The Career Atlas:
Mathematical Notation

Cao Bittencourt

[a] *B. Sc. in Economics from EPGE (FGV), RJ, Brazil.*
[b] *Statistician at Atlas Career Guide Inc., FL, USA.*

Abstract

This is a brief document to define statistical methods for data-drive career choice and development. It deals with topics such as: career matching (i.e. vocational choice); estimation of competence, or overall skill level; estimation of skill set generality; versatility; skill set profitability; employability; labor market competitiveness; labor market taxonomy; optimal human resources acquistion and allocation; and so on and so forth. Each concept shall be explained at length in separate articles.

Keywords: Career choice; Career development; Matching algorithms; Competence; Similarity.

1. Basic Definitions

1.1. Skill Sets

The i-th professional attribute, or competency, of a person k is defined as:

$$a_i^k \in [0, 100], \tag{1}$$

where the interval $[0, 100]$ determines the bounds for every competency.[1]

The skill set, or career profile, of a person k is defined as the vector of their m attributes:

$$\boldsymbol{a_k} = (a_1^k, \dots, a_m^k). \tag{2}$$

A skill set matrix, or career profile matrix, is the collection of all n skill sets in the economy:

$$\mathbf{A} = \begin{bmatrix} a_1^1 & \dots & a_m^1 \\ \vdots & \ddots & \vdots \\ a_1^n & \dots & a_m^n \end{bmatrix}. \tag{3}$$

[1] More generally, these could be defined as a_{lb} (the lower bound) and a_{ub} (the upper bound). Here, the interval $[0, 100]$ is used because of its ease of interpretation.

1.2. Skill Set Normalization

Normalization by the scale bounds is defined by the tilde operator:

$$\tilde{a}_i^k = \frac{a_i^k - 0}{100 - 0} = \frac{a_i^k}{100} \in [0, 1]; \tag{4}$$

$$\tilde{\boldsymbol{a}}_{\boldsymbol{k}} = (\tilde{a}_1^k, \dots, \tilde{a}_m^k); \tag{5}$$

$$\tilde{\mathbf{A}} = \begin{bmatrix} \tilde{a}_1^1 & \dots & \tilde{a}_m^1 \\ \vdots & \ddots & \vdots \\ \tilde{a}_1^n & \dots & \tilde{a}_m^n \end{bmatrix}. \tag{6}$$

Normalization by a skill set's highest attribute is defined by the hat operator:

$$\hat{a}_i^k = \frac{a_i^k}{\max a_j^k} \in [0, 1]; \tag{7}$$

$$\hat{\boldsymbol{a}}_{\boldsymbol{k}} = (\hat{a}_1^k, \dots, \hat{a}_m^k); \tag{8}$$

$$\hat{\mathbf{A}} = \begin{bmatrix} \hat{a}_1^1 & \dots & \hat{a}_m^1 \\ \vdots & \ddots & \vdots \\ \hat{a}_1^n & \dots & \hat{a}_m^n \end{bmatrix}. \tag{9}$$

2. Basic Skill Set Models

The generality of a skill set is the mean of its maxima-normalized attributes:

$$\gamma_k = \left(\frac{1}{m}\right) \sum_{i=1}^{m} \hat{a}_i^k. \tag{10}$$

Generalists have high γ_k scores. Specialists have low γ_k scores. The generality vector of all n skill sets in the economy is:

$$\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_n). \tag{11}$$

The attribute equivalence of a particular attribute in a skill set measures the importance of that attribute relative to the skill set's highest attribute, using the skill set's generality as a midpoint and scaling parameter. The attribute equivalence of an attribute is denoted by the umlaut operator:

$$\ddot{a}_i^k = \text{aeq}(\hat{a}_i^k, \gamma_k) = \hat{a}_i^k \left[1 + \gamma_k (1 - \hat{a}_i^k) \exp\left(\frac{\hat{a}_i^k - \gamma_k}{\gamma_k - 1}\right)\right]^{-\frac{\gamma_k}{\hat{a}_i^k}}. \tag{12}$$

Attributes with high levels of attribute equivalence $\ddot{a}_i^k$ are said to be equivalent to the skill set's most importante attribute. These attributes are called *core* attributes. The attribute equivalence vector of a skill set is given by the collection of their m umlauted attributes:

$$\ddot{\boldsymbol{a}}_{\boldsymbol{k}} = (\ddot{a}_1^k, ..., \ddot{a}_m^k). \tag{13}$$

The attribute equivalence matrix is the collection of all attribute equivalence vectors in the economy:

$$\ddot{\mathbf{A}} = \begin{bmatrix} \ddot{a}_1^1 & \dots & \ddot{a}_m^1 \\ \vdots & \ddots & \vdots \\ \ddot{a}_1^n & \dots & \ddot{a}_m^n \end{bmatrix}. \tag{14}$$

The overall competence of a skill set is the mean of its scale-normalized attributes, weighted by each attribute's importance (i.e. its attribute equivalence):

$$c_k = \frac{\sum_{i=1}^m \ddot{a}_i^k \tilde{a}_i^k}{\sum_{i=1}^m \ddot{a}_i^k}. \tag{15}$$

The competence vector of all n skill sets in the economy is:

$$\boldsymbol{c} = (c_1, \dots, c_n). \tag{16}$$

3. Comparative Models

3.1. Matching Models

The most basic comparative model is that of Euclidean matching with linear weights:

$$s_{k,q} = s(\boldsymbol{a_k}, \boldsymbol{a_q}) = 1 - \tilde{d}(\boldsymbol{a_k}, \boldsymbol{a_q}) \in [0,1], \tag{17}$$

where

$$\tilde{d}_{k,q} = \tilde{d}(\boldsymbol{a_k}, \boldsymbol{a_q}) = \sqrt{\frac{\sum_{i=1}^m a_i^q (a_i^k - a_i^q)^2}{\sum_{i=1}^m a_i^q \max(100 - a_i^q, a_i^q)^2}} \in [0,1]. \tag{18}$$

In this model, we compare a skill set $\boldsymbol{a_k}$ to a skill set $\boldsymbol{a_q}$ by calculating the weighted Euclidean distance from $\boldsymbol{a_k}$ to $\boldsymbol{a_q}$ normalized by the maximum theoretical distance to $\boldsymbol{a_q}$.

Other weighting systems can be employed in this type of matching model. We could, for instance, substitute the linear weights with either quadratic weights,

$$a_i^{q\,2} \in [0,1], \tag{19}$$

or speciality-root weights,

$$a_i^{q\,\frac{1}{1-\gamma_k}} \in [0,1]. \tag{20}$$

But the best and most interpretable results are obtained using attribute equivalence as the weighting function:

$$\tilde{d}_{k,q} = \tilde{d}(\boldsymbol{a_k}, \boldsymbol{a_q}) = \sqrt{\frac{\sum_{i=1}^m \ddot{a}_i^q (a_i^k - a_i^q)^2}{\sum_{i=1}^m \ddot{a}_i^q \max(100 - a_i^q, a_i^q)^2}} \in [0,1]. \tag{21}$$

We could also employ other matching models instead of the "baseline" weighted Euclidean method. [detail each method later]:

1. logit regression matching
2. probit regression matching
3. bvls regression matching
4. tobit regression matching
5. pearson correlation matching
6. kendal nonparametric correlation matching
7. spearman nonparametric correlation matching

At last, similarity and normalized distance metrics determine the respective vectors and matrices, as follows:

$$\boldsymbol{s_k} = (s_{k,1}, \ldots, s_{k,n}); \tag{22}$$

$$\boldsymbol{\tilde{d}_k} = (\tilde{d}_{k,1}, \ldots, \tilde{d}_{k,n}); \tag{23}$$

$$\mathbf{S} = \begin{bmatrix} s_{1,1} & \ldots & s_{n,1} \\ \vdots & \ddots & \vdots \\ s_{1,n} & \ldots & s_{n,n} \end{bmatrix} = \begin{bmatrix} 1 & \ldots & s_{k,1} & \ldots & s_{n,1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{1,k} & \ldots & 1 & \ldots & s_{n,k} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{1,n} & \ldots & s_{k,n} & \ldots & 1 \end{bmatrix}; \tag{24}$$

$$\mathbf{D} = \begin{bmatrix} \tilde{d}_{1,1} & \ldots & \tilde{d}_{n,1} \\ \vdots & \ddots & \vdots \\ \tilde{d}_{1,n} & \ldots & \tilde{d}_{n,n} \end{bmatrix} = \begin{bmatrix} 0 & \ldots & \tilde{d}_{k,1} & \ldots & \tilde{d}_{n,1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{d}_{1,k} & \ldots & 0 & \ldots & \tilde{d}_{n,k} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{d}_{1,n} & \ldots & \tilde{d}_{k,n} & \ldots & 0 \end{bmatrix}. \tag{25}$$

3.2. Qualification Models

A closely related concept to matching is the qualification comparative model. In this family of functions, however, Euclidean matching is mandatory, as other matching methods do not make sense for this specific type of calculation. The reason for this is at that, here, we are not only interested in matching (i.e. a typical classification problem), but rather in the actual distances between comparison skill sets.

To define these models, we first have to define the gap function, which measures only positive competency gaps:

$$\delta^i_{k,q} = \delta(a^k_i, a^q_i) = \max(a^k_i - a^q_i, 0) \in [0, 100]. \tag{26}$$

Now that we have defined the gap function, we write the underqualification model:

$$\tilde{\delta}^<_{k,q} = \mathrm{uqa}(\boldsymbol{a_k}, \boldsymbol{a_q}) = \sqrt{\frac{\sum_{i=1}^m \ddot{a}^q_i \delta(a^q_i, a^k_i)^2}{\sum_{i=1}^m \ddot{a}^q_i \delta(a^q_i, 0)^2}} = \sqrt{\frac{\sum_{i=1}^m \ddot{a}^q_i \delta(a^q_i, a^k_i)^2}{\sum_{i=1}^m \ddot{a}^q_i {a^q_i}^2}}. \tag{27}$$

Analogously, the overqualification model is given by:

$$\tilde{\delta}^{\geq}_{k,q} = \mathrm{oqa}(\boldsymbol{a_k}, \boldsymbol{a_q}) = \sqrt{\frac{\sum_{i=1}^{m} \ddot{a}_i^q \delta(a_i^k, a_i^q)^2}{\sum_{i=1}^{m} \ddot{a}_i^q \delta(a_i^q, 100)^2}} = \sqrt{\frac{\sum_{i=1}^{m} \ddot{a}_i^q \delta(a_i^k, a_i^q)^2}{\sum_{i=1}^{m} \ddot{a}_i^q (100 - a_i^q)^2}}. \quad (28)$$

Finally, it is evident that “sufficient qualification” is the complement of the underqualification model:

$$s^{\geq}_{k,q} = \mathrm{sqa}(\boldsymbol{a_k}, \boldsymbol{a_q}) = 1 - \sqrt{\frac{\sum_{i=1}^{m} \ddot{a}_i^q \delta(a_i^q, a_i^k)^2}{\sum_{i=1}^{m} \ddot{a}_i^q \delta(a_i^q, 0)^2}} = 1 - \mathrm{uqa}(\boldsymbol{a_k}, \boldsymbol{a_q}). \quad (29)$$

As with the similarity and normalized distance statistics described above, all these three qualification models are bounded to the $[0, 1]$ interval. They also determine the following vectors:

$$\boldsymbol{\tilde{\delta}^{<}_{k}} = (\tilde{\delta}^{<}_{k,1}, \ldots, \tilde{\delta}^{<}_{k,n}); \quad (30)$$

$$\boldsymbol{\tilde{\delta}^{\geq}_{k}} = (\tilde{\delta}^{\geq}_{k,1}, \ldots, \tilde{\delta}^{\geq}_{k,n}); \quad (31)$$

$$\boldsymbol{s^{\geq}_{k}} = (s^{\geq}_{k,1}, \ldots, s^{\geq}_{k,n}); \quad (32)$$

and matrices

$$\tilde{\Delta}_{<} = \begin{bmatrix} \tilde{\delta}^{<}_{1,1} & \ldots & \tilde{\delta}^{<}_{n,1} \\ \vdots & \ddots & \vdots \\ \tilde{\delta}^{<}_{1,n} & \ldots & \tilde{\delta}^{<}_{n,n} \end{bmatrix} = \begin{bmatrix} 0 & \ldots & \tilde{\delta}^{<}_{k,1} & \ldots & \tilde{\delta}^{<}_{n,1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{\delta}^{<}_{1,k} & \ldots & 0 & \ldots & \tilde{\delta}^{<}_{n,k} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{\delta}^{<}_{1,n} & \ldots & \tilde{\delta}^{<}_{k,n} & \ldots & 0 \end{bmatrix}; \quad (33)$$

$$\tilde{\Delta}_{\geq} = \begin{bmatrix} \tilde{\delta}^{\geq}_{1,1} & \ldots & \tilde{\delta}^{\geq}_{n,1} \\ \vdots & \ddots & \vdots \\ \tilde{\delta}^{\geq}_{1,n} & \ldots & \tilde{\delta}^{\geq}_{n,n} \end{bmatrix} = \begin{bmatrix} 0 & \ldots & \tilde{\delta}^{\geq}_{k,1} & \ldots & \tilde{\delta}^{\geq}_{n,1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{\delta}^{\geq}_{1,k} & \ldots & 0 & \ldots & \tilde{\delta}^{\geq}_{n,k} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{\delta}^{\geq}_{1,n} & \ldots & \tilde{\delta}^{\geq}_{k,n} & \ldots & 0 \end{bmatrix}; \quad (34)$$

$$\mathbf{S}_{\geq} = \begin{bmatrix} s_{1,1} & \ldots & s_{n,1} \\ \vdots & \ddots & \vdots \\ s_{1,n} & \ldots & s_{n,n} \end{bmatrix} = \begin{bmatrix} 1 & \ldots & s_{k,1} & \ldots & s_{n,1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s^{\geq}_{1,k} & \ldots & 1 & \ldots & s^{\geq}_{n,k} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s^{\geq}_{1,n} & \ldots & s^{\geq}_{k,n} & \ldots & 1 \end{bmatrix}. \quad (35)$$