

Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

Contributions of J.N.K. Rao to Complex Survey Multilevel Models and Composite Likelihood

by Mary E. Thompson

Release date: September 8, 2025



Statistics
Canada

Statistique
Canada

Canada

Contributions of J.N.K. Rao to Complex Survey Multilevel Models and Composite Likelihood

Mary E. Thompson¹

Abstract

In the setting of multilevel models to be estimated using data from surveys with complex sampling designs, this paper outlines some contributions of the landmark paper by Rao, Verret and Hidiroglou (Survey Methodology, 2013) and subsequent related work.

Key Words: Complex sampling design; Multistage sampling; Analytic use of survey data; Variance components; Estimating function system.

1. Introduction: Multilevel Models and Variance Components

One of the many directions of Professor J. N. K. Rao's research concerns his work on multilevel models and composite likelihood. This article will discuss his 2013 *Survey Methodology* paper with Verret and Hidiroglou and its extensions.

What is a multilevel model? A classic example (see for example Singer, 1998) is the *MATHACH* [*math achievement*] model, where the elementary units of the population are secondary school students in a certain grade in some jurisdiction, and the response variable Y represents a suitable transformation of their scores on a standardized mathematics test. In (1) below, i indexes their school and j indexes the students within a school. The independent variable \mathbf{x}_{ij} might be something like the row vector $(1, SES_{ij})$, where SES is a socioeconomic status variable. The model specification can be written as

$$Y_{ij} | \mathbf{x}_{ij}, \mathbf{V}_i = \mathbf{v}_i :_{ind} N(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{x}_{ij}^T \mathbf{v}_i, \sigma_e^2), \quad (1)$$

$$\mathbf{V}_i :_{iid} N_p(0, \boldsymbol{\Sigma}_v). \quad (2)$$

The school is the higher level, sometimes called the "level 2" unit, while the student is the lower level, sometimes called a "level 1" unit. The $\boldsymbol{\beta}$ coefficient vector would include an intercept and be a so-called *fixed effect*, while \mathbf{V}_i denotes a school-level *random effect* of the same or a lesser dimension that is multivariate normally distributed.

The Ministry of Education would be interested in the $\boldsymbol{\beta}$ parameters, and how well the system is doing overall. If they wanted means for individual schools they could get estimates of the $\boldsymbol{\beta}$ parameters and the random effects \mathbf{v}_i , which would give them small area mean estimates.

However, one of the main aims of multilevel modeling is to assess variation, for example estimating how much of the variation in scores is due to variation between schools. If we suppose for simplicity that $\boldsymbol{\Sigma}_v$ is diagonal, then the relevant objects of inference are the variances of the \mathbf{V} components, i.e. "school level variation", and how much they contribute to the overall variation in Y – in particular whether they would be very small in comparison with the student level variance σ_e^2 .

¹Mary E. Thompson, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1

The problem has points of difficulty: in particular, if the magnitudes of the variances on the diagonal of Σ_v are very small it may be difficult to estimate them well.

The math achievement example is formulated in terms of a hypothetical population model, with no sampling of units involved, but there are corresponding cases in survey populations. Consider a national health survey with an area sampling design, where Y is a suitable measure of Body Mass Index (BMI) for a certain subpopulation, and the variability of Y at the area level as compared with individual level variation is of interest.

The classic multilevel model is Gaussian, but what if the Y variable is not continuous but binary, or continuous but not transformable to something Gaussian? Consider a sample of votes cast in a probability sample of polling stations in a state part way through election day, where the main problem is to forecast the final election result in that state. A supplementary question might be to assess whether the variation in polling station results reflects underlying local effects. Let Y_{ij} be the indicator of vote for candidate A by voter j at polling station i . A model for the binary variable Y_{ij} could be

$$Prob(Y_{ij} = 1 | \mathbf{x}_{ij}, \mathbf{V}_i = \mathbf{v}_i) = \text{expit}(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{x}_{ij}^T \mathbf{v}_i), \mathbf{V}_i :_{iid} N_p(0, \boldsymbol{\Sigma}_v).$$

Multi-level modelling became increasingly popular in epidemiology and the social sciences with the democratization of statistical computing from about the mid-1990s onward. This was also a time of increasing development of complex survey software and more computer-intensive approaches to analysis. A typical solution for analytic point estimation from a survey sample was to consider a census-based estimating equation (or system) for a superpopulation estimand, assuming the survey population to be generated by SRS from the superpopulation; then estimate that estimating equation (or system) from the sample using inflation survey weights and solve the resulting sample estimating equation, to obtain a design-consistent estimator of the population level estimator of a model parameter. The *pseudo-likelihood method*, put forward by Binder (1983) and formalized by Skinner (1989), was a straightforward extension. This approach turned out to be problematic for variance component estimation in a two-level problem because, although it is natural to think of the number of higher level units in the sample becoming large, it is appropriate in many applications to think of the number of lower level units in each higher level unit as being fixed. In that asymptotic framework, the basic method yields a bias that does not go away as the number of higher level units grows. Thus a problem identified early on (see e.g. Pfeffermann et al., 1998) was to formulate an appropriate framework for design-consistency of survey-based estimators of the variance components. The first techniques proposed were ways of rescaling the survey weights, particularly at the higher level.

In September 2004 there took place an opening meeting of a program on statistics and the social sciences at SAMSI. Survey methods was one of the subthemes, and several developers of statistical software for social sciences were represented, reporting adaptations to complex survey data for Mplus, Stata and other platforms. The question receiving most attention was how to handle multilevel modelling, and that turned out to be a major focus for the survey sampling group in the SAMSI project year 2004-2005.

Stata GLLAMM – generalized linear latent and mixed models – was being developed by Sophia Rabe-Hesketh, and she and Anders Skrondal tackled the problem of estimating parameters of multi-level logistic regression parameters, including using a pseudo-likelihood method when the response variable was binary. In their 2006 *JRSS-B* paper they evaluated various weight rescaling methods, and in the end concluded that with small sizes of higher level units, even with weight rescaling, the pseudo-maximum-likelihood method should be used with caution.

2. The Paper by Rao, Verret and Hidiroglou (2013)

In 2010 at the Statistical Society of Canada meeting at Université Laval, Professor Rao unveiled a new approach, based on estimation of estimating functions, and this eventually became the paper by Rao-Verret-Hidiroglou (2013), or RVH. This method and theoretical framework would extend greatly the set of scenarios where consistent estimation of variance components was possible. It was a very exciting moment.

2.1 Model and formulae

From that paper, here is the general two-level superpopulation model, at the finite population level; there are general densities for Y given \mathbf{x} and \mathbf{V} , and for \mathbf{V} itself:

$$Y_{ij} | \mathbf{x}_{ij}, \mathbf{V}_i = \mathbf{v}_i : \text{ind } f(y_{ij} | \mathbf{x}_{ij}, \mathbf{v}_i, \boldsymbol{\theta}_1), j = 1, \dots, M_i$$

and

$$\mathbf{V}_i : \text{ind } f(\mathbf{v}_i | \boldsymbol{\theta}_2), i = 1, \dots, N.$$

Standard methods for multi-level models that ignore the sampling design and assume that this model holds for the sample can lead to asymptotically biased estimators of the model parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.

Suppose a two stage sample is taken with unequal probability sampling, where the first stage is at the higher level and the second stage is at the lower level.

Special cases of a multilevel model with the same units are the simple nested error mean (Gaussian) model where the V_i are real-valued, and the parameters are μ , σ_e^2 and σ_v^2 :

$$Y_{ij} | v_i : \text{ind } N(\mu + v_i, \sigma_e^2), V_i : \text{iid } N(0, \sigma_v^2)$$

$$\boldsymbol{\theta} = (\sigma_e^2, \sigma_v^2);$$

and the linear two-level model, where the intercept and other regression coefficients have a random component with variance-covariance matrix $\boldsymbol{\Sigma}_v$, σ_e^2 is the residual variance, and T denotes transpose:

$$Y_{ij} | \mathbf{x}_{ij}, \mathbf{v}_i : \text{ind } N(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{x}_{ij}^T \mathbf{v}_i, \sigma_e^2), \mathbf{V}_i : \text{iid } N_p(0, \boldsymbol{\Sigma}_v)$$

$$\boldsymbol{\theta} = (\sigma_e^2, \boldsymbol{\Sigma}_v);$$

and two-level logistic regression, where again the regression coefficient has a random component with variance-covariance matrix $\boldsymbol{\Sigma}_v$:

$$\text{Prob}(Y_{ij} = 1 | \mathbf{x}_{ij}, \mathbf{v}_i) : \text{ind } \text{expit}(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{x}_{ij}^T \mathbf{v}_i), \mathbf{V}_i : \text{iid } N_p(0, \boldsymbol{\Sigma}_v)$$

$$\boldsymbol{\theta} = \boldsymbol{\Sigma}_v.$$

Suppose the following sampling scheme: at the first stage, n higher level units are selected with inclusion probabilities π_i , and at the second stage, m_i lower level units are selected with inclusion probabilities. Suppose the sampling scheme is informative, in the sense that the sampling or inclusion probabilities depend on the distribution of Y given \mathbf{x} . Can we formulate maximum likelihood estimating equations at the population level, and replace population sums at the two levels with sample sums, weighted by inflation weights, and obtain consistent estimators?

As in Section 1, let Y_{ij} denote the response variable for second-stage unit j in first-stage unit i for $i = 1, \dots, n$, and $j = 1, \dots, m$. We use lower case letter y_{ij} to represent realized values of Y_{ij} . Let $\mathbf{y}(n) = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ denote the sample data with $\mathbf{y}_i = (y_{i1}, \dots, y_{im})^T$ for $i = 1, \dots, n$.

As already noted, what had been done up to that point in statistical software was to apply rescaling to the usual inflation weights, particularly at the second stage, so that the software would not be deceived into thinking that the sample second stage sample sizes were much larger than they actually were. See Pfeffermann et al. (1998); Asparouhov (2006); and Grilli and Pratesi (2004) and Rabe-Hesketh and Skrondal (2006) for the context of binary models.

Professor Rao's idea was to start with linear estimating functions for the both the means *and* the variance component parameters. The system for point estimation for the simple nested error mean model is given below, where the terms are u_1 , u_2 and u_3 , and the estimating function equations with inverse inclusion and joint inclusion probability weights follow. In asymptotic frameworks where the sample sizes at the lower level remain fixed, while the number of higher

level units tends to infinity, the estimators are design-model consistent for θ and design-consistent for its census-based estimate θ_N .

Let us consider point estimation for the simple nested error mean model. The elementary estimating functions are:

- $u_1(y_{ij}, \boldsymbol{\theta}) = y_{ij} - \mu$
- $u_2(y_{ij}, \boldsymbol{\theta}) = (y_{ij} - \mu)^2 - (\sigma_v^2 + \sigma_e^2)$
- $u_3(y_{ij}, y_{ik}, \boldsymbol{\theta}) = [(y_{ij} - \mu) - (y_{ik} - \mu)]^2 - 2\sigma_e^2, j \neq k.$

The design-weighted estimating equations (WEE) for the model parameters are given by:

$$\begin{aligned} \sum_{i \in \mathcal{S}} w_i \sum_{j \in \mathcal{S}(i)} w_{ji} u_1(y_{ij}, \boldsymbol{\theta}) &= 0 \\ \sum_{i \in \mathcal{S}} w_i \sum_{j \in \mathcal{S}(i)} w_{ji} u_2(y_{ij}, \boldsymbol{\theta}) &= 0 \\ \sum_{i \in \mathcal{S}} w_i \sum_{j < k \in \mathcal{S}(i)} w_{jki} u_3(y_{ij}, y_{ik}, \boldsymbol{\theta}) &= 0, \end{aligned}$$

where the last sum is a sum over pairs (j, k) of second stage units in cluster i . Estimators from these estimating equations are design-model consistent for $\boldsymbol{\theta}$ and design-consistent for $\boldsymbol{\theta}_N$.

3. The Composite Likelihood Formulation

It was soon realized that these method-of-moment-like estimating functions at the population level were examples of composite likelihood estimation, an approach that would be applicable to both linear and generalized linear models. Pairwise composite likelihood functions can take several different forms, one of which is given below, where the inner sum is a double sum over pairs of units within a higher level unit, and the outer sum is taken *over* the higher level units. The function f is the marginal joint density of y_{ij} and y_{ik} in the same upper level unit i in the two-level model:

$$l_c(\boldsymbol{\theta}) = \sum_{i=1}^N \sum_{j < k=1}^{M_i} \log f(y_{ij}, y_{ik} | \boldsymbol{\theta})$$

where

$$f(y_{ij}, y_{ik} | \boldsymbol{\theta}) = \int f_{y|v}(y_{ij} | \mathbf{x}_{ij}, \mathbf{v}_i) f_{y|v}(y_{ik} | \mathbf{x}_{ik}, \mathbf{v}_i) f_v(\mathbf{v}_i) d\mathbf{v}_i$$

is the marginal joint density of y_{ij} and y_{ik} .

The sample (design-weighted) version of the log pairwise likelihood is given here:

$$l_{wc}(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}} w_i \sum_{j < k \in \mathcal{S}(i)} w_{jki} \log f(y_{ij}, y_{ik} | \boldsymbol{\theta}).$$

To perform point estimation, we solve the weighted composite score equations

$$\hat{U}_{wc}(\boldsymbol{\theta}) = \partial l_{wc}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} = \mathbf{0}$$

to obtain a weighted composite likelihood estimator $\hat{\boldsymbol{\theta}}_{wc}$.

One advantage of this formulation was to make it possible to extend the technique beyond the two-level linear model to the two-level generalized linear model. Professor Rao worked with collaborators Grace Yi and Haocheng Li on a general framework, resulting a paper that appeared in *Statistica Sinica* in 2016. A marginal census pairwise composite likelihood was formed by multiplying together marginal density terms for lower level individuals and marginal joint density terms for pairs, ignoring the dependence that comes through the higher level units i . Let

$$L_{ij} = \int f_{y_{ij}|v}(y_{ij} | \mathbf{x}_{ij}, \mathbf{v}_i) f_v(\mathbf{v}_i) d\mathbf{v}_i,$$

$$L_{ijk} = \int f_{y_{ij}|v}(y_{ij} | \mathbf{x}_{ij}, \mathbf{v}_i) f_{y_{ik}|v}(y_{ik} | \mathbf{x}_{ik}, \mathbf{v}_i) f_v(\mathbf{v}_i) d\mathbf{v}_i$$

for $j \neq k$. Then the census composite likelihood is of the form:

$$C(\boldsymbol{\theta}) = \prod_{i=1}^N \prod_{j < k} L_{ijk}^{B_{jk}} L_{ij}^{B_j} L_{ik}^{B_k}.$$

The B_j and B_{jk} can be chosen in various ways.

The consistency theorem in the Yi, Rao and Li (2016) paper takes B_j and B_k to be 0, and uses only the pairwise terms for pairs of units within the same higher level unit. The function l_{wc} below is the sample log composite likelihood:

$$l_{wc}(\boldsymbol{\theta}) = \sum_{i \in S} w_i \sum_{j < k \in S(i)} w_{jk|i} B_{jk} l_{ijk}(\boldsymbol{\theta}),$$

where $w_{jk|i} = \pi_{jk|i}^{-1}$ and $l_{ijk}(\boldsymbol{\theta}) = \log(L_{ijk}(\boldsymbol{\theta}))$. The sample score vector is U_{wc} , the vector of derivatives of the log composite likelihood with respect to each of the parameters, and U_{iwc} is the derivative vector within the i -th higher level unit:

$$U_{wc}(\boldsymbol{\theta}) = \frac{\partial l_{wc}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i \in S} U_{iwc}(\boldsymbol{\theta}).$$

The consistency result is that under regularity conditions, the solution of $\hat{\boldsymbol{\theta}}_w$ of $\mathbf{U}_{wc}(\boldsymbol{\theta}) = \mathbf{0}$ tends in probability to $\boldsymbol{\theta}$ as $n \rightarrow \infty$, where

$$\mathbf{U}_{wc}(\boldsymbol{\theta}) = \frac{\partial l_{wc}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i \in S} U_{iwc}(\boldsymbol{\theta}) = \mathbf{0}$$

and

$$\mathbf{U}_{iwc}(\boldsymbol{\theta}) = \sum_{j < k \in S(i)} w_{jk|i} B_{jk} \partial l_{ijk} / \partial \boldsymbol{\theta}.$$

The paper also contains simulation studies. One of these looks at a logistic regression mixed model where Y_{ij} is a binary response with conditional mean $\mu_{ij} = E(Y_{ij} | \mathbf{x}_i, v_i)$ and where in particular $\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 X_{ij} + v_i$; the v_i are i.i.d. $N(0, \sigma^2)$; and the survey weights are informative. It is found that when the sample sizes within clusters are small, maximizing the weighted pairwise likelihood estimates the β parameters and the variance parameter σ well, better than maximizing either the unweighted log-likelihood or the pseudo-log likelihood.

Dumitrescu, Qian and Rao (2021) worked to extend the theory for this same context, where, again, as they put it, “the design structure (two stage sampling) matches the hierarchy (levels)”. They took the theory beyond point estimate consistency to properties of testing hypotheses on the multilevel model parameters, proving an asymptotic normality result for the weighted composite likelihood estimators, and establishing asymptotic distributions for score test and likelihood ratio test statistics. The findings were supported by simulation studies.

Thompson et al. (2022) (without rigorous theoretical justification) applied the technique of RVH to a study of the continuous variable measuring Post-Traumatic Stress Disorder for a longitudinal survey in the area of Galveston Bay, Texas. In this case there was a three-level model structure, where the lowest level was survey waves within individuals, the middle level was individuals, and the highest level was Primary Sampling Units (PSUs) or “clusters”. There was expected to be a random effect at the cluster level that had both an independent component and a spatially correlated component, and both the mean function and the variance component parameters were of interest. The technique of RVH was adapted to joint estimation of mean function parameters and variance components. The sampling design was incorporated through weighting of the estimating functions, and it turned out in this case to be quite practicable

to reconstruct the weights at the highest and middle levels (they were equal to 1 at the lowest level). Results supported a non-negligible cluster effect with spatial autocorrelation.

The sampling design information was included among the regression covariates, to bring the model close to what would be assumed in model-based analyses, as well as being incorporated through the weighting of the estimating functions. A Bayesian analysis using WinBugs and noninformative priors gives similar results for the regression coefficients and for the variance component estimation.

4. Conclusion

The RVH method is a frequentist method, through which there are good design-based properties and model-design consistency. It could be expected to work well in situations of larger and more structured sets of survey data than the one considered by Thompson et al. (2022). There are two recent papers by Lumley and Huang (2023 and 2024), who have implemented the method in R and investigated its properties. Among the conclusions of the first paper is that the pairwise likelihood estimator has the potential to be “extended to settings where the model groups and design clusters are independent”.

The second follows this suggestion up, with an implementation where in the two-level linear model the hierarchy need not match the design. There is an application to the case where the PSUs are geographic and the higher level groupings are genetic relationships.

The results are mixed, but among the conclusions of this second paper is a confirmation that “We recommend weighted pairwise likelihood estimation when sampling is expected to be informative and the variance components are of substantive interest, and design variables are either non-available or not appropriate for inclusion in the model”.

The RVH work and the composite likelihood approach, along with their extensions and offshoots, constitute a very important contribution not only to methods but to foundations: a frequentist method for complex analytic inference, and foundational principles applied to a difficult problem in data analysis, where conclusions are important and often elusive.

Acknowledgement

The research is partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

- Asparouhov, T. (2006), “Generalized multi-level modeling with sampling weights”, *Communications in Statistics - Theory and Methods*, 35, pp. 439-460.
- Binder, D.A. (1983), “On the variances of asymptotically normal estimators from complex surveys”, *International Statistical Review*, 51, pp. 279-292.
- Dumitrescu, L., Qian, W., and Rao, J.N.K. (2021), “A weighted composite likelihood approach to inference from clustered survey data under a two-level model”, *Sankhya*, 83, pp. 814-843.
- Grilli, L. and Pratesi, M. (2004), “Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs”, *Survey Methodology*, 30(1), pp. 93-103.

- Lumley, T. and Huang, X. (2023), “Weighted composite likelihood for linear mixed models in complex samples”, arXiv:2311.13048.
- Lumley, T. and Huang, X. (2024), “Linear mixed models for complex survey data: Implementing and evaluating pairwise likelihood”, *Stat* 13(1). DOI:10.1002/sta4.657.
- Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998), “Weighting for unequal selection probabilities in multi-level models”, *Journal of the Royal Statistical Society, Series B*, 60, pp. 23-56.
- Rabe-Hesketh, S. and Skrondal, A. (2006), “Multilevel modeling of complex survey data”, *Journal of the Royal Statistical Society, Series A*, 169, pp. 805-827.
- Rao, J.N.K., Verret, F. and Hidirolou, M.A. (2013), “A weighted composite likelihood approach to inference for two-level models from survey data”, *Survey Methodology*, 39(2), pp. 263-282.
- Singer, J.D. (1998), “Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models”, *Journal of Educational and Behavioral Statistics*, 24(4), pp. 323-355.
- Skinner, C.J. (1989), “Domain means, regression and multivariate analysis”, in *Analysis of Complex Surveys*, Eds. Skinner, C.J., Holt, D. and Smith, T.M.F. Chichester: John Wiley & Sons, Inc., pp. 59-87.
- Thompson, M.E., Meng, G., Sedransk, J., Chen, Q. and Anthopolos, R. (2022), “Spatial multilevel modelling in the Galveston Bay Recovery Study Survey”, in *Advances and Innovations in Statistics and Data Science*, Eds. He, W., Wang, L., Chen, J. and Lin, C.D. Springer Cham, pp. 275-293.
- Yi, G. Y., Rao, J. N. K., and Li, H. (2016), “A weighted composite likelihood approach for analysis of survey data under two-level models” *Statistica Sinica*, 26, pp. 569-587.