Teorie & Modelli, n.s., XII, 1-2, 2007 (211-221)

A general method for parameter estimation of averaging models

Giulio Vidotto & *Marco Vicentini*[*] (Padua)

Averaging models have been proposed within the framework of Information Integration Theory (Anderson, 1965). Nowadays, these models are applied in several psychological areas. However, statistical problems such as divergence, uniqueness, and goodness of fit arise from parameter identification. To deal with these issues, Anderson (1981, 1982) suggested the method of sub-designs, capable of providing theoretically valid estimations. Using the empirical results obtained by this approach, Zalinski (1984, 1987) offered a statistical procedure for the estimation of weights and scale values in averaging models. This procedure can be primarily applied when a full factorial design is followed by all its sub-designs; it considers a subset of the possible averaging models, and the software permits to manage data from one subject and one session at a time. Moreover, Wang and Yang (1998) noted that, although this procedure was carefully designed to allow for proper tests of statistical significance especially involving the Chi-square test, it does not involve validation criteria for the estimated weights and scale values as, for example, the information criterion.

The present paper proposes a new method which allows considering the results of all the acceptable sub-design experimental conditions with a new statistical procedure to improve the estimation of weights and scale values and to provide validation criteria.

Averaging models

Let R denote the integrated response produced by an individual to the different scale values, s_i, of a single stimulus variable and let w_i denote the weight representing the importance of these values. The weighted averaging model states that

[*] Department of General Psychology, University of Padua.

$$R = \frac{w_0\, s_0 + \sum w_i\, s_i}{w_0 + \sum w_i} \ , \qquad (1)$$

that is, R is a weighted sum of values divided by the sum of weights.

In Equation 1 the effect of stimuli may depend on the amount of the initial state, with the same stimulus having opposite effects depending on whether its scale value is larger or less than the present response. The initial state of the process is represented by w_0 and s_0. The initial state enables the model to take account of the set-size effect in which pieces of added information of equal value can produce a more extreme response. As a consequence of the initial state, the response to a single stimulus is not usually a linear function of its scale value. That is, Equation 1 implies that R is not usually a linear function of the scale value s_i.

If all levels of a factor have the same weight, then this factor is said to be equally-weighted. The sum of weights in the denominator has the same value in each cell of the design and can be absorbed into an arbitrary scale unit. Accordingly, this model has a linear form. Essentially, all the well known methods to analyze linear models apply directly to the equal weight averaging model (EAM). Thus, the averaging model is easy to test when the equal-weight condition is met.

In some situations, evidences indicate that some up to almost all levels of a given factor may have different weights. A family of several models can be described starting from the EAM up to a model in which all the levels differ in weight from one another.

The EAM model may not be the best solution to explain crossover effects, i.e., the observed interactions among experimental factors. This problem can be solved by adding one parameter to the model, that is, by changing the importance of a single weight. The introduction of a new parameter makes the model more complex, generally providing better goodness of fit indexes especially for the residual sum of squares, but it decreases the degrees of freedom of the model itself. This model is usually called differential weighting model. By adding other parameters, the overall fit may not be improved, although the complexity increases while the degrees of freedom decrease. The most complex situation is the complete differential weight averaging model (complete DAM) in which all weights differ for each level and for each factor.

An important complication arises with unequal weighting within one single attribute dimension (Oden & Anderson, 1971). The complete DAM allows each stimulus to have its own weight as well as its own scale value. The sum of absolute weights in the denominator of Equation 1 is variable

from cell to cell and the model becomes inherently non linear. It is now necessary to estimate weights and scale values for each stimulus level of one or more factors. That is, instead of a single weight for each factor, a weight is estimated for each level of each factor. While weights and scale values are still identifiable, this non-linearity introduces statistical problems concerning bias, convergence, reliability, and goodness of fit. It requires a suitable methodology (Zalinski & Anderson, 1990).

Method of sub-designs

A general problem in estimation concerns identifiability and uniqueness. In a linear model applied to a factorial design, for example, weights are not generally identifiable since they are usually confounded with scale units. With a suitable design, averaging models can provide ratio-scale estimates of weight and linear-scale estimates of scale values. On the basis of these scaling results, valid statistical comparisons can be made among estimated weights and estimated scale values (Zalinski, 1987).

A proper experimental design is necessary for a unique parameter estimation. According to Anderson (2001), the general method to obtain complete identifiability of all parameters is to adjoin selected sub-designs to a full factorial design. This method allows us to obtain more observations from the factorial design, and parameter estimation can be performed on several observations. In general, providing more observed data increases the degrees of freedom, the estimation is more stable, and the measurement error is reduced.

Procedures

Average

Starting from the RECIPE procedure (Birnbaum, 1976), Zalinski (1984, 1987) presented a procedure suitable for the estimation of weights and scale values for the EAM and the complete DAM, providing reliable estimations with complete data from the full-factorial design when accompanied by all the sub-designs.

For each subject, the AVERAGE program (Zalinski & Anderson, 1986) generates an absolute weight and a scale value for each level of factors, and a single weight and a single scale value for the initial impression (complete DAM case). Nevertheless, it is also possible to let the program generate only one weight for each factor (EAM case).

The parameter estimates are obtained by adjusting iteratively the scale values to find out those that best fit the observed data by a least squares criterion. The iterative adjustments are handled by the STEPIT function (Chandler, 1969), a general algorithm for multivariate minimization that does not require derivatives, using only the function values.

R-*Average*

We have implemented the *R*-Average procedure[1], capable of providing reliable estimates for each trial as well for all the replications of a subject, both from the full factorial design and from the sub-designs. It could take simultaneously into account many replications to achieve a more reliable estimation.

Our first aim is to provide efficient parameter estimations. The weight-and-value parameters are estimated by minimizing the residual sum of squares of the non-linear model. This step is performed with the L-BFGS-B algorithm, implemented by Byrd, Lu, Nocedal, and Zhu (1995). We prefer this function for its native capability to deal with bounded parameters, as suggested by Zalinski and Anderson (1990). Reliable estimations and theoretically valid weights are provided especially when bounds are provided to the minimization function. This function provides the fundamental Residual Sum of Squares index (RSS), from which further goodness-of-fit indexes are computed: Adjusted *R*-square, Akaike Information Criteria (AIC; Akaike 1976), and Bayesian Information Criteria (BIC; Raftery, 1995). These last two indexes are computed starting from the logarithm of RSS weighted for the number of observations, and by penalizing the models with additional parameters. These indexes differ in the penalty function, since BIC imposes a larger penalty for complex models than AIC does. The RSS index is defined within the range of zero to infinity, where the zero value indicates no discrepancy between observed and estimated data. The AIC and BIC goodness-of-fit indexes are especially useful in model comparison, where the better model has the lower value of the index. A difference of $\Delta BIC > 2$ is generally accepted as a positive evidence for a difference between models.

Besides, our focal goal is to identify the most suitable model in the averaging models family. The choice of the optimal model is made according to the so-called "principle of parsimony". In this approach, given a speci-

[1] The *R*-Average procedure is implemented as a computer-library within the *R*-project (R Development Core Team, 2007). The *R*-Average library is well integrated in the *R* framework and it is specifically designed to easily manage data with several subjects and replications. The most time spending functions are optimized by calling external *C* code.

fied number of parameters, an information criterion is given by penalizing the models with additional parameters, following a selection criteria based on parsimony. Our aim is to identify which are the relevant weight parameters according to the overall goodness-of-fit indexes and the complexity of the design. That is, with a Bayesian approach (Burnham & Anderson, 2004), the function analyzes both these conditions: testing whether each single weight is important for the overall fit of the model and selecting the fundamental weights which can differ from the others. In fact, the best model is not the linear nor the full non-linear, rather the best is the "simplex" model which better explains the data by using the smallest number of different weight parameters.

To select the optimal averaging model, we use a method which represents a compromise between efficiency and performance. While it is possible to select the best model only considering the whole family of averaging models, we estimate an optimal model with the following algorithms.

In the first, we start the analysis considering EAM as a baseline. Step-by-step we change one single weight parameter looking for an improvement. We select this model for the next step whenever the evidence for this new model is at least positive. Different criteria can be set. According to the selection criterion, if the model with more parameters is better than the previous one, it is set as a new baseline. This procedure goes on iteratively until any improvement is found.

The second algorithm follows an alternative procedure. The baseline is represented by the complete DAM in which each weight differs from the others. Progressively, we reduce the number of different weights, i.e., the parameters, equalizing the weights that are not significantly different until any improvement is obtained by the procedure.

At last, we offer a third algorithm, which tests all the subsets of averaging models, considering each combination of DAM. This last procedure is very time-consuming and in general does not provide solutions different from those of the other two procedures.

Monte Carlo simulation for procedures comparison

We used a Monte Carlo simulation technique to investigate the properties of the averaging model parameter estimations with two numerical procedures. The aim of this analysis was to verify the capability of the estimated parameters to account accurately for the data, and to compare the properties of the two minimization algorithms for the averaging model: that implemented in the AVERAGE procedure and that used by *R*-Average.

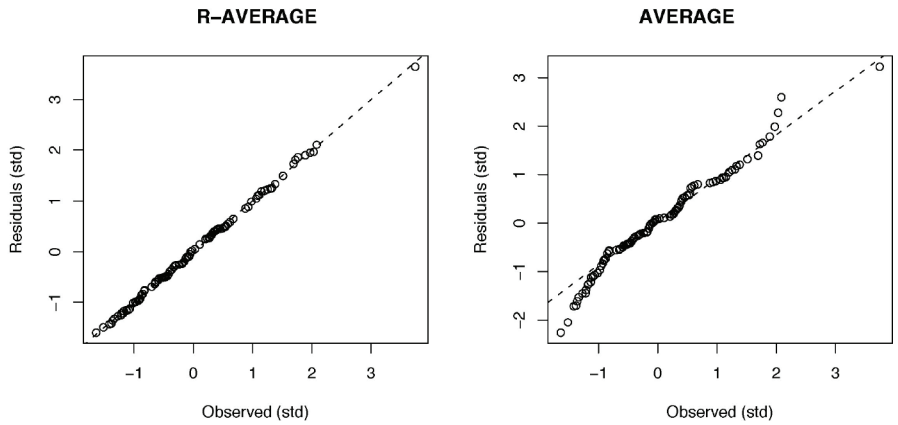

Figure 1. First simulation run (Errors *SD* = 1): QQ-Plot of the residuals, i.e., the difference between observed and expected data based on the parameters estimated by the *R*-Average and AVERAGE procedure.

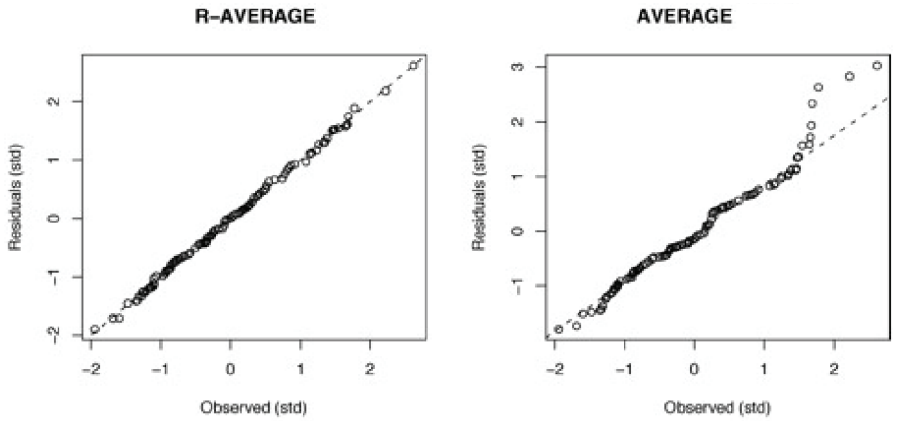

Figure 2. Second simulation run (Errors *SD* = 1.5): QQ-Plot of the residuals, i.e., the difference between observed and expected data based on the parameters estimated by the *R*-Average and AVERAGE procedure.

We follow a method similar to the one proposed by Zalinski (1987). That is, we use Monte Carlo techniques to estimate the averaging model parameters using simulated (i.e., synthetic) data from a standard factorial design. This simulation shows that accurate estimations of averaging model

parameters can be consistently obtained from realistically simulated experiments. Monte Carlo runs are realized by specifying the design size and true parameter values, by generating error-free data, by adding random normal error to simulate real data, and then by estimating the model parameters from these data.

We analyzed a 3 × 3 design in which two independent factors were compounded in accordance with the averaging model. For the error-free averaging model, the set of averaging model parameters is defined by the six values that represent the simulated subjects' responses, ranging from zero to 20, and by the six weights, ranging from one to three. The weight w_0 for the initial state was set to zero. Parameters are reported in Table 1. Random errors are obtained by generating two collections of independent normal random numbers (*SD* = 1 and *SD* = 1.5); each collection is defined by replication (*n* = 100) × parameters of normal random numbers.

Each simulation runs by estimating 100 separate sets of the averaging model parameters with both AVERAGE and *R*-Average. We estimated the parameters by fitting the model to the data of each replication using the AVERAGE program as indicated by Zalinski and Anderson (1990). Besides, we estimated the set of parameters with *R*-Average by fitting the complete DAM to each set of replications of each collection. The same set of bound constraints was used for the two different procedures.

To obtain a measure of data variability, we calculated the RSS provided by the set of real parameters. Then, we calculated the RSS and the other goodness-of-fit indexes provided by these two sets of parameters.

Figure 1 shows the normal QQ-plot of residuals for the first simulation run, in which the standard deviation of errors is unity (*SD* = 1). Analysis of residuals shows that *R*-Average produces unbiased estimates of the expected values for all the examined response conditions. AVERAGE appears to be more inaccurate in the parameter estimation not only for extreme observations but also for observations outside one standard deviation. Figures 2 shows similar findings for the other condition (error *SD* = 1.5).

As shown in Table 1, results show that the two functions offer similar performances in parameter estimation. However, *R*-Average estimated more reliable parameters with the lowest RSS and with the best goodness-of-fit indexes. All the weight and value parameters estimated with *R*-Average are within a range of 1% from the real value. The RSS is always smaller than the one provided for the real data, and the BIC indexes provide strong evidences for this procedure, showing the efficacy of the minimization algorithm that better accounts for the data variability. The weights-and-values and goodness-of-fit indexes get worse by using the STEPIT algorithm. Especially the RSS increases up to twice than data variability (RSS +84%).

Table 1A

	Values						Weights								
	Factor A			Factor B			Factor A			Factor B			RSS	BIC	AIC
	low	med	high	low	med	high	low	med	high	low	med	high			
Error-free parameters	2.0	8.0	18.0	1.0	9.0	18.5	1.0	3.0	2.0	3.0	1.0	2.0	1546.1	48.5	-50.8
Average															
Mean (sd)	0.2 (0.6)	8.4 (1.5)	19.7 (0.7)	1.0 (0.9)	8.6 (1.4)	16.7 (1.5)	1.0 (1.0)	3.1 (1.0)	3.0 (1.0)	4.4 (1.0)	2.9 (1.0)	3.0 (1.0)	2858.3	1101.0	1003.0
R-Average															
Mean (sd)	2.1 (1.0)	8.3 (0.9)	18.3 (0.9)	1.0 (0.8)	9.1 (0.9)	18.5 (0.9)	1.0 (1.8)	2.9 (3.3)	2.3 (2.5)	3.4 (3.5)	1.0 (1.6)	2.1 (2.9)	367.8	-1770.0	1868.0

Table 1B

	Values						Weights								
	Factor A			Factor B			Factor A			Factor B			RSS	BIC	AIC
	low	med	high	low	med	high	low	med	high	low	med	high			
Error-free parameters	2.0	8.0	18.0	1.0	9.0	18.5	1.0	3.0	2.0	3.0	1.0	2.0	3845.8	1506.5	1407.2
Average															
Mean (sd)	0.5 (1.4)	8.0 (2.1)	19.5 (1.2)	0.9 (1.1)	8.4 (1.8)	17.2 (1.9)	1.0 (0.9)	3.1 (0.9)	2.9 (0.8)	3.8 (0.9)	2.8 (0.9)	2.9 (0.8)	5059.0	1900.0	1803.0
R-Average															
Mean (sd)	2.1 (1.4)	8.0 (1.3)	18.2 (1.2)	1.1 (1.0)	9.0 (1.6)	18.6 (1.1)	1.0 (1.9)	3.2 (3.7)	2.4 (2.8)	3.5 (3.5)	1.1 (2.2)	2.1 (3.0)	1096.9	-240.0	-338.0

Table 1. Comparison of the estimated parameters and goodness of fit indexes for the two collections of data. The parameters of the averaging model estimated by AVERAGE and *R*-Average are compared with the error-free parameters. The RSS, BIC, and AIC indexes are indicated: the lower is the value of the index the better is the goodness of the estimated set of parameters. The results of the first simulation are reported in Table 1A (normal random error = 1 *SD*) and those of the second simulation in Table 1B (error = 1.5 *SD*).

Conclusion

The present study allows to conclude that the new procedure correctly estimates the weight and value parameters for the averaging models, as the AVERAGE procedure does. The proposed procedure has the capability to estimate better model parameters, especially coping with extreme data. We found that the previous procedure does not provide robust estimations when observations are outside one standard deviation from the model, although we have specified the parameter bounds in order to achieve a theoretically better estimation. For this reason, we consider the new procedure as a valid substitute to AVERAGE.

The difference in estimation between the two procedures can be alternatively explained. The minimization routine used for *R*-Average can be superior to STEPIT, which was at that time identified by Zalinski as the best among several minimization routines. A second possibility is that these results can be peculiar to this small 3×3 design, which estimates 12 parameters from 15 data points. Using larger designs, more constraints are provided and the two minimization functions could offer different performance.

R-Average provides several goodness-of-fit indexes in addition to the Chi-Square test; these indexes are especially useful for model comparison. Moreover, we have improved the capability to manage replications and to deal accurately with incomplete factorial designs. Besides, the proposed theoretical method permits the selection and estimation of the optimal number of parameters for the DAM.

An open issues concerns the theoretical validity and the correctness of the methodology for group analysis. This special issue was previously considered by Zalinski and Anderson (1990). Group analysis can cause bias in the estimated weights and scale values. Even if this is not a problem with the current simulation, since all replications had the same true parameters, it constitutes an actual problem with groups of real subjects because within-subject and between-subject variability can be confounded.

To better understand the statistical proprieties of the parameters estimated by these and other techniques, more analyses and data generations are needed, varying the numbers of factors and levels, the standard error, and the starting weight and value parameters. Incomplete responses to the full factorial design could also be considered.

References

Akaike, H. (1976). Canonical correlation analysis of time series and the use of an information criterion. In R. K. Mehra & D. G. Lainotis (Eds.), *System identification: Advances and case studies* (pp. 52-107). New York: Academic Press.

Anderson, N. H. (1965). Averaging versus adding as a stimulus combination rule in impression formation. *Journal of Experimental Psychology*, *70*, 394-400.

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

Anderson, N. H. (1982). *Methods of information integration theory*. New York: Academic Press.

Anderson, N. H. (2001). *Empirical direction in design and analysis*. Mahwah, NY: Erlbaum.

Birnbaum, M. H. (1976). Intuitive numerical prediction. *American Journal of Psychology*, *89*, 417-429

Burnham K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261-304.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *Journal Scientific Computing*, *16*, 1190-1208.

Chandler, J. P. (1969). STEPIT finds local minima of a smooth function of several parameters. *Behavioral Science*, *14*, 81-82.

Oden, G. C., & Anderson, N. H. (1971). Differential weighting in integration theory. *Journal of Applied Psychology*, *89*, 152-161.

R Development Core Team (2007). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111-163.

Wang, M., & Yang, J. (1998). A multi-criterion experimental comparison of three multi-attribute weight measurement methods. *Journal of Multicriteria Decision Analysis*, *7*, 340-350.

Zalinski, J. (1984). Parameter estimation: An estimated variance-covariance matrix for averaging model parameters in information integration theory. *Behavior Research Methods, Instruments, & Computer*, *16*, 557-558.

Zalinski, J. (1987). *Parameter estimation for the averaging model of information integration theory*. Unpublished doctoral dissertation, University of California, San Diego, La Jolla, CA.

Zalinski, J., & Anderson, N. H. (1986). *Average: A user-friendly fortran-77 program for parameter estimation for the averaging model of information integration theory* [computer software]. San Diego, CA.

Zalinski, J., & Anderson, N. H. (1990). Parameter estimation for averaging theory. In N. H. Anderson (Ed.), *Contributions to information integration theory. Vol. 1: Cognition* (pp. 353-394). Hillsdale, NJ: Erlbaum.

Abstract

Anderson and Zalinski proposed a procedure of parameter estimation for equal and complete differential averaging weight models. This procedure can analyze the data from one subject and one session at a time, and it is designed to allow for proper tests of statistical significance. In the present study we describe an alternative general procedure for parameter estimation and for model selection. It allows selecting an optimal number of parameters, providing reliable parameter estimation for averaging models. To test the goodness of the estimation, the procedure considers the AIC and the BIC validation criteria, which are reliable goodness-of-fit indices and can also be used for concurrent model comparison. We have compared the two procedures by a Monte Carlo simulation. Results show that these procedures offer similar performances in parameter estimation. However, analysis of residuals shows that the new procedure produces always unbiased estimates of the expected values for all the examined response conditions. This improvement could depend on the minimization routine.

Riassunto

Anderson e Zalinski hanno proposto una procedura per la stima dei parametri dei modelli Averaging con pesi uguali e completamente differenziati. Questa procedura analizza principalmente i dati raccolti da un soggetto e da una sessione per volta; è stata sviluppata anche per permettere un'adeguata verifica della significatività delle stime. Nel presente studio viene descritta una procedura alternativa per la stima dei parametri e per la selezione del modello. Questa procedura permette di selezionare il numero ottimale di parametri, fornendo una stima affidabile dei parametri per i modelli Averaging. Per verificare la bontà dei parametri stimati, la procedura impiega gli indici AIC e BIC come criteri di validità: tali indici di bontà di adattamento del modello ai dati possono essere ulteriormente utilizzati per la selezione del modello ottimale. Sono state poste a confronto le due procedure attraverso una simulazione con il metodo di Monte Carlo. I risultati mostrano come entrambi le procedure offrano risultati equivalenti nella stima dei parametri. Comunque, l'analisi dei residui mostra che la nuova procedura produce sempre stime corrette dei valori attesi per tutte le condizioni di risposta esaminate. Questo miglioramento potrebbe dipendere dalla routine di minimizzazione.

Acknowledgments. We thank Prof. Norman H. Anderson for useful comments.

Addresses. Giulio Vidotto (giulio.vidotto@unipd.it), Marco Vicentini (marco.vicentini@gmail.com), University of Padua, Department of General Psychology, via Venezia 8, I-35131 Padova.