**A simulation-based comparison of approaches to significance testing using generalised additive mixed models**

Márton Sóskuthy, *University of York*

There is increasing interest in dynamic properties of speech in phonetics and related fields, even in areas where static analyses used to be the norm (e.g. formant measurements for monophthongs; Watson & Harrington, 1999; Sóskuthy et al., 2015). Dynamic analyses of speech typically focus on trajectories or contours (e.g. formants, tongue contours), which are series of measurements with a clear temporal or spatial structure. This type of data presents two analytical challenges: (i) phonetic trajectories and contours are rarely linear and (ii) individual observations such as vowel tokens are represented by multiple data points, introducing hierarchical relationships in the data. Ignoring non-linearity and hierarchical structure in regression models leads to unreliable estimates and *p*-values (see e.g. Barr et al., 2013). Generalised additive mixed models (GAMMs) offer a solution to these issues by allowing the inclusion of *smooth terms* and *random smooths* in addition to linear terms (Wood, 2006). Smooth terms capture non-linear effects through curve-fitting. Random smooths extend the same principle to random effects, fitting separate curves at each value of a grouping variable. However, GAMMs also present the analyst with difficult choices: What methods of significance testing should be used? What random structures should be included? This talk addresses these questions using results from type I and II error simulations.

The simulations look at a simple scenario with 100 randomly generated trajectories that are similar in shape but also show some variation. The trajectories are assigned to two groups. Various model structures and significance tests are used to test for differences between the groups. Type I error simulations model a scenario where the trajectories all come from the same underlying trajectory, and are assigned randomly to groups (Figure 1). Type II error simulations model a scenario where the trajectories in the two groups come from different underlying trajectories (Figure 2). Large batches of simulations are run for different combinations of random structure specification and method of significance testing. The proportion of significant results (at $\alpha = 0.05$) in each batch represents the type I or type II error rate depending on the type of simulation. The following methods of significance testing were compared: looking at the significance of the parametric and/or the smooth term in the model summary; comparison of nested models using likelihood-ratio tests (models fitted with ML or REML); and visual testing using the confidence interval around the difference smooth. The random structure specifications include: no random effects; random intercepts by trajectory; random intercepts and slopes by trajectory; and random smooths by trajectory.

Table 1 presents the results of the type I error simulations (the description of the table provides more detail about the methods). Models without random smooths by trajectory invariably yield high rates of false positives, even when random intercepts and slopes are both included. The intuitively attractive way of simultaneously checking the significance of both the parametric and the smooth terms in the model summary (third row in the table) also leads to inflated type I error rates. Visual tests based on the difference smooth can also yield high type I error rates when a significant difference along a short portion of the trajectory is deemed sufficient to declare overall significance. Finally, as noted in many places (e.g. Zuur et al. 2009), models with different fixed effects cannot be compared when fitted with REML. Based on the type I error rates, the best method appears to be the comparison of nested models fitted with ML – or looking at the model summary in cases where one's hypothesis is based specifically on a difference in trajectory height or shape. Table 2 presents type II error rates that largely support this conclusion: model comparison appears to have the highest power out of the methods with acceptable type I error rates.

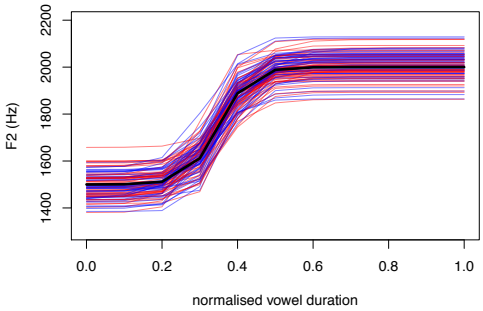


Figure 1: Underlying trajectory in type I error simulations (black) and simulated trajectories after assignment to groups (red and blue).

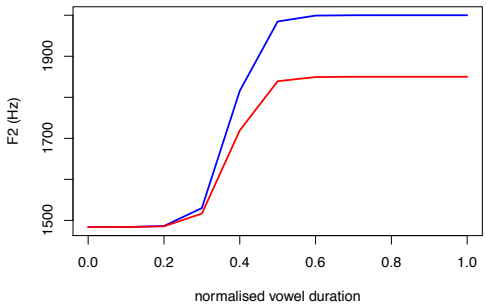


Figure 2: Underlying trajectories in type II error simulations.

| | no rnd | rnd smooth | rnd intercept | rnd intercept+slope |
|---|---|---|---|---|
| summary: parametric | 0.436 | 0.049 | – | – |
| summary: smooth | 0.236 | 0.053 | – | – |
| summary: either | 0.561 | 0.100 | – | – |
| model comp: ML | 0.396 | 0.036 | 0.270 | 0.116 |
| model comp: REML | 1.000 | 1.000 | – | – |
| visual: > 10% diff | 0.599 | 0.120 | – | – |
| visual: > 20% diff | 0.577 | 0.089 | – | – |
| visual: > 50% diff | 0.423 | 0.029 | – | – |

Table 1: Type I error rates. Columns = random structures. Rows = methods of significance testing. Cells in red show higher-than-nominal false positive rates. Some combinations were not included in the simulations. Tests in "summary: …" rows are based on *p*-values for parametric/smooth difference terms from the model summary. "summary: either" refers to a method where significance is claimed when either the parametric or the smooth term is significant. For model comparisons, the nested model excludes both the parametric and the smooth difference terms. Visual tests are based on checking whether the confidence interval around the difference smooth includes 0. The number indicates the proportion of the trajectory where the confidence interval excludes 0, e.g. "> 20% diff" refers to a case where more than 20% of the trajectory shows a significant difference.

| | no rnd | rnd smooth |
|---|---|---|
| summary: parametric | 0.912 | 0.524 |
| summary: smooth | 0.753 | 0.520 |
| summary: either | 0.968 | 0.728 |
| model comp: ML | 0.945 | 0.594 |

Table 2: Type II error rates. The table is set up the same way as Table 1.

*References*

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Sóskuthy, M., Foulkes, P., Hughes, V., Hay, J., Haddican, B. (2015). Word-level distributions and structural factors codetermine GOOSE fronting. *Proceedings of the 18th International Congress of Phonetic Sciences*, 10-14 August 2015, Glasgow.

Watson, C. I., & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *The Journal of the Acoustical Society of America*, *106*(1), 458–468.

Wood, S. (2006). *Generalized additive models: an introduction with R*. Boca Raton: CRC press.

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). Mixed effects models and extensions in ecology with R. New York: Springer.