Modeling psycholinguistic effect timecourses with deconvolutional time series regression

Cory Shain and William Schuler, Ohio State

This work proposes a new approach to modeling psycholinguistic time series — deconvolutional time series regression (DTSR) — and applies it to existing datasets in order to show that temporal diffusion of effects may play a major role in subject behavior that existing modeling approaches do not fully account for. Although time series are abundant in psycholinguistic data, popular modeling techniques like linear mixed-effects regression (LME) [3, 4, 10, 8] and generalized additive models (GAM) [9] make strong temporal independence assumptions. A standard approach for handling the possibility of temporal diffusion or delay is to use spillover regressors, where a regressor value is used to predict the dependent variable at some later point in the time series. But this strategy has several undesirable properties. First, the modeled spillover positions are assumed as part of the model architecture rather than learned from data. Second, the use of relative event indices obscures potentially important details about the actual amount of time passed between events. And third, including multiple spillover positions per predictor quickly leads to a parametric explosion, especially if fully-articulated random effects structures are used.

DTSR avoids these pitfalls by learning *impulse response functions* (IRF) of the history of independent variable values rather than linear weights for the values at a particular timestep. Specifically, given an assumed parameterizable functional form for the IRF, DTSR learns best-fit convolutional parameters that mediate the relationship between two signals: the independent variable(s) on the one hand, and the dependent variable on the other. This procedure produces fine-grained estimates of the timecourse of each regressor's influence on the dependent measure.[1]

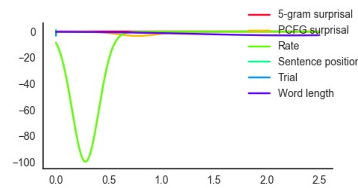
This work applies DTSR to an existing self-paced reading (SPR) corpus (Natural Stories [6]) and two existing eye-tracking corpora (Dundee [7] and UCL [5]). Both Natural Stories and Dundee contain long time series (stories/articles), while UCL contains short time series (isolated sentences). Modeled variables include *sentence position*, *series position*, *word length*, *5-gram surprisal*, and *PCFG surprisal*. The DTSR models also include a *rate* predictor equal to 1 at each fixation[2] and apply a Gaussian convolution to all predictors but *sentence position* and *series position*. Results are shown in Figure 1. Loss improvement on unseen data of DTSR over LME and GAM models, pooled across corpora, is highly significant by permutation test ($p \leq 1e-4$ for all six competitors, the smallest detectible value using 10,000 resampling iterations).

Four important generalizations emerge from these results. First, for Natural Stories and Dundee (but not UCL), DTSR produces a substantially better predictive model of reading than either LME or GAM, suggesting that it is indeed history modeling that drives the performance improvement. Second, the non-linguistic variables *rate* and *saccade length* dominate model estimates and are assigned larger influences than those typically observed in broad-coverage LME estimates [8]. Third, the large negative estimates for *rate* are likely capturing inertial characteristics (fast reading in the past engenders fast reading in the future). And fourth, while *5-gram surprisal* emerges as a strong predictor in eye-tracking (Dundee/UCL), the large prediction improvement for Natural Stories is driven almost entirely by *rate*. DTSR might thus be revealing inertial patterns in SPR (habituation to repeated button presses) to which eye-tracking may be less susceptible. While it is possible that different design decisions might change the relative importance of predictors, these results support the hypothesis that temporal diffusion is a major potential confound in psycholinguistic time series that, if left uncontrolled, might lead researchers to draw incorrect conclusions.

[1] The goal of controlling for temporal diffusion is separate from but complementary to recent work [1, 2] on controlling for auto-correlation and non-stationarity in psycholinguistic time series.

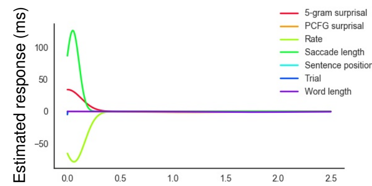
[2] Note that this variable is nonsensical for non-deconvolution regression approaches, since it is identical to the intercept term. For this reason it was not included in competing models.

System	∅		$\mathrm{I}_{\mathrm{subj}}$		$\mathrm{S}_{\mathrm{subj}}$	
	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**



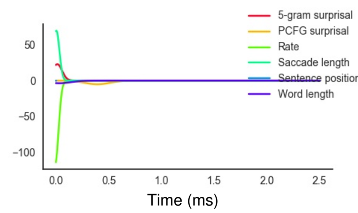
(a) Natural Stories

System	∅		$\mathrm{I}_{\mathrm{subj}}$		$\mathrm{S}_{\mathrm{subj}}$	
	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

System	∅		$\mathrm{I}_{\mathrm{subj}}$		$\mathrm{S}_{\mathrm{subj}}$	
	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

Figure 1: 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 (∅), by-subject random intercepts ($\mathrm{I}_{\mathrm{subj}}$), and by-subject random intercepts and slopes ($\mathrm{S}_{\mathrm{subj}}$). Daggers indicate convergence failure. Missing GAM cells are because prediction from mixed models is not implemented in `mgcv`. Plots show learned IRF from the $\mathrm{S}_{\mathrm{subj}}$ 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).

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

[4] 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.

[5] 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.

[6] Richard Futrell, Edward Gibson, Hal Tily, Anastasia Vishnevetsky, Steve Piantadosi, and Evelina Fedorenko. Natural stories corpus. in prep.

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

[8] 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.

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

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