Generalised additive mixed modelling for dynamic formant analysis

Márton Sóskuthy
University of York

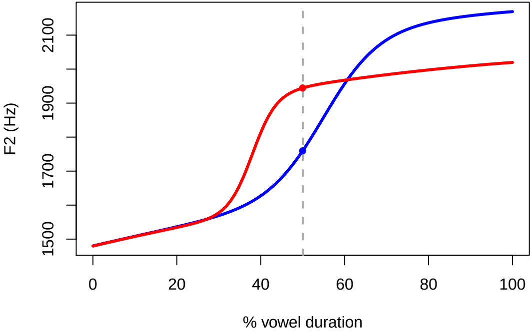
13 July 2016
LabPhon 15, Ithaca

Introduction

- today's focus: comparing vocalic sounds
- two different strategies:
  - single-point (e.g. Peterson and Barney 1952; Labov et al. 2005; Hay et al. 2015)
  - dynamic (e.g. Watson and Harrington 1999; Fox and Jacewicz 2009; Cardoso 2015)

Introduction

- single-point analysis: mid-point



Introduction

- dynamic analysis: a lot of choices
  - duration differences
  - timing differences
  - degree of diphthongisation, e.g. Euclidean distance
  - re-parameterising curves, e.g. polynomials, Discrete Cosine Transform
  - comparing entire trajectories using visual methods / regression
    - Linear (Mixed Effects) Regression
    - Smooth-Spline ANOVA
    - Generalised Additive (Mixed) Models

Introduction

- what this talk is not:
  - a comparison of single-point vs. dynamic
  - an argument for a specific method of analysis
- what this talk is:
  - an exploration of the statistical properties of an increasingly popular method: GAMMs (Wood 2006; Baayen 2015, 2016)

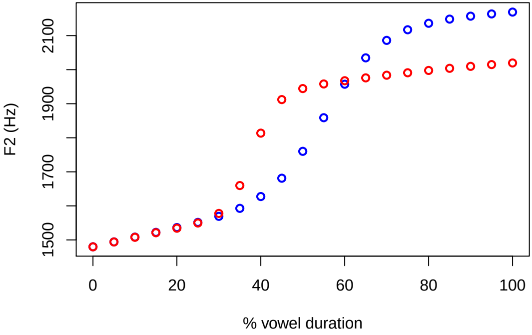
Introduction

- specifically:
  - how to test for significant differences using GAMMs?
  - how to avoid false positives?
  - how to correctly specify random smooths?
- format of talk:
  - fake-data simulations for investigating false positive and false negative rates
  - a very brief case study using data from Stuart-Smith et al. (2015)

What are GAM(M)s?

- traditional regression models:

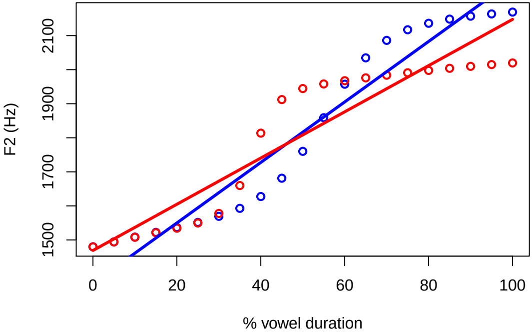
$$\mathbf{y} = \alpha + \beta_1 \mathbf{x_1} + \beta_2 \mathbf{x_2} + \ldots + \epsilon \tag{1}$$



What are GAM(M)s?

- traditional regression models:

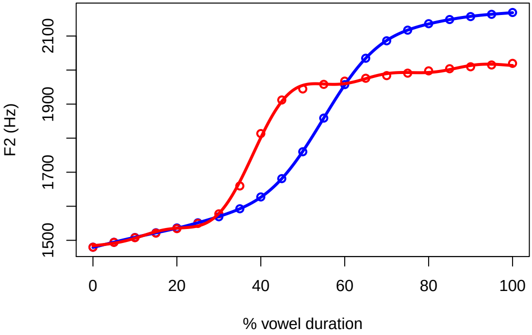
$$\mathbf{y} = \alpha + \beta_1 \mathbf{x_1} + \beta_2 \mathbf{x_2} + \ldots + \epsilon \tag{2}$$



What are GAM(M)s?

- (generalised) additive models:

$$\mathbf{y} = \alpha + \beta_1 \mathbf{x_1} + \ldots + f_1(\mathbf{x_2}) + f_2(\mathbf{x_3}) + \ldots + \epsilon \qquad (3)$$



What are GAM(M)s?

- what can GAM(M)s be used for?
  - testing for overall differences between curves
  - testing for changes in curves as a function of other predictors
  - locating differences – where are two curves different?

What are GAM(M)s?

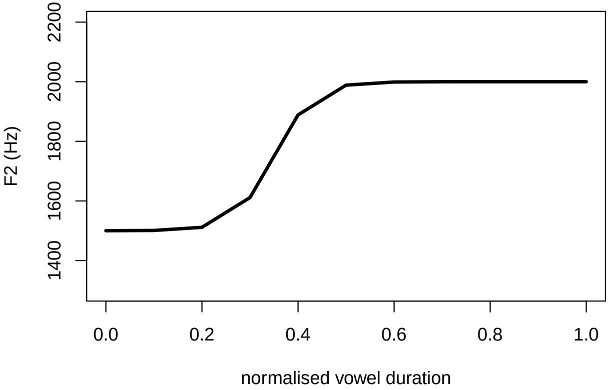
- focus on overall and local differences here
- (not looking at e.g. multidimensional smooths)
- how should we perform significance testing using GAM(M)s?
  1. model structure
  2. type of test
- goal: keeping false positives under 5%, while also keeping the rate of false negatives low

Simulations: false positives

- fake data simulations along the lines of Barr et al. (2013)
- scenario:
  - two words realised with same vowel
  - a speaker reads each of them 50 times
  - 11 evenly spaced points along F2 trajectories
  - repeat this experiment thousands of times
  - appropriate statistical test: 5% false positive rate

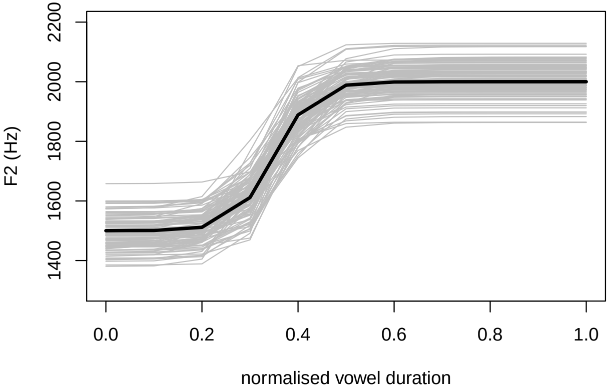
Simulations: false positives

- the underlying curve:



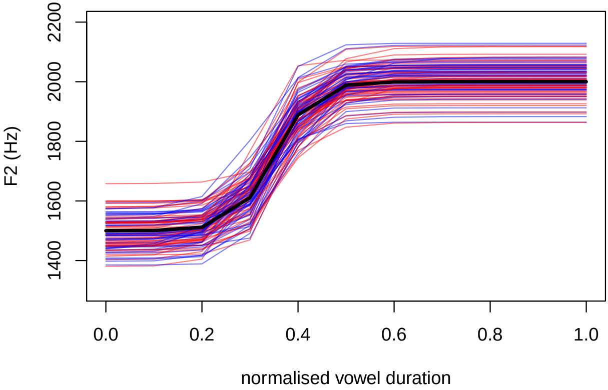
Simulations: false positives

- sample 100 random curves with some variation:



Simulations: false positives

- assigned to two words randomly:



Simulations: false positives

- an example GAMM:

  y ∼ word + s(x) + s(x, by=word)

- parametric coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1826.996      2.420 754.812  < 2e-16 ***
wordB        -12.896      3.423  -3.767 0.000174 ***
```

- smooth terms:

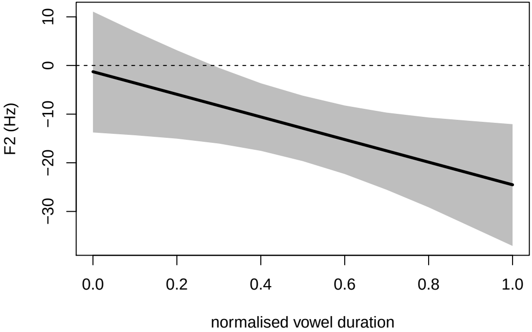
```
              edf Ref.df        F p-value
s(xs)       8.786  8.987 1165.128  <2e-16 ***
s(xs):wordB 1.000  1.000    4.583  0.0325 *
```

Simulations: false positives

- multiple ways to test for significance:
  1. parametric (P) term significant
  2. difference smooth (S) significant
  3. one or both of P and S significant
  4. model comparison: both P and S excluded from nested model
  5. visual inspection of difference plot based on posterior simulations...

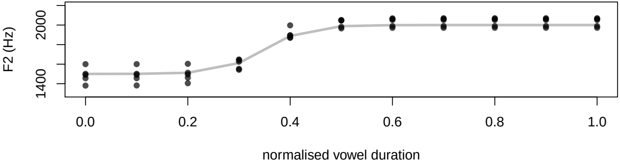
Simulations: false positives

- difference plot based on posterior simulations:

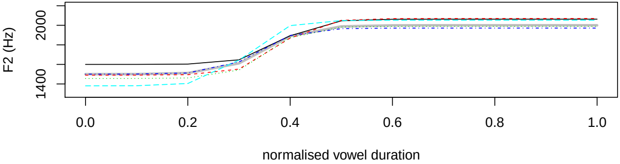


Simulations: false positives

- two different model types tested:
  1. without random smooths



  2. with random smooths



Simulations: false positives

- false positive rates

	no rnd smooths	rnd smooths
parametric	0.436	0.049
smooth	0.236	0.053
param. / smooth	0.561	0.100
model comparison	0.396	0.036
visual: $<$ 10% diff.	0.599	0.120
visual: $<$ 20% diff.	0.577	0.089
visual: $<$ 50% diff.	0.423	0.029

Simulations: false positives

- a closer look at the relationship between the parametric / smooth significances vs. the model comparison:

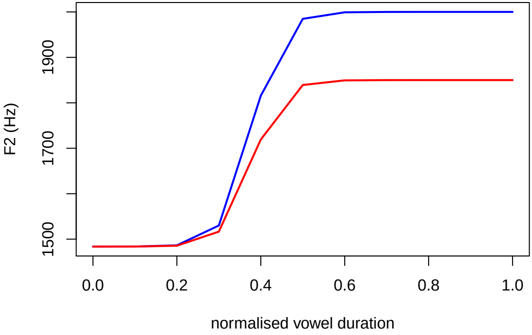
	ANOVA $\neg$sig.	ANOVA sig.
P. & S. $\neg$sig.	0.900	0.000
P. or S. sig.	0.064	0.036

Simulations: false positives

- random smooths are necessary when treatment predictor varies *between* items
- significance values for individual parametric / smooth components can *only* be used if there are prior hypotheses about them
- claiming significant differences based on visual inspection when only a few points are different produces high false positive rates
- recommended method: model comparison where nested model excludes *both* parametric and smooth terms

Simulations: false negative rates

- very similar simulations
- underlying trajectories for two words are different near the final 50%



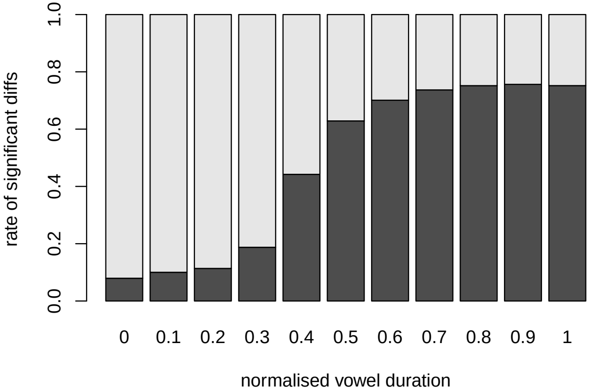
Simulations: false negative rates

- power (1 − false negative rate)

	no rnd smooths	rnd smooths
parametric	0.912	0.524
smooth	0.753	0.520
param. / smooth	0.968	0.728
model comp.	0.945	0.594

Simulations: false negative rates

- looking at significant differences along trajectory: with random smooths



Simulations: false negative rates

- models without random smooths have more power, but it comes at the cost of a very high rate of false positives
- model comparison has reasonably high power
- pointwise comparison is still problematic

Simulations: nested random smooths

- a twist on the first simulation:
  - a single underlying trajectory (i.e. no significant diff)
  - 100 words randomly distributed between two groups
  - 5 trajectories per word
- what random smooths should we include?
  - random smooths by words?
  - random smooths by trajectories?
  - random smooths by both?

Simulations: nested random smooths

- false positive rates (model comparison):

no rnd smooths:	0.699
by-traj rnd smooths:	0.406
by-word rnd smooths:	0.023
by-word & by-traj:	0.038

Simulations: nested random smooths

- ideal scenario: by-word and by-trajectory random smooths both included
- but it is sufficient to only include by-word trajectories if by-trajectory smooths are too costly

Recommendations based on simulations

- use ANOVA based on model comparison unless your hypothesis is specifically about parametric / smooth terms
- (if none of the individual terms are significant, ANOVA probably won't be either)
- random smooths are necessary when treatment predictor varies *between* items
- when there is a nested group structure, highest-level random smooth might be enough
- visual comparison should only be done *after* significance testing via ANOVA

A very brief case study

- data from Stuart-Smith et al. (2015):
  - Vr sequences
  - F3 trajectories for male speakers
  - older speakers, recorded in 1970, 1980, 1990, 2000
  - many different preceding/following vowels
  - many different words
- question: *has there been a significant change over time?*

A very brief case study

- included in the model:
  - parametric term for decade (factor)
  - four smooths by decade (not random)
  - smooth term for duration
  - random smooth by speakers
  - random smooths by preceding vowel
  - random smooths by following sound

A very brief case study

- parametric summary:

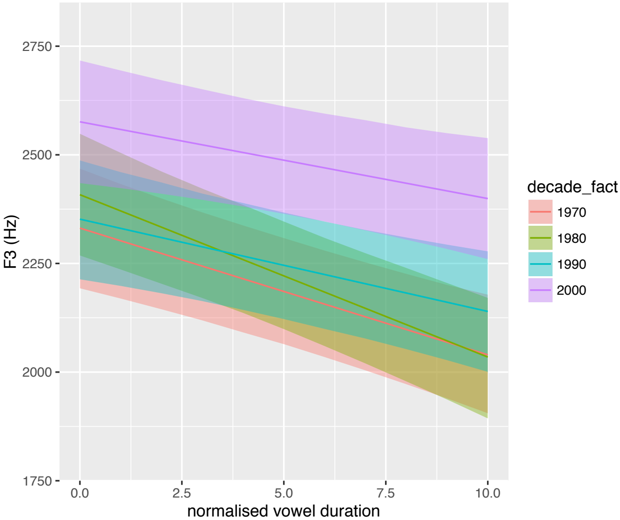
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2290.30	49.49	46.275	< 2e-16	***
decade_fact.L	207.81	44.42	4.678	2.99e-06	***
decade_fact.Q	103.23	44.42	2.324	0.0202	*
decade_fact.C	51.46	44.43	1.158	0.2468	

- smooth summary:

	edf	Ref.df	F	p-value	
s(measurement_no)	1.00	1.00	19.609	9.74e-06	***
s(duration)	8.52	8.52	17.638	< 2e-16	***
s(measurement_no):decade_fact1980	1.00	1.00	1.162	0.281	
s(measurement_no):decade_fact1990	1.00	1.00	1.112	0.292	
s(measurement_no):decade_fact2000	1.00	1.00	2.306	0.129	
s(measurement_no,speaker)	27.61	NA	NA	NA	
s(measurement_no,preceding)	15.14	NA	NA	NA	
s(measurement_no,following)	54.65	NA	NA	NA	

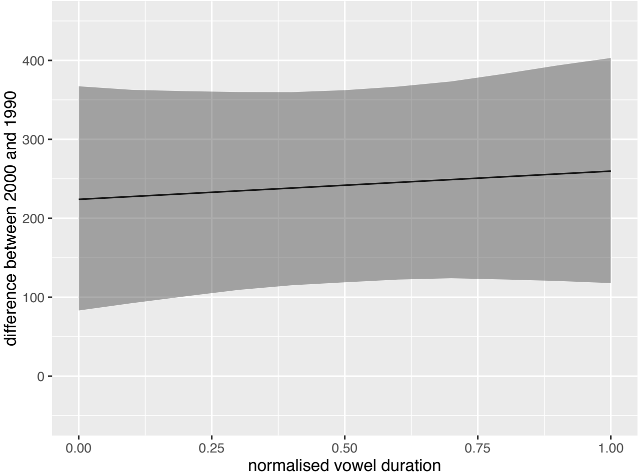
A very brief case study

- ANOVA: effect of decade is significant at $p < 0.012$



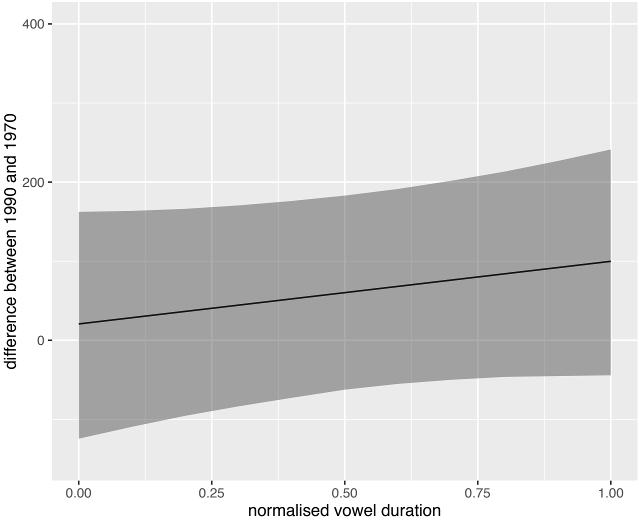
A very brief case study

- difference between 2000 and 1990:



A very brief case study

- difference between 1990 and 1970:



Cardoso, A. (2015). Variation in nasal-obstruent clusters and its influence on PRICE and MOUTH in Scouse. *English Language and Linguistics*, 19(3):505–532.

Fox, R. A. and Jacewicz, E. (2009). Cross-dialectal variation in formant dynamics of American English vowels. *The Journal of the Acoustical Society of America*, 126(5):2603–2618.

Hay, J. B., Pierrehumbert, J. B., Walker, A. J., and LaShell, P. (2015). Tracking word frequency effects through 130 years of sound change. *Cognition*, 139:83–91.

Labov, W., Ash, S., and Boberg, C. (2005). *The Atlas of North American English: Phonetics, phonology and sound change*. Mouton de Gruyter, Berlin.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in the study of vowels. *Journal of the Acoustical Society of America*, 24:175–184.

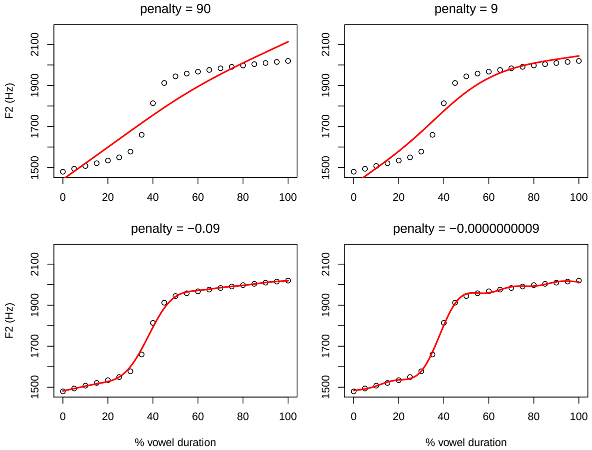
Stuart-Smith, J., Lennon, R., Macdonald, R., Robertson, D., Sóskuthy, M., José, B., and Evers, L. (2015). A dynamic acoustic view of real-time change in word-final liquids in spontaneous glaswegian. *Proceedings of the 18th International Congress of Phonetic Sciences*, 10-14 August 2015, Glasgow.

Watson, C. I. and Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in australian english vowels. The Journal of the Acoustical Society of America, 106:458–468.

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

Additional stuff

- 'wiggliness' of smooths estimated from data through penalised regression



Additional stuff

- looking at significant differences along trajectory: without random smooths