Specification of linear and rank-scaled expert estimations using measured data

Vadim Strijov

Computing Center of the Russian Academy of Sciences

EURO-2009, 5-8 July, Bonn

What is an index?

- There is a set of objects, i.e. power plants:
  - Beckjord
  - East Bend
  - Miami Fort
  - Zimmer
- Index a measure of an object's quality,
  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	Short tons per month	Short tons per month	Short tons 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.

Supervised way-1,

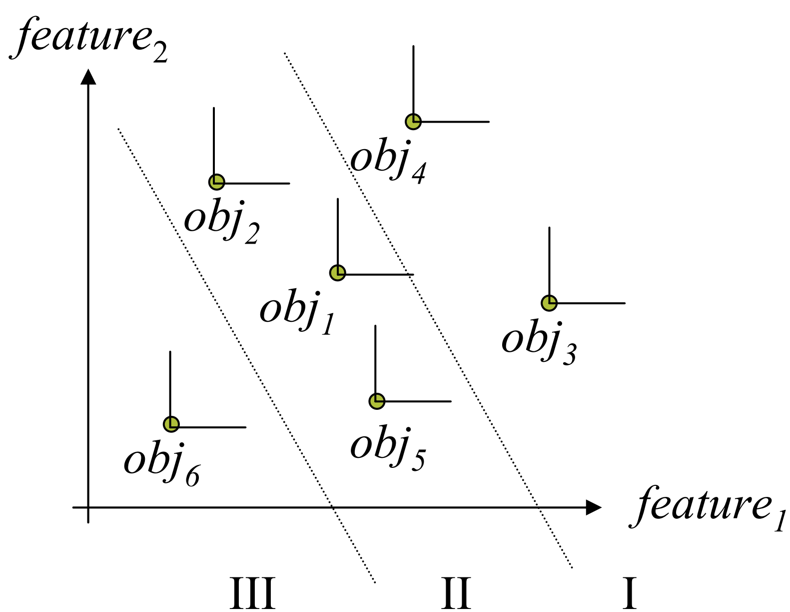
the Weighted sum

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

	w_1	$\dots$	w_{1n}
q_1	a_{11}	$\dots$	a_{1n}
$\dots$	$\dots$	$\dots$	$\dots$
q_m	a_{m1}	$\dots$	a_{mn}

The first method, Pareto slicing

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

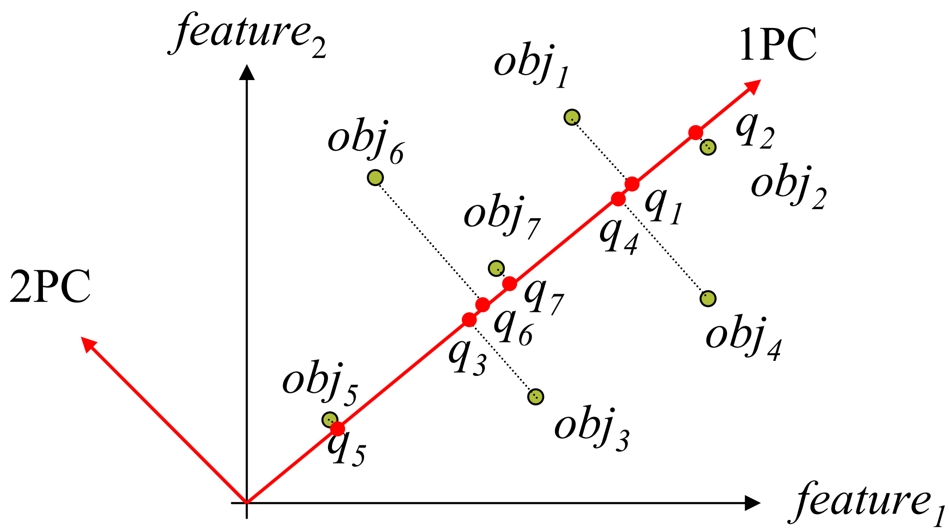


Unsupervised way,

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.

Supervised way-2,

the Expert-Statistical Technique

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

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

or

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

$A^+ = V\Lambda^{-1}U^T,$ where $A = U\Lambda V^T$ – singular values decomposition.

The problem of specification

- **We have**

the data table A,

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

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

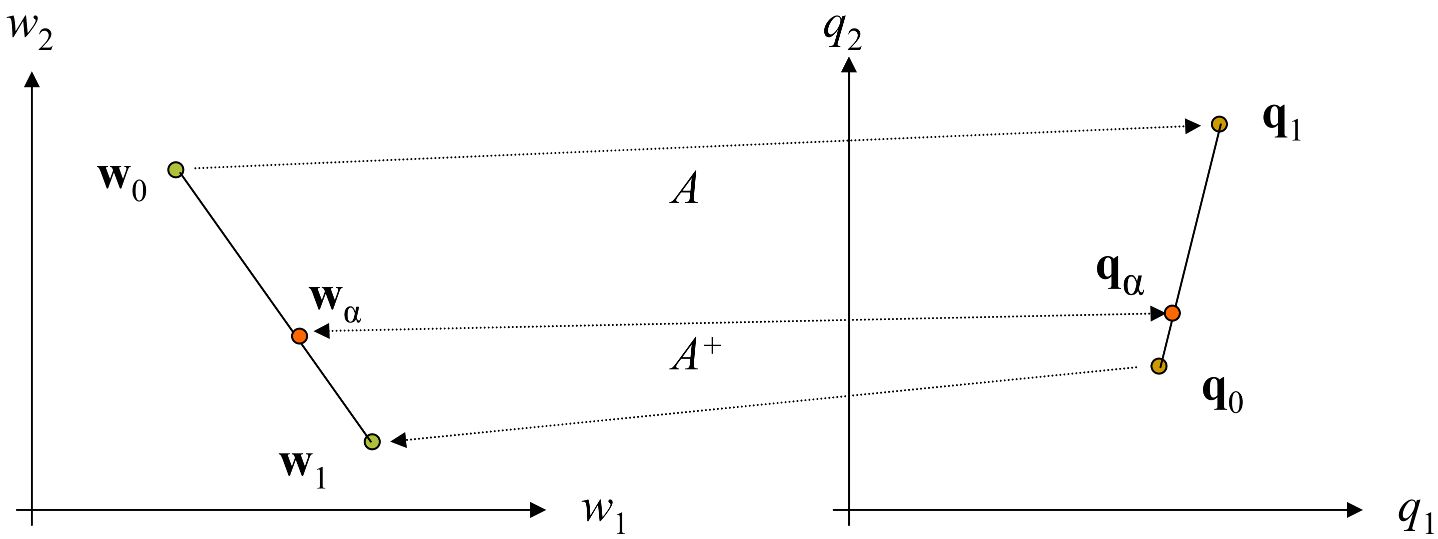
- **Contradiction**

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, neither $\mathbf{q}_0 \neq A\mathbf{w}_0$, nor $\mathbf{w}_0 \neq A^{+}\mathbf{q}_0$.

Linear specification



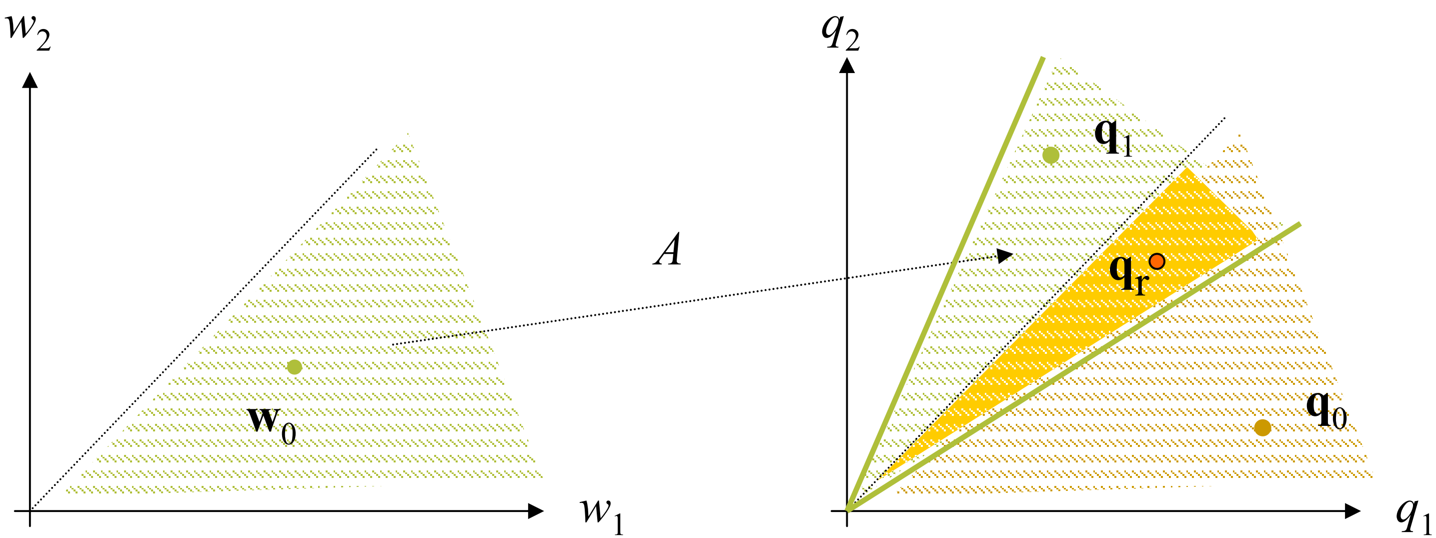
$$\mathbf{w}_\alpha = \alpha A^+ \mathbf{q}_0 + (\mathbf{1}\text{-}\alpha)\mathbf{w}_0, \qquad \mathbf{q}_\alpha = (1\text{-}\alpha)A\mathbf{w}_0 + \alpha\mathbf{q}_0.$$

Parameter α is in [0,1].

α = 0, we trust expert estimations of the weights,

α = 1, we trust expert estimations of the indices.

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

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

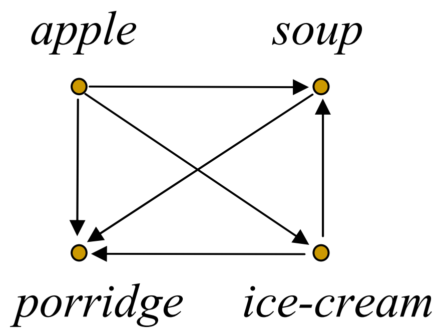
or not, then specify

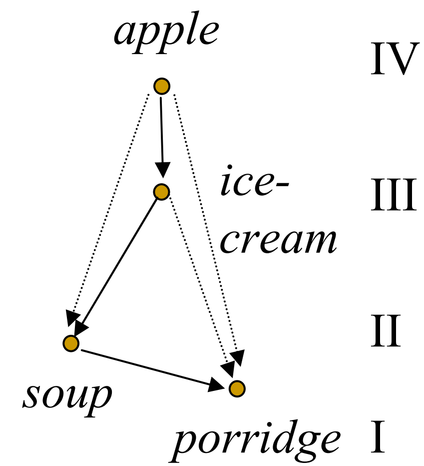
$$\mathbf{q}_\alpha = (1-\alpha)A\mathbf{w}' + \alpha\mathbf{q}', \quad \text{where}$$

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

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.

strijov@ccas.ru