Methods of Model Selection and Dimensionality Reduction

Vadim STRIJOV

Computing Center of the
Russian Academy of Sciences

Ankara, October 06, 2009
Institute of Applied Mathematics, METU

Dimensionality reduction and index construction problem

- There is a set of objects, i.e. power plants:
  - Beckjord
  - East Bend
  - Miami Fort
  - Zimmer
- The **index** is a measure of an object's quality.
  It is a scalar, corresponded to an object.
- Expert estimation of an object's quality
  could be an index, too.

Examples

Index name	Objects	Features	Model
TOEFL	Students	Tests	Sum of scores
Eurovision	Singers	Televotes, Jury votes	Linear (weighted sum)
S&P500, NASDAQ	Time ticks	Shares (prices, volumes)	Non-linear
Bank ratings	Banks	Requirements	By an expert commission
Kyoto-index	**Power plants**	**Greenhouse gases**	**Linear**

There are lots of ways to construct indices. However, when algorithms are chosen and some results obtained, the following question arises:

- **How to show adequacy of the calculated indices?**

To answer the question analysts invite experts. The experts express their opinion and then the second question arises:

- **How to show that expert estimations are valid?**

How to construct an index?

- Assign a comparison criterion.
- Gather a set of comparable objects.
- Gather features of the objects.
- Make a data table: objects/features, i.e.

#	Plant Name	Plant Type	**Total Net Generation**	**CO_2 emission**	**NO_X emission**	**SO_X emission**	**Population density**
			10^6 KWHours	Shorttons per month	Shorttons per month	Shorttons per month	Qty per sq.mile
1	**Beckjord**	Coal	458505	191	16	45	23
3	**East Bend**	Coal	356124	147	16	43	34
4	**Miami Fort**	Coal	484590	204	6	23	45
5	**Dark Creek**	Coal	818435	329	5	64	34
		Optimal value	**max**	**min**	**min**	**min**	**min**

The criterion could be: **Ecological footprint of a plant**

Notations

$A=\{a_{ij}\}$ – $(n \times m)$ real matrix, **data set**,

$\mathbf{q}=[q_1, \ldots, q_m]^T$ – vector of **object indices**,

$\mathbf{w}=[w_1, \ldots, w_n]^T$ – vector of **feature importance weights**,

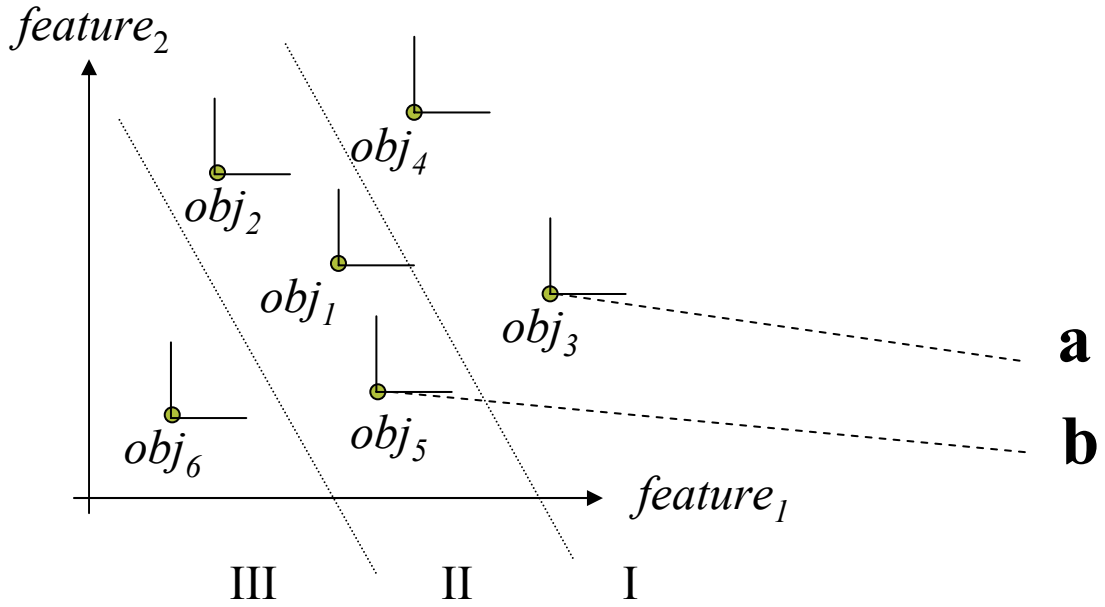
$\mathbf{q}_0$, $\mathbf{w}_0$ – **expert estimations of indices and weights**.

Usually, data prepared so that

1. the minimum of each feature equals 0, while the maximum equals 1;
2. the bigger value of each implies better quality of the index.

The first method, Pareto slicing

An easiest method to obtain indices in ordinal scales is to find non-dominated objects at each slicing level.



The object **a** is non-dominated if there is no $\mathbf{b}_i$ such that $b_{ij} \geq a_i$ for all features j.

Supervised way-1,

the Weighted sum

$$\mathbf{q}_1 = A\ \mathbf{w}_0.$$

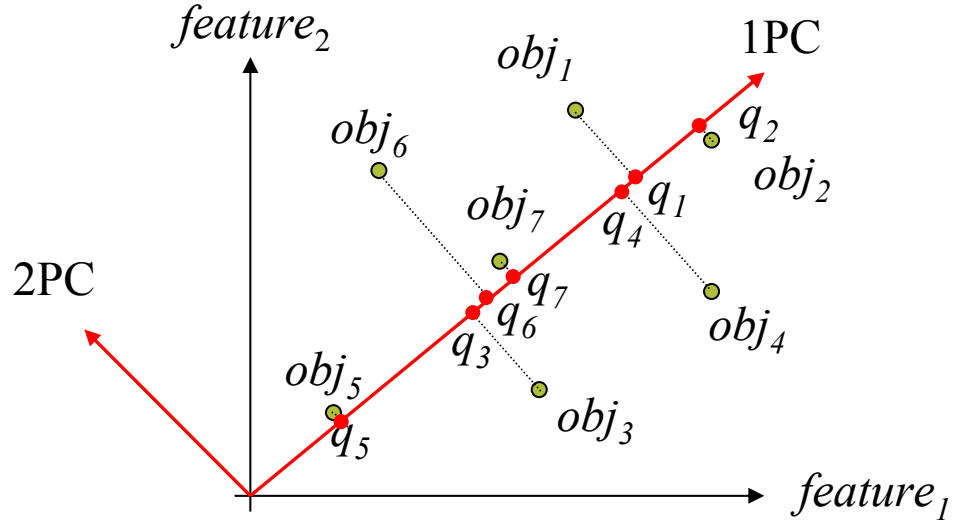
	w_1	…	w_n
q_1	a_{11}	…	a_{1n}
…	…	…	…
q_m	a_{m1}	…	a_{mn}

Unsupervised way,

Principal Components Analysis

$Q=AW$, where W—rotation matrix of the principal components.

$\mathbf{q}_{PCA}=A\mathbf{w}_{1PC}$, where $\mathbf{w}_{1PC}$ is the 1st column vector of W.



PCA gives minimal mean square error between objects and their projections.

Unsupervised way,

useful tool for PCA

$$A = ULW^T$$

$$A^T A = WLU^T ULW^T$$

$$A^T AW = WL^2$$

Supervised way-2,

the Expert-Statistical Technique

$$\mathbf{w}_1 = \arg\min \|\mathbf{q}_0 - A\,\mathbf{w}\|^2,$$

least squares, $\mathbf{w}_1 = (A^T A)^{-1} A^T \mathbf{q}_0.$

The problem of specification

- **We have**

the data table A,

expert estimations $\mathbf{q}_0, \mathbf{w}_0$,

calculated weights and indices $\mathbf{q}_1, \mathbf{w}_1$.

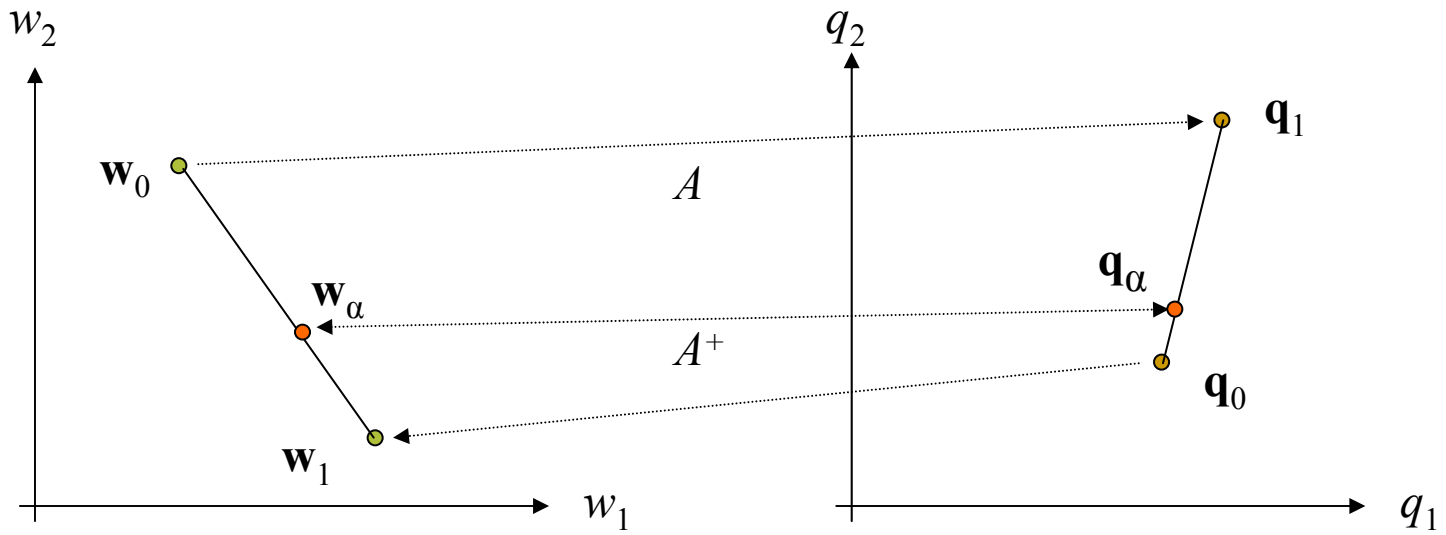
- **Contradiction**

Calculated indices are not the same as the expert estimations for the indices;

as well, calculated weights are not the same as the expert estimations of the weights:

in general,

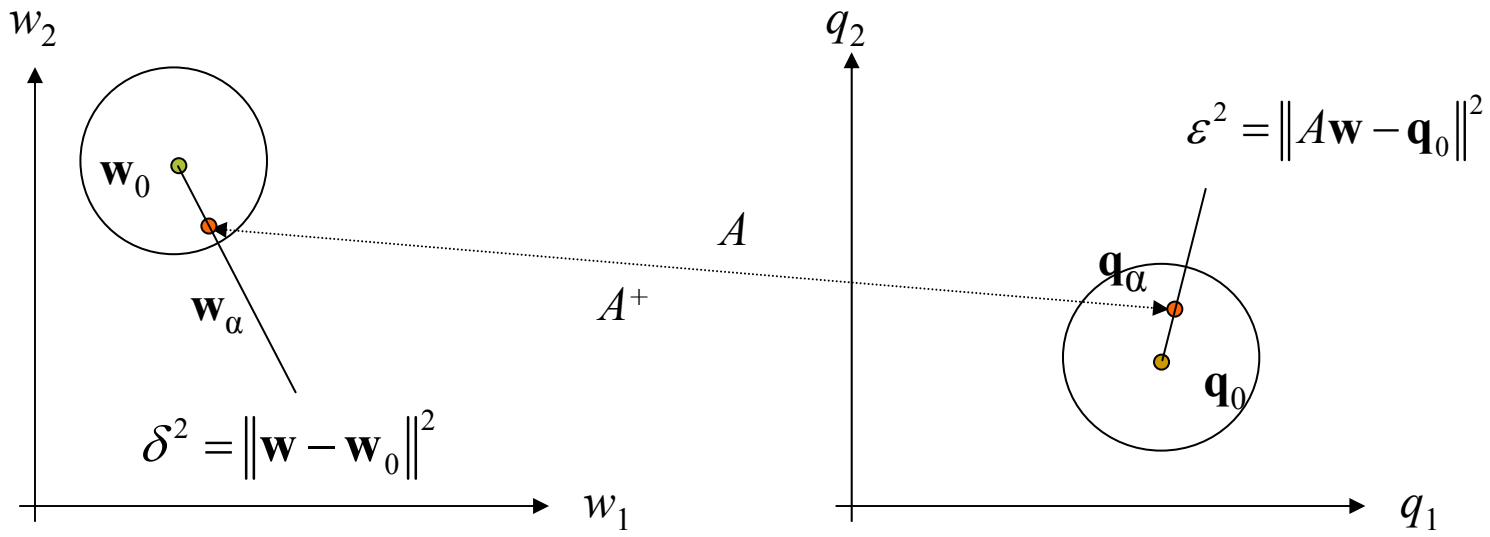
Linear specification



$$\mathbf{w}_\alpha = \alpha A^+ \mathbf{q}_0 + (\mathbf{1}\text{-}\alpha)\mathbf{w}_0, \qquad \mathbf{q}_\alpha = (1\text{-}\alpha) A \mathbf{w}_0 + \alpha \mathbf{q}_0.$$

Parameter α is in [0,1].
α = 0, we trust expert estimations of the weights,
α = 1, we trust expert estimations of the indices.

Quadratic specification

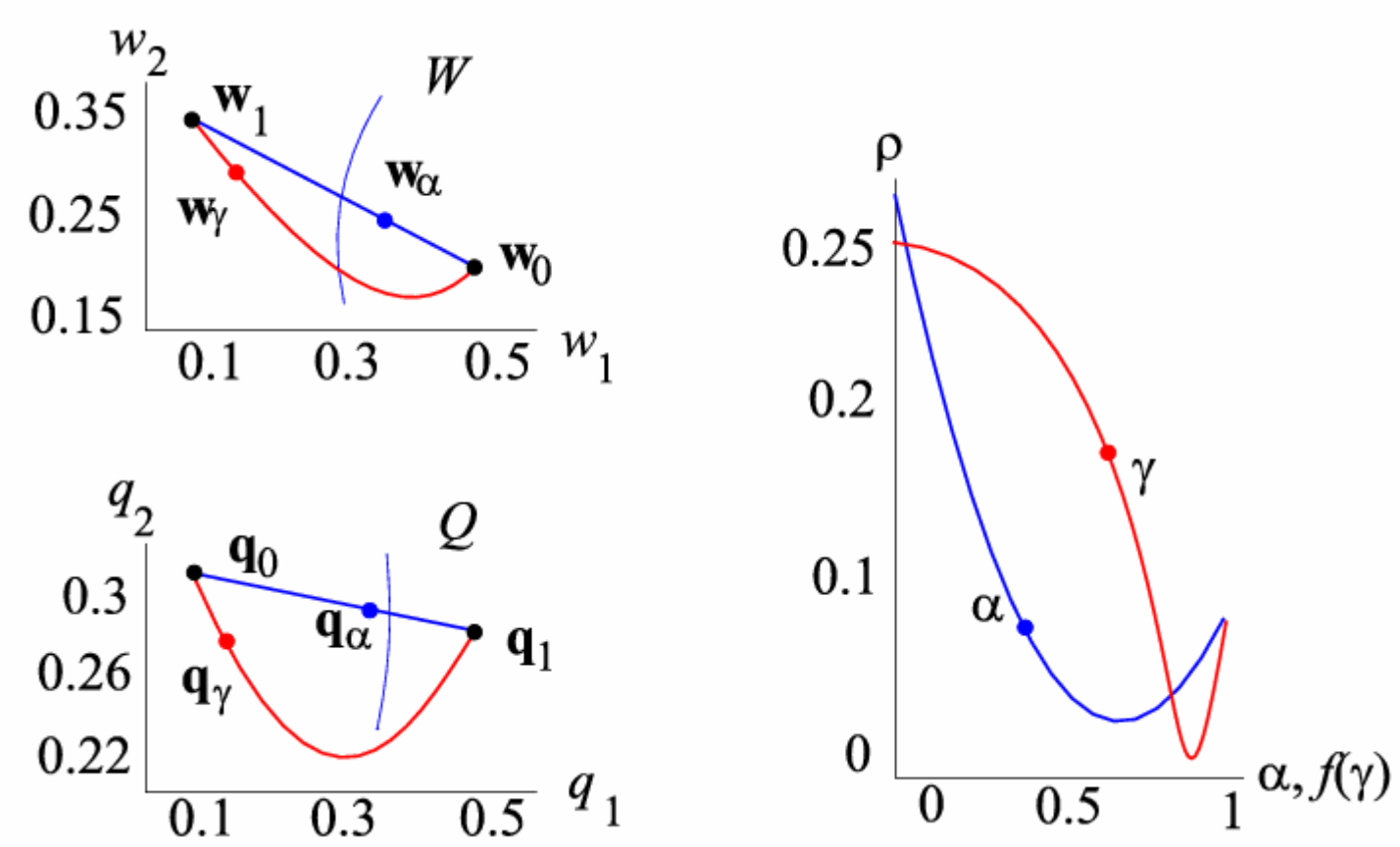


$$\mathbf{w}_\gamma = \arg\min_{\mathbf{w}\in W}(\varepsilon^2 - \gamma^2\delta^2), \quad \mathbf{w}_\gamma = (A^T A + \gamma^2 I)^{-1}(A^T \mathbf{q}_0 + \gamma^2 \mathbf{w}_0).$$

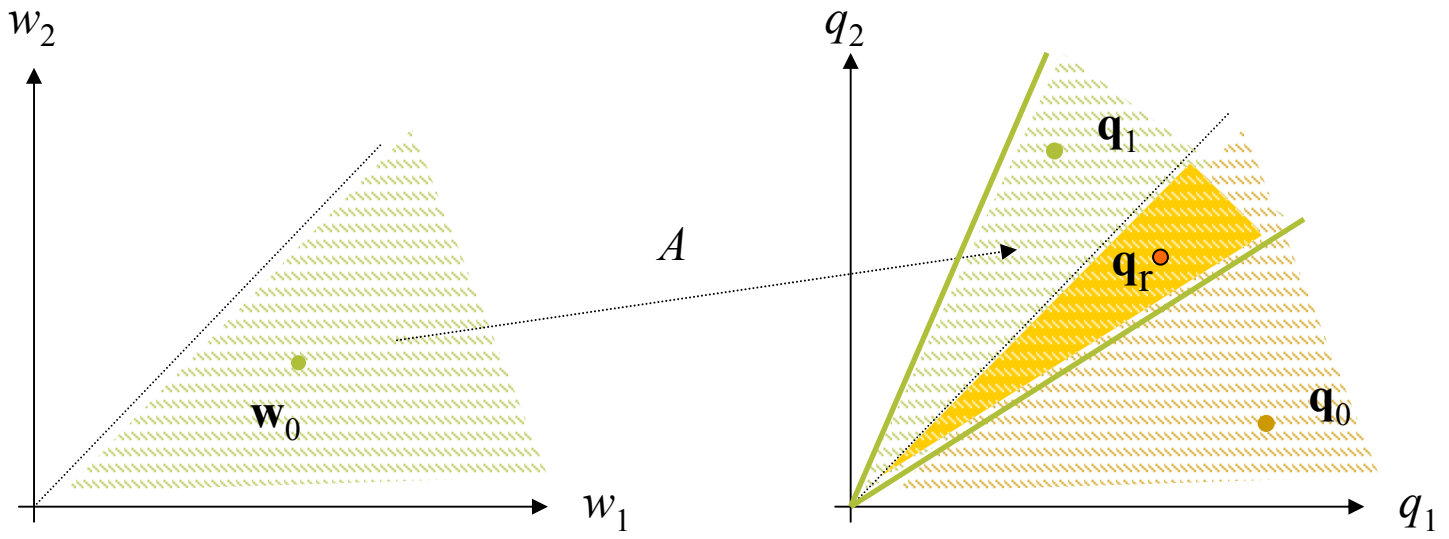
If parameter γ^2 is 0, then we trust expert estimations of the indices.

Comparison of the methods,

what is the difference?



Ordinal specification



$$\mathbf{w}_0 = [w_1 \geq w_2 \geq \ldots \geq w_n \geq 0]^T, \mathbf{q}_0 = [q_1 \geq q_2 \geq \ldots \geq q_m \geq 0]^T.$$

Rank-scaled expert estimations

$$\mathbf{w}_0 = [w_1 \geq w_2 \geq \ldots \geq w_n \geq 0]^T, \mathbf{q}_0 = [q_1 \geq q_2 \geq \ldots \geq q_m \geq 0]^T.$$

$$Q_0 = \{\mathbf{q}_0 \mid J_m \mathbf{q}_0 \geq \mathbf{0}\},$$
$$W_0 = \{\mathbf{w}_0 \mid J_n \mathbf{w}_0 \geq \mathbf{0}\}.$$

$$J = \begin{pmatrix} 1 & -1 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$

The cones intersection exists

$$\mathbf{q}_1 \in AW_0 \bigcap Q_0,$$

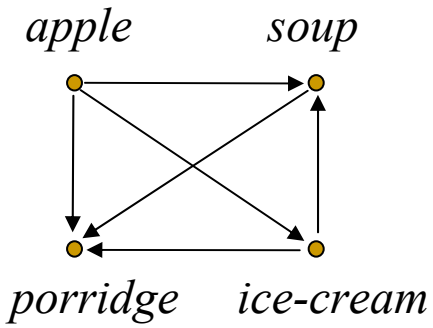
or not, then specify

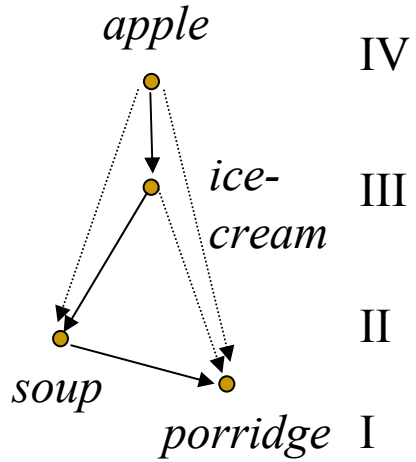
$$\mathbf{q}_\alpha = (1-\alpha)A\mathbf{w}' + \alpha\mathbf{q}', \quad \text{where}$$

$$\mathbf{w}', \mathbf{q}' = \arg \min_{\substack{\mathbf{w}\in W_0, \|\mathbf{w}\|^2=1 \\ \mathbf{q}\in Q_0, \|\mathbf{q}\|^2=1}} \left\| A\mathbf{w} - \mathbf{q} \right\|^2 .$$

Pair-wise comparison

	a	*s*	*p*	*i-c*
apple	○	+	+	+
soup		○	+	–
porridge			○	–
ice-cream				○





If an object in a row is better than the other one in a column then put "+", otherwise "-".

Make a graph, *row* + *column* means *row* ○⟶○ *column*.

Find the top and remove extra nodes.

The results of the specification are

- adequate indices,
- reasoned expert estimations.

We know why our expert valued each object
and what contribution each feature makes to the index.

Model selection for (generalized) linear models

Let there be given

1. Sample set:
$\{(\mathbf{x}_i, y_i)|i = 1, \ldots, \ell\}$, where $\mathbf{x}_i \in \mathbb{R}^P$, $y_i \in \mathbb{R}^1$, $P = |N|$
   - and $N \subset \mathbb{N}$.

s

2. Linear model:

$$y = f(\mathbf{w}, \mathbf{x}) + \varepsilon,$$
$$y = \langle \mathbf{w}, \mathbf{x} \rangle + \varepsilon.$$

3. Data generation hypothesis:
distribution of the random variable ε_i is in the exponential family.
4. Target function:
minimum of the residual vector norm

$$SSE = \sum^{\ell} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 \to \min .$$

One must to

find a subset $\mathcal{A} \subset N$ of the indices $\hat{\mathbf{x}} = \{x_i^j | j \in \mathcal{A}\}$, such that the model $f(\mathbf{w}, \hat{\mathbf{x}})$ brings the optimum to the given criterion.

For example to the Colin Mallows' C_P:

$$C_P = \frac{SSE_P}{RMS} - \ell + 2P,$$

where

$$RMS = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - f(\mathbf{w}, \mathbf{x}_i))^2.$$

Or to another criterion from the following list.

Criterions for model selection

1. Information criterions
   - Akaike Information Criterion, $\mathrm{AIC} = 2P - 2\ln(S)$
   - Bayesian Information Criterion, $\mathrm{BIC} = P\ln(\ell) - 2\ln(S)$
2. Cross-validation
   - Retrospective Forecasting
   - Leave One Out
   - Random Split, etc.
3. Model Comparison
   - Bayesian Comparison
   - Minimum Description Length

Superposition construction

Let there be given

- $\Xi = \{\xi^u\}_{u=1}^{U}$ — set of measured (nongenerated) independent variables,
- $G = \mathrm{id} \cup \{g_v\}_{v=2}^{V}$ — finite set of primitive functions.

Consider Cartesian product $G \times \Xi$. An element (g_v, ξ^u) corresponds to the superposition $g_v(\xi^u)$ and defined by indices v, u.
Denote $s_\iota = g_v(\xi^u)$, where the index $\iota = (v-1)U + u$.

Consider $S \times S \times \ldots \times S$ — Cartesian product $\mathcal{N}$ of the sets $S = \{s_\iota\}$. Each element of $\mathcal{N}$ bijectively corresponds to the superposition $a_i = s_\iota^1 \circ s_\iota^2 \circ \ldots \circ s_\iota^{\mathcal{N}}$.

Kolmogorov-Gabor polynomial

The basic model of the feature generation is

$$y = w_0 + \sum_{i=1}^{UV} w_i a_i + \sum_{i=1}^{UV}\sum_{j=1}^{UV} w_{ij} a_i a_j + \cdots + \sum_{i=1}^{UV} \ldots \sum_{z=1}^{UV} w_{i\ldots z} a_i \ldots a_z,$$

where the coefficients

$$\mathbf{w} = (w_0, w_i, w_{ij}, \ldots, w_{i\ldots z})_{i,j,\ldots,z=1,\ldots,UV}.$$

Represent this series as

$$y = \sum_{j \in N} w_j x^j.$$

The variables $\{x^j\}$ bijectively correspond to monomials of the polynomial.

The model selection algorithms

Exhaustive search and modifications

1. Exhaustive search of 2^P models
2. Method of group data handling, $K \cdot C_P^2$ models
3. Genetic algorithms
4. Add (append a feature), $P(P-1)/2$ models
5. Del (eliminate a feature)
6. Add-del or stepwise regression, $\sim P^2$ models

Parameter space analysis

1. Least angle regression (LARS), Lasso
2. Optimal brain surgery

Exhaustive search algorithm

The basic linear model includes all independent variables

$$y = w_0 + \alpha_1 w_1 x_1 + \alpha_2 w_2 x_2 + \ldots + \alpha_R w_P x_P.$$

The hyperparameter $\alpha \in \{0, 1\}$ is included for the model. The exhaustive search

α_1	α_2	$\ldots$	α_P
1	0	$\ldots$	0
0	1	$\ldots$	0
$\ldots$	$\ldots$	$\ldots$	$\ldots$
1	1	$\ldots$	1

.

Add (append a feature)

Step 0.
The active set $\mathcal{A}_0 = \emptyset$, and N is the set if feature indices, $P = |N|$.
Step $k = 1, \ldots, P$.
Select the next best feature index

$$\hat{j} = \arg \min_{j \in P \setminus \mathcal{A}_k} \min_{\mathbf{w} \in \mathbb{W}_k} \|(X_{\mathcal{A}_k} \vdots \mathbf{x}_j)\mathbf{w} - \mathbf{y}\|_2^2,$$

then

$$\mathcal{A}_{k+1} = \mathcal{A}_k \cup \hat{j}.$$

Assume the following

The column vectors

$$\mathbf{x}^j = \{x_i^j | i \in 1, \ldots, \ell\} \quad \text{and} \quad \mathbf{y} = \{y_i | i \in 1, \ldots, \ell\}.$$

The model

$$\mathbf{y} = w_1 \mathbf{x}^j + \ldots + w_P \mathbf{x}^P + \varepsilon,$$

in the other words,

$$\mathbf{y} = X\mathbf{w} + \varepsilon.$$

Assume for all $j \in N$

$$\|\mathbf{x}^j\|_1 = 0, \quad \|\mathbf{x}^j\|_2 = 1 \quad \text{and} \quad \|\mathbf{y}\|_1 = 0, \quad \|\mathbf{y}\|_2 = 1.$$

For all $j, k \in N$, $j \neq k$ the vectors $\mathbf{x}^j, \mathbf{x}^k$ are linear independent. Then the vector of correlation coefficients

$$\mathbf{c} = X^T \mathbf{y}.$$

Fast orthogonal search

Step 0.
The residuals $\varepsilon_0 = \mathbf{0}$, the active set $\mathcal{A}_0 = \emptyset$.

Step $k = 1, \ldots, P$.

$$\mathcal{A}_k = \mathcal{A}_{k-1} \cup \hat{j},$$

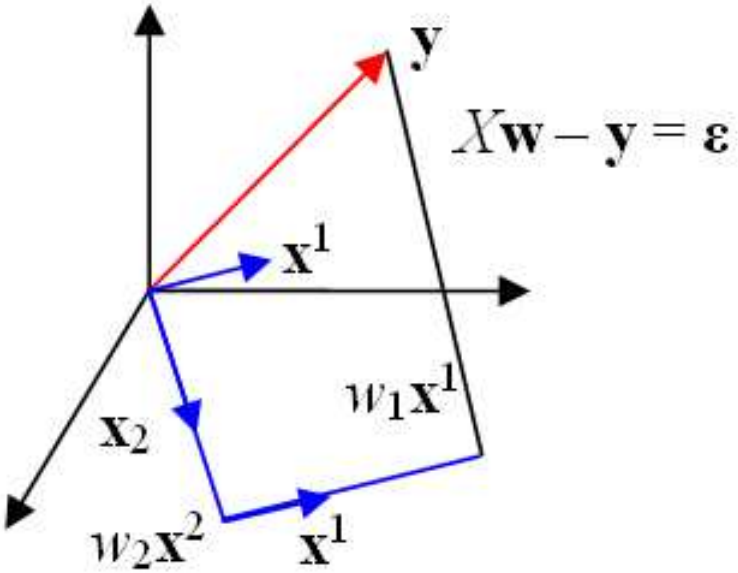
where $\hat{j}$ — feature, which has maximum correlation with ε_k:

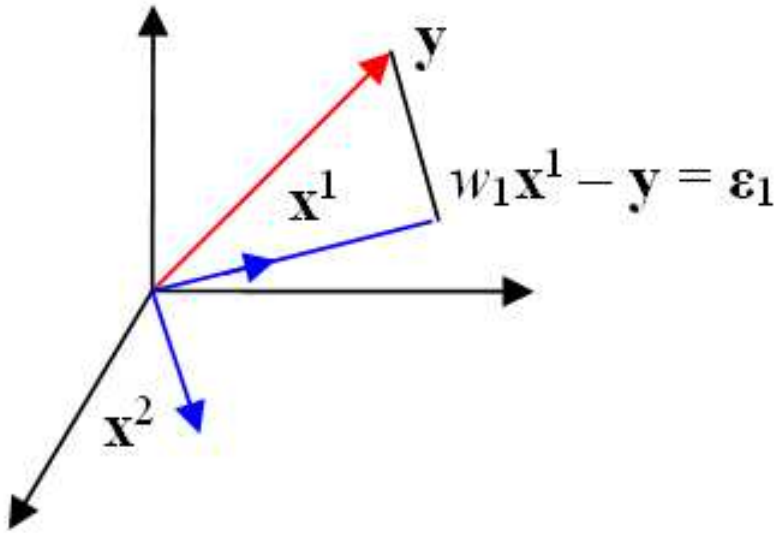
$$\hat{j} = \arg \max_{j \in \{N \setminus \mathcal{A}_k\}} \frac{\langle \mathbf{w}, \mathbf{x}^j \rangle}{\|\mathbf{x}\| \|\varepsilon_k\|},$$

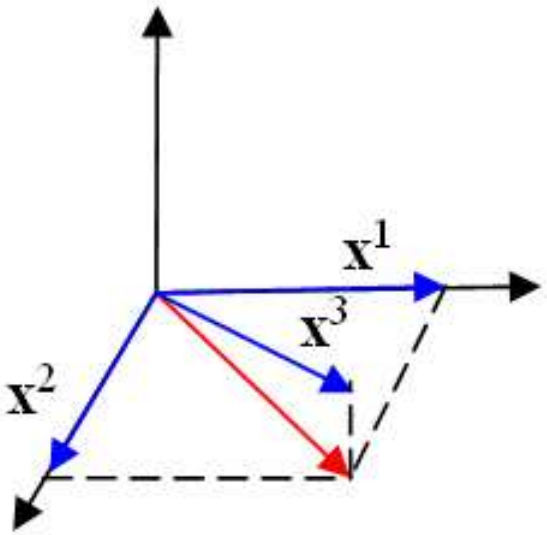
and

$$\varepsilon_k = X_{\mathcal{A}} \mathbf{w}_{\mathcal{A}} - \varepsilon_{k-1}.$$

Fast orthogonal search







Least angle regression, LARS

Denote $\mu = X\mathbf{w}$.

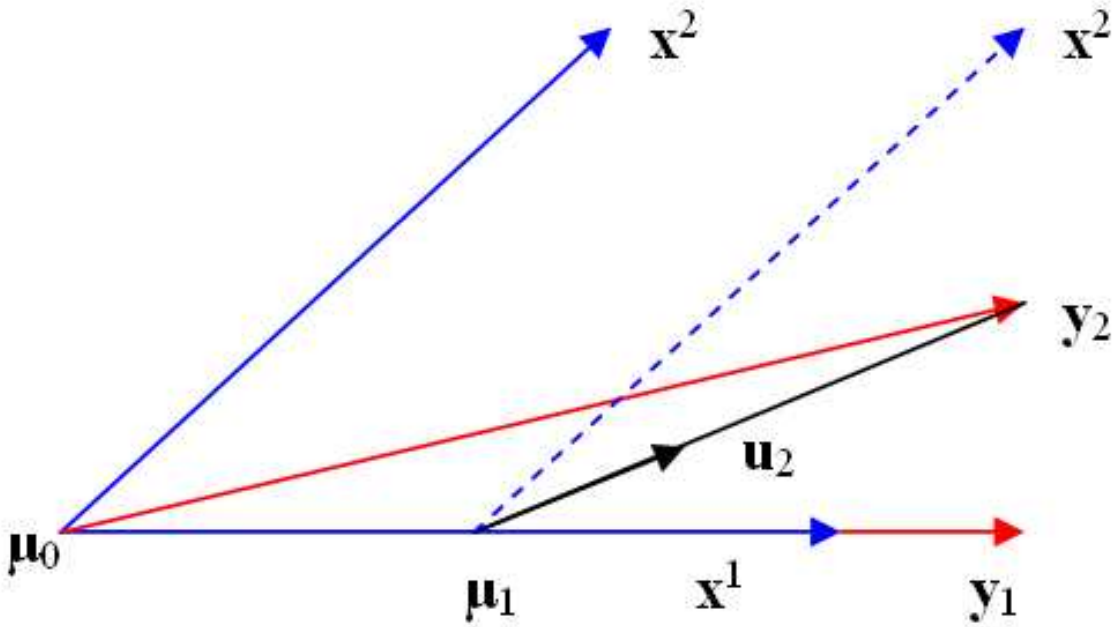
Step 0.
$\mu_0 = \mathbf{0}$, residual vector $\varepsilon_0 = \mathbf{y} - \mu_0$.

Step 1.
Let $\mathbf{y}$ has greater correlation with $\mathbf{x}^1$ than with $\mathbf{x}^2$. Then the new value of $\mu_1 = \mu_0 + w_1\mathbf{x}^1$, where w_1 is chosen so, that the vector $\mathbf{y}_2 - \mu$ — is a bisector for the vectors $\mathbf{x}^1, \mathbf{x}^2$.

Step 2.
For the unit bisector $\mathbf{u}_2$ calculate w_2:

$$\mu_2 = \mu_1 + w_2\mathbf{u}_2 = \mathbf{y}_2 \quad \text{for P=2.}$$

Least angle regression, LARS



Lasso

Minimize the error

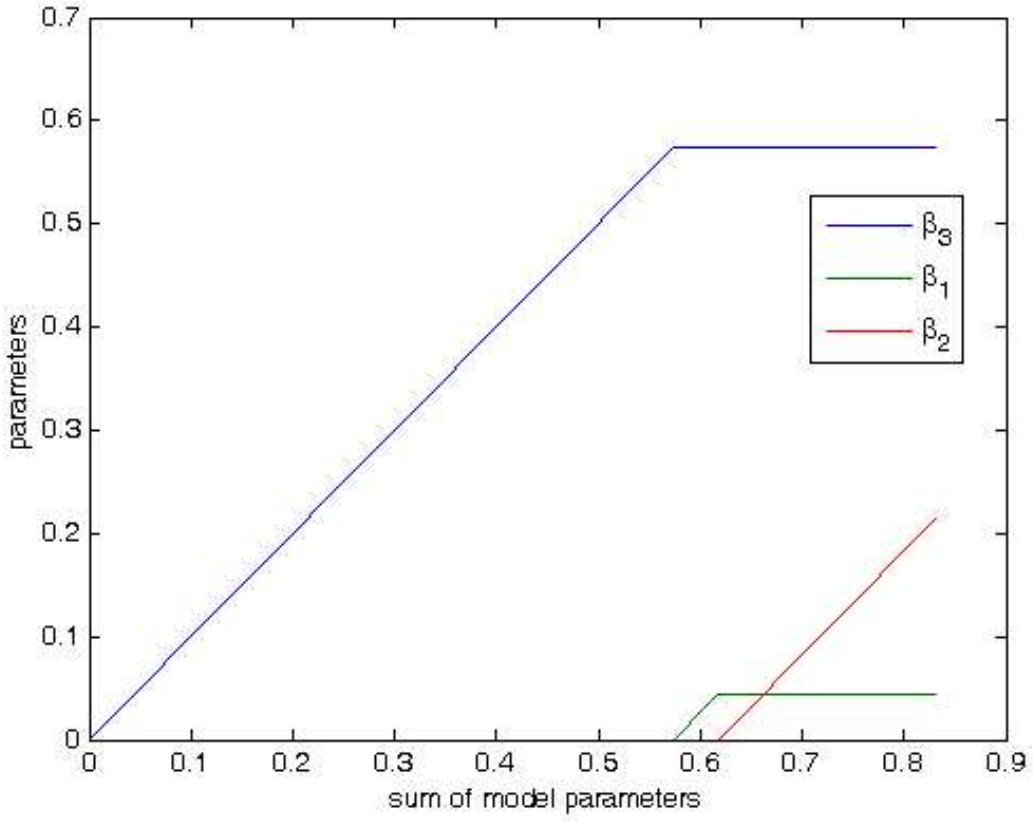
$$\|X\mathbf{w} - \mathbf{y}\|_2^2 \to \min,$$

subject to

$$\sum_{j \in N} \|w_j\|_1 \leqslant T.$$

Theorem (Efron et al., 2004).
Assuming the «one at time» condition, the LARS algorithm yields all Lasso solutions.

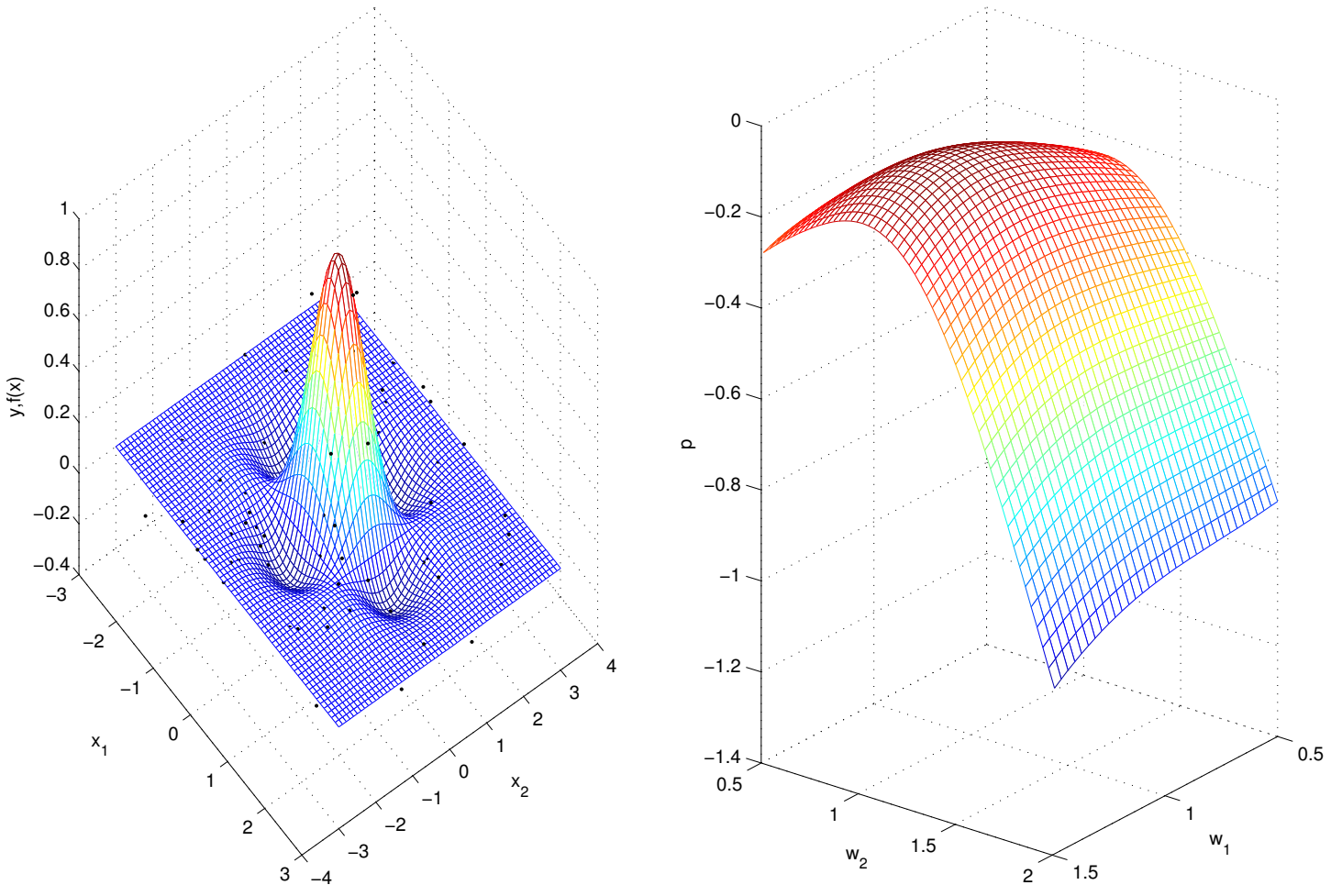
Lasso and LARS



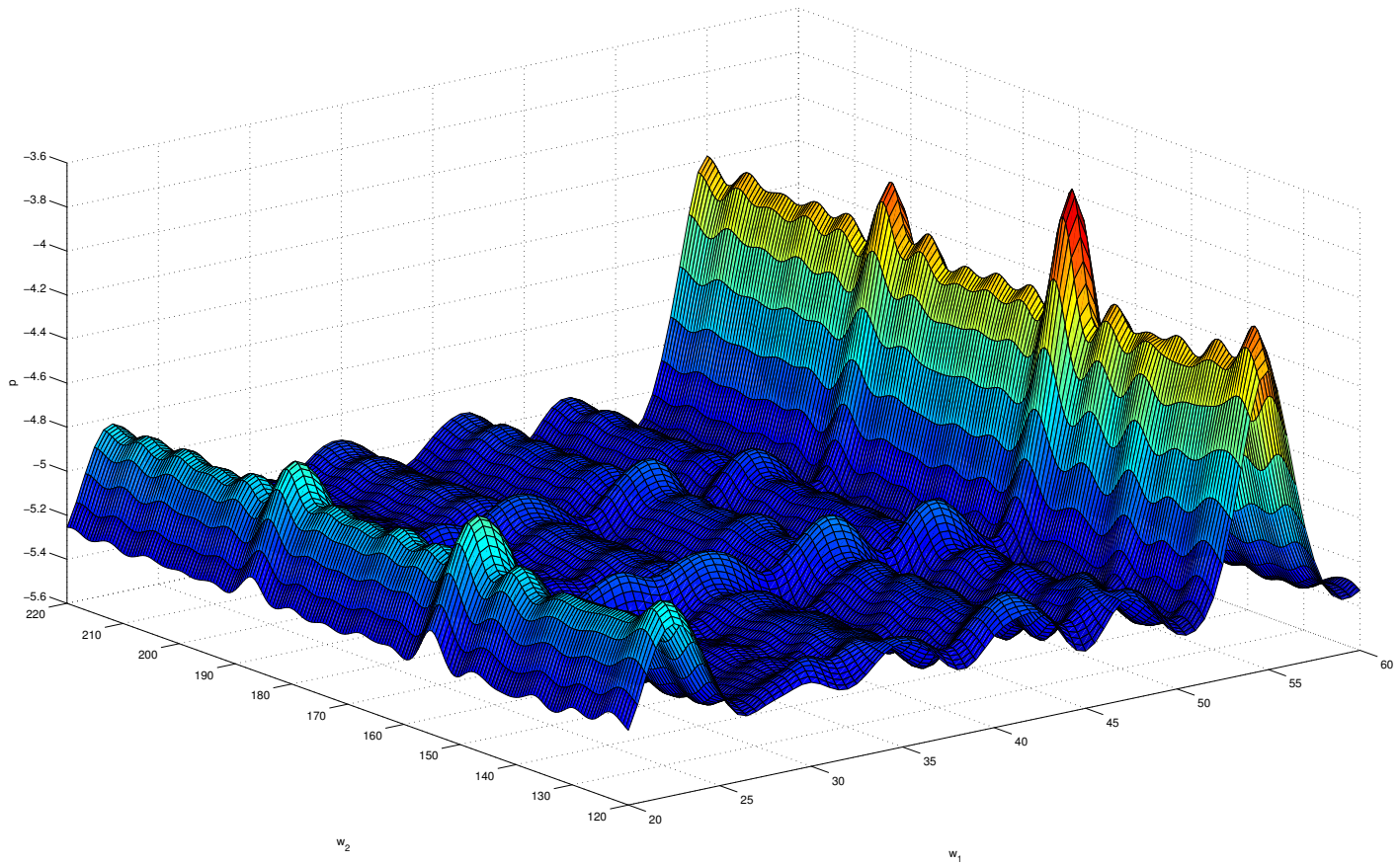
Optimal brain surgery

- Approximate $SSE = S(\mathbf{w})$:
  $S(\mathbf{w} + \Delta\mathbf{w}) = S(\mathbf{w}) + \mathbf{g}^T(\mathbf{w})\Delta\mathbf{w} + \frac{1}{2}\Delta\mathbf{w}^T H \Delta\mathbf{w} + o(\|\mathbf{w}\|^3)$.
- Elimination a feature is equivalent to $\mathbf{e}_i^T \Delta\mathbf{w} + w_i = 0$.
- Minimize the quadratic form $\Delta\mathbf{w}^T H \Delta\mathbf{w}$ subject to $\mathbf{e}_i^T + w_i = 0$, for all i.
- The index of the eliminated feature is $i = \arg\min_i(\min_{\Delta\mathbf{w}}(\Delta\mathbf{w}^T H \Delta\mathbf{w} | \mathbf{e}_i^T + w_i = 0))$.
- Introduce Lagrange function $S = \Delta\mathbf{w}^T H \Delta\mathbf{w} - \lambda(\mathbf{e}_i^T + w_i)$.
- For all i $\Delta\mathbf{w} = -\frac{w_i}{[H^{-1}]_{ii}} H^{-1}\mathbf{e}_i$.
- The salience of the target function is $L_i = \frac{w_i^2}{2[H^{-1}]_{ii}}$.

Optimal brain surgery



Optimal brain surgery



Ventia non sunt multiplicanda praeter necessitatem

William of Ockham
1285-1349

Occam's rasor: entities (model elements)

must not be multiplied beyond necessity