An Information-Loss-Minimizing Approach to Multinominal Alignment in Microsimulation Models

Peter Stephensen
Danish Rational Economic Agents Model, DREAM

**DREAM Conference Paper
for the European Meeting of
the International Microsimulation Association
(preliminary version)**
October 2014

Abstract
There exist a number of alignment methods. They typically address the binary case and they typically put special emphasis on one of the alternatives rather than treating all alternatives in a symmetric way (for a review, see Li & O'Donoghue, 2014). In this paper we propose a more general mathematical foundation of multinominal alignment. Meancorrection alignment is defined as a method that minimizes the information loss in the adjustment process. The analytical solution to the alignment problem is characterized and furthermore applied in deriving various properties for the method. It is demonstrated that there exits an algorithm called Bi-Proportional Scaling that converge to the solution of the problem. This is tested against two versions of the Newton-Raphson-algoritm, and it is demonstrated that it is at least twice as fast as these methods. Finally, it is argued, that the method is very easy to implement.

DREAM, Danish Rational Economic Agents Model. Amaliegade 44, 1256 København K
www.dreammodel.dk

An Information-Loss-Minimizing Approach to Multinominal Alignment in Microsimulationmodels*

Peter Stephensen (Version 0.2, Preliminary)†

October 10, 2014

Abstract

There exist a number of alignment methods. They typically address the binary case and they typically put special emphasis on one of the alternatives rather than treating all alternatives in a symmetric way (for a review, see Li & O'Donoghue, 2014). In this paper we propose a more general mathematical foundation of multinominal alignment. Mean-correction alignment is defined as a method that minimizes the information loss in the adjustment process. The analytical solution to the alignment problem is characterized and furthermore applied in deriving various properties for the method. It is demonstrated that there exits an algorithm called *Bi-Proportional Scaling* that converge to the solution of the problem. This is tested against two versions of the Newton-Raphson-algoritm, and it is demonstrated that it is at least twice as fast as these methods. Finally, it is argued, that the method is very easy to implement.

*Presented at European Meeting of the International Microsimulation Association, Maastricht, 23-24 October

†Research director, Ph.D. The economic model DREAM, Amaliegade 44, DK-1256, Copenhagen, Denmark. psp@dreammodel.dk

1 Introduction

Alignment is the core facility for controlling microsimulation models. Microsimulation models are large, elaborate models that need such a tool at the macro level. Alignment basically works by manipulating the probabilities of the model, aiming to do *model calibration*, *comparative analysis* (static or dynamic) and/or to *eliminate Monte-Carlo noise*.

In a more technical sense, alignment can be said to do one of two things: *mean-correction* or *variance-elimination*[1]. The latter is equivalent to the above mentioned elimination of Monte-Carlo noise. In microsimulation models, events are typically modeled by drawing a uniformly distributed number and compering it to a probability that the event happens. This Monte-Carlo technique introduces noise into the model. Such noise can be eliminated form the model, using alignment techniques (Baekgaard, 2002). This paper is *not* dealing with this kind of alignment. We concentrate on *mean-correction*. Mean correction is the process by which you manipulate the probabilities of the individuals such that i) the manipulation is as gentle as possible (discussed below), and ii) you hit a predefined mean/average propensity for the entire group. This is a very important facility to have in a model. It makes it possible for you to calibrate the model, and to do comparative analysis. *Calibration* is necessary for many reasons[more to come...base year, region etc.].

Once you have a well-calibrated model and a reasonable base run the next step is to do some analysis. Here mean-correction alignment is important as well. In most economic models (CGE, DSGE, SEM etc.) comparative analysis (static or dynamic) is done by changing a few parameters or exogenous variables (tax rate, labor supply, public demand etc.) and then compare the new outcome with the base run. Typically this is not the way to do it in a microsimulation

[1] Or a more general term for both: *moment correction*, as it corrects the first two moments of the probability distribution.

model. The reason is that agent behavior is modeled differently in microsimulation models, compared to the typical economic model. In a typical economic model, behavior is modeled either by some estimated structural equations (dependent on prices, tax rates, wages etc.) or by calibrated demand systems (based on utility/profit-maximizing behavior). In a microsimulation model, behavior is defined by the estimated transition probabilities. These probabilities will mostly not depend on economic variables like prices, tax rates and wages (all though they can), but will be based on characteristics such as gender, age, education, region etc. Therefore the if/then-questions are different in the different models. Questions you might ask in such a microsimulation model are: what happens if the education level increases? If the structural unemployment decreases? In the first case you should somehow align the education system to produce more students (bigger student uptake, less dropouts). In the second case you should align the average unemployment propensity.

There exist a number of alignment methods. They typically address the binary case and they typically put special emphasis on one of the alternatives rather than treating all alternatives in a symmetric way (for a review, see Li & O'Donoghue, 2014). In this paper we propose a more general mathematical foundation of multinominal alignment. Mean-correction alignment is defined as a method that minimizes the information loss in the adjustment process. The analytical solution to the alignment problem is characterized and furthermore applied in deriving various properties for the method. It is demonstrated that there exits an algorithm called Bi-Proportional Scaling that converge to the solution of the problem. This is tested against two versions of the Newton-Raphson-algoritm, and it is demonstrated that it is at least twice as fast as these methods. Finally, it is argued, that the method is very easy to implement.

2 What is a good alignment method?

The following is a nice-to-have-list for mean-correction alignment methods:

1. Adjust correctly to the target means.
2. Retain the original shape of the probability distribution.
3. Retain zero probabilities.
4. Symmetric formulation.
5. Multinominal alignment. Should easily generalize to a situation with more than 2 alternatives.
6. Ability to work well in a logit environment.
7. Computer efficiency.
8. Easy implementation.

Point 1) and 2) are properly need-to-have's and are similar to the first 2 points in Li & O'Donoghue(2014). The basic idea of mean-correction alignment is to be able to change the mean of the probability distribution. Therefore the method should be precise when it comes to adjusting to a new target (point 1). Let's try to make this more precise. Let p_{ia} be individual i's probability of entering alternative a ($i = 1,...,N$ and $a = 1,...,A$). In a binary model $A = 2$. Because we are working with probabilities we should necessary have that

$$\sum_a p_{ia} = 1,\ i = 1,...,N$$

Let X_a be the number of persons in alternative a (such that $\sum_a X_a = N$). The expected number of persons in alternative a is given by:

$$E[X_a] = \sum_i p_{ia}, a = 1, ..., A$$

We can now define mean-correction alignment:

Definition 1. Let p_{ia}^0 ($i = 1, ..., N$ and $a = 1, ..., A$) be the initial probabilities in a population. A *mean-correction alignment* to a target $\left(\overline{X}_1, ..., \overline{X}_A\right)$ is a set of new probabilities p_{ia}, such that:

1. $\sum_i p_{ia} = \overline{X}_a, a = 1, ..., A$
2. $\sum_a p_{ia} = 1, i = 1, ..., N$
3. $p_{ia} \geq 0, i = 1, ..., N, a = 1, ..., A$

According to this definition there is no connection between the initial probabilities and the aligned probabilities. This is not satisfying and that is why we need 2). We need in some sense to try to retain the original distribution. To be as gentle and non destructive as possible. A way to handle this is to define some kind of distribution measure, that measures the difference between the initial probabilities and the aligned probabilities. Let P be a $N \times A$ matrix containing the $p_{ia}-$elements. Let $D\left(P, P^0\right)$ be our measure. A precise version of 2) would be to minimize $D\left(P, P^0\right)$ given the restrictions in Definition 1. We will return to this optimization problem later.

The third point deals with zero probabilities. Saying that a probability is zero is a very strong statement. It means that something is impossible, and that it will never happen. An alignment

method should therefore retain zeroes such that what is impossible before the alignment is also impossible after the alignment.

The fourth point on the list is *symmetric formulation*. A good example of an asymmetric alignment method is *multiplicative scaling* (Li & O'Donoghue, 2014). According to this kind of alignment you choose one of the alternatives (let's say alternative 1) and scale the probabilities to the desired level:

$$p_{i1} = \lambda p_{i1}^0$$

where the scaling factor λ is defined by

$$\sum_i p_{i1} = \sum_i \lambda p_{i1}^0 = \overline{X}_1$$

If we have two alternatives, the probability of the other alternative is given residually

$$p_{i2} = 1 - p_{i1}$$

This is clearly asymmetric. If we had chosen to scale the probabilities of the second alternative instead, we would have had another alignment. This is not satisfying. The only way to make sure that the alignment formulation is symmetric, is to make a definition based on all alternatives. A good alignment method should therefore depend on *all* the initial probabilities and some parameter vector θ_a

$$p_{ia} = f_a\left(p_{i1}^0, ..., p_{iA}^0; \theta_a\right), a = 1, ..., A$$

This is typically not the case in the most used alignment methods (Li & O'Donoghue, 2014).

A good alignment technique should generalize easily to the *multinominal* case ($A > 2$). As mentioned above, most alignment methods are binary. Multinominal problems are often solved by a sequence of binary alignments. It is quite difficult to know precisely what is going on if this approach is used. It is preferable to have a method that is designed to the general multinominal case.

3 A multinominal alignment method

In this section we present an alignment method that satisfies all points 1-8 in the last section. We automatically satisfy 1-3 and 5 by defining the alignment method as a optimization problem. Define the *information loss*[2] D by using P instead of P^0:

$$D\left(P,P^0\right) = \sum_i \sum_a p_{ia} log\left(\frac{p_{ia}}{p^0_{ia}}\right) \tag{3.1}$$

This measure is derived from information theory (Shannon, 1948) and it is suitable on measuring how much the distribution of probabilities are changed by the alignment process. Nowadays it is often used in machine learning as an alternative to the Gini-measure, when evaluating decision trees (Raileanu & Stoffel, 2004).

We would like to minimize the information loss D given the restrictions from definition 1:

$$\sum_i p_{ia} = \overline{X}_a, a = 1,...,A \tag{3.2}$$

$$\sum_a p_{ia} = 1, i = 1,...,N \tag{3.3}$$

[2]Also called Kullback–Leibler divergence, information divergence, information gain, relative entropy and cross entropy

$$p_{ia} \geq 0, i = 1,...,N, a = 1,...,A \tag{3.4}$$

The restriction (3.2) makes sure that we satisfy 1) in our list (ability to adjust correctly to the target means). The fact that we minimize the information loss satisfies 2) in the list (ability to retain the original shape of the probability distribution). 'Symmetric formulation' and 'multi-nominal alignment' come automatically (point 4 and 5).

We can now derive our new alignment method:

Definition 2. Let p_{ia}^0 ($i = 1,...,N$ and $a = 1,...,A$) be the initial probabilities in a population. An *Information-Loss-Minimizing Mean-Correction Alignment* to a target $\left(\overline{X}_1,...,\overline{X}_A\right)$ is a set of new probabilities p_{ia} that minimize (3.1), given the restrictions (3.2), (3.3) and (3.4).

And we can prove the theorem:

Theorem 3. *Let p_{ia}^0 ($i = 1,...,N$ and $a = 1,...,A$) be the initial probabilities in a population. In an Information-Loss-Minimizing Mean-Correction Alignment the probabilities p_{ia} are given by:*

$$p_{ia} = \frac{e^{\varphi_a} p_{ia}^0}{\sum_s e^{\varphi_s} p_{is}^0}$$

where the parameters φ_a satisfies:

$$\sum_a \varphi_a = 0$$

Proof. Define the Lagrange-function L:

$$L = \sum_i \sum_a p_{ia} \left(log\left(p_{ia}\right) - log\left(p_{ia}^0\right)\right) - \sum_i \lambda_i \left(\sum_a p_{ia} - 1\right) - \sum_a \varphi_a \left(\sum_i p_{ia} - \overline{X}_a\right)$$

The first order derivatives are given by

$$\frac{\partial L}{\partial p_{ia}} = \left(log\left(p_{ia}\right) - log\left(p_{ia}^0\right)\right) + 1 - \lambda_i - \varphi_a = 0$$

such that:

$$p_{ia} = e^{\lambda_i - 1} e^{\varphi_a} p_{ia}^0 \tag{3.5}$$

Sum over the alternatives a (using (3.3)):

$$\sum_a p_{ia} = \sum_a e^{\lambda_i - 1} e^{\varphi_a} p_{ia}^0 = 1$$

such that

$$e^{\lambda_i - 1} = \left[\sum_a e^{\varphi_a} p_{ia}^0\right]^{-1}$$

Inserting this into (3.5) yields:

$$p_{ia} = \frac{e^{\varphi_a} p_{ia}^0}{\sum_s e^{\varphi_s} p_{is}^0}$$

Observe that the Lagrange-parameters φ_a are not unique, as

$$\frac{e^{\varphi_a} p_{ia}^0}{\sum_s e^{\varphi_s} p_{is}^0} = \frac{e^{x+\varphi_a} p_{ia}^0}{\sum_s e^{x+\varphi_s} p_{is}^0}$$

We can therefore assume that

$$\sum_a \varphi_a = 0$$

□

Observe that $p_{ia} \in [0,1]$ and that 3) is satisfied (retaining zero).

3.1 Logit Scaling

The alignment method could alternatively be called *Logit Scaling*. The method fits perfectly to a situation where the initial probabilities are supplied from a logit-model (point 6 in the nice-to-have-list). Assume the probabilities are given by

$$p_{ia}^{0} = \frac{e^{V_{ia}}}{\sum_s e^{V_{is}}}$$

where V_{ia} are some logit-utilities. Then according to theorem 3:

$$p_{ia} = \frac{e^{\varphi_a + V_{ia}}}{\sum_s e^{\varphi_s + V_{is}}}$$

The alignment is done by adding alternative-specific constants to the logit-utilities. As a corollary to this, if the taste of the agent are changed in an additive way, the probabilities can be corrected with alignment. This is similar to the result in Li & O'Donoghue(2014) where it is shown that a "biased intercept" in a logit model can be corrected with some of the aliment methods ("Sidewalk hybrid with non linear adjustment" and "Sort by the difference between logistic adjusted predicted probability and random number"). Here we give a mathematical explanation.

3.2 Bi-proportional scaling: An algorithmic solution

The alignment method seemed to have some nice properties, but how about the practical side of the equation? We need to solve the following problem: choose parameters φ_a, $a = 1,...,A$, such that:

$$\sum_a \varphi_a = 0$$

$$\sum_i \frac{e^{\varphi_a} p_{ia}^0}{\sum_s e^{\varphi_s} p_{is}^0} = \overline{X}_a, a = 1, ..., A-1$$

This is A non-linear equations with A unknown. As A typically are rather small (2 to 10), this is not a big equation system. But from a computational point of view it is a problem that we sum over the number of individuals i. This can be a very big number (up to millions). We therefore need a fast algorithm.

One approach is to set up a Newton-Raphson algorithm. This implies the calculation of the so called $A \times A$ Jacobi-matrix in each iteration. Each element in this matrix is a sum over all individuals. This can be done, and actually work pretty well.

But there is another way that is considerably faster: *Bi-proportional scaling*. Look at the original problem: minimize (3.1) given (3.2) and (3.3). This is actually a *Matrix Balancing Problem* (Schneider & Zenios, 1990). Instead of the initial matrix P^0 we should choose another matrix P, such that the difference between the matrices are as small as possible (3.1), and such that the row sums are given (3.3) and the column sum are given (3.2). It is well known (McDougall, 1999) that if the measure of difference between the matrices is Information-Loss (or gain as it typically is defined as), then there exists a simple algorithm that converge to the solution to the matrix balancing problem. This algorithm is called bi-proportional scaling: First you scale the rows such that the row sums are correct. Then you scale the columns such that the column sums are correct (but then you destroyed the row sums). You repeat this until the system converge. In real world problems the convergence is very fast (often less than 10 iterations).

The method was tested against two versions of the Newton-Raphson algorithm, a standard version and an optimized version (see table 1). In the standard version, the Jacobi matrix was calculated based on an abstract differentiation of the non-linear equation system. In each iteration the elements in the matrix was calculated by an average over all individuals. The matrix

	Seconds used
Newton-Raphson (Standard)	3.94
Newton-Raphson (Optimized)	1.45
Bi-Proportional Scaling	0.60

Table 1: Speed test. Comparing Bi-proportional scaling with two Newton-Raphton algorithms ($N = 1.000.000, A = 4$).

was inverted using the C#-library Math.Net. In the optimized version, the procedure was the same except for the calculation of the Jacobi matrix. The elements in the matrix was averaged over a sample of 10 pct. of the individuals. This improved the speed with only very modest effect on the precision. The artificial population was 1.000.000 individuals and there was 4 alternatives. Bi-proportional scaling ran twice as fast as the optimized Newton-Raphson, and mere than 6 times faster than the standard Newton-Raphson.

The final point on the list is ease of implementation. It took more than a week to develop and test the Newton-Raphson algorithms. When the idea came to the author, it took half an hour to code bi-proportional scaling, showing the power of the approach.

References

Baekgaard, H. (2002). Micro-macro linkage and the alignment of transition processes : some issues, techniques and examples. National Center for Social and Economic Modeling (NATSEM) Technical paper No. 25.

Li, Jinjing and Cathal O'Donoghue (2014) "Evaluating Binary Alignment Methods in Microsimulation Models" Journal of Artificial Societies and Social Simulation 17 (1) 15

McDougall, Robert A. (1999) "Entropy Theory and RAS are Friends". GTAP Working Paper 5-14-1999

Neufeld, C. (2000). Alignment and variance reduction in DYNACAN. In A. Gupta & V. Kapur (Eds.), Microsimulation in Government Policy and Forecasting (pp. 361–382). Amsterdam: North-Holland.

Raileanu, Laura Elena and Kilian Stoffel (2004) "Theoretical Comparison between the Gini Index and Information Gain Criteria" Annals of Mathematics and Artificial Intelligence May 2004, Volume 41, Issue 1, pp 77-93

Schneider, Michael H. and Zenios, Stavros A. (1990) "A Comparative Study of Algorithms for Matrix Balancing". Operations Research, Vol. 38, No. 3 (May - Jun., 1990), pp. 439-455.

Shannon, C.E. (1948) "A mathematical theory of communication". Bell System Technical Journal, 27:379–423, 623–659.