Analysis of the Regression Model Parameters

Vadim STRIJOV

Computing Center of the
Russian Academy of Sciences

Ankara, October 08, 2009
Institute of Applied Mathematics, METU

William of Ockham, 1285-1349

Ventia non sunt multiplicanda praeter necessitatem.

Occam's razor: entities (model elements) must not be multiplied beyond necessity.

Coherent Bayesian Inference

Coherent Bayesian Inference is a method of the model comparison. This method uses Bayesian inference two times:

1. to estimate the posterior probability of the model itself and
2. to estimate the posterior probability of the model parameters.

Bayesian Comparison

Consider a finite set of models $f_1, \ldots, f_M$ that fit the data D.
Denote prior probability of i-th model by $P(f_i)$. After the data have come, the posterior probability of the model

$$P(f_i|D) = \frac{P(D|f_i)P(f_i)}{\sum_{j=1}^{M} P(D|f_j)P(f_j)}.$$

The probability $P(D|f_i)$ of data D, given model f_i is called the evidence of the model f_i.
Since the denominator for all models from the set is the same,

$$P(D) = \sum_{j=1}^{n} P(D|f_j)P(f_j),$$

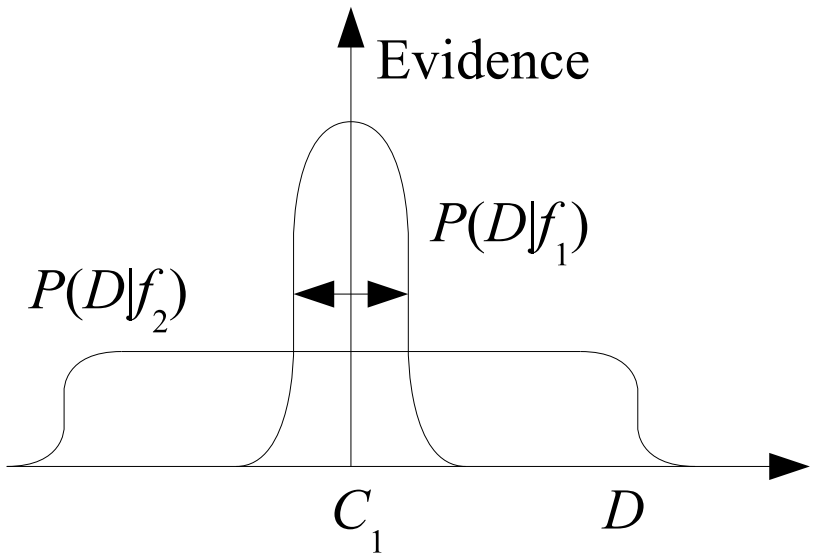
then

$$\frac{P(f_i|D)}{P(f_j|D)} = \frac{P(f_i)P(D|f_i)}{P(f_j)P(D|f_j)}.$$

Assume the prior probabilities to be equal, $P(f_i) = P(f_j)$.

The Occam's razor

If f_2 — is more complex model, then its distribution $P(D|f_2)$ has smaller values (variance has greater values). If the errors of both models are equal, then the simple model f_1 is more probable than the complex model f_2.



A toy example of evidence computation

Let there be given the series $\{-1, 3, 7, 1\}$. One must to forecast the next two elements.
The model f_a:

$$x_{i+1} = x_i + 4$$

gives the next elements $15, 19$.
The model f_c:

$$x_{i+1} = -\frac{x_i^3}{11} + \frac{9x_i^2}{11} + \frac{23}{11}$$

gives the next elements $-19.9, 1043.8$.

Let the prior probabilities be equal or comparable.
Let each parameter of the models is in the set

$$\{-50, \ldots, 0, \ldots, 50\}.$$

A toy example, continued

The parameters $(n = 4, x_1 = -1)$ brings the proper model with zero-error.
The evidence of the model f_a is

$$P(D|f_a) = \frac{1}{101}\frac{1}{101} = 0.00010.$$

Let the denominators of the second models are in the set $\{0, \ldots, 50\}$.
Take account of $c = -1/11 = -2/22 = -3/33 = -4/44$.
The evidence of the model f_c is

$$P(D|f_c) = \left(\frac{1}{101}\right)\left(\frac{4}{101}\frac{1}{50}\right)\left(\frac{4}{101}\frac{1}{50}\right)\left(\frac{2}{101}\frac{1}{50}\right) = 2.5 \times 10^{-12}.$$

The result of the model comparison is

$$\frac{P(D|f_a)}{P(D|f_c)} = \frac{0.00010}{2.5 \times 10^{-12}}.$$

The 1st level of the inference

At the first level one must to estimate the model parameters $\mathbf{w}$, given data D,

$$P(\mathbf{w}|D, f_i) = \frac{P(D|\mathbf{w}, f_i)P(\mathbf{w}|f_i)}{P(D|f_i)}.$$

The model evidence $P(D|f_i)$ is not considered at this level.
To estimate parameters, approximate logarithm of the posterior distribution of $P(\mathbf{w}|D, f_i)$ by Taylor power series,

$$P(\mathbf{w}|D, f_i) \approx P(\mathbf{w}_{MP}|D, f_i) \exp(-\frac{1}{2}\Delta\mathbf{w}^T A \Delta\mathbf{w})),$$

where $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{MP}$.
Here the matrix A is the covariance matrix at the neighborhood of $\mathbf{w}_{MP}$.

The 2nd level of the inference

The second level of the Bayesian inference defines what the model is more adequate for the given data. The posterior probability of i-th model is given by

$$P(f_i|D) \propto P(D|f_i)P(f_i).$$

Here $P(D|f_i)$ is the evidence of the model and the denominator at the 1st level:

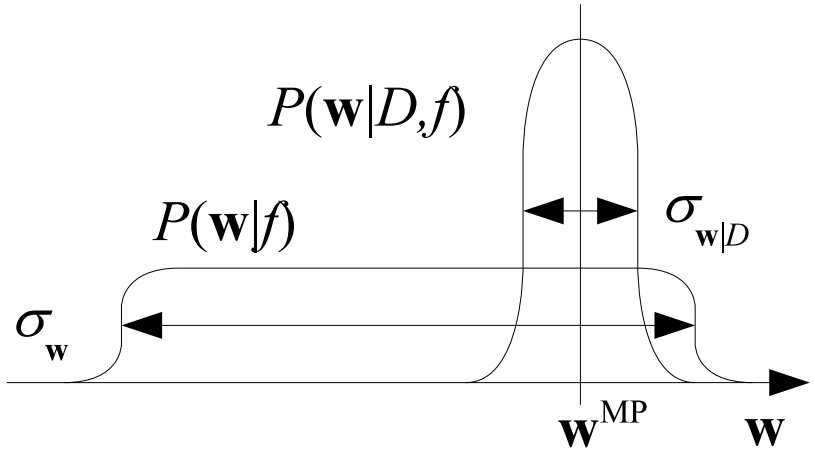
$$P(D|f_i) = \int P(D|\mathbf{w}, f_i)P(\mathbf{w}|f_i)d\mathbf{w}.$$

Assume the distribution $P(\mathbf{w}|D, f_i) \propto P(D|\mathbf{w}, f_i)P(\mathbf{w}|f_i)$ has a peak at $\mathbf{w}_{MP}$. According to the Laplace approximation,

$$P(D|f_i) \approx P(D|\mathbf{w}_{MP}, f_i)P(\mathbf{w}_{MP}|f_i) \times \sigma_{w|D},$$

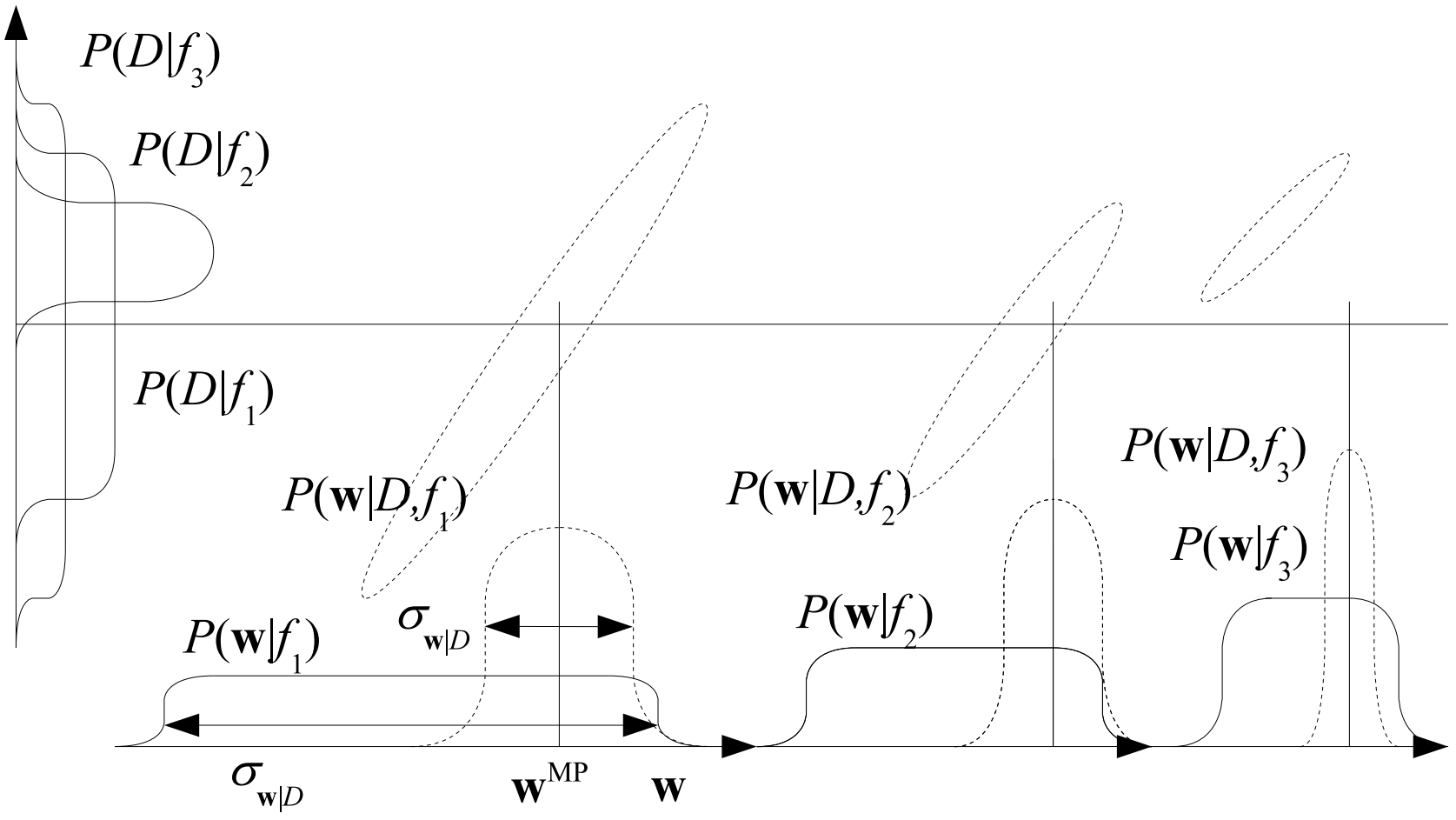
$$\text{evidence} \approx \text{maximum likelihood} \times \text{Occam factor}.$$

Occam factor

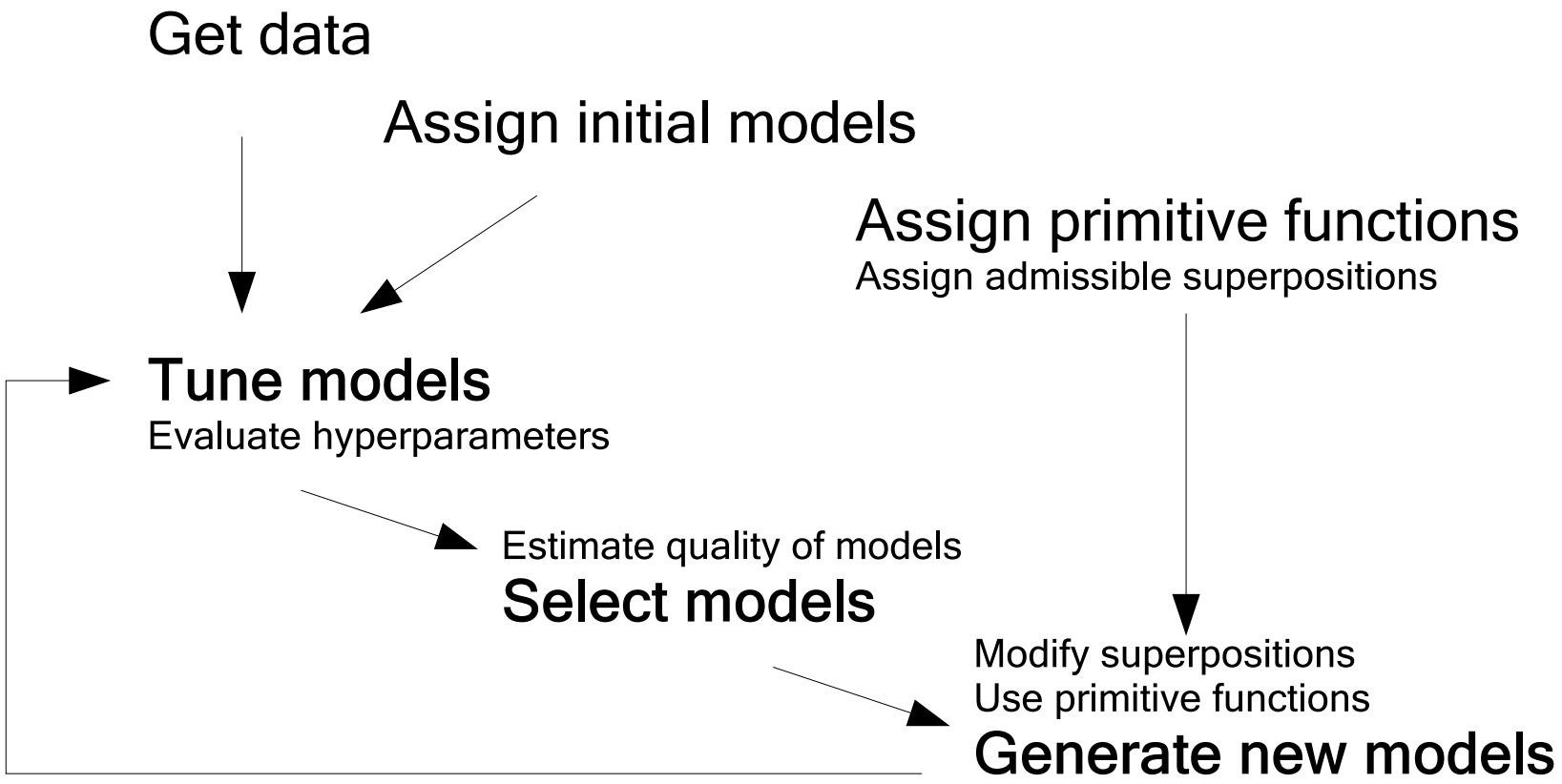


The Occam factor is given by the variance of the model parameters.
The variable $\sigma_{w|D}$ depends on the posterior distribution of the parameters **w**.
The Occam factor shows the "compression"of the parameter space when the data have come.

An example of the method



The process of the model construction



Let there be given

The samples:

$\{\mathbf{x}_1, ..., \mathbf{x}_N | \mathbf{x} \in \mathbb{R}^P\}$ the independent variables,

$\{y_1, ..., y_N | y \in \mathbb{R}\}$ the corresponding depended variables.

Denote by D the sample set $\{(\mathbf{x}_n, y_n)\}$.

The primitive functions:

$G = \{g | g : \mathbb{R} \times ... \times \mathbb{R} \longrightarrow \mathbb{R}\}$ parametric functions,

$g = g(\mathbf{b}, \cdot, \cdot, ..., \cdot)$.

G defines the set of admissible superpositions $\mathcal{F} = \{f_i\}$ inductively by its elements g.

$f_i = f_i(\mathbf{w}, \mathbf{x})$,

where $\mathbf{w} = \mathbf{b}_1 \vdots \mathbf{b}_2 \vdots ... \vdots \mathbf{b}_r$.

A regression model of the optimal structure is to be found

$$y = f_i(\mathbf{w}, \mathbf{x}) + \nu$$

One must find a model $f_i \in \mathcal{F}$, which brings the maximum to the target function $p(\mathbf{w}|D, \alpha, \beta, f_i)$,

$f_i \in \mathcal{F}$ — the set of competitive models,

$\mathbf{w}$ — model parameters,

D — sample set (data),

α, β — regularization parameters.

Target function

Given $\nu \sim \mathcal{N}(0, \frac{1}{\beta^2})$ does not depend on $\mathbf{x}$ (homoscedacity),

$$p(y|\mathbf{x}, \mathbf{w}, \beta, f) \equiv p(D|\mathbf{w}, \beta, f) = \frac{\exp(-\beta E_D)}{Z_D(\beta)},$$

$$E_D = \frac{1}{2}\sum_{n=1}^{N}(f(x_n) - y_n)^2, \qquad Z_D(\beta) = \left(\frac{2\pi}{\beta}\right)^{\frac{N}{2}}.$$

Given A — diagonal covariance matrix of the model parameters $\mathbf{w}$,

$$p(\mathbf{w}|A, f) = \frac{\exp(-E_{\mathbf{w}})}{Z_{\mathbf{w}}(A)},$$

$$E_{\mathbf{w}} = \frac{1}{2}\mathbf{w}^T A\mathbf{w}, \qquad Z_{\mathbf{w}}(A) = (2\pi)^{\frac{W}{2}}|A|^{\frac{1}{2}}.$$

The diagonal of the covariance matrix A is $\alpha_1, \alpha_2, \ldots, \alpha_W$. Each hyperparameter corresponds to its own parameter.

Error function and hyperparameters

According to the Bayesian rule, the target function

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

the error function

$$S(\mathbf{w}|A, \beta) = \frac{1}{2}\mathbf{w}^T A\mathbf{w} + \beta E_D,$$

and

$$p(\mathbf{w}|D, A, \beta, f) \propto \exp(-S(\mathbf{w})).$$

How to estimate the hyperparameters?

Maximize the model evidence $p(D|A,\beta)$ according to A and β

$$p(D|A,\beta) = \int p(D|\mathbf{w},A,\beta)p(\mathbf{w}|A)d\mathbf{w} \to \max.$$

Use the Laplace approximation,

$$p(D|A,\beta) = \frac{1}{Z_{\mathbf{w}}(A)}\frac{1}{Z_D(\beta)}\int \exp(-S(\mathbf{w}))d\mathbf{w}.$$

Substitute $Z_{\mathbf{w}}(A)$, $Z_D(\beta)$ and $S(\mathbf{w})$ and find the logarithm of it:

$$p(D|A,\beta) = \frac{1}{Z_{\mathbf{w}}(A)}\frac{1}{Z_D(\beta)}\exp(-S(\mathbf{w}_0))(2\pi)^{\frac{W}{2}}|H|^{-\frac{1}{2}}.$$

$$\ln p(D|A,\beta) = \underbrace{-\frac{W}{2}\ln 2\pi - \frac{1}{2}\ln|A|}_{Z_{\mathbf{w}}^{-1}(A)} \underbrace{-\frac{N}{2}\ln 2\pi + \frac{N}{2}\ln\beta}_{Z_D^{-1}(\beta)} \underbrace{-S(\mathbf{w}_0) + \frac{W}{2}\ln 2\pi - \frac{1}{2}\ln|H|}_{Z_S} =$$

$$= -\frac{1}{2}\ln|A| - \frac{N}{2}\ln 2\pi + \frac{N}{2}\ln\beta \underbrace{-\beta E_D - E_{\mathbf{w}}}_{-S(\mathbf{w}_0)} - \frac{1}{2}\ln|H|.$$

How to estimate the hyperparameters?

As the result of the evidence maximization we obtain

$$2\alpha_j E'_{\mathbf{w}} = W - \gamma_j, \quad \text{where} \quad \gamma_j = \frac{\alpha_j}{\lambda_j + \alpha_j}$$

and

$$2\beta E'_D = N - \sum_{j=1}^{W} \gamma_j.$$

Estimate the hyperparameters α and β_i iteratively,

$$\alpha_j^{\text{new}} = \frac{W - \gamma_j}{2E'_{\mathbf{w}}}, \qquad \beta^{\text{new}} = \frac{N - \sum_{j=1}^{W} \gamma_j}{2E'_D}.$$

Parameter optimization

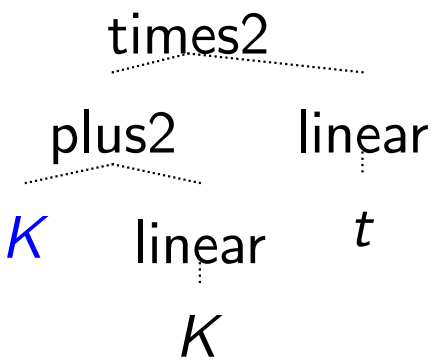
The model generation algorithm contains three steps and runs iteratively.

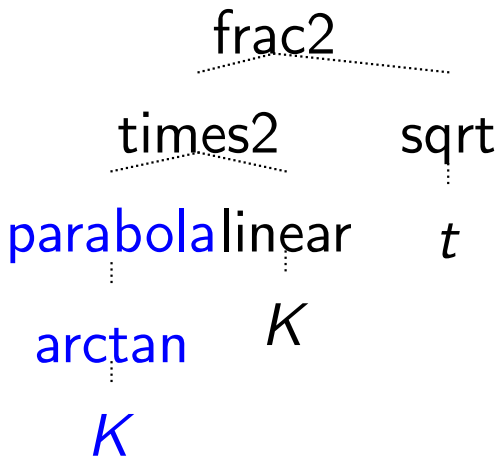
1. Optimize parameters and hyperparameters of every model from the generated set $\mathcal{F} = \{f_1, \ldots, f_M\}$:

$$\mathbf{w}_i^{\mathrm{MP}} = \arg\min_{\mathbf{w}} S(\mathbf{w}|D, A, \beta, f_i).$$

Element exchange

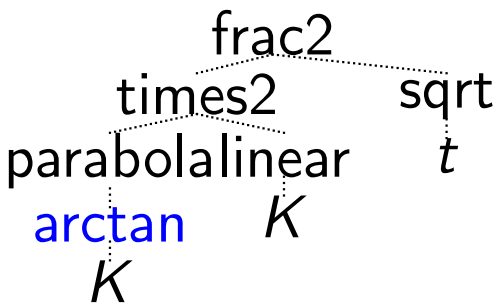
2. Exchange elements of two models:
   1. select randomly a pair of model indexes $i, j \in \{1, \ldots, M\}$,
   2. select from the models f_i and f_j the elements g_{ik} and g_{jl},
   3. create new models f_i' и f_j'.





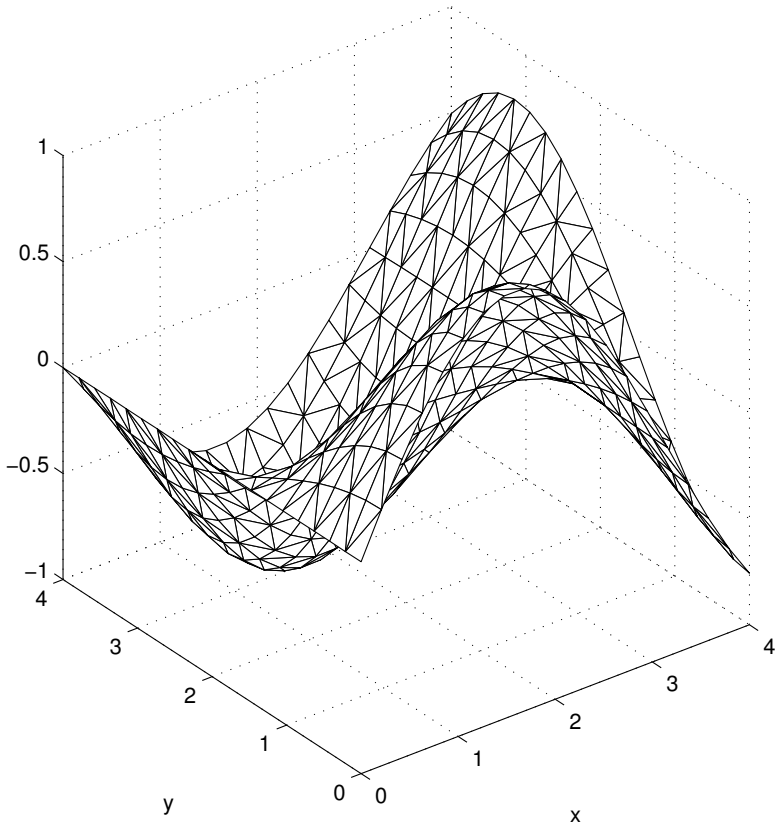
Element modification

3. Modify elements of the new models $\{f_i'\}$:
   1. select a model element g_{ik} from the model f_i,
   2. select from set G an element g_s (it must have the same number of arguments as g_{ik}),
   3. g_{ik} change the model element g_{ik} for the primitive g_s.



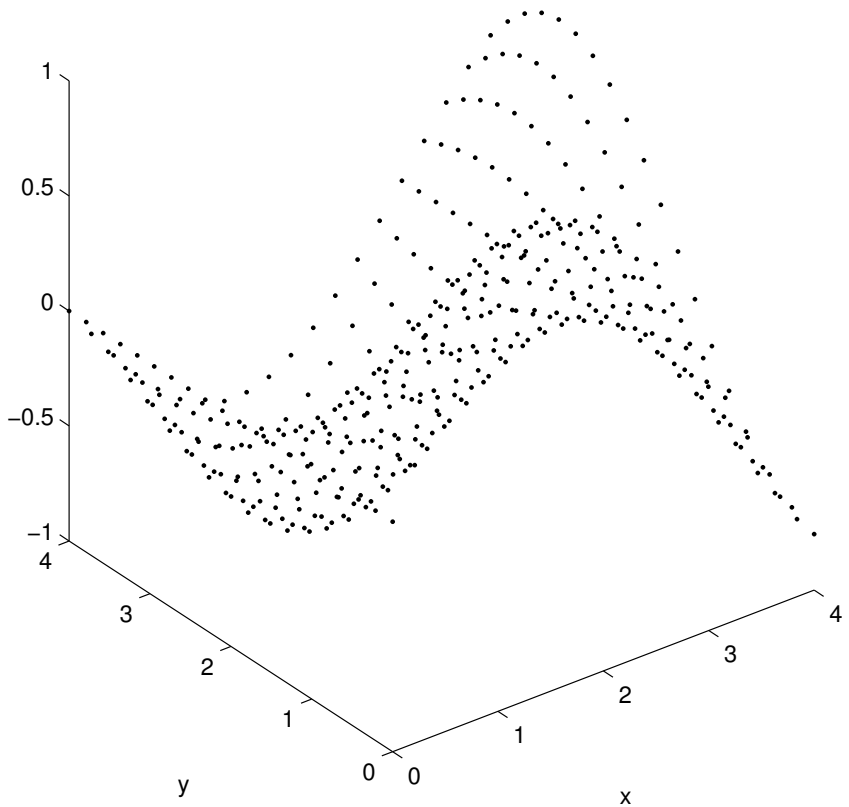
Think of a model

Let it be $y = f(\mathbf{w}, \mathbf{x}) = sin(x_1) * sin(w_1 x_2 + w_2)$.



Given data

The corresponded sample set is shown; it has 380 samples.



Given primitive functions

Function	Description	Parameters
	$g(\mathbf{b}, x_1, x_2)$	
plus	$y = x_1 + x_2$	–
times	$y = x_1 x_2$	–
	$g(\mathbf{b}, x_1)$	
divide	$y = 1/x$	–
multiply	$y = ax$	a
add	$y = x + a$	a
normal	$y = \frac{\lambda}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right) + a$	λ, σ, ξ, a
linear	$y = ax + b$	a, b
parabolic	$y = ax^2 + bx + c$	a, b, c
sin	$y = sin(x)$	–
logsig	$y = \frac{\lambda}{1+\exp(-\sigma(x-\xi))} + a$	λ, σ, ξ, a

Set of the generated models

Let the generated models $\mathcal{F} = \{f_i\}$ be a set
of admissible superpositions
of the primitive functions $G = \{g\}$.

Expert information

Experts assign the initial models

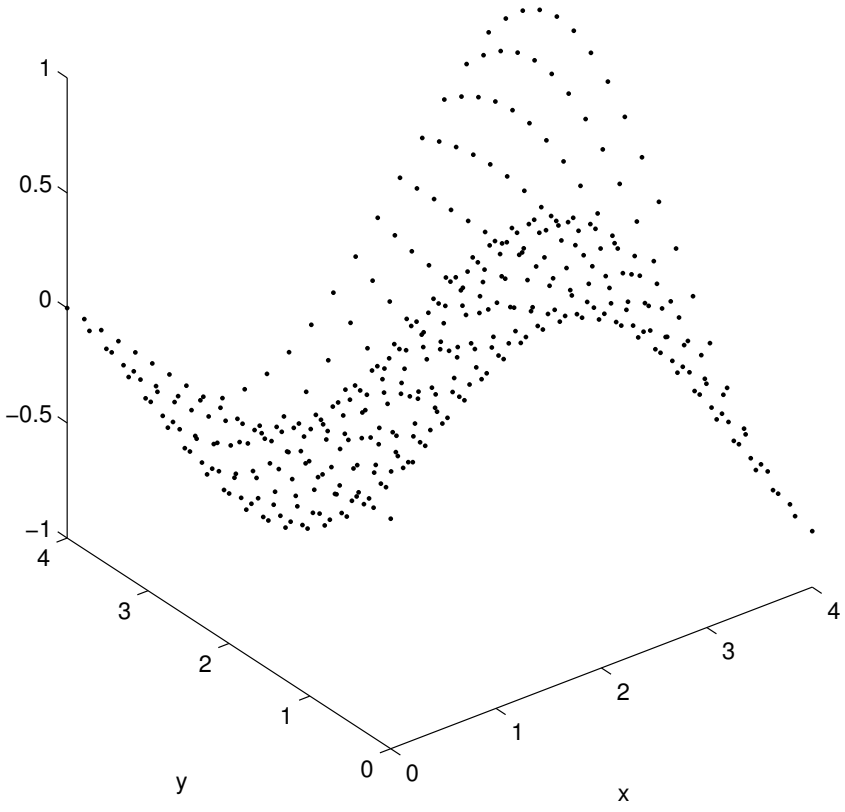
$$\begin{aligned} f_1: &\quad y = \text{linear}(x_1), \\ f_2: &\quad y = \text{normal}(x_2). \end{aligned}$$

And the initial conditions

1. the model complexity:
   $\begin{cases} \text{number of primitives in a superposition } g \text{ no more than } 8, \\ \text{number of parameters } w \text{ no more than } 10; \end{cases}$
2. the target function is sum of squared errors, SSE.

Competitive models

Given data

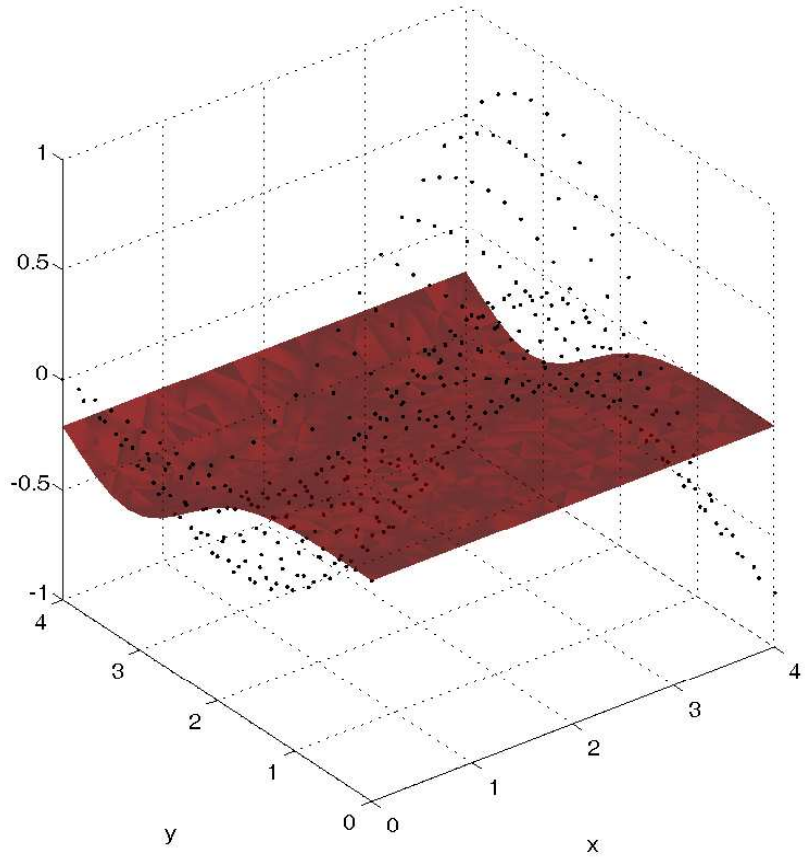


Competitive models

normal($w_{1:3}$,x_2)

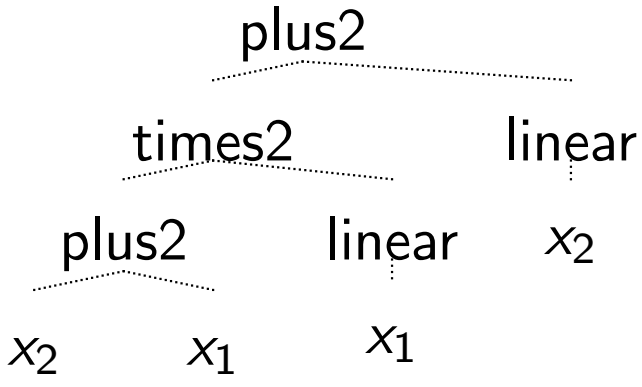
normal

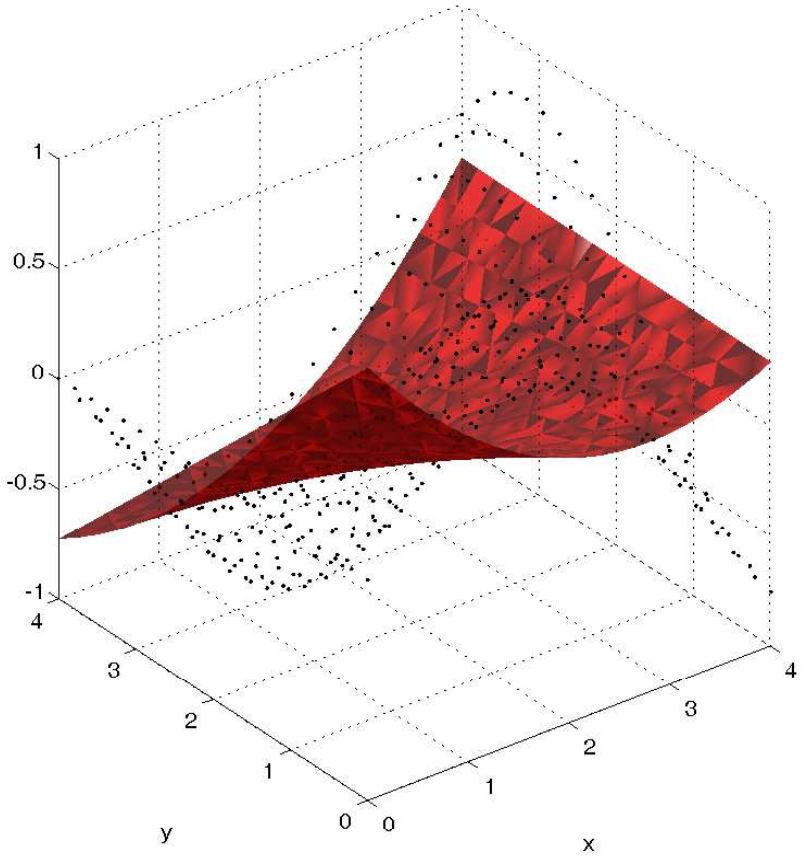
x_2



Competitive models

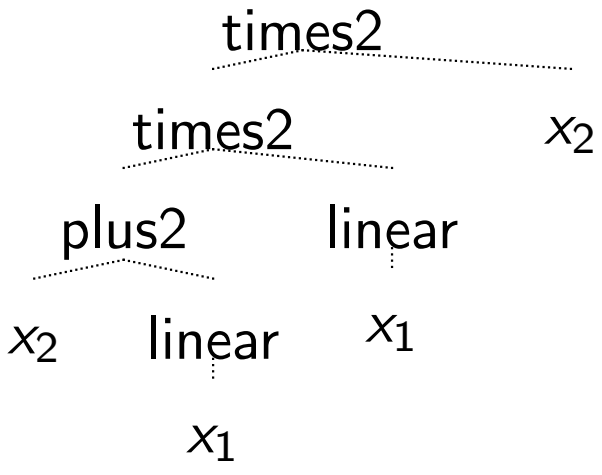
$$\text{plus2}(\emptyset,\text{times2}(\emptyset,\text{plus2}(\emptyset,x_2,x_1),\text{linear}(w_{1:2},x_1)),\text{linear}(w_{3:4},x_2))$$

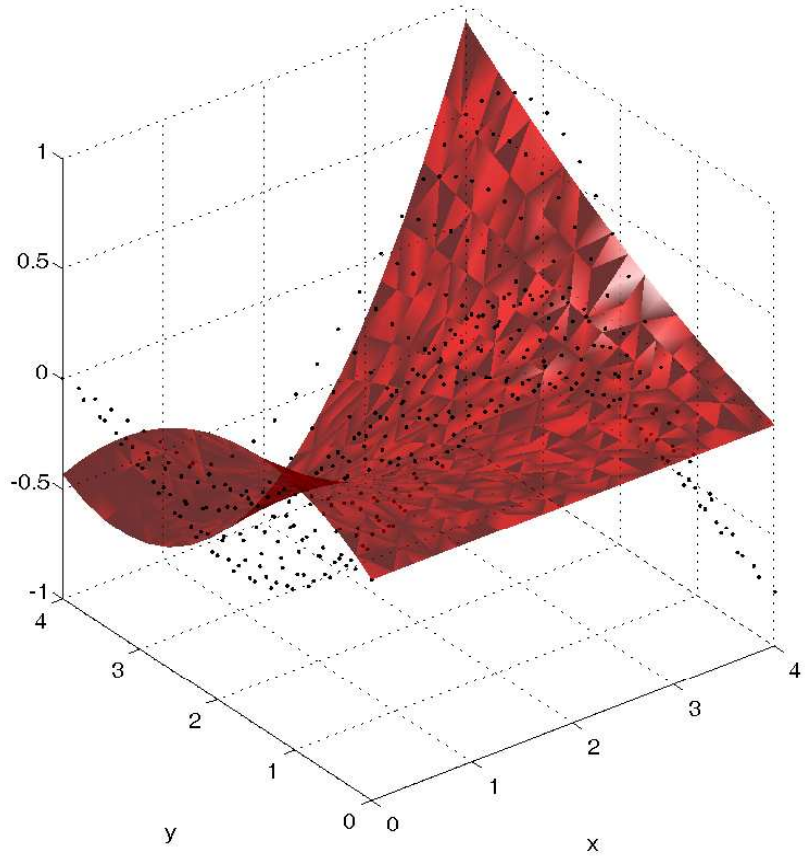




Competitive models

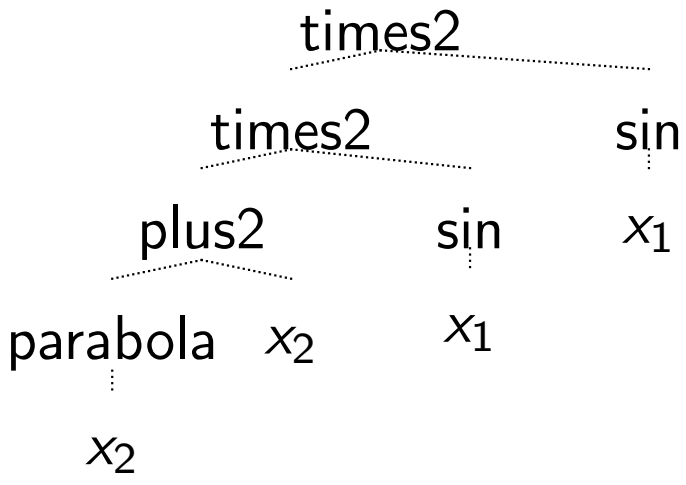
$$\text{times2}(\emptyset,\text{times2}(\emptyset,\text{plus2}(\emptyset,x_2,\text{linear}(w_{1:2},x_1)),\text{linear}(w_{3:4},x_1)),x_2)$$

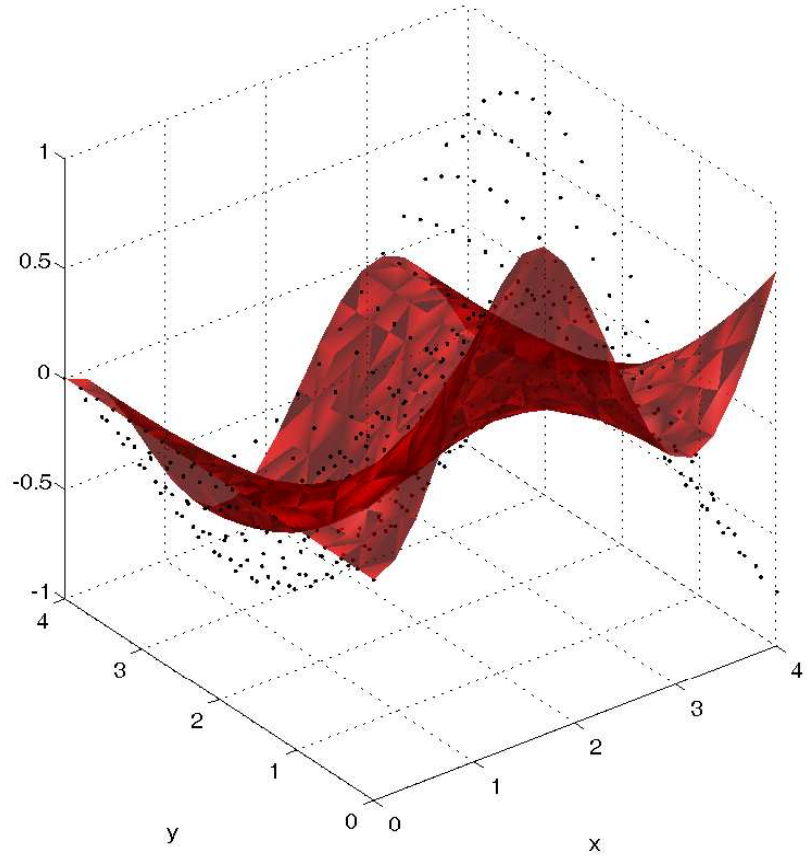




Competitive models

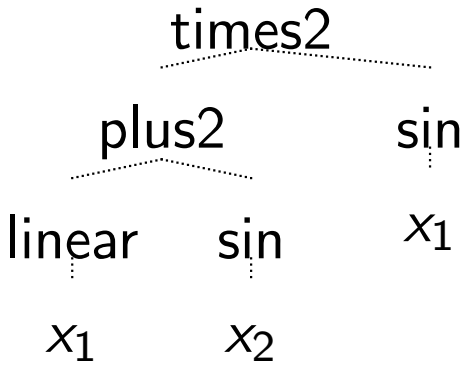
$$\text{times2}(\emptyset,\text{times2}(\emptyset,\text{plus2}(\emptyset,\text{parabola}(w_{1:3},x_2),x_2),\sin(\emptyset,x_1)),\sin(\emptyset,x_1))$$

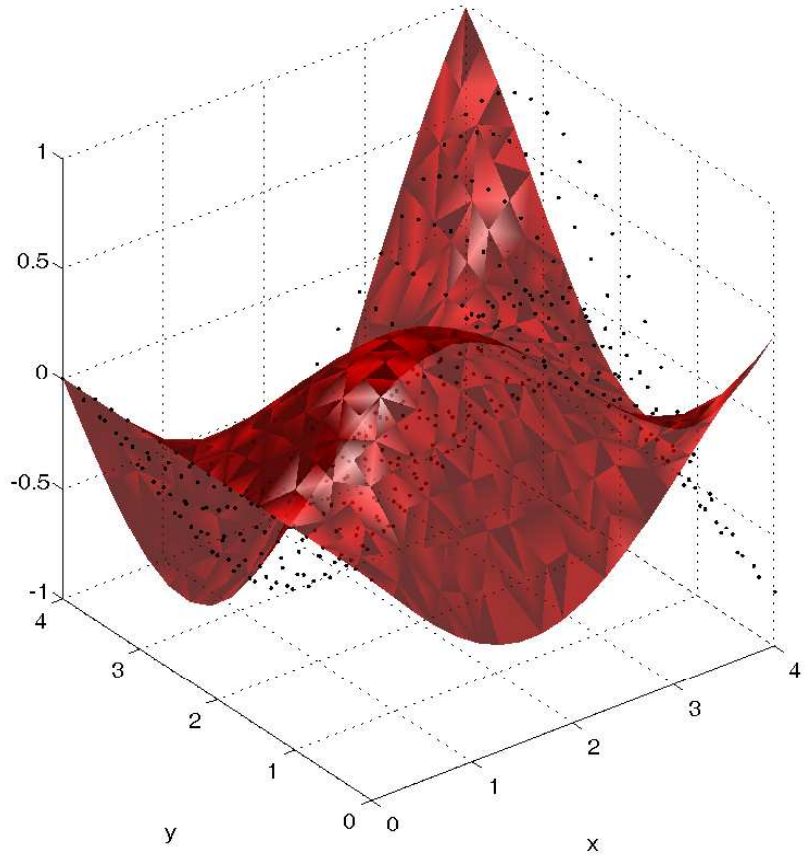




Competitive models

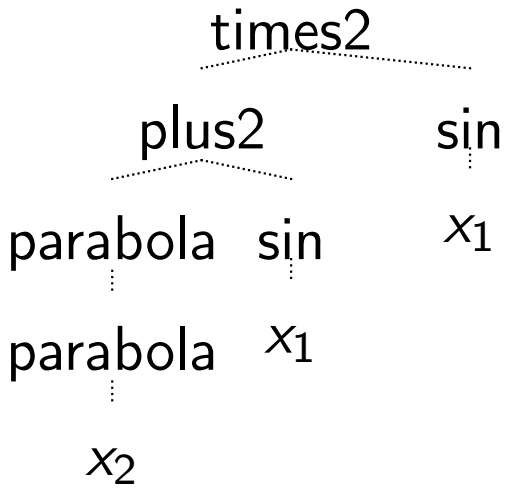
$$\text{times2}(\emptyset,\text{plus2}(\emptyset,\text{linear}(w_{1:2},x_1),\sin(\emptyset,x_2)),\sin(\emptyset,x_1))$$

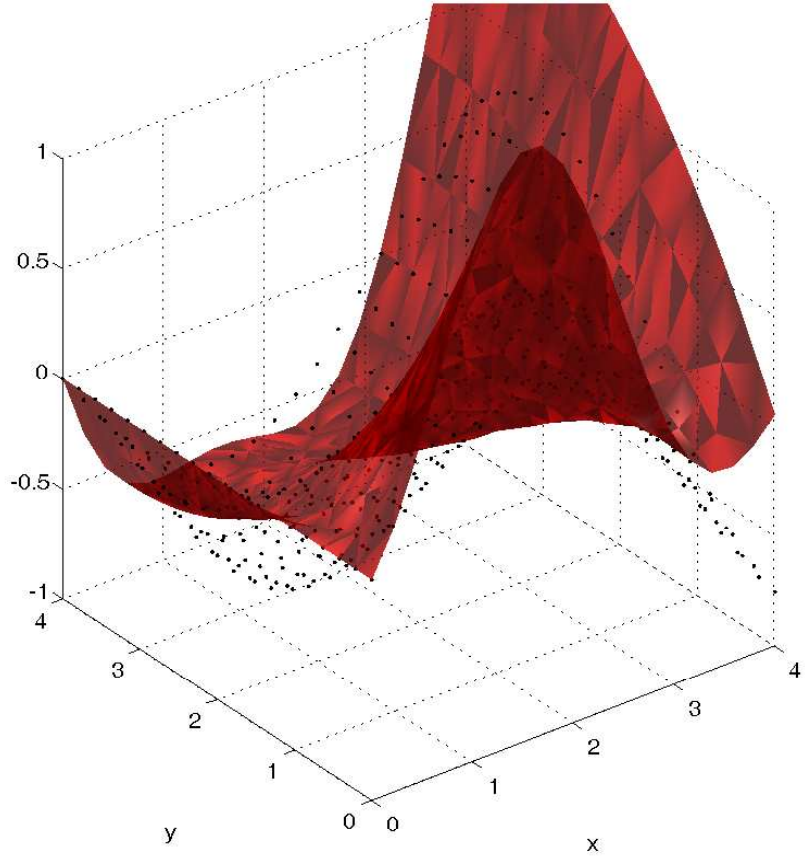




Competitive models

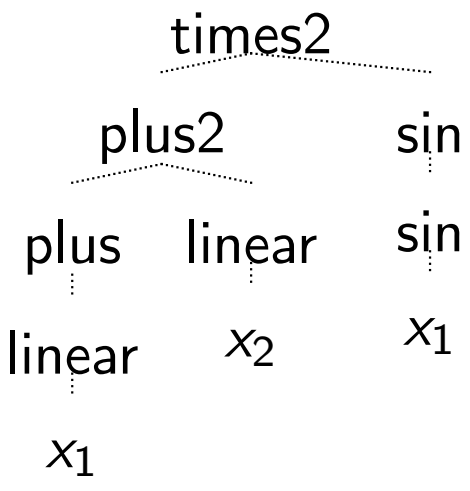
$$\text{times2}(\emptyset,\text{plus2}(\emptyset,\text{parabola}(w_{1:3},\text{parabola}(w_{4:6},x_2)),\sin(\emptyset,x_1)),\sin(\emptyset,x_1))$$

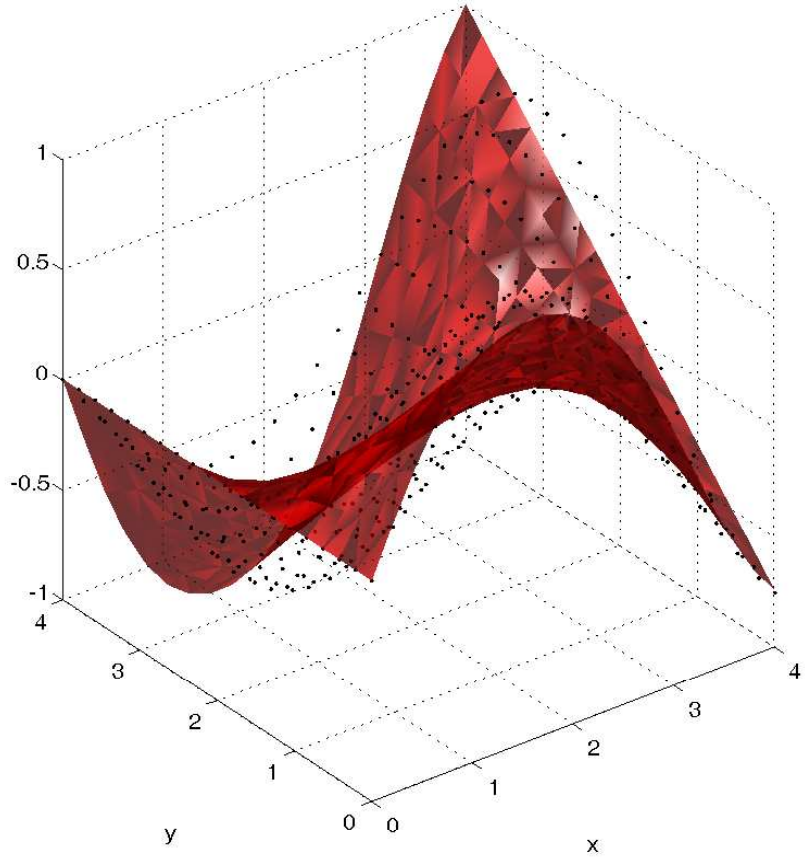




Competitive models

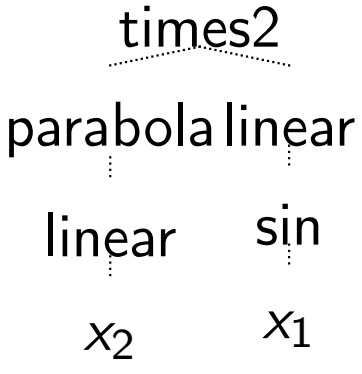
$$\text{times2}(\emptyset,\text{plus2}(\emptyset,\text{plus}(w_1,\text{linear}(w_{2:3},x_1)),\text{linear}(w_{4:5},x_2)),\sin(\emptyset,\sin(\emptyset,x_1)))$$

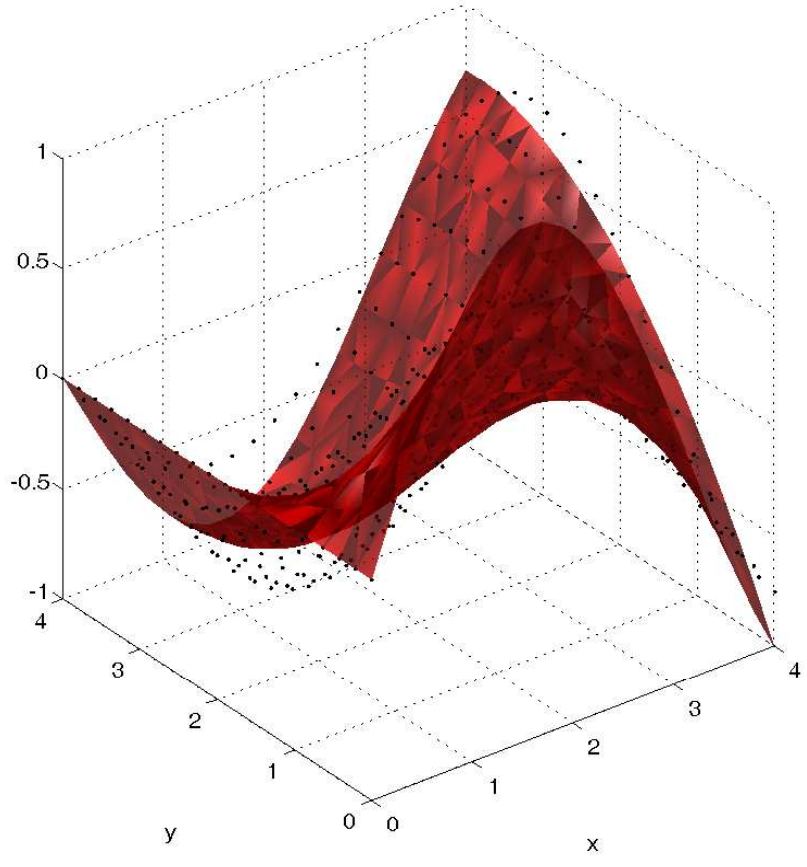




Competitive models

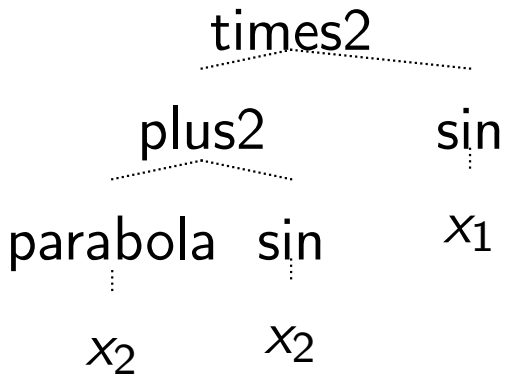
$$\text{times2}(\emptyset,\text{parabola}(w_{1:3},\text{linear}(w_{4:5},x_2)),\text{linear}(w_{6:7},\sin(\emptyset,x_1)))$$

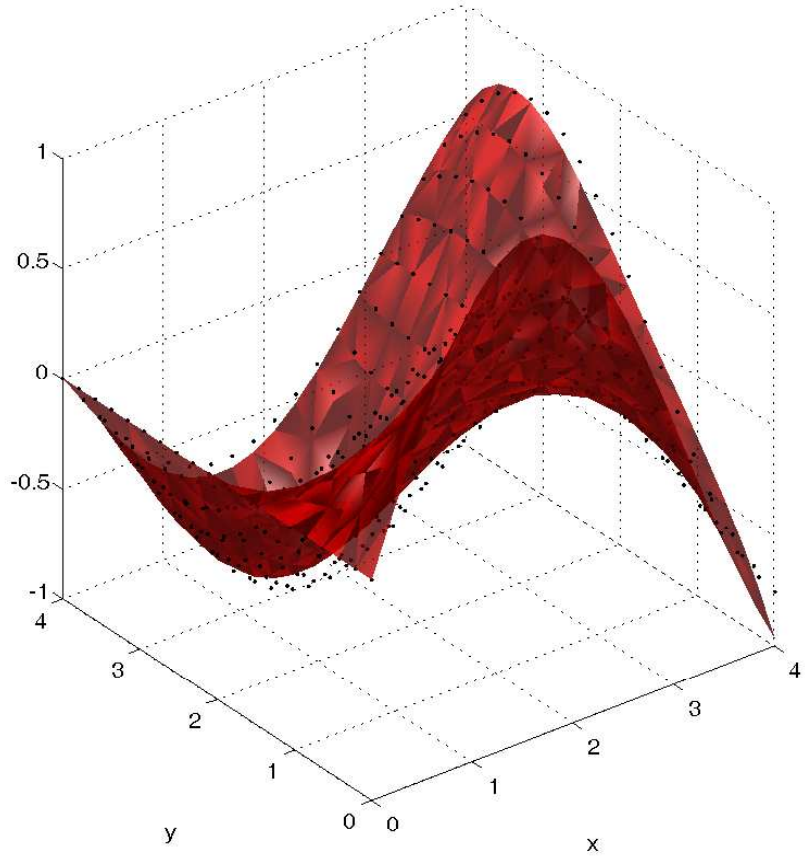




Competitive models

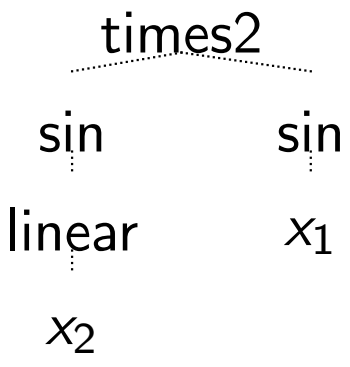
$$\text{times2}(\emptyset,\text{plus2}(\emptyset,\text{parabola}(w_{1:3},x_2),\sin(\emptyset,x_2)),\sin(\emptyset,x_1))$$

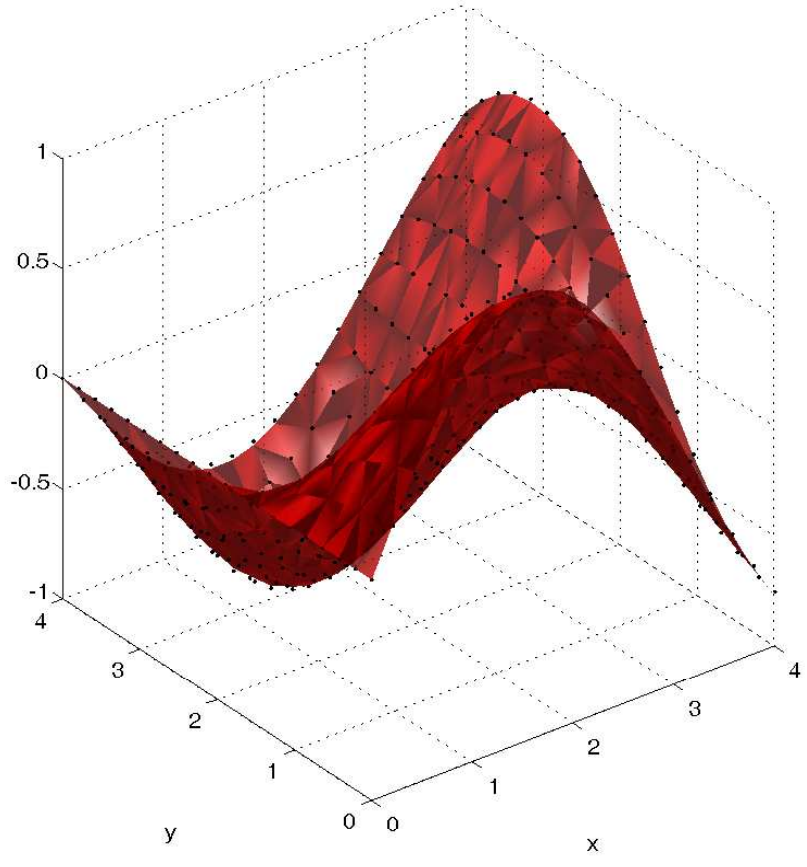




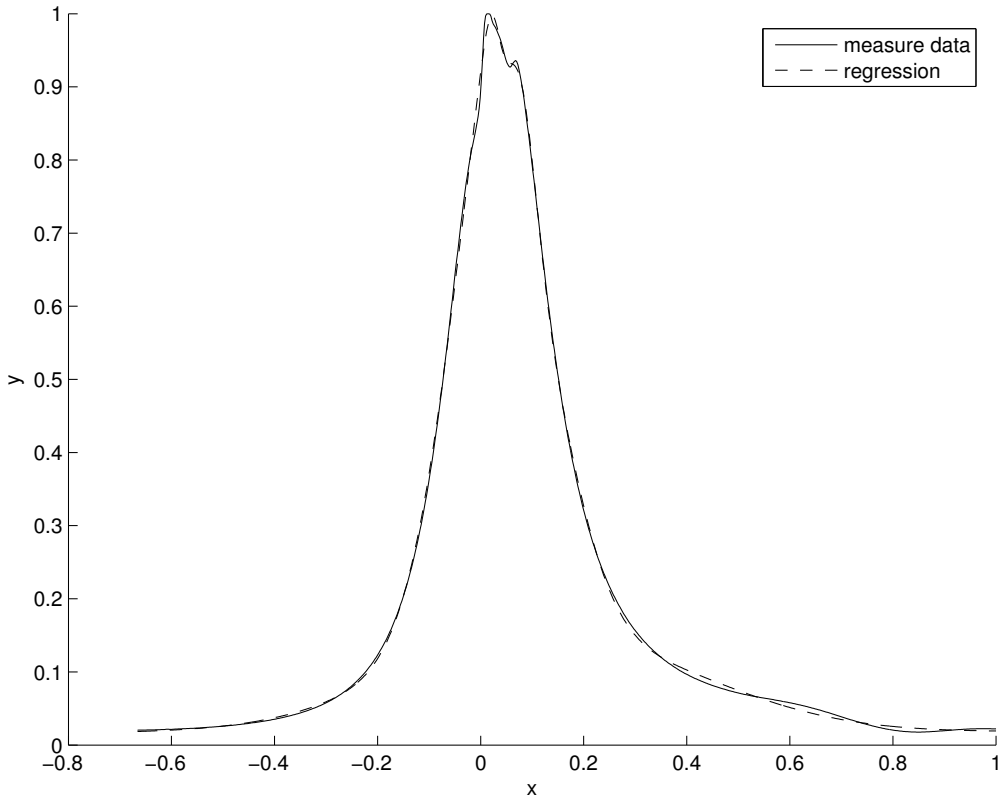
Competitive models

$\text{times2}(\emptyset,\sin(\emptyset,\text{linear}(w_{1:2},x_2)),\sin(\emptyset,x_1))$





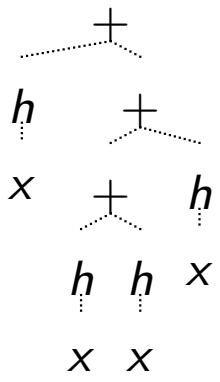
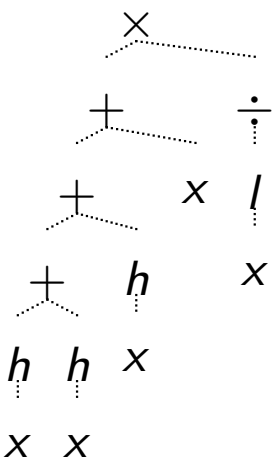
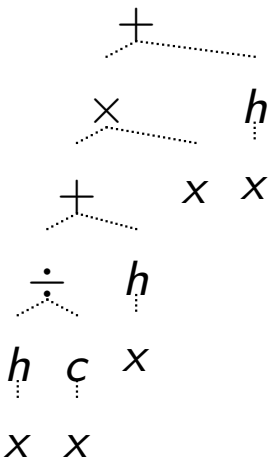
Practical example: automotive



The pressure in the combusting camera of the diesel engine:

x — crankshaft rotation angle, normalized,
y — pressure, normalized,
the data set contain 4000 samples.

The selected models

Model 1	Model 2	Model 3
		

Legend: h — gaussian $y = \lambda(2\pi\sigma^{-1/2})\exp(-(x-\xi)^2(2\sigma^{-2}) + a)$,
c — cubic $y = ax^3 + bx^2 + cx + d$, l — linear $y = ax + b$.

$$f_2 = g_1(g_2(g_3(g_4(g_5(x), g_6(x)), g_7(x)), x), g_8(x)).$$

The full representation of the Model 2 is

$$y = (ax+b)^{-1}\left(x + \sum_{i=1}^{3} \frac{\lambda_i}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(x-\xi_i)^2}{2\sigma_i^2}\right) + a_i\right).$$

Conclusion

- Hyperparameters depend on the variance of model parameters.
- If the variance is large the model parameter and corresponded element could be eliminated.

Outline

- Algorithms of inductive model generation use expert-defined set of primitives, specially designed for an application.
- Experts could explain obtained models in terms of the application.
- Initial expert models could be advanced by the model generation algorithms.