Statistical Modelling of Reviewer Scores for Abstracts

Page	Title
1	Introduction — the problem
2	Data illustration I
3	Statistical models
4	Data illustration II $\sim$ problems
5	Results I (original data)
6	Evaluation of methods (simulation)
7	Results II (augmented data)
8	Conclusions/Discussion

INTRODUCTION — THE PROBLEM

Task at hand:

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

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

- 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 to base ranking on:

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

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

(1x) expanded simple: request extra review for selected (15) abstracts, and then use (1) with simple average across all reviewers per abstract.

Aim of this exploration: determine the feasibility of (2) and compare its results with (1) and (1x).

Data Illustration I

Possible data layout for 10 abstracts and 5 reviewers:

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

an incomplete two-way classification

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

- ∗ equal number of reviews per reviewer (4),
- ∗ 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 layout,

- ∘ simple means and adjusted means for reviewers are not the same,
- ∘ similarly nice properties do not hold for abstracts (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,

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

– reviewer effects ($\boldsymbol{\alpha_i}$) and abstract effects ($\boldsymbol{\beta_j}$) are taken as either fixed or random (drawn from $\mathrm{N}(\mathbf{0}, \boldsymbol{\sigma}^2)$ distribution(s)),

– assuming all $\boldsymbol{\alpha_i} = \mathbf{0}$ corresponds to a one-way ANOVA, method (1).

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 (next slide),
- (mixed): fixed effects for raters (reviewers) is common in item response models (Skrondal & Rabe-Hesketh, 2004), and may be preferable for a small number of raters or with reviewer effects not approximated well by $\mathrm{N}(\mathbf{0}, \boldsymbol{\sigma}^2)$.

Initial focus: method (1) vs. random effects model vs. mixed model.

Data Illustration II ~ Problems

Modified data layout with extra abstracts and reviewers, in two scenarios (A) and (B):

	Abstract										
Reviewer	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),
* these types of collinearities can typically be detected from data summaries.

Note: estimation would still be possible in a random effects model!

Data Illustration II ∼ Problems

Modified data layout with extra abstracts and reviewers, in two scenarios (A) and (B):

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

— also (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.

Take-away message: it is probably useful to try a fixed effects model, in order to detect such potential collinearities so as to be aware of their impact on results.

RESULTS I (ORIGINAL DATA)

Good news: no collinearity (despite 2 reviewers with only 1 review).

Comparison of results from 3 approaches:
(1) simple means, (2) random effects model, (2m) mixed model (fixed reviewer effects).

Summary table for absolute rank differences between methods:
mean (sd) below diagonal \ interquartile range (full range) above diagonal

Method	simple	random	mixed
simple	-	4.5-21 (0-59)	6-25.5 (0-65)
random	13.9 (11.3)	-	1-4 (0-22)
mixed	16.7 (12.8)	3.2 (3.3)	-

Disagreements in classifications between methods, when splitting the abstracts 51:68:
simple vs random: $7 + 7$; simple vs mixed: $9 + 9$; random vs mixed: $2 + 2$.

Findings:

- differences substantial between simple and model-based, but relatively minor between the 2 models,
- inspection of the data reveals that simple and model-based ranks differ most when the 2 reviewers are both extreme in the same direction (both low, or both high).

Evaluation of Methods (Simulation)

Aim: explore performance of methods (simple, random effects model, mixed model),

- random variation across simulations: errors only, or errors + reviewer effects,
- error variance: high ($\sim$ data) or low (about one-tenth),
- measures: mean (sd) of absolute rank differences, mean (sd) of misclassification counts (for 51:68 split).

Results from 100 simulations for each setting:

Setting	Measure	simple	random	mixed
error only $\sigma^2 = 223$	ranking diff.	21.7 (1.5)	18.8 (1.6)	19.0 (1.6)
	misclassif.	29.5 (3.7)	25.5 (4.1)	26.0 (4.1)
error only $\sigma^2 = 22$	ranking diff.	16.7 (0.6)	7.1 (0.6)	7.1 (0.6)
	misclassif.	20.0 (2.4)	7.1 (2.3)	7.5 (2.4)
error + rev. $\sigma^2 = 223$	ranking diff.	21.1 (1.8)	18.4 (1.5)	18.8 (1.6)
	misclassif.	28.8 (4.5)	25.1 (4.0)	26.1 (4.1)
error + rev. $\sigma^2 = 22$	ranking diff.	16.4 (2.1)	7.2 (0.5)	7.2 (0.5)
	misclassif.	21.5 (5.0)	8.1 (2.5)	8.2 (2.4)

- model-based methods perform clearly better with low error variance, but only slightly better with actual error variance,
- very minimal differences between random and mixed models, and random settings.

Results II (Augmented Data)

Augmented data: additional review obtained for abstracts considered to be close to relevant cut-off and with clear disagreement in its two reviewer scores
$\Rightarrow$ 253 reviews (still 27 reviewers and 119 abstracts).

Impact of augmentation for the 15 abstracts involved: overall minor (average rank differences within $(-8.3, 5)$), and 5 changes in classification.[1]

Summary table for absolute rank differences between methods:
mean (sd) below diagonal \ interquartile range (full range) above diagonal

Method	simple	random	mixed
simple	-	5-20 (0-45)	6-25 (0-51)
random	13.9 (10.6)	-	1-4 (0-22)
mixed	16.7 (12.2)	3.2 (3.3)	-

Disagreements in classifications between methods, when splitting the abstracts 49:70:
simple vs random: $7 + 7$; simple vs mixed: $9 + 9$; random vs mixed: $2 + 2$
(same numbers as before, but not quite the same abstracts).

- very similar results to those for the original data.

[1] Actual classification split for the augmented data was 49:70 instead of 51:68, so some differences may also be due the small change in proportions.

Conclusion/Discussion

Some cautious conclusions:

- the model-based approach(es) seemed to work reasonably well, and only a simple scale change was required to meet model assumptions,
- some clear differences in rankings and classifications between model-based approach(es) and simple means, but simulation study showed the error variance to be too large to demonstrate one as clearly "more correct" for the data at hand,
- augmentation of data with 15 additional reviews did not have much of an impact.

Additional methodological considerations:

- Bayesian modelling/estimation possible as well, but seems to agree well with (RE)ML estimation, and does not help with diagnosis of problems (results not shown here),
- only 2 reviews per abstract limits the options for more complex models:
  * congeneric measurement model (Skrondal & Rabe-Hesketh, 2004) is not identifiable,
  * attempts to incorporate unequal variances across reviewers were not successful (convergence problems).

The $1000 question is (of course):
What is the best approach to rank the next (and much larger) pool of abstracts?