eta = \eta(\hat{\mathbf{w}})$ is equal to the condition number of inverse covariance matrix $\mathbf{A}$ of $\mathbf{w}$:

$$\eta(\hat{\mathbf{w}}) = \frac{\lambda_{\max}}{\lambda_{\min}},$$

where $\lambda.$ stand for eigenvalues of $\mathbf{A}$.
Robustness increases with decrease of η: the best case is $\lambda_{\min} = \lambda_{\max},\ \eta = 1.$

Precision S is measured as error $S(\mathbf{w}|\mathcal{L})$.

Model interpretation



Generated models in complexity-robustness coordinates.

Dataset: Energy consumption, an example

Algorithm	$S_{\mathcal{L}}$	$S_{\mathcal{C}}$	AIC	BIC	C_p	$\lg \kappa$	k
Genetics	0,073	0,107	-1152	-1072	337	13	26
GMDH	0,146	0,194	-1076	-1045	745	6	10
Stepwise	0,128	0,154	-1092	-1055	644	7	12
Ridge	0,111	0,146	-819	-330	832	33	160
Lasso	0,121	0,147	-1089	-1034	611	5	18
Stagewise	0,071	0,096	-1157	-1077	324	9	26
FOS	0,106	0,135	-1105	-1044	527	7	20
LARS	0,098	0,095	-1102	-1017	492	7	28
Consequent	0,097	0,123	-1118	-1054	469	5	21