Analysis of variance[*]

Andrew Gelman[†]

February 25, 2005

Abstract

Analysis of variance (ANOVA) is a statistical procedure for summarizing a classical linear model—a decomposition of sum of squares into a component for each source of variation in the model—along with an associated test (the *F-test*) of the hypothesis that any given source of variation in the model is zero. When applied to generalized linear models, multilevel models, and other extensions of classical regression, ANOVA can be extended in two different directions. First, the F-test can be used (in an asymptotic or approximate fashion) to compare nested models, to test the hypothesis that the simpler of the models is sufficient to explain the data. Second, the idea of variance decomposition can be interpreted as inference for the variances of batches of parameters (sources of variation) in multilevel regressions.

1 Introduction

Analysis of variance (ANOVA) represents a set of models that can be fit to data, and also a set of methods that can be used to summarize an existing fitted model. We shall first consider ANOVA as it applies to classical linear models (the context for which it was originally devised; Fisher, 1925) and then discuss how ANOVA has been extended to generalized linear models and multilevel models. Analysis of variance is particularly effective tool for analyzing highly structured experimental data (in agriculture, multiple treatments applied to different batches of animals or crops; in psychology, multi-factorial experiments manipulating several independent experimental conditions and applied to groups of people; industrial experiments in which multiple factors can be altered at different times and in different locations).

2 ANOVA for classical linear models

2.1 ANOVA as a family of statistical methods

When formulated as a statistical model, analysis of variance refers to an additive decomposition of data into a grand mean, main effects, possible interactions, and an error term. For example, Gawron et al. (2003) describe a flight-simulator experiment that we summarize as a 5 × 8 array of measurements under 5 treatment conditions and 8 different airports. The corresponding two-way ANOVA model is $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$. The data as described here have no replication, and so the two-way interaction becomes part of the error term.[1]

[*] For the New Palgrave Dictionary of Economics, second edition. We thank the National Science Foundation for financial support.

[†] Department of Statistics and Department of Political Science, Columbia University, New York, gelman@stat.columbia.edu, www.stat.columbia.edu/∼gelman

[1] If, for example, each treatment × airport condition were replicated three times, then the 120 data points could be modeled as $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$, with two sets of main effects, a two-way interaction, and an error term.

	Degrees of	Sum of	Mean		
Source	freedom	squares	square	F-ratio	p-value
Treatment	4	0.078	0.020	0.39	0.816
Airport	7	3.944	0.563	11.13	< 0.001
Residual	28	1.417	0.051		

Figure 1: *Classical two-way analysis of variance for data on 5 treatments and 8 airports with no replication. The treatment-level variation is not statistically distinguishable from noise, but the airport effects are statistically significant. This and the other examples in this article come from Gelman (2005) and Gelman and Hill (2006).*

This is a linear model with $1+4+7$ coefficients, which is typically identified by constraining the $\sum_{i=1}^{5} \alpha_i = 0$ and $\sum_{j=1}^{8} \beta_j = 0$. The corresponding ANOVA display is shown in Figure 1. We shall describe each column in turn:

- For each source of variation, the degrees of freedom represent the number of effects at that level, minus the number of constraints (the 5 treatment effects sum to zero, the 5 airport effects sum to zero, and each row and column of the 40 residuals sums to zero).
- The total sum of squares—that is, $\sum_{i=1}^{5} \sum_{j=1}^{8} (y_{ij} - \bar{y}_{..})^2$—is $0.078 + 3.944 + 1.417$, which can be decomposed into these three terms corresponding to variance described by treatment, variance described by airport, and residuals.
- The mean square for each row is the sum of squares divided by degrees of freedom. Under the null hypothesis of zero row and column effects, their mean squares would, in expectation, simply equal the mean square of the residuals.
- The F-ratio for each row (excluding the residuals) is the mean square, divided by the residual mean square. This ratio should be approximately 1 (in expectation) if the corresponding effects are zero; otherwise we would generally expect the F-ratio to exceed 1. We would expect the F-ratio to be less than 1 only in unusual models with negative within-group correlations (for example, if the data y have been renormalized in some way, and this had not been accounted for in the data analysis.)
- The p-value gives the statistical significance of the F-ratio with reference to the F_{ν_1,ν_2}, where ν_1 and ν_2 are the numerator and denominator degrees of freedom, respectively. (Thus, the two F-ratios in Figure 1 are being compared to $F_{4,28}$ and $F_{7,28}$ distributions, respectively.) In this example, the treatment mean square is lower than expected (an F-ratio of less than 1), but the difference from 1 is not statistically significant (a p-value of 82%), hence it is reasonable to judge this difference as explainable by chance, and consistent with zero treatment effects. The airport mean square, is much higher than would be expected by chance, with an F-ratio that is highly statistically-significantly larger than 1; hence we can confidently reject the hypothesis of zero airport effects.

More complicated designs will have correspondingly complicated ANOVA models, and complexities arise with multiple error terms. Figure 2 illustrates in the context of an agricultural experiment in which five different treatments (A,B,C,D,E) were applied to a 5×5 grid of plots in a latin-square[2], and then treatment varieties 1,2 were randomly applied to 2 sub-plots within each main plot. The ANOVA table is divided into main-plot and sub-plot effects, with different residual mean squares for each part. We do not intend to explain such hierarchical designs and analyses here, but we wish to alert the reader to such complications. Textbooks such as Snedecor and Cochran (1989) and Kirk (1995) provide examples of analysis of variance for a wide range of designs.

[2] See, for example, Cochran and Cox (1957), for descriptions and motivations of latin squares and other factorial and fractional-factorial experimental designs.

Source	Degrees of freedom	Sum of squares	Mean square	F-ratio	p-value
Row	4	288.5	72.1	2.81	0.074
Column	4	389.5	97.4	3.79	0.032
A,B,C,D,E	4	702.3	175.6	6.84	0.004
Main-plot residual	12	308.0	25.67		
1,2	1	332.8	332.8	18.59	0.001
Row × 1,2	4	74.1	18.5	1.03	0.429
Column × 1,2	4	96.7	24.2	1.35	0.308
A,B,C,D,E× 1,2	4	57.1	14.3	0.80	0.550
Sub-plot residual	12	214.8	17.9		

Figure 2: *Classical split-plot analysis of variance for an experiment with a latin square design of 5 treatments (A,B,C,D,E) applied on a 5×5 grid of plots, each divided into 2 sub-plots which are randomly assigned the treatment varieties 1,2. In computing the F-ratios and p-values, the main-plot and sub-plot effects are compared to the main-plot and sub-plot residuals, respectively.*

2.2 ANOVA to summarize a model that has already been fitted

We have just demonstrated ANOVA as a method of analyzing highly structured data by decomposing variance into different sources, and comparing the explained variance at each level to what would be expected by chance alone. Any classical analysis of variance corresponds to a linear model (that is, a regression model, possibly with multiple error terms as in Figure 2); conversely, ANOVA tools can be used to summarize an existing linear model.

The key is the idea of "sources of variation," each of which corresponds to a batch of coefficients in a regression. Thus, with the model $y = X\beta + \epsilon$, the columns of X can often be batched in a reasonable way (for example, from the previous section, a constant term, 4 treatment indicators, and 7 airport indicators), and the mean squares and F-tests then provide information about the amount of variance explained by each batch.

The same idea occurs in hierarchical models: the split-plot latin-square experiment considered earlier has a model of the form, $y = X\beta + \eta + \epsilon$, with main-plot and sub-plot error terms, and columns of X that can be batched into four row indicators, four column indicators, four main-plot treatment indicators, an indicator for the sub-plot treatment, and the sub-plot treatment interacted with rows, columns, and main-plot treatment.

Our point here is that such models could be fit without any reference to ANOVA, but ANOVA tools could then be used to make some sense of the fitted models, and to test hypotheses about batches of coefficients.

2.3 Balanced and unbalanced data

In general, the amount of variance explained by a batch of predictors in a regression depends on which other variables have already been included in the model. With *balanced data*, however, in which all groups have the same number of observations (for example, each treatment applied exactly eight times, and each airport used for exactly five observations), the variance decomposition does not depend on the order in which the variables are entered. ANOVA is thus particularly easy to interpret with balanced data. The analysis of variance can also be applied to unbalanced data, but then the sums of squares, mean squares, and F-ratios will depend on the order in which the sources of variation are considered.

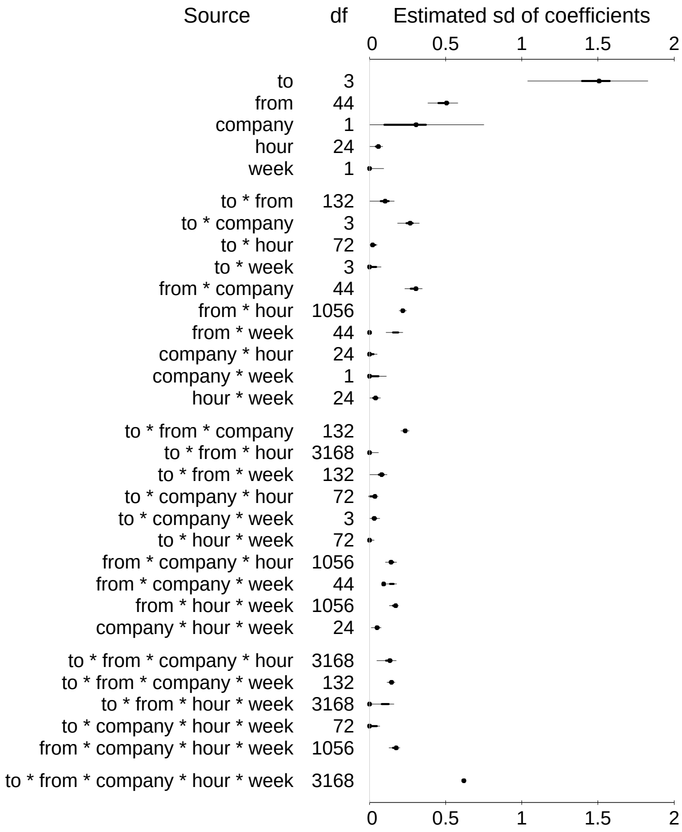

Figure 3: *ANOVA display for a five-way factorial dataset. The bars indicate 50% and 95% intervals for the standard deviation of the batch of coefficients at each level (source of variation) of the model. The display makes apparent the magnitudes and uncertainties of the different components of variation. Since the data are on the logarithmic scale, the standard deviation parameters can be interpreted directly. For example,* $s_m = 0.20$ *corresponds to a coefficient of variation of* $\exp(0.2) - 1 \approx 0.2$ *on the original scale, and so the exponentiated coefficients* $\exp(\beta_j^{(m)})$ *in this batch correspond to multiplicative increases or decreases in the range of 20%. (The dots on the bars show simple classical estimates of the variance components that were used as starting points in the Bayesian simulation.)*

3 ANOVA for more general models

Analysis of variance represents a way of summarizing regressions with large numbers of predictors that can be arranged in batches, and a way of testing hypotheses about batches of coefficients. Both these ideas can be applied in settings more general than linear models with balanced data.

3.1 F tests

In a classical balanced design (as in the examples of the previous section), each F-ratio compares a particular batch of effects to zero, testing the hypothesis that this particular source of variation is not necessary to fit the data.

More generally, the F test can be considered as a comparison of two nested models, testing the hypothesis that the smaller model fits the data adequately and (so that the larger model is unnecessary). In a linear model context, the F-ratio is $\frac{(\mathrm{SS}_2 - \mathrm{SS}_1)/(\mathrm{df}_2 - \mathrm{df}_1)}{\mathrm{SS}_1/\mathrm{df}_1}$, where SS_1, df_1 and

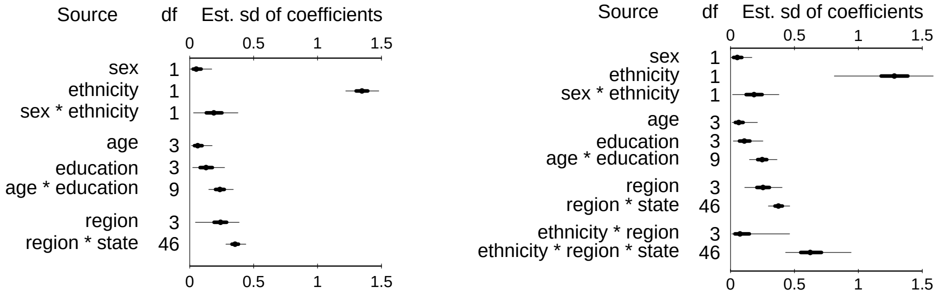

Figure 4: *ANOVA display for two logistic regression models of the probability that a survey respondent prefers the Republican candidate for the 1988 U.S. Presidential election, based on data from seven CBS News polls. Point estimates and error bars show median estimates, 50% intervals, and 95% intervals of the standard deviation of each batch of coefficients. The large coefficients for ethnicity, region, and state suggest that it might make sense to include interactions, hence the inclusion of ethnicity* × *region and ethnicity* × *state interactions in the second model.*

SS_2, df_1 are the residual sums of squares and degrees of freedom from fitting the larger and smaller models, respectively.

For generalized linear models, formulas exist using the *deviance* (the log-likelihood multiplied by -2) that are asymptotically equivalent to F-ratios. In general, such models are not balanced, and the test for including another batch of coefficients depends on which other sources of variation have already been included in the model.

3.2 Inference for variance parameters

A different sort of generalization interprets the ANOVA display as inference about the variance of each batch of coefficients, which we can think of as the relative importance of each source of variation in predicting the data.

Even in a classical balanced ANOVA, the sums of squares and mean squares do not exactly do this, but the information contained therein can be used to estimate the variance components (Cornfield and Tukey, 1956, Searle, Casella, and McCulloch, 1992). Bayesian simulation can then be used to obtain confidence intervals for the variance parameters. Figure 3 illustrates for a dataset of logarithms of internet connect times, classified by five different variables, each with two or more levels. We display inferences for standard deviations (rather than variances) because these are more directly interpretable. Compared to the classical ANOVA display, this plot emphasizes the estimated variance parameters rather than testing the hypothesis that they are zero.

3.3 Generalized linear models

The idea of estimating variance parameters applies directly to generalized linear models as well as unbalanced datasets. All that is needed is that the parameters of a regression model are batched into "sources of variation." Figure 4 illustrates with a multilevel logistic regression model, predicting vote preference given a set of demographic and geographic variables.

3.4 Multilevel models and Bayesian inference

Analysis of variance is closely tied to multilevel (hierarchical) modeling, with each source of variation in the ANOVA table corresponding to a variance component in a multilevel model (see Gelman, 2005). In practice, this can mean that we perform ANOVA by fitting a multilevel model, or that we use ANOVA ideas to summarize multilevel inferences. Multilevel modeling is inherently Bayesian

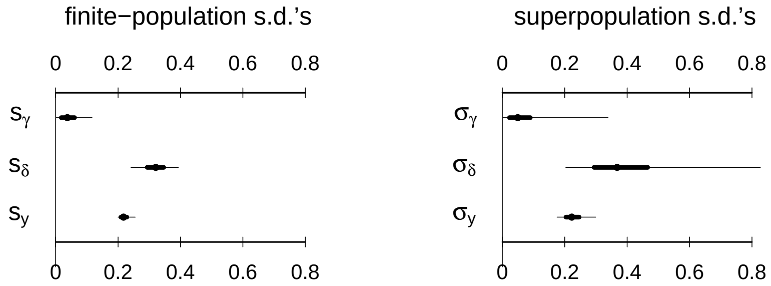

Figure 5: *Median estimates, 50% intervals, and 95% intervals for (a) finite-population and (b) superpopulation standard deviations of the treatment-level, airport-level, and data-level errors in the flight-simulator example from Figure 1. The two sorts of standard deviation parameters have essentially the same estimates, but the finite-population quantities are estimated much more precisely. (We follow the general practice in statistical notation, using Greek and Roman letters for population and sample quantities, respectively.)*

in that it involves a potentially large number of parameters that are modeled with probability distributions (see, for example, Goldstein, 1995, Kreft and De Leeuw, 1998, Snijders and Bosker, 1999). The differences between Bayesian and non-Bayesian multilevel models are typically minor except in settings with many sources of variation and little information on each, in which case some benefit can be gained from a fully-Bayesian approach which models the variance parameters.

4 Related topics

4.1 Finite-population and superpopulation variances

So far in this article we have considered, at each level (that is, each source of variation) of a model, the standard deviation of the corresponding set of coefficients. We call this the *finite-population* standard deviation. Another quantity of potential interest is the standard deviation of the hypothetical *superpopulation* from which these particular coefficients were drawn. The point estimates of these two variance parameters are similar—with the classical method of moments, the estimates are identical, because the superpopulation variance is the expected value of the finite-population variance—but they will have different uncertainties. The inferences for the finite-population standard deviations are more precise, as they correspond to effects for which we actually have data.

Figure 5 illustrates the finite-population and superpopulation inferences for the standard deviations at each level of the model for the flight-simulator example. We know much more about the 5 treatments and 8 airports in our dataset than for the general populations of treatments and airports. (We similarly know more about the standard deviation of the 40 particular errors in out dataset than about their hypothetical superpopulation, but the differences here are not so large, because the superpopulation distribution is fairly well estimated from the 28 degrees of freedom available from these data.)

There has been much discussion about fixed and random effects in the statistical literature (see Eisenhart, 1947, Green and Tukey, 1960, Plackett, 1960, Yates, 1967, LaMotte, 1983, and Nelder, 1977, 1994, for a range of viewpoints), and unfortunately the terminology used in these discussions is incoherent (see Gelman, 2005, Section 6). Our resolution to some of these difficulties is to always fit a multilevel model but to summarize it with the appropriate class of estimand—superpopulation or finite-population—depending on the context of the problem. Sometimes we are interested in the particular groups at hand; other times they are a sample from a larger population of interest. A change of focus should not require a change in the model, only a change in the inferential summaries.

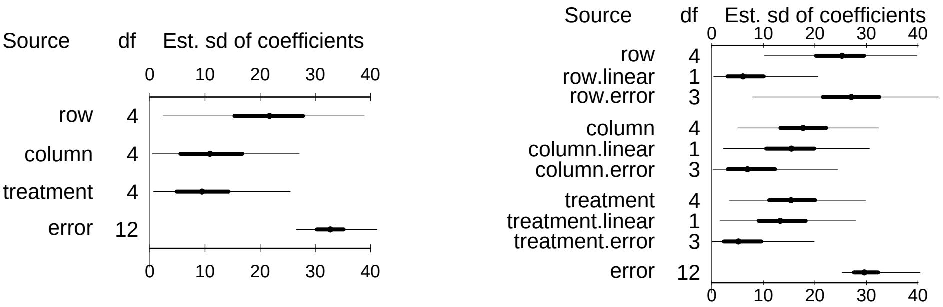

Figure 6: *ANOVA displays for a* 5 × 5 *latin square experiment (an example of a crossed three-way structure): (a) with no group-level predictors, (b) contrast analysis including linear trends for rows, columns, and treatments. See also the plots of coefficient estimates and trends in Figure 7.*

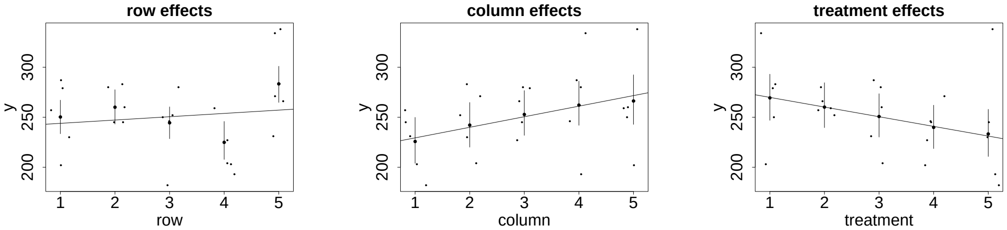

Figure 7: *Estimates* ±1 *standard error for the row, column, and treatment effects for the latin square experiment summarized in Figure 6. The five levels of each factor are ordered, and the lines display the estimated linear trends.*

4.2 Contrast analysis

Contrasts are a way to structuring the effects within a source of variation. In a multilevel modeling context, a contrast is simply a group-level coefficient. Introducing contrasts into an ANOVA allows a further decomposition of variance. Figure 6 illustrates for a 5 × 5 latin square experiment (this time, not a split plot): the left plot in the figure shows the standard ANOVA, and the right plot shows a contrast analysis including linear trends for the row, column, and treatment effects. The linear trends for the columns and treatments are large, explaining most of the variation at each of these levels, but there is no evidence for a linear trend in the row effects.

Figure 7 shows the estimated effects and linear trends at each level (along with the raw data from the study), as estimated from a multilevel model. This plot shows in a different way that the variation among columns and treatments, but not among rows, is well explained by linear trends.

4.3 Nonexchangeable models

In all the ANOVA models we have discussed so far, the effects within any batch (source of variation) are modeled exchangeably, as a set of coefficients with mean 0 and some variance. An important direction of generalization is to nonexchangeable models, such as in time series, spatial structures (Besag and Higdon, 1999), correlations that arise in particular application areas such as genetics (McCullagh, 2005), and dependence in multi-way structures (Aldous, 1981, Hodges et al., 2005). In these settings, both the hypothesis-testing and variance-estimating extensions of ANOVA become more elaborate. The central idea of clustering effects into batches remains, however. In this sense, "analysis of variance" represents all efforts to summarize the relative importance of different components of a complex model.

References

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis* **11**, 581–598.

Besag, J., and Higdon, D. (1999). Bayesian analysis of agricultural field experiments (with discussion). *Journal of the Royal Statistical Society B* **61**, 691–746.

Cochran, W. G., and Cox, G. M. (1957). *Experimental Designs*, second edition. New York: Wiley.

Cornfield, J., and Tukey, J. W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics* **27**, 907–949.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics* **3**, 1–21.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Gawron, V. J., Berman, B. A., Dismukes, R. K., and Peer, J. H. (2003). New airline pilots may not receive sufficient training to cope with airplane upsets. *Flight Safety Digest* (July–August), 19–32.

Gelman, A. (2005). Analysis of variance: why it is more important than ever (with discussion). *Annals of Statistics* **33**, 1–53.

Gelman, A., and Hill, J. (2006). *Applied Regression and Multilevel (Hierarchical) Models.* Cambridge University Press.

Gelman, A., Pasarica, C., and Dodhia, R. M. (2002). Let's practice what we preach: using graphs instead of tables. *The American Statistician* **56**, 121–130.

Goldstein, H. (1995). *Multilevel Statistical Models*, second edition. London: Edward Arnold.

Green, B. F., and Tukey, J. W. (1960). Complex analyses of variance: general problems. *Psychometrika* **25** 127–152.

Hodges, J. S., Cui, Y., Sargent, D. J., and Carlin, B. P. (2005). Smoothed ANOVA. Technical report, Department of Biostatistics, University of Minnesota.

Kirk, R. E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, third edition. Brooks/Cole.

Kreft, I., and De Leeuw, J. (1998). *Introducing Multilevel Modeling.* London: Sage.

LaMotte, L. R. (1983). Fixed-, random-, and mixed-effects models. In *Encyclopedia of Statistical Sciences*, ed. S. Kotz, N. L. Johnson, and C. B. Read, **3**, 137–141.

McCullagh, P. (2005). Discussion of Gelman (2005). *Annals of Statistics* **33**, 33–38.

Nelder, J. A. (1977). A reformulation of linear models (with discussion). *Journal of the Royal Statistical Society A* **140**, 48–76.

Nelder, J. A. (1994). The statistics of linear models: back to basics. *Statistics and Computing* **4**, 221–234.

Plackett, R. L. (1960). Models in the analysis of variance (with discussion). *Journal of the Royal Statistical Society B* **22**, 195–217.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components.* New York: Wiley.

Snedecor, G. W., and Cochran, W. G. (1989). *Statistical Methods*, eighth edition. Iowa State University Press.

Snijders, T. A. B., and Bosker, R. J. (1999). *Multilevel Analysis.* London: Sage.

Yates, F. (1967). A fresh look at the basic principles of the design and analysis of experiments. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **4**, 777–790.