On the likelyhood for finite mixture models and Kirill Kalinin's paper "Validation of the Finite Mixture Model Using Quasi-Experimental Data and Geography"

Alexander Shen

On the likelyhood for finite mixture models and Kirill Kalinin's paper "Validation of the Finite Mixture Model Using Quasi-Experimental Data and Geography"

Alexander Shen*

June 28, 2018

The analysis presented in [1] is based on an approach suggested by Walter Mebane et al. in [3] and later used by Mebane in [4]. In this approach (following Klimek et al [2]) the voting process is modeled by a mixture of several distributions that correspond to "fair counting", "incremental fraud" and "extreme fraud". Authors then try to apply the expectation-maximization (EM) algorithm to find the parameters of the model that maximize the finite mixture likelyhood. This makes perfect sense if done correctly. However, Mebane [3, page 10] writes:

> The likelyhood for a finite mixture model can be written as
>
> $$\mathscr{F}(\mathbf{W}, \mathbf{A} \mid \mathbf{N}; \Psi) = \sum_{j \in \{0,\mathrm{i},\mathrm{e}\}} f_j \prod_{i=1}^{n} g_{jW}(W_i \mid N_i; \Psi) g_{jA}(A_i \mid N_i, \Psi) \tag{7}$$
>
> where f_0, f_i, f_e are probabilites with $f_0 + f_\mathrm{i} + f_\mathrm{e} = 1$.

This expression (reproduced also as (6) in [4] and in Section 3 of [1]) does not represent the likelyhood for finite mixture model correctly (it assumes that the choice between three possibilities is done once, while in the model it is done independently for each point), and its maximization makes no sense. Indeed, the expression being maximized, considered as a function of $f_0, f_\mathrm{i}, f_\mathrm{e}$, is a linear function on the triangle $f_0 + f_\mathrm{i} + f_\mathrm{e} = 1$, $f_0, f_\mathrm{i}, f_\mathrm{e} \geqslant 0$. Such a function always reaches maximum at some vertex of this triangle. If a maximization algorithm is applied to this function and gives a point that is an internal point of the triangle (all three parameters are positive, as it happens in [1, 4]), this means that

- either maximization algorithm does not work correctly and produces an internal point of triangle where the expression is *not* maximal,
- or the function is constant on the triangle, and every point of the triangle gives the same maximal value.

In both cases the values of $f_0, f_\mathrm{i}, f_\mathrm{f}$ provided by the maximization algorithm do not make sense for obvious reasons.

It is theoretically possible that the actual software implementation of the EM-algorithm was correct, and the incorrect formula was used just for publication. One can only guess whether it was the case; still the significant difference between the results obtained in [4, 1] and the initial estimates given in [2] (using a much simpler alternative approach) makes this conjecture less plausible.

References

[1] Kalinin, K., *Validation of the Finite Mixture Model Using Quasi-Experimental Data and Geography*, this volume.

[2] Klimek, Peter; Yegorov, Yuri; Hanel, Rudolf; Thurner, Stephan, Statistical detection of systematic election irregularities, *Proceedings of the National Academy of Sciences of the United States of America*, **109**(41), 16469–16473 (2012), https://doi.org/10.1073/pnas.1210722109

[3] Mebane, W.R., Jr.; Egami, Naoki; Klaver, Joseph; Wall, Jonathan, *Positive Empirical Models of Election Fraud (that May Also Measure Voters' Strategic Behavior).* Prepared for presentation at the 2014 Summer Meeting of the Political Methodology Society, University of Georgia, July 24–26, 2014. Available as http://www-personal.umich.edu/~wmebane/pm14.pdf (downloaded June 5, 2018)

[4] Mebane, W.R., Jr. *Election Forensics: Frauds Tests and Observation-level Frauds Probabilities.* Prepared for presentation at the 2016 Annual Meeting of the Midwest Political Science Association, Chicago, IL, April 7–10, 2016. Available as http://www-personal.umich.edu/~wmebane/mw16.pdf (downloaded June 5, 2018)

*LIRMM, University of Montpellier, CNRS. alexander.shen@lirmm.fr