Modeling psycholinguistic effect timecourses with deconvolutional time series regression

Cory Shain and William Schuler

shain.3@osu.edu

Ohio State University

Question

How do effects on sentence processing difficulty unfold over time during naturalistic reading?

Hypothesis

Human responses to linguistic variables during reading are temporally diffuse and can be better explained by directly modeling temporal diffusion.

Problem

- Linear Mixed Effects models (LME) [3, 4, 5, 11, 9] and Generalized Additive Models (GAM) [12, 10] make implausible temporal independence assumptions.
- *Spillover* has several undesirable properties:
  - Spillover positions must be assumed in advance
  - Actual time is ignored
  - Can lead to parametric explosion
- Newer work on dealing with autocorrelation and non-stationarity in psycholinguistic time series [1, 2] still does not address *temporal diffusion*

Our approach

Use deconvolutional time series regression (DTSR) to find best-fit *impulse response functions* (IRF) from independent to dependent variables.

References

[1] Harald Baayen, Shravan Vasishth, Reinhold Kliegl, and Douglas Bates. The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, 94(Supplement C):206 – 234, 2017.

[2] R. Harald Baayen, Jacolien van Rij, Cecile de Cat, and Simon Wood. Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. In Dirk Speelman, Kris Heylen, and Dirk Geeraerts, editors, *Mixed Effects Regression Models in Linguistics*. Springer, Berlin, 2018.

[3] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[4] Vera Demberg and Frank Keller. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210, 2008.

[5] Victoria Fossum and Roger Levy. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of CMCL 2012*. Association for Computational Linguistics, 2012.

[6] Stefan L. Frank, Irene Fernandez Monsalve, Robin L. Thompson, and Gabriella Vigliocco. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45:1182–1190, 2013.

[7] Richard Futrell, Edward Gibson, Hal Tily, Anastasia Vishnevetsky, Steve Piantadosi, and Evelina Fedorenko. The natural stories corpus. *arXiv*, (1708.05763), 2017.

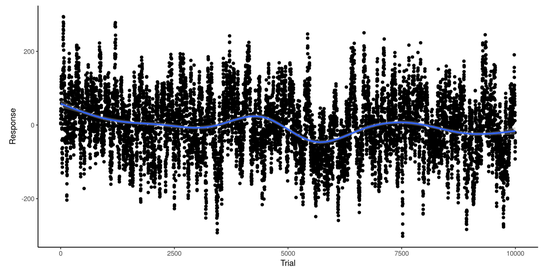
[8] Alan Kennedy, James Pynte, and Robin Hill. The Dundee corpus. In *Proceedings of the 12th European conference on eye movement*, 2003.

[9] Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Computational Linguistics for Linguistic Complexity Workshop*, pages 49–58. Association for Computational Linguistics, 2016.

[10] Nathaniel J. Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319, 2013.

[11] Marten van Schijndel and William Schuler. Hierarchic syntax improves reading time prediction. In *Proceedings of NAACL-HLT 2015*. Association for Computational Linguistics, 2015.

[12] Simon N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, Boca Raton, 2006.



Synthetic responses produced by a randomly sampled stationary impulse response function. Note the undulating GAM smooth in blue, suggesting non-stationarity that is not in fact present in the underlying generative process.

Mathematical description

Name	Explanation
$\mathbf{X}$	Fixed effects design matrix
$\mathbf{Z}$	Random effects design matrix
$\mathbf{y}$	Dependent variable observation vector
$\hat{\mu}$	Intercept
$\mathbf{I}$	Random intercept indicator matrix
$\hat{\mathbf{z}}$	Random intercepts vector
$\hat{\mathbf{b}}$	Fixed effects vector
$\hat{\mathbf{u}}$	Random effects vector
o	Number of fixed effect parameters
p	Number of random effect parameters
q	Number of parameters of g
$\hat{\mathbf{A}}_{\text{fixed}}$	$q \times o$ dimensional matrix of fixed convolution parameters
$\hat{\mathbf{A}}_{\text{ran}}$	$q \times p$ dimensional matrix of random convolution parameters
c	Function from index in y to last preceding index in $\mathbf{X}, \mathbf{Z}$
$t_{\mathbf{y}}$, $t_{\mathbf{X}}$, $t_{\mathbf{Z}}$	Timestamp functions mapping observation indices in $\mathbf{y}$, $\mathbf{X}$ and $\mathbf{Z}$ respectively to their timestamps.

$$\hat{\mathbf{y}} \stackrel{\text{def}}{=} \hat{\mu} + \mathbf{I}\,\hat{\mathbf{z}} + \hat{\mathbf{X}}_{\text{conv}}\,\hat{\mathbf{b}} + \hat{\mathbf{Z}}_{\text{conv}}\,\hat{\mathbf{u}} \quad (1)$$

$$\hat{\mathbf{X}}_{\text{conv}[i,j]} \stackrel{\text{def}}{=} \sum_{k=1}^{c(i)} \mathbf{X}_{[k,j]} \cdot g(\ t_{\mathbf{y}}(i) - t_{\mathbf{X}}(k);\ \hat{\mathbf{A}}_{\text{fixed}[*,j]}\ ) \quad (2)$$

$$\hat{\mathbf{Z}}_{\text{conv}[i,j]} \stackrel{\text{def}}{=} \sum_{k=1}^{c(i)} \mathbf{Z}_{[k,j]} \cdot g(\ t_{\mathbf{y}}(i) - t_{\mathbf{Z}}(k);\ \hat{\mathbf{A}}_{\text{ran}[*,j]}\ ) \quad (3)$$

Use SGD to minimize MSE loss, assuming Gaussian kernel g:

$$\hat{\mu}, \hat{\mathbf{z}}, \hat{\mathbf{b}}, \hat{\mathbf{u}}, \hat{\mathbf{A}}_{\text{fixed}}, \hat{\mathbf{A}}_{\text{ran}} = \underset{\hat{\mu},\hat{\mathbf{z}},\hat{\mathbf{b}},\hat{\mathbf{u}},\hat{\mathbf{A}}_{\text{fixed}},\hat{\mathbf{A}}_{\text{ran}}}{\text{argmin}} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_{[i]} - \hat{\mathbf{y}}_{[i]})^2 \quad (4)$$

Data

We modeled reading times in the Natural Stories [7], Dundee [8], and UCL [6] corpora. These three corpora provide a useful 2-way manipulation of series length and modality:

Long series		Short series
Natural Stories	**Dundee**	**UCL**
Self-paced reading	Eye-tracking	

Results

	Random effects structure					
	$\emptyset$		I_{subj}		S_{subj}	
System	Train	Test	Train	Test	Train	Test
LMEnoS	27761	27899	20546	20850	20187†	20620†
LMEoptS	27708	27849	20496	20797	20132†	20566†
LMEfullS	27674	27819	20461	20763	19979†	20505†
GAMnoS	27728	27838	20484	20816	—	—
GAMoptS	27692	27795	20446	20774	—	—
GAMfullS	27636	27767	20395	20733	—	—
DTSR	**20748**	**21046**	**19300**	**19608**	**18493**	**18820**

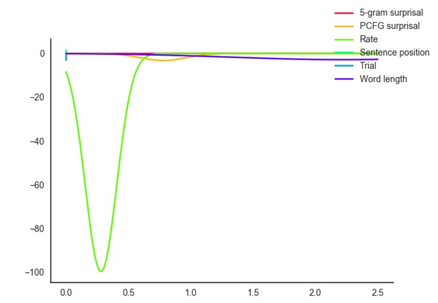
(a) Natural Stories

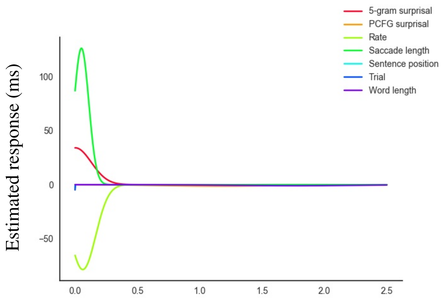
	Random effects structure					
	$\emptyset$		I_{subj}		S_{subj}	
System	Train	Test	Train	Test	Train	Test
LMEnoS	49068	48003	47871	46793	47565	46515
LMEoptS	48392	47654	47293	46533	46465†	45825†
LMEfullS	48000	47190	46852	46006	45909†	45422†
GAMnoS	48649	47590	47576	46531	—	—
GAMoptS	48822	47907	47691	46844	—	—
GAMfullS	48452	47611	47314	46521	—	—
DTSR	**46437**	**44307**	**46195**	**44008**	**45451**	**43114**

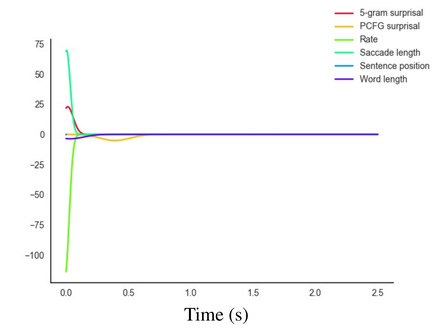
(b) Dundee

	Random effects structure					
	$\emptyset$		I_{subj}		S_{subj}	
System	Train	Test	Train	Test	Train	Test
LMEnoS	70738	64539	67583	61882	66034†	60930†
LMEoptS	70425	64314	67362	61765	65779	**60912**
LMEfullS	70084	63902	67151	61454	**64498**†	60982†
GAMnoS	68764	63195	66100	60982	—	—
GAMoptS	69259	63657	66242	61106	—	—
GAMfullS	**68505**	**62955**	**65993**	**60782**	—	—
DTSR	70139	64117	68852	62925	65994	61311

(c) UCL







Mean squared prediction error from DTSR vs. LME/GAM baseline models on the Natural Stories, Dundee, and UCL corpora. Train and test losses are presented for various random effects structures, since differences can be diagnostic of overfitting. For history modeling, each baseline system was tried with no spillover (noS), optimized spillover (optS) and full spillover 0-3 (fullS) of all predictors. Each system was evaluated with no random effects ($\emptyset$), by-subject random intercepts (I_{subj}), and by-subject random intercepts and slopes (S_{subj}). Daggers indicate convergence failure. Missing GAM cells are because prediction from mixed models is not implemented in `mgcv`. Plots show learned IRF from the S_{subj} DTSR model for each dataset. Each curve shows the estimated influence on reading latency over time of a single observation of 1 SD of the independent variable (e.g. for Dundee, 1 SD of 5-gram surprisal will incur about 35ms of slowdown at the current word and about 10ms of slowdown at a word observed 200ms later).

Discussion

- DTSR can provide a better predictive model of reading across modalities (Natural Stories and Dundee), but short series make IRF estimation less reliable (UCL).
- Learned IRF heavily emphasize low-level variables like rate (convolution of 1's at every stimulus) and saccade length. Linguistic variables appear to be of negligible value in SPR, suggesting a large influence of inertia.
- Temporal diffusion is mostly restricted to the first second after stimulus onset across datasets.
- Overall improvement of DTSR predictions over competitors (pooled across corpora) is highly significant by 2-tailed permutation test ($p = 0.0001$).
- Results support temporal diffusion as an important potential confound in psycholinguistic data and suggest DTSR as a tool to address it.