Metaheuristic-algorithm-based technique to build cost model from non-complete database

Luca Visconti

ESA ESTEC

16/09/2022

Agenda

1. Parametric cost modelling general features

2. General introduction about missing Data:
    - 2.1 missing data pattern
    - 2.2 missing data mechanisms
    - 2.3 Methodologies and details on EM

3. Metaheuristic algorithm: Brief description of GA

4. Application of GA to EM method: EMGA

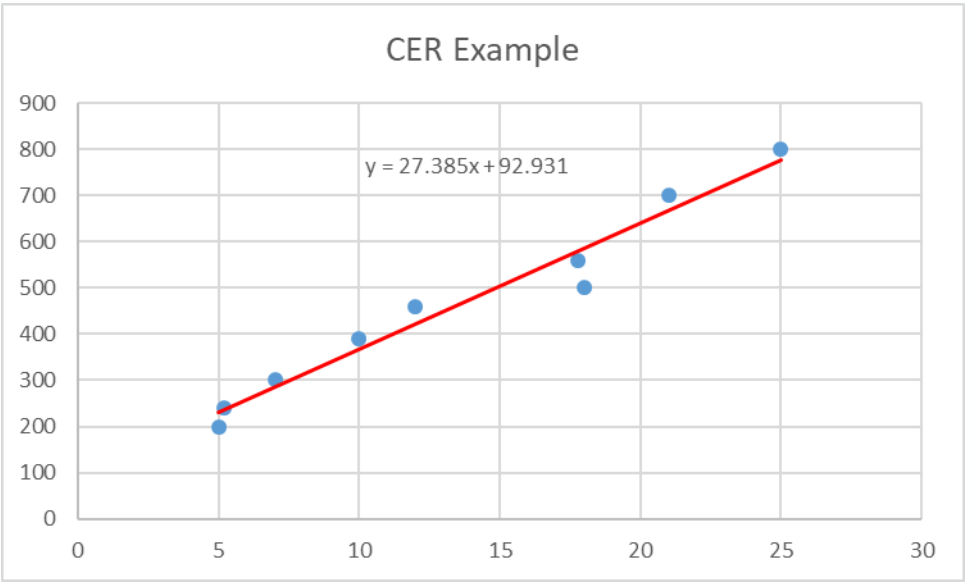
5. Study case

6. Results and conclusions

Parametric Cost Model

Parametric cost model are based on regression analysis over a set of observed data to develop cost estimating relationship (CER).

	Cost [k$ @2020EC]	Power [kW]
Project 1	390	10
Project 2	200	5
Project 3	240	5.2
Project 4	300	7
Project 5	460	12
Project 6	560	17.8
Project 7	700	21
Project 8	800	25
Project 9	500	18



Parametric Cost Model

Cost	Par. X
700	2.5
500	3.2
400	
300	6.3
500	7.3
600	
300	7.9
100	
350	8.4
550	
500	5.5
700	2.1
900	10.3
400	2.7

$$y = \alpha + \beta x$$

Cost	Par. X
700	2.5
500	3.2
400	
300	6.3
500	7.3
600	
300	7.9
100	
350	8.4
550	
500	5.5
700	2.1
900	10.3
400	2.7

$$\widehat{\alpha} = \bar{y} - (\widehat{\beta}\,\bar{x}),$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

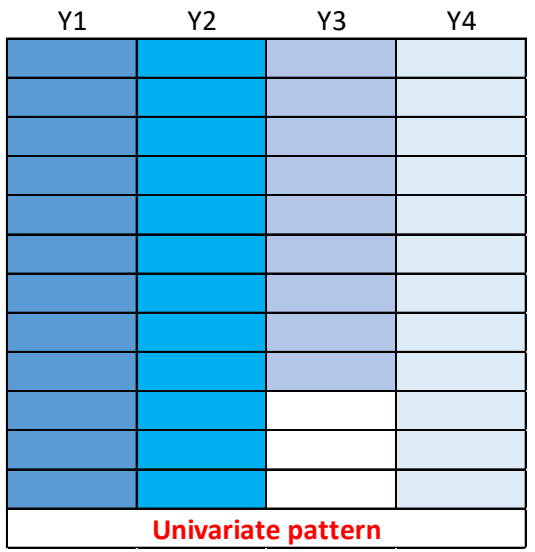
$$= \frac{s_{x,y}}{s_x^2}$$

$$= r_{xy}\frac{s_y}{s_x}.$$

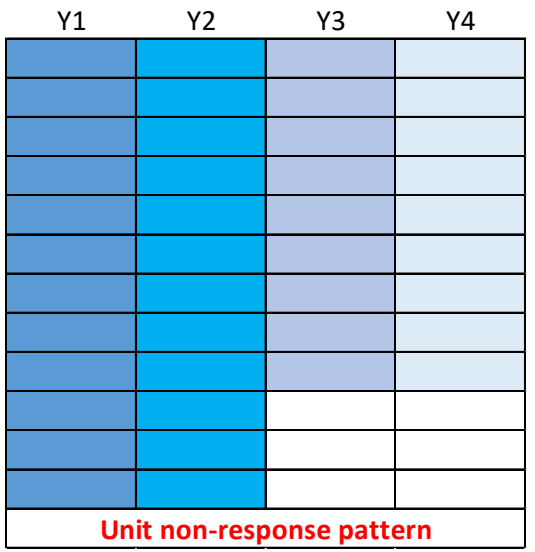
Cost	Par. X	Par. Y	Par. Z
700	2.5	22.1	3.9
500	3.2	28.4	5.0
400		13.1	
300	6.3	56.3	9.8
500	7.3		11.3
600		56.3	9.8
300	7.9		12.2
100		80.6	
350	8.4	75.2	13.0
550		23.9	4.2
500	5.5		8.5
700	2.1	18.5	3.3
900	10.3	92.3	16.0
400	2.7	23.9	

The importance is not to retrieve missing data but to build a model

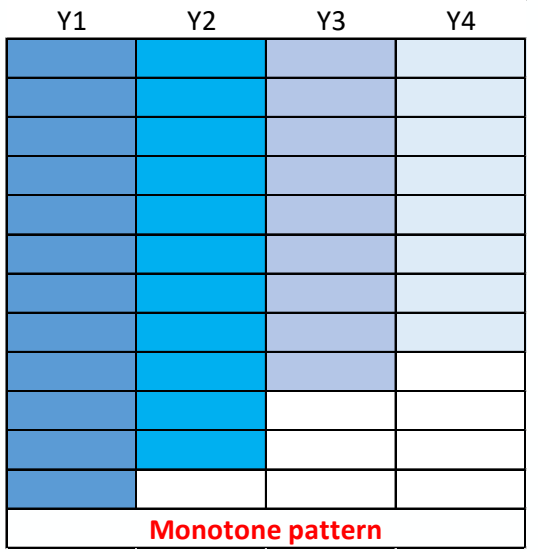
Missing data pattern typologies



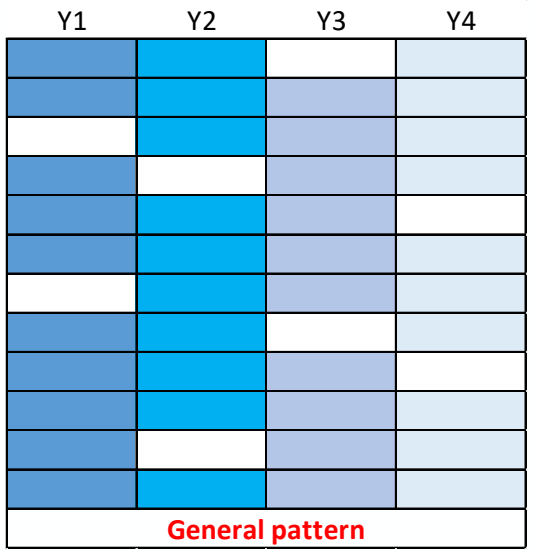
Missing values are isolated to a single variable



Missing values are relevant to the case where for some of the variables, respondents refuse to answer



A variable is missing for a particular observation implies that all subsequent variables are missing for that observation.



Most common pattern

the missing data pattern describes the location of the missing values and not the reasons for missingness.

Missingness mechanism typology

Missing completely at random (**MCAR**): missingness is unrelated of any observed and unobserved data, meaning that the probability of a missing data value is independent of any observation in the data set. In this case, missing and observed observations are generated from the same distribution, means there is no systematic mechanism that makes the data to be missing more than others.

$$p(R \mid \phi)$$

Missing at random (**MAR**): missingness is systematically related to the observed but not the unobserved data. For example, a registry examining depression may encounter data that are MAR if male participants are less likely to complete a survey about depression severity than female participants. That is, if probability of completion of the survey is related to their sex (which is fully observed) but not the severity of their depression, then the data may be regarded as MAR.

$$p(R \mid Y_{obs}, \phi)$$

Missing not at random (**MNAR**): missingness is systematically related to the unobserved data, that is, the missingness is related to events or factors which are not measured by the CE. For example, the depression registry may encounter data that are MNAR if participants with severe depression are more likely to refuse to complete the survey about depression severity.

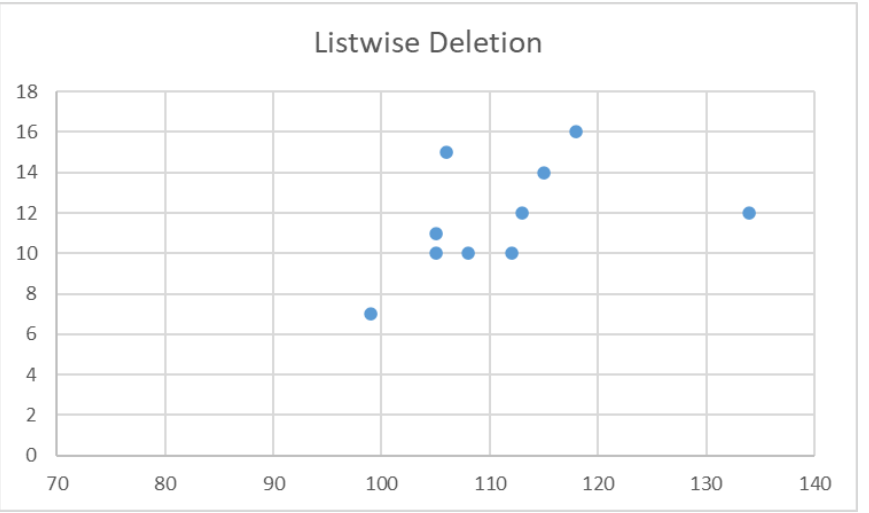
$$p(R \mid Y_{obs}, Y_{mis}, \phi)$$

Methodologies to deal with missing data

Deletion methods:

Pros: extremely easy to implement

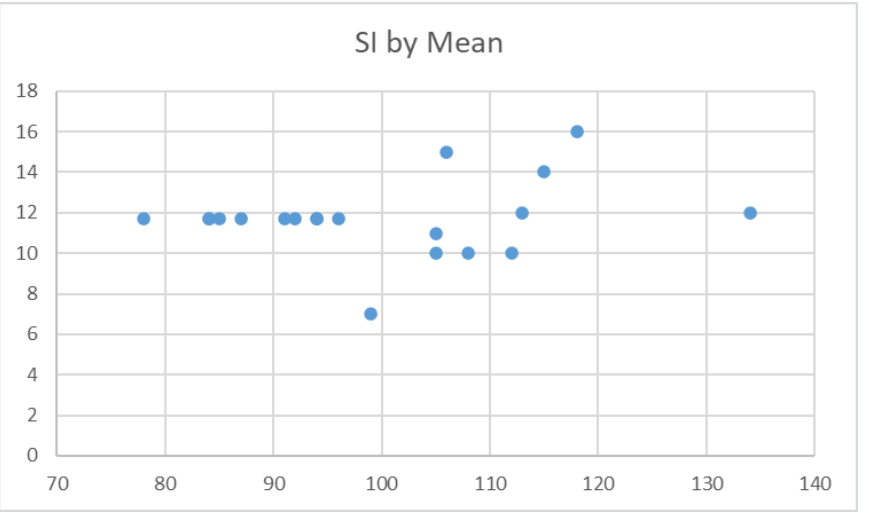
Cons: useful only for MCAR, eliminating observation is wasteful and reduce power of the dataset



Single imputation by mean:

Pros: easy to implement and guarantees the full data set

Cons: one single values for each missing variable attenuating variability of the data

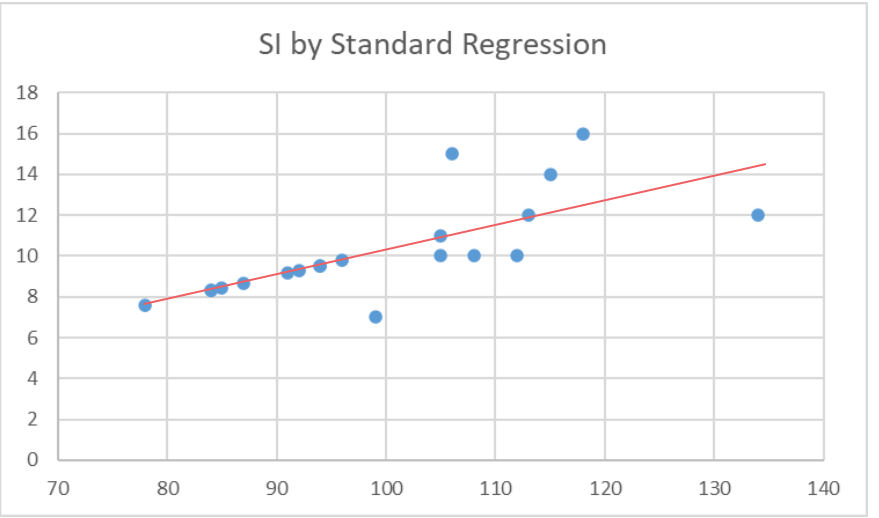


Methodologies to deal with missing data

Regression imputation:

Pros: full data set, borrowing information from observed data

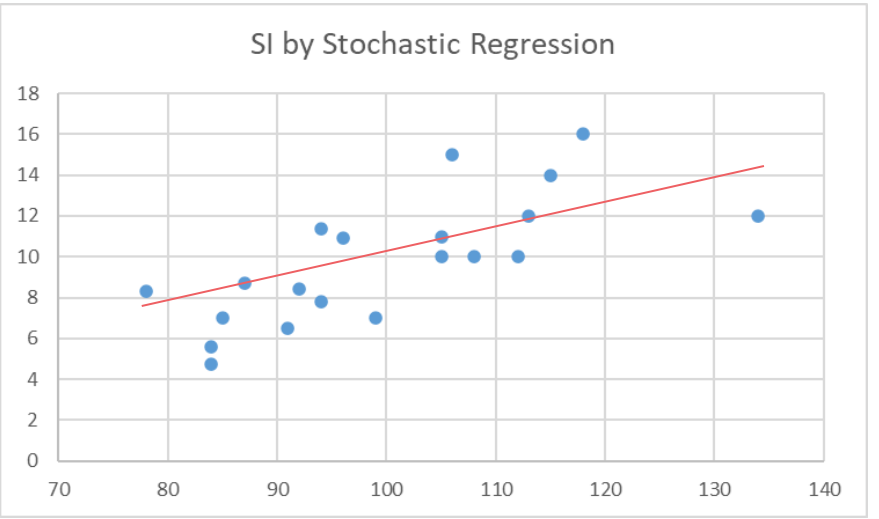
Cons: overestimation of correlation and R^2



Stochastic regression imputation:

Pros: full data set, eliminates bias associated to regression imputation by adding residuals z~N(0,σres)

Cons: attenuation of the standard error



Expectation Maximisation Method (EM)

Intuitive description of the **EM** based on a simple example:

CASE 1 - [1, 2, x] drawn from N(1,1). What is the best guess from the missing value?

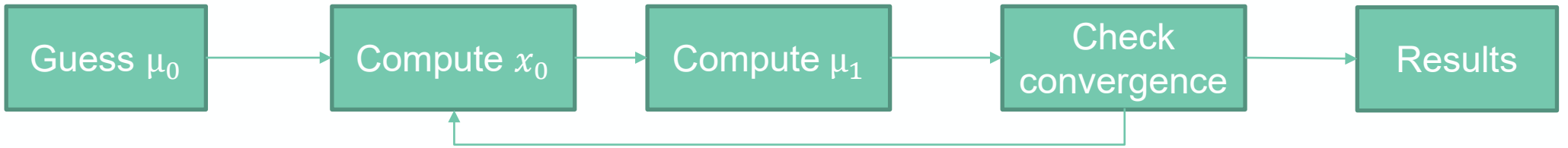
TAKE THE DISTRIBUTION MEAN

CASE 2 - [0, 1, 2] drawn from N(μ,1). What is the best guess for the distribution mean?

COMPUTE THE MEAN FROM THE DATA

CASE 3 - [1, 2, x] drawn from N(μ,1). What is the best guess for the missing value and the distribution mean?

ITERATIVE METHOD IS NEEDED:



Expectation Maximization Method (EM)

The EM is an iterative method for performing maximum likelihood estimation in the presence of latent variables.

"A.P. Dempster, N.M. Laird, D.B. Rubin Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. (1977)"

Given a set of observed data X and unobserved data Z, given also a vector of unknown parameter **θ** and the likelihood function, EM consists of 2 steps that are repeated iteratively until convergence:

E-step (Expectation) = Use the best guess for the parameters of the data model to get the log-likelihood for all the data (observed and latent), then consider the marginal log-likelihood with respect to the unobserved data:

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(t)}} \left[\log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})\right]$$

M-step (Maximisation) = Find the parameter θ that maximizes Q and call it θ^{t+1}:

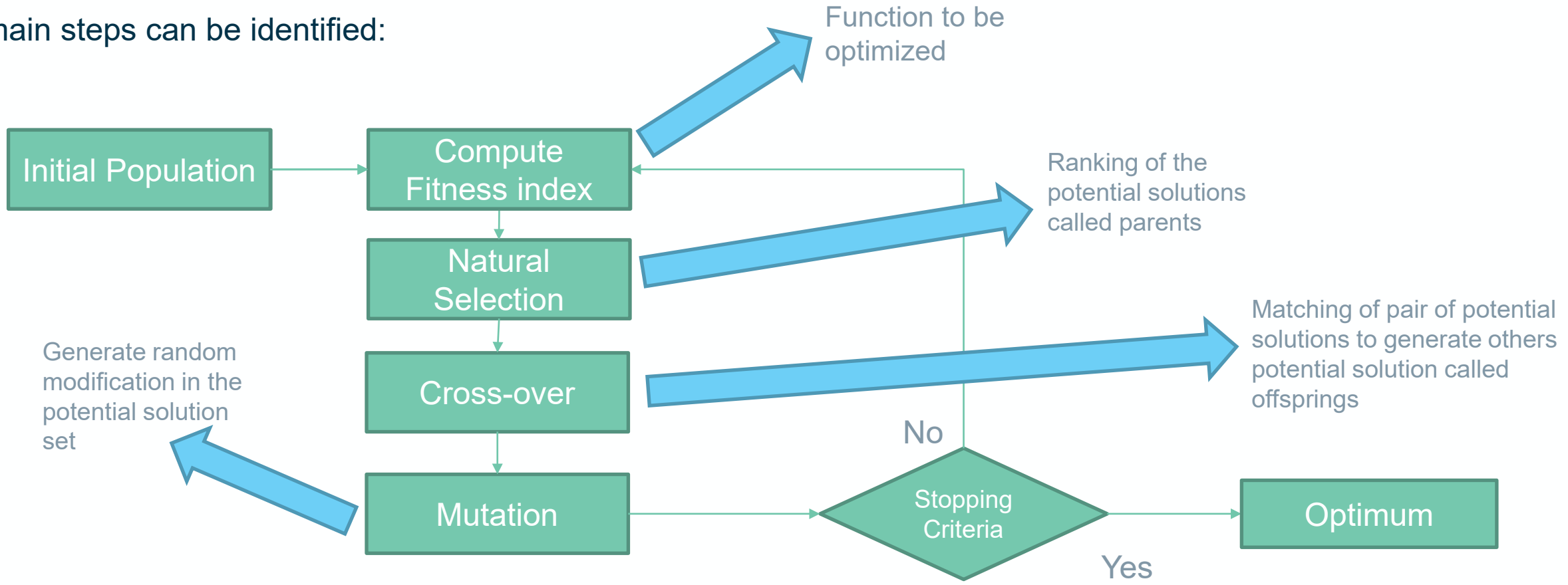
$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$$

Δ Not easy to find the global maximum, often convergence to local maximum; slow convergence; handling boundary constraints.

Genetic Algorithm (GA) brief description

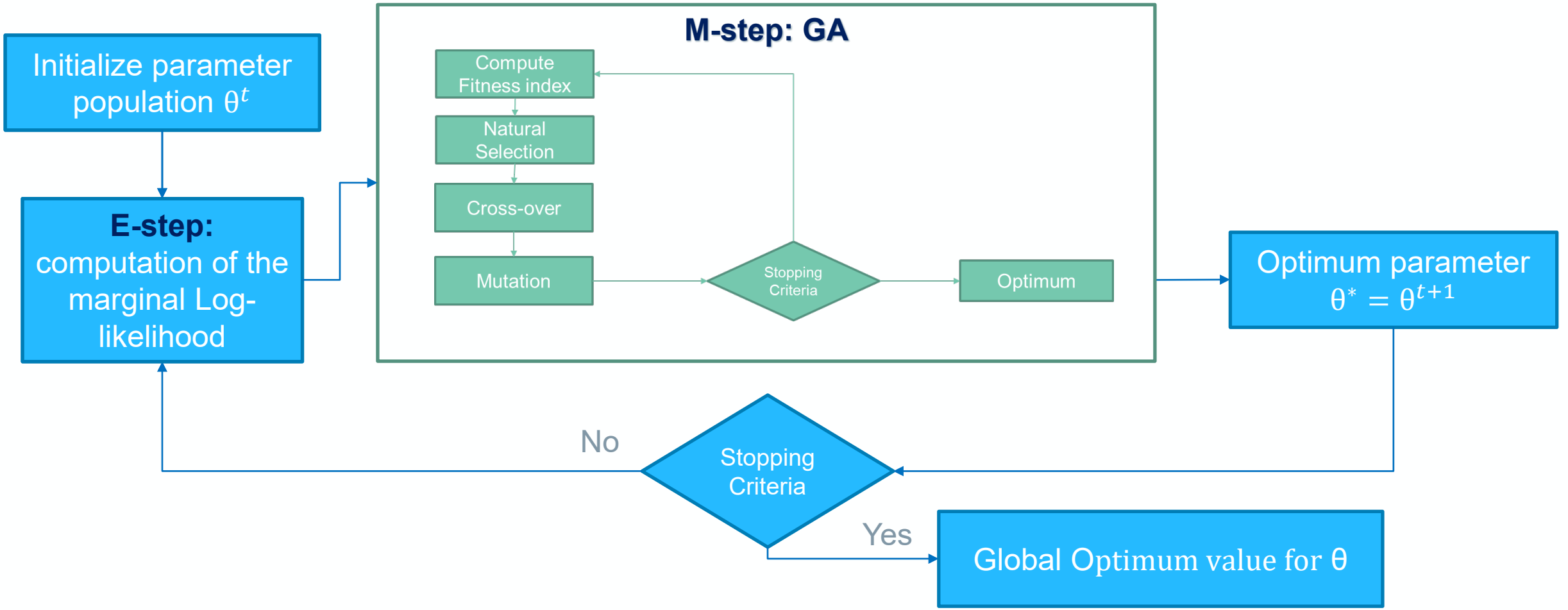
GA is a metaheuristic optimization technique based on the principle of genetics and natural evolution (survival of the fittest).

5 main steps can be identified:



EMGA method description

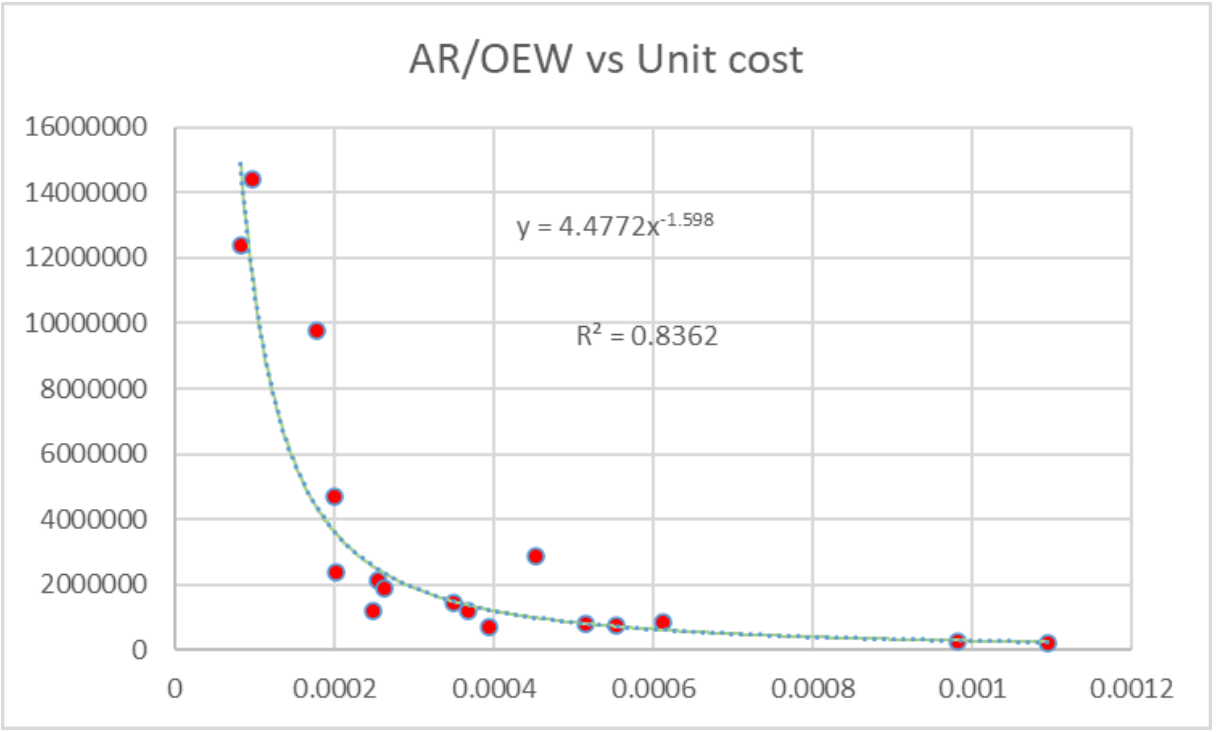
The EMGA method implements the EM algorithm replacing the standard technique used to solve the M-step with GA.



Study case

The study case shows the unitary cost for a set of military aircraft w.r.t. a parameter defined as the aspect ratio (AR) divided by the operative empty weight (OEW). The model is clearly a power CER of the type $y = Ax^{-b}$

AR/OEW	Unit cost	Model
0.000981128	237247	F-84
0.000553683	769330	F-84F
0.001095	219457	F-86
0.00051483	801602	F-89
0.000394057	697029	F-100
0.00036801	1200000	F-101C
0.000249832	1200000	F-102
0.000349617	1420000	F-104
0.000254352	2140000	F-105
0.00019985	4700000	F-106
0.000202056	2400000	F4
0.000178276	9800000	F-111
0.000612177	860000	A-4
0.000452814	2860000	A-7
0.000262921	1900000	B-47
9.71434E-05	14430000	B-52B
8.28792E-05	12400000	B-58



Study case

The algorithm has been tested considering missing observation equal 47% of the overall observed data.

Missing data pattern is general. Missing mechanism is assumed to be MCAR. All the variables are assumed to follow a normal distribution. Each observation is independent from the others. Data have been normalised to avoid numerical instability.

AR/OEW	Unit cost	Model
	237247	F-84
	769330	F-84F
0.001095	219457	F-86
0.00051483	801602	F-89
0.000394057	697029	F-100
0.00036801	1200000	F-101C
0.000249832	1200000	F-102
0.000349617		F-104
0.000254352	2140000	F-105
0.00019985	4700000	F-106
0.000202056	2400000	F4
0.000178276		F-111
0.000612177	860000	A-4
	2860000	A-7
0.000262921		B-47
	14430000	B-52B
8.28792E-05		B-58

Algorithm parameters:

minX=minW=0.5

maxX=maxW=1.05

Starting parameter vector=[$\mu_{x\ inc}$ $\mu_{w\ inc}$ σ_x $\sigma_{xw}\sigma_w$] =[-0.75 0.85 0.5 -0.1 0.6]

Npop=size of the population=4000

Nselection=Number of kept potential solution=0.5xNpop

Cross-over performed by using blending method

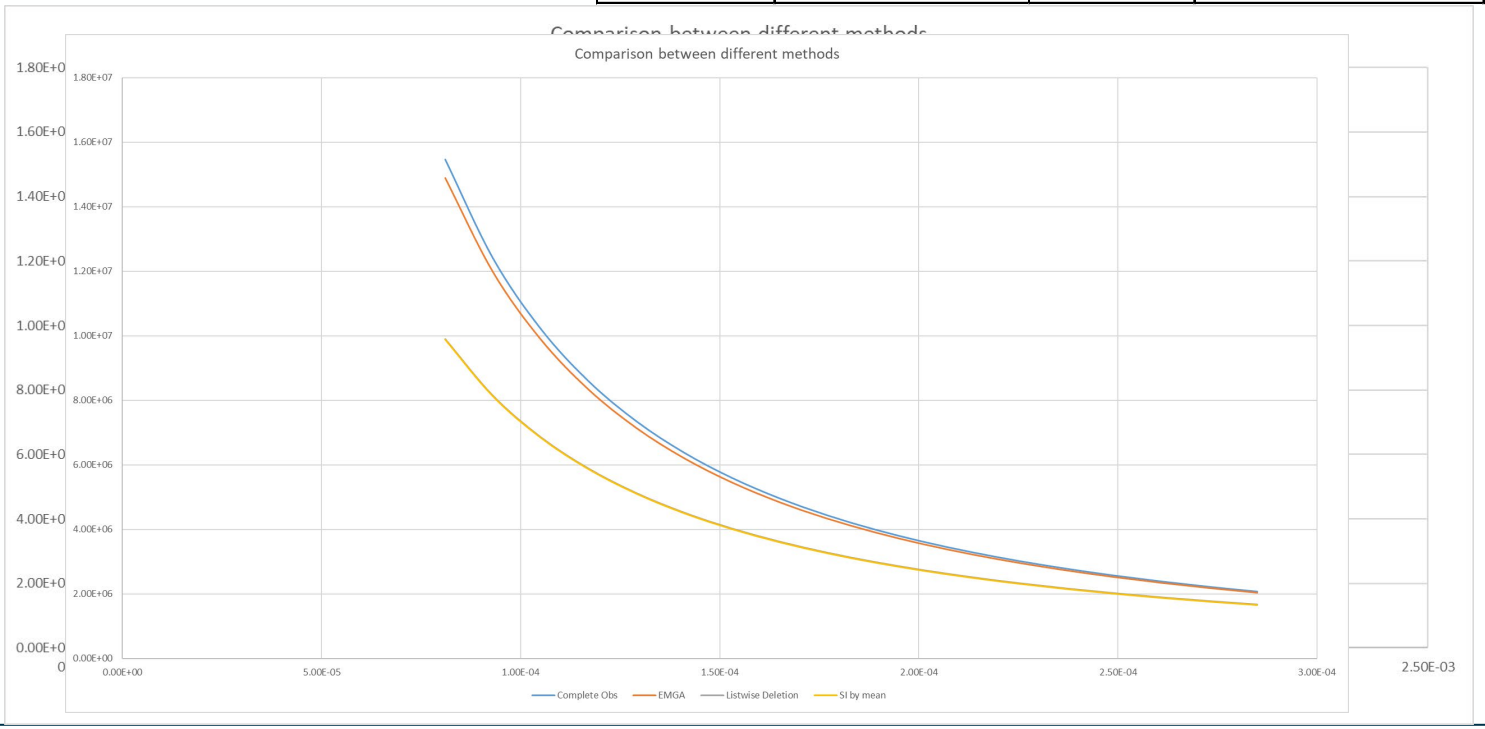
Rate of mutation is chosen randomly between 0 and 1, at each iteration

Accuracy of GA <1E-4

Results and conclusions

A new hybrid method that makes use of classic standard EM technique and metaheuristic algorithm based on genetic algorithm has been implemented mainly to cope with inherent EM issues, like difficulty in convergence to the global optimum. Accurate results have been found already after 23 iterations (θ^{23}). Results are compared with other techniques discussed previously:

	Complete Data	EMGA	Listwise Deletion	Imputation by mean
A	4.4772	5.3033	16.073	16.057
b	-1.598	-1.576	-1.415	-1.415



Differences for the 3 models w.r.t. the complete case are much more evident for small value of the parameter, producing the following percentage error w.r.t. the value itself:

AR/OEW	small value	high value
EMGA	3.72%	3.65%
Listwise Deletion	35.98%	18.23%
SI by mean	36.04%	14.47%