A Random Effects Model for Sparse Cross-Classification Data

Henrik Stryhn

CVER - Centre for Veterinary Epidemiological Research
University of Prince Edward Island

SSC 2022
Annual
Meeting
Online

Context/Problem: Ranking Conference Abstracts

Practical task at hand:

- based on reviewer scores, rank abstracts from highest to lowest,
- make decisions about "acceptance" (for oral pres.) of abstracts at suitable cut-off.

Data at hand (first round of abstract submissions for ISVEE 16[1]):

- 119 abstracts, each scored twice ($0-100$ scale) by two of 27 reviewers,
- reviewers assessed $1-15$ abstracts (average $238/27 = 8.8$).

Approaches considered (post-hoc; (1x) was actually used) to base the ranking on:

(1) simple: average score for two reviewers per abstract,

(1x) expanded simple: request extra review for selected abstracts (with strong reviewer disagreement), and then use (1) with simple average across all reviewers/abstract,

(2) model-based: estimate abstract levels from a statistical model.

First aim: determine the feasibility of (2) and compare its results with (1) and (1x).

[1] 16th International Symposium of Veterinary Epidemiology and Economics, August $7-12$, 2022, Halifax, NS.

DATA ILLUSTRATION

Possible data layout for 10 abstracts and 5 reviewers:

	Abstract									
Reviewer	1	2	3	4	5	6	7	8	9	10
1	$\surd$	-	$\surd$	-	-	-	$\surd$	-	-	$\surd$
2	-	$\surd$	-	-	-	$\surd$	$\surd$	-	$\surd$	-
3	$\surd$	$\surd$	-	-	$\surd$	-	-	$\surd$	-	-
4	-	-	-	$\surd$	$\surd$	$\surd$	-	-	-	$\surd$
5	-	-	$\surd$	$\surd$	-	-	-	$\surd$	$\surd$	-

an incomplete two-way classification

— unrealistically “nice” design (actually a balanced incomplete block design) with

* equal number (4) of reviews per reviewer ($\sim$ “treatment”),
* each pair of reviewers share exactly one abstract,

which gives nice (equal precision) comparisons between reviewers in a model accounting for both abstracts and reviewers; but even in this nice design,

- simple means and adjusted means for reviewers (“treatments”) are not the same,
- similarly nice properties do not hold for abstracts ($\sim$ “blocks”), due to too little replication.

Take-away message: we cannot compensate for the incompleteness by a clever design, and dependence on the other classification variable is unavoidable.

Statistical Models

The data layout invites a two-way ANOVA,

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \text{where}$$

– reviewer effects (α_i) and abstract effects (β_j) are taken as either fixed or random (drawn from respective $N(0, \sigma^2)$ distribution(s)),

– extensions (e.g., unequal reviewer variances[2]) are possible, but not discussed here.

Fixed or random effects? in two-way ANOVA:

- (random): assumes effects drawn from a population, avoids estimation of a large number of individual parameters, exerts smoothing on estimation, balances abstract and reviewer effects against their respective distribution assumptions,
- (fixed): no assumptions on effects — estimated freely to achieve best fit, requires estimation of a very large number of parameters with potential (near-)collinearity between them,
- (mixed): fixed effects for raters (reviewers) is common in item response models[2], and may be preferable for a small number of raters or with reviewer effects not approximated well by $N(0, \sigma^2)$.

Our focus (here): method (1) vs. random effects (reviewers and abstracts) model.

[2] An introductory review of models for the two-way layout is in Skrondal & Rabe-Hesketh (2004): *Generalized Latent Variable Modelling*, Section 3.3, including the unequal-variances *congeneric measurement model*.

Results: Original Data + Simulation

Variance components: σ^2(abstr.) = **99.6**, σ^2(rev.) = **179.3**, σ^2(error) = **223.6**.

Comparison between simple and random effects model results (original data):

- absolute rank differences: mean = **13.9**, sd = **11.3**, IQR : **4.5–21**, full range : **0–59**,
- classification differences: **7 + 7** abstracts (when split as 51:68).

Simulation study for two error level settings (100 simulations each, including random variation among reviewers and errors):

mean (sd)	error $\sigma^2 = 223$		error $\sigma^2 = 22$	
Statistic/Method	simple	random	simple	random
absolute ranking error	21.1 (1.8)	18.4 (1.5)	16.4 (2.1)	7.2 (0.5)
count misclassified	28.8 (4.5)	25.1 (4.0)	21.5 (5.0)	8.1 (2.5)

- substantial differences between simple and model-based results; inspection of the data reveals that simple and model-based ranks differ most when the two reviewers are both extreme in the same direction (both low, or both high)[3],
- model-based rankings perform clearly better with low error variance, but only slightly better with actual error variance.

[3] Expanding the data with 15 extra reviews (where the two reviewers strongly disagreed) did not change that; nor did it change the disagreements substantially: absolute rank differences: 13.9 (10.6), and 8 + 8 misclassifications.

Second Part: How Does the Model Deal with "Problems"?

Back in earlier days, such data (highly unbalanced and sparse cross-classified[4]) could be termed as "messy" and be subject to special scrutiny[5].

Estimation (ML or REML, or Bayesian MCMC) did not seem to experience problems: all analyses converged nicely and agreed between software implementations[6].

What about the model's sensitivity to added difficulties, such as

- actual collinearity between reviewers and abstracts (in a fixed parameter model sense); two examples to follow (none of which occurred in the actual data),
- unequal error variances by reviewers $\sim$ reviewer-dependent utilization of scale (may be plausible in this context).

Second aim: explore performance of the random effects model under these circumstances.

[4] Recall that only two reviews were obtained per abstract, and that the 27 reviewers handled between 1 and 15 abstracts, of which two reviewers only scored one abstract.

[5] For example, Milliken & Johnson (1992): *Analysis of Messy Data*, or Aitken (1978), *J. Royal Statist. Soc. A* 141, 195 − 223.

[6] All results shown are based on Stata's implementation in the `mixed` command.

Data Illustration ∼ Problem I

Modified data layout with extra abstracts and reviewers, in scenario (I):

Reviewer	Abstract 1	2	3	4	5	6	7	8	9	10	11
1	√	-	√	-	-	-	√	-	-	√	-
2	-	√	-	-	-	√	√	-	√	-	-
3	√	√	-	-	√	-	-	√	-	-	-
4	-	-	-	√	√	√	-	-	-	√	-
5	-	-	√	√	-	-	-	√	√	-	-
6	-	-	-	-	-	-	-	-	-	-	√
7	-	-	-	-	-	-	-	-	-	-	√

— a (fixed effects) collinearity between added reviewer and abstract effects:

∗ effects of reviewers 6−7 and abstract 11 cannot be separated from each other,

∗ effectively, 3 added parameters but only 2 extra observations (and essentially the same problem would occur with reviewer 6 only),

∗ such collinearities can typically be detected from data summaries.

Note: estimation is still possible in the random effects model!

Data Illustration ∼ Problem II

Modified data layout with extra abstracts and reviewers, in scenario (II):

	Abstract												
Reviewer	1	2	3	4	5	6	7	8	9	10	11	12	13
1	√	-	√	-	-	-	√	-	-	√	-	-	-
2	-	√	-	-	-	√	√	-	√	-	-	-	-
3	√	√	-	-	√	-	-	√	-	-	-	-	-
4	-	-	-	√	√	√	-	-	-	√	-	-	-
5	-	-	√	√	-	-	-	√	√	-	-	-	-
6	-	-	-	-	-	-	-	-	-	-	√	-	√
7	-	-	-	-	-	-	-	-	-	-	√	√	-
8	-	-	-	-	-	-	-	-	-	-	-	√	√

— also a (fixed effects) collinearity between added reviewer and abstract effects:

* reviewers 6−8 and abstracts 11−13 are separated from rest of design ⇒ abstracts cannot be compared to other abstracts without including effects of reviewers,
* this type of collinearity may be less obvious visually, but can be detected in a fixed effects model; it seems quite plausible within the conference abstracts context (reviewers may form groups based on their expertise).

Impact of Data Collinearities (I – II)

Methods: create desired design collinearities by minimal rearrangement of reviewers in actual data, and simulate datasets with random variation for reviewer and error terms, in order to compare distributions of estimates and ranks with those of abstracts not affected by collinearities.

Summary of results for collinearity I – II:

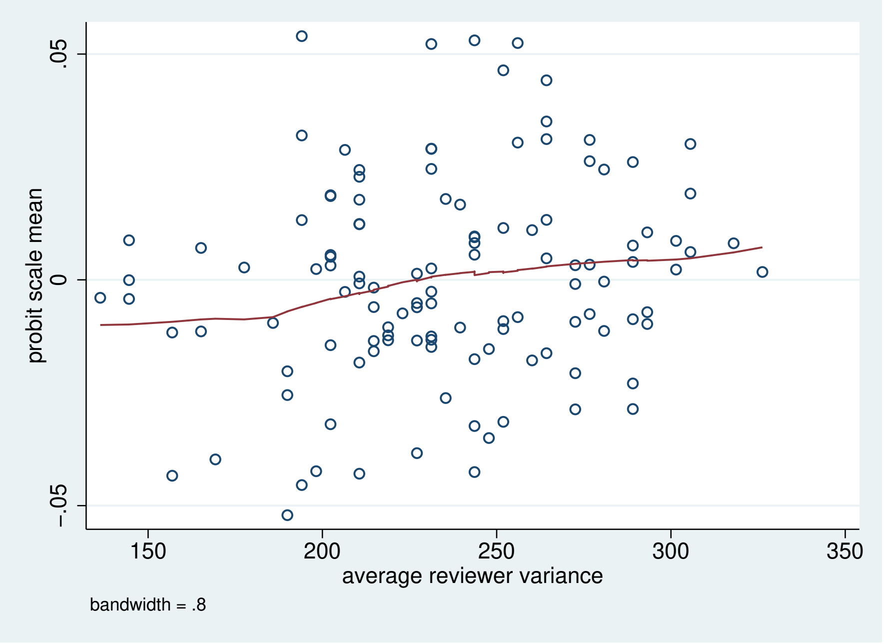
- estimates for affected abstract(s) show higher/highest variability, both on the modelling scale and for the ranks, when compared to similarly ranked (± 5) abstracts; findings consistent across:
  - magnitudes of error variance (as before, $\sigma^2 = 223$ or 22), with more dramatic effects seen for lower error variance,
  - magnitudes of targeted abstract rank(s) (percentiles ranging $\approx 10-50\%$),
  - sizes of simulation (100 or 1000 replicates),

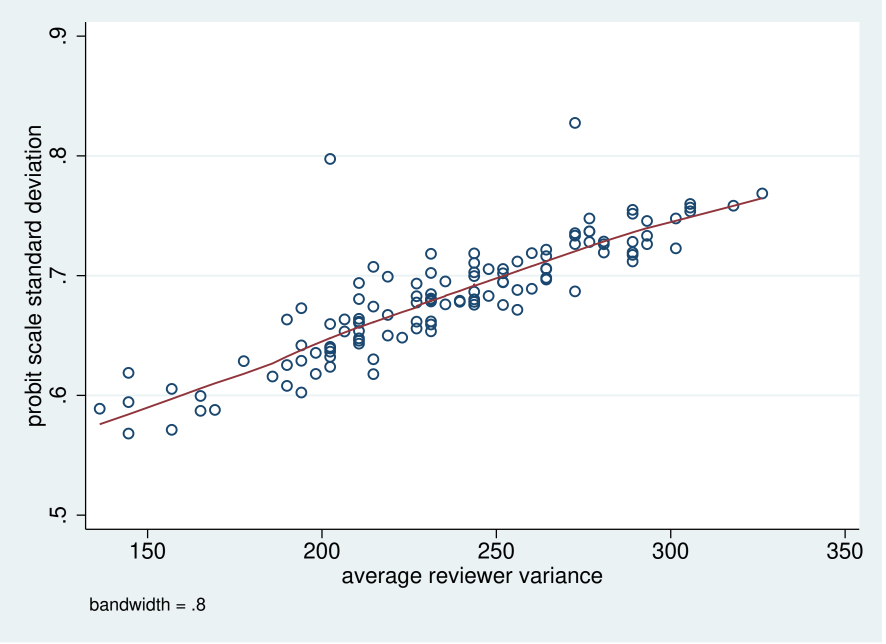
  — possible interpretation: added reviewer variability in the abstract estimate/rank when there is no calibration with larger pool of reviewers.
- also some instances of bias in the ranks towards the center, in particular when the abstract rank is extreme or the error variance is low.

Impact of Unequal Variances

Methods: similar (without collinearities, 1000 replicates), and

- unequal error variances for reviewers (range: **116 − 330**) created in simulations,
- abstract variation included in simulations, smoothing out abstract configurations,
- quantification of rank deviations (estimated minus true) on probit scale[7].





Interpretation of variance-dependence: limited impact on bias, clear impact on spread.

[7] The boundedness of ranks make an assessment on original scale difficult.

Conclusion/Discussion

Some cautious conclusions:

- the model-based approach seemed to apply[8] and work reasonably well and better than simple averaging, even if the improvement in practice might be limited with a large error variance,
- the random effects model seemed reasonably robust to (fixed effects) collinearities, but it might still be useful to detect them (e.g., in a fixed effects model), so as to pay attention to their impact on results,
- the random effects model showed only little bias from unequal variances, but these did lead to different spread in estimates.

Additional methodological considerations:

- Bayesian modelling/estimation possible as well, but seems to agree well with (RE)ML estimation, and did not help with diagnosis of problems (results not shown here),
- only two reviews per abstract limits the options for more complex modeling, for example to account for reviewer heterogeneity[9] (beyond random intercepts).

Practical "happy ending": the study convinced the conference organizers to use a model-based ranking for the full ($\approx$ 600 abstracts) submission!

[8] The (formal) model assumptions could be met to a satisfactory degree after a simple scale change, not shown.
[9] The congeneric measurement model is not identified; Rabe-Hesketh et. al (2001), *The Stata Journal* 1.