Deconvolutional Time Series Regression: A Technique for Modeling Temporally Diffuse Effects

Cory Shain
Department of Linguistics
The Ohio State University
shain.3@osu.edu

William Schuler
Department of Linguistics
The Ohio State University
schuler.77@osu.edu

Abstract

Researchers in computational psycholinguistics frequently use linear models to study time series data generated by human subjects. However, time series may violate the assumptions of these models through temporal diffusion, where stimulus presentation has a lingering influence on the response as the rest of the experiment unfolds. This paper proposes a new statistical model that borrows from digital signal processing by recasting the predictors and response as convolutionally-related signals, using recent advances in machine learning to fit latent impulse response functions (IRFs) of arbitrary shape. A synthetic experiment shows successful recovery of true latent IRFs, and psycholinguistic experiments reveal plausible, replicable, and fine-grained estimates of latent temporal dynamics, with comparable or improved prediction quality to widely-used alternatives.

1 Introduction

Time series are abundant in many naturally-occurring phenomena of interest to science, and they frequently violate the assumptions of linear modeling and its generalizations. One confound that may be widespread in psycholinguistic data is *temporal diffusion*: the dependent variable may evolve slowly in response to its inputs, with the result that a particular predictor observed at a particular time may continue to exert an influence on the response as the rest of the process unfolds. If not properly controlled for, such a confound could have a detrimental impact on parameter estimation, model interpretation, and hypothesis testing.

The problem of temporal diffusion remains largely unsolved in the general case.[1] A standard approach for handling the possibility of temporally diffuse relationships between the predictors and the response is to use spillover or lag regressors, where the observed predictor value is used to predict subsequent observations of the response (Erlich and Rayner, 1983). But this strategy has several undesirable properties. First, the choice of spillover position(s) for a given predictor is difficult to motivate empirically. Second, in experiments with variably long trials the use of relative event indices obscures potentially important details about the actual amount of time that passed between events. And third, including multiple spillover positions per predictor quickly leads to parametric explosion on realistically complex models over realistically sized data sets, especially if random effects structures are included.

As a solution to the problem of temporal diffusion, this paper proposes deconvolutional time series regression (DTSR), a technique that directly models diffusion by learning parametric *impulse response functions* (IRFs) of the predictors that mediate their relationship to the response variable over time. Parametric deconvolution is difficult in the general case because the likelihood surface depends on the choice IRF kernel, requiring the user to re-derive estimators for each unique model structure. Furthermore, arbitrary IRF kernels are not guaranteed to afford analytical estimator functions or unique real-valued solutions. However, recent advances in machine learning have led to libraries like Tensorflow (Abadi et al., 2015) — which uses auto-differentiation to support optimization of arbitrary computation graphs — and Edward (Tran et al., 2016) — which enables black box variational inference (BBVI) on Tensorflow graphs. While these libraries are typically used to build and train deep networks, DTSR uses them

[1] Although recent work in computational psycholinguistics has begun to address separate but related problems in time series modeling (auto-correlation and non-stationarity) using generalized additive models (GAM) with a particular structure (Baayen et al., 2017, 2018).

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2679–2689
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

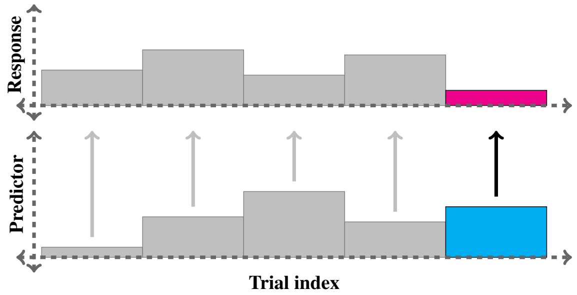

Figure 1: Effects in a linear time series model

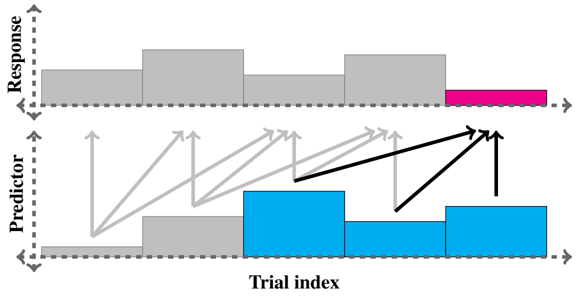

Figure 2: Linear time series model with spillover

to overcome the aforementioned difficulties with general-purpose temporal deconvolution by eliminating the need for hand-derivation of estimators and sampling distributions for each model.

The IRFs learned by DTSR are interpretable as estimates of the temporal shape of predictors' influence on the response variable. By convolving predictors with their IRFs, DTSR is able to consider arbitrarily long histories of independent variable observations in generating a given prediction, and (in contrast to spillover) model complexity is constant on the length of the history window. DTSR is thus a parsimonious technique for directly measuring temporal diffusion.

Figures 1–3 illustrate the present proposal and how it differs from linear time series models. As shown in Figure 1, a standard linear model assumes conditional independence of the response from all preceding observations of the predictor. This independence assumption can be weakened by including additional spillover predictors (Figure 2), at a cost of requiring additional parameters. In both cases, only the relative order of events is considered, not their actual distance in time. By contrast, DTSR recasts the predictor and response vectors as streams of impulses and responses (respectively) localized in time. It then fits latent IRFs that govern the influence of each predictor value on the response as a function of time (Figure 3).

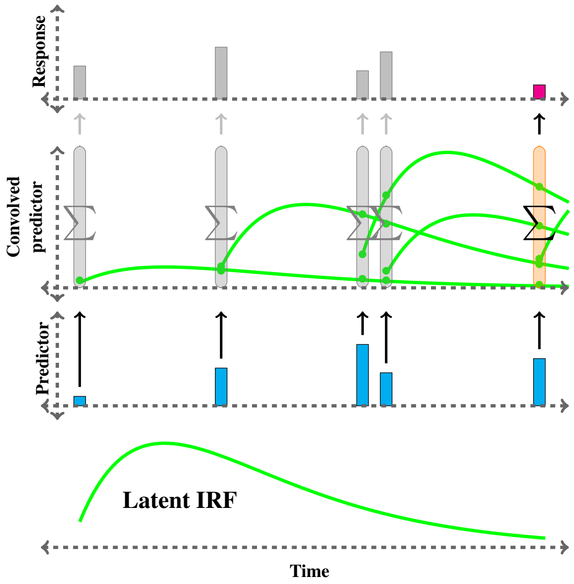

Figure 3: Effects of predictors in DTSR

This paper presents evidence that DTSR can (1) recover known underlying IRFs from synthetic data, (2) discover previously unknown temporal structure in human data (psycholinguistic reading time experiments), (3) provide support for the *absence* of temporal diffusion in settings where it might exist in principle, and (4) provide comparable (or in some cases improved) prediction quality to standard linear mixed-effects (LME) and generalized additive (GAM) models.

2 Related work

2.1 Non-deconvolutional time series modeling

The two most widely used tools for analyzing psycholinguistic time series are linear mixed effects regression (LME) (Bates et al., 2015) and generalized additive models (GAM) (Hastie and Tibshirani, 1986; Wood, 2006). LME learns a linear combination of the predictors that generates a given response variable. GAM generalizes linear models by allowing the response variable to be computed as the sum of smooth functions of one or more predictors.

In both approaches, responses are modeled as conditionally independent of preceding observations of predictors unless spillover terms are added, with the attendant drawbacks discussed in Section 1. To make this point more forcefully, take for example Shain et al. (2016), who find significant effects of constituent wrap-up ($p = 2.33\text{e-}14$) and dependency locality ($p = 4.87\text{e-}10$) in the

Natural Stories self-paced reading corpus (Futrell et al., 2018). They argue that this constitutes the first strong evidence of memory effects in broad-coverage sentence processing. However, it turns out that when one baseline predictor — probabilistic context free grammar (PCFG) surprisal — is spilled over one position, the reported effects disappear: $p = 0.816$ for constituent wrap-up and $p = 0.370$ for dependency locality. Thus, a reasonable but ultimately inaccurate assumption about baseline effect timecourses can have a dramatic impact on the conclusions supported by the statistical model. DTSR offers a way forward by bringing temporal diffusion under direct statistical control.

2.2 Deconvolutional time series modeling

Deconvolutional modeling has long been used in a variety of scientific fields, including economics (Ramey, 2016), epidemiology (Goldstein et al., 2011), and neuroimaging (Friston et al., 1998). Non-parametric deconvolutional models quantize the time series and fit estimates for each time point within some window, similarly to the spillover approach discussed above. These estimates can be unconstrained, as in finite impulse response models (FIR) (H. Glover, 1999; Ward, 2006), or smoothed with some form of regularization (Goutte et al., 2000; Pedregosa et al., 2014). Additional post-hoc interpolation is necessary in order to obtain a closed-form continuous IRF. These non-parametric approaches are prone to parametric explosion as well as sparsity problems when trials are variably spaced in time.

Parametric deconvolutional approaches (i.e. specific instantiations of DTSR) have evolved in certain fields (e.g. fMRI modeling) to solve particular problems, generally with some independently-motivated IRF kernel like the hemodynamic response function (HRF) (Friston et al., 1998; Lindquist and Wager, 2007; Lindquist et al., 2009). However, to our knowledge DTSR constitutes the first mathematical formulation and software implementation of general-purpose mixed effects parametric deconvolutional regression for arbitrary impulse response kernels. DTSR also supports Bayesian inference, enabling quantification of uncertainty in the absence of analytic formulae for standard errors. With these properties, DTSR expands the range of possible applications of parametric deconvolution beyond those fields for which appropriate formulations have already been developed.

3 Model definition

This section presents the mathematical definition of DTSR. For readability, only a fixed effects model is presented below, since mixed modeling substantially complicates the equations. The full model definition is provided in Appendix A. Note that the full definition is used to construct all reading time models reported in subsequent sections, since they contain random effects.

Let $\mathbf{X} \in \mathbb{R}^{M \times K}$ be a design matrix of M observations for K predictor variables and $\mathbf{y} \in \mathbb{R}^N$ be a vector of N responses, both of which contain contiguous temporally-sorted time series. DTSR models the relationship between $\mathbf{X}$ and $\mathbf{y}$ using parameters consisting of:

- a scalar intercept $\mu \in \mathbb{R}$
- a vector $\mathbf{u} \in \mathbb{R}^K$ of K coefficients
- a matrix $\mathbf{A} \in \mathbb{R}^{R \times K}$ of R IRF kernel parameters for K fixed impulse vectors
- a scalar variance $\sigma^2 \in \mathbb{R}$ of the response

To define the convolution step, let g_k for $k \in \{1, 2, \dots, K\}$ be a set of parametric IRF kernels, one for each predictor; let $\mathbf{a} \in \mathbb{R}^M$ and $\mathbf{b} \in \mathbb{R}^N$ be vectors of timestamps associated with each observation in $\mathbf{X}$ and $\mathbf{y}$, respectively; and let $\mathbf{c} \in \mathbb{N}^M$ and $\mathbf{d} \in \mathbb{N}^N$ be vectors of series ID's associated with each observation in $\mathbf{X}$ and $\mathbf{y}$, respectively. A filter $\mathbf{F} \in \mathbb{R}^{N \times M}$ admits only those observations in $\mathbf{X}$ that precede $\mathbf{y}_{[n]}$ in the same time series:

$$\mathbf{F}_{[n,m]} \overset{\text{def}}{=} \begin{cases} 1 & \mathbf{c}_{[m]} = \mathbf{d}_{[n]} \wedge \mathbf{a}_{[m]} \leq \mathbf{b}_{[n]} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The inputs $\mathbf{X}$ can be convolved with each IRF g_k by premultiplication with sparse matrix $\mathbf{G}_k \in \mathbb{R}^{N \times M}$ for $k \in \{1, 2, ..., K\}$ as defined below:

$$\mathbf{G}_k = g_k\left(\mathbf{b}\mathbf{1}^\top - \mathbf{1}\mathbf{a}^\top; \mathbf{A}_{[*,k]}\right) \odot \mathbf{F} \tag{2}$$

The convolution that yields the design matrix of convolved predictors $\mathbf{X}' \in \mathbb{R}^{N \times K}$ is then defined using products of the $\mathbf{G}$ matrices and the design matrix $\mathbf{X}$:[2]

$$\mathbf{X}'_{[*,k]} \overset{\text{def}}{=} \mathbf{G}_k\,\mathbf{X}_{[*,k]} \tag{3}$$

[2] This implementation of convolution is only exact when the predictors fully describe a discrete impulse signal. Exact convolution of samples from continuous signals is generally not possible because the signal is generally not analytically integrable. For continuous signals, DTSR can approximate the convolution as long as the predictor is interpolated between sample points at a fixed frequency prior to fitting.

Following convolution, DTSR is simply a linear model. The full model mean is the sum of (1) the intercept μ and (2) the product of the convolved predictor matrix $\mathbf{X}'$ and the coefficient vector $\mathbf{u}$:

$$\mathbf{y} \sim \mathcal{N}\left(\mu + \mathbf{X}'\mathbf{u}, \sigma^2\right) \tag{4}$$

4 Implementation

The present implementation defines the aforementioned equations[3] as a Bayesian computation graph in Tensorflow and Edward and trains it with black box variation inference (BBVI) using the Nadam optimizer (Dozat, 2016)[4] with a constant learning rate of 0.01 and minibatches of size 1024. For computational efficiency, histories are truncated at 128 timesteps. Prediction from the network uses an exponential moving average of parameter iterates with a decay rate of 0.998. Convergence was visually diagnosed.

The present experiments use a ShiftedGamma IRF kernel:

$$f(x; \alpha, \beta, \delta) = \frac{\beta^{\alpha}(x-\delta)^{\alpha-1}e^{-\beta(x-\delta)}}{\Gamma(\alpha)} \tag{5}$$

This is simply the PDF of the Gamma distribution augmented with a shift parameter δ allowing the lower bound of the support of the distribution to deviate from 0. We constrain δ to be strictly negative, thereby allowing the model to find a nonzero instantaneous response. We also constrain k to be strictly greater than 1, which deconfounds the shape and shift parameters. All bounded variables are constrained using the softplus bijection:

$$\text{softplus}(x) = \log(e^x + 1) \tag{6}$$

The ShiftedGamma kernel is used here because it can fit a wide range of response shapes and has precedent in the fMRI literature, where HRF kernels are often assumed to be Gamma-shaped (Lindquist et al., 2009).[5]

All parameters are given normal priors with unit variance. Prior means for the fixed IRF kernel parameters are domain-specific and discussed in the experiments sections below. To center the prior at an intercept-only model,[6] prior means for the intercept μ and variance σ^2 are set (respectively) to the empirical mean and variance of the response, and prior means for both fixed coefficients and random effects[7] are set to 0. Although the Bayesian implementation of DTSR is used for this study because it provides quantification of uncertainty, placing priors on the IRF kernel parameters is not crucial to the success of the system. In all experiments reported below, the MLE implementation arrives at similar solutions and achieves slightly better error.

In the interests of enabling the use of DTSR by the scientific community, the implementation of DTSR used here is offered as a documented open-source Python package with support for (1) Bayesian, variational Bayesian, and MLE inferences and (2) a variety model structures and impulse response kernels. The Tensorflow backend also enables GPU acceleration where available. Source code and links to documentation are available at `https://github.com/coryshain/dtsr`.

5 Experiment 1: Synthetic data

An initial experiment fits DTSR estimates to synthetic data to determine whether the model can recover known ground truth IRFs. Synthetic data were synthesized using the following procedure. First, 20 input vectors of size 10,000 were drawn from a standard normal distribution. These values synthesize an impulse stream containing 20 covariates, each with 10,000 observations. A ShiftedGamma IRF was then drawn for each of the 20 covariates. Coefficients were drawn from a uniform distribution $\mathcal{U}(-50, 50)$, and IRF parameters were drawn from the following distributions: $\alpha \sim \mathcal{U}(1, 6)$, $\beta \sim \mathcal{U}(0, 5)$, $\delta \sim \mathcal{U}(-1, 0)$. The prior means for the corresponding IRF kernel parameters are placed at the centers of these ranges. The stream of responses was generated by convolving the covariates with their corresponding IRFs. Gaussian noise with standard deviation 20 was injected into the response following generation. The 10,000 trials were spaced 100ms apart. As shown in Figure 4, the DTSR estimates for the

[3] As noted above, for expository purposes the definition in Section 3 only supports fixed-effects models. The full definition for mixed-effects DTSR models is provided in Appendix A. Mixed models are used throughout the experiments reported below.

[4] The Adam optimizer (Kingma and Ba, 2014) with Nesterov momentum (Nesterov, 1983)

[5] Other IRF kernels, including spline functions and composition of convolutions, are supported by the current implementation of DTSR but are not explored in these experiments. More details are provided in the software documentation.

[6] A model in which the response is insensitive to the model structure.

[7] See Appendix A for the definition of the mixed-effects DTSR model, which includes random effects.

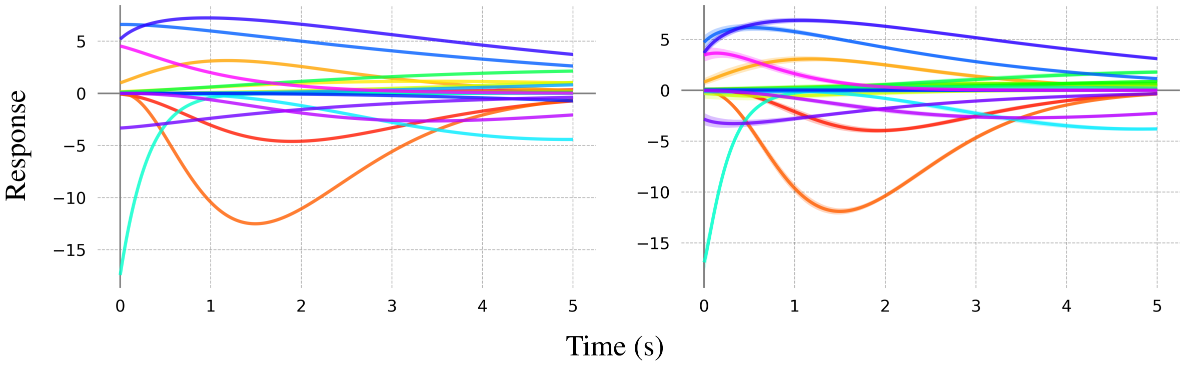

Figure 4: *Synthetic data*. True IRFs (left) and estimated IRFs with 95% credible intervals (right).

synthetic data are very similar to the ground truth, confirming that when the data-generating model matches the assumptions of DTSR, DTSR can recover its latent structure with high fidelity.

6 Experiment 2: Human reading times

6.1 Background and experimental design

The main interest of DTSR is the potential to better understand real-world dynamical systems like the human sentence processing response. Therefore, Experiment 2 applies DTSR to three existing datasets of naturalistic reading: Natural Stories (Futrell et al., 2018), Dundee (Kennedy et al., 2003), and UCL (Frank et al., 2013).

Natural Stories is a self-paced reading (SPR) corpus consisting of narratives designed to provide context-rich, fluent-sounding stimuli that nonetheless contain many grammatical constructions that rarely occur naturally in texts. The public release of the corpus contains data collected from 181 subjects. The stimulus set contains 10 stories with a total of 485 sentences and 10,245 tokens, for a total 848,768 fixation events.

Dundee is an eye-tracking (ET) corpus containing newspaper editorials read by 10 subjects, with incremental eye fixation data recorded during reading. The stimulus set contains 20 editorials with a total of 2,368 sentences and 51,502 tokens, for a total of 260,065 fixation events.

UCL is a reading corpus containing individual sentences that were extracted from novels written by amateur authors. The sentences were shuffled and presented in isolation to 42 subjects. The eye-tracking portion of the UCL corpus used in these experiments contains 205 sentences with a total of 1,931 tokens, for a total of 53,070 fixation events.

In all experiments, the response variable is log fixation duration (go-past duration for ET). Models use the following set of predictor variables in common use in psycholinguistics: *Sentence position* (index of word in sentence), *Trial* (index of trial in series),[8] *Saccade Length* (in words, ET only), *Word Length* (in characters), *Unigram Logprob*, and *5-gram Surprisal*. *Unigram Logprob* and *5-gram Surprisal* are computed by the KenLM toolkit (Heafield et al., 2013) trained on Gigaword 4 (Parker et al., 2009). In addition, DTSR enables fitting of a *Rate* predictor, which is simply a vector of ones, one for each observation, that is convolved using a latent IRF. *Rate* thus measures the response to density of stimulus presentation in the recent past. Since without deconvolution *Rate* is identical to the intercept, it is excluded from non-deconvolutional baseline models. Following standard practice in psycholinguistics, by-subject random coefficients for each of these predictors are included in all models (baseline and DTSR).[9]

ShiftedGamma IRFs are fitted to all predictors except *Sentence Position*, which is assigned a Dirac delta IRF (i.e. a linear coefficient) since it increases linearly within the sentence and is not expected to have a diffuse response. In plots, the *Sentence Position* estimate is shown as a stick function at time 0s. Prior means used for the IRF kernel parameters are $\alpha = 2$, $\beta = 5$, and $\delta = -0.5$. Together, these priors define an expected exponential-like IRF which decays to near-zero in about 1s, which seems plausible for human reading times. In practice they do not appear to be very constraining, since posterior means of fitted

[8] Except UCL, which contains isolated sentences, in which case *Trial* is identical to *Sentence Position*.

[9] By-subject IRF parameters were not used for this study because they substantially complicate the model and initial experiments using them showed little benefit on training data.

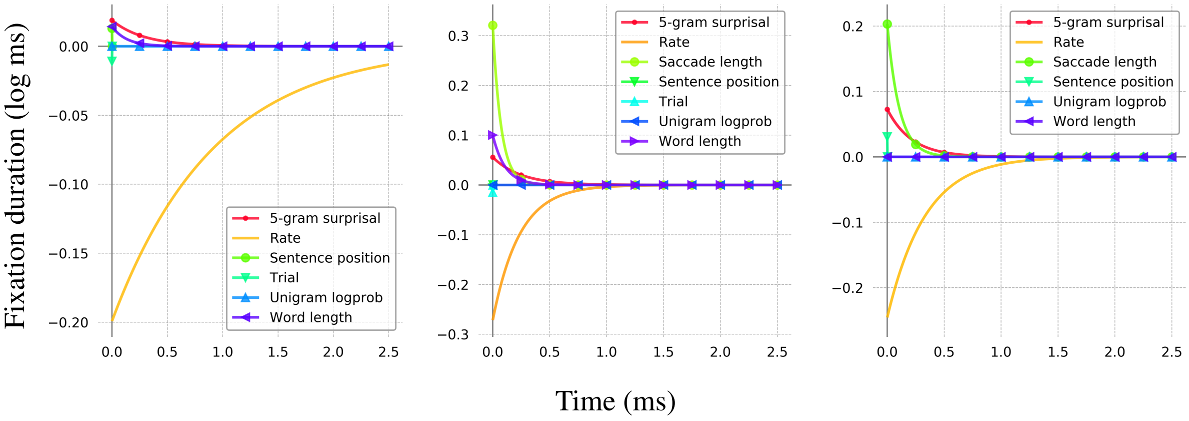

Figure 5: *Human data*. Estimated IRFs with 95% credible intervals for Natural Stories (left), Dundee (center) and UCL (right). Intervals are too tight to be seen.

models often deviate quite far from these values.

Existing work provides some expectations about the relationships of these variables to reading time. Processing difficulty is expected to increase with *Saccade Length*, *Word Length*, and *5-gram Surprisal*, and positive linear relationships have been shown experimentally (Demberg and Keller, 2008). *Unigram Logprob* is expected to be negatively correlated with reading times, since more frequent words are expected to be easier to process. *Sentence Position*, *Trial*, and *Rate* index different kinds of change in the response over time and their relationship has not been carefully studied, in part for lack of deconvolutional regression tools. Although reading times tend to decrease over the course of the experiment (Baayen et al., 2018), suggesting an expected negative effect of *Trial*, this may be partially explained by temporal diffusion. For the present study, all predictors are rescaled by their standard deviations.[10]

In all reading experiments, data are partitioned into training (50%), development (25%) and test (25%) sets. Outlier filtering is also performed. For Natural Stories, following Shain et al. (2016), items are excluded if they have fixations shorter than 100ms or longer than 3000ms, if they start or end a sentence, or if subjects missed 4 or more subsequent comprehension questions. For Dundee, following van Schijndel and Schuler (2015), unfixated items are excluded as well as (1) items following saccades longer than 4 words and (2) starts and ends of sentences, screens, documents, and lines. For UCL, unfixated items are excluded as well as (1) items following saccades longer than 4 words and (2) sentence starts and ends. Partitioning and filtering are applied only to the response series. The entire predictor history remains visible to the model.

From a modeling perspective, the primary results of interest in Experiment 2 are the IRFs themselves and the insights they provide into human sentence processing. However, to check the reliability of the DTSR estimates, prediction quality on unseen data is compared to that of non-deconvolutional baseline models fitted with LME and GAM.[11] Both baselines are fitted with and without three preceding spillover positions for each predictor (baselines with spillover are designated throughout this paper with the suffix -*S*).[12]

6.2 Results

The fitted IRFs for Natural Stories, Dundee, and UCL are shown in Figure 5. Effect sizes by corpus — computed here as the integral of each IRF over the first 10s — are shown in Table 1, along with

[10] Except *Rate*, which has no variance and therefore cannot be scaled by its standard deviation of 0.

[11] Formulae used to construct each model reported in this study are available in the associated code repository.

[12] This number of spillover positions is among the largest attested in the psycholinguistic literature because model complexity in LME and GAM increases substantially with each spillover position added, especially when by-subject random slopes are included for each spillover position for each variable. Indeed, many of the baseline models run for these experiments are already at the limits of tractability, as shown by the non-convergence reported in certain cells of Table 2. An advantage of the DTSR approach is that it can consider arbitrarily long histories at no cost to model complexity. While this permits DTSR to consider longer histories than its competitors (in these experiments, 128 timepoints vs. 4), DTSR is more constrained in its use of history since it must apply the same set of IRFs to all datapoints, while the baselines essentially fit separate models for each spillover position.

	Natural Stories			Dundee			UCL		
Predictor	Mean	2.5%	97.5%	Mean	2.5%	97.5%	Mean	2.5%	97.5%
Trial	-0.0053	-0.0057	-0.0049	-0.0085	-0.0010	-0.0071	—	—	—
Sent pos	0.0154	0.0148	0.0160	0.0004	-0.0013	0.0022	0.0340	0.0301	0.0379
Rate	-0.1853	-0.1858	-0.1848	-0.0649	-0.0659	-0.0640	-0.0806	-0.0832	-0.0781
Sac len	—	—	—	0.0249	0.0216	0.0207	0.0217	0.0209	0.0225
Word len	0.0020	0.0019	0.0021	0.0107	0.0105	0.0109	-8e-07	-1.7e-5	1.4e-5
Unigram	2.6e-6	-5e-6	2.2e-5	-2.0e-6	-3.9e-5	2.8e-5	1e-06	-4e-6	1.2e-5
5-gram	0.0057	0.0056	0.0059	0.0139	0.0134	0.0145	0.0159	0.0148	0.0171

Table 1: Effect sizes by corpus with 95% credible intervals based on 1024 posterior samples

95% credible intervals (CI). The IRFs (curves) in these plots represent the expected change in the response over time from observing a unit impulse of the predictor. For example, the Dundee model estimates that observing a standard deviation of *5-gram surprisal* engenders a slowdown of about 0.05 log ms instantaneously and a slowdown of about 0.03 log ms 250 ms after stimulus presentation. Because the response is reading time, positive IRFs represent inhibition and negative IRFs represent facilitation. Detailed interpretation of these curves is provided below in Section 6.3.

Table 2 shows prediction error from DTSR vs. baselines fitted to the same feature set. As shown, DTSR provides comparable or improved prediction performance to the baselines, even against the *-S* models which are more heavily parameterized. DTSR outperforms LME models on unseen data across all corpora and generally improves upon or closely matches the performance of GAM (with no spillover). Compared to GAM-S (with three additional spillover positions), there is a clear advantage of DTSR for Natural Stories but not for the eye-tracking (ET) datasets. This is likely due to more pronounced temporal confounds in Natural Stories (especially of *Rate*, which the baseline models cannot estimate) compared to the other corpora.[13] However, even in the absence of sufficiently diffuse effects to afford prediction improvements, the ability to measure diffusion directly is a major advantage of the DTSR model, since it can be used to detect the *absence* of diffusion in settings where it might in principle exist. Further discussion of the DTSR IRF estimates themselves is provided in Section 6.3.

As shown in Table 3, pooling across corpora, permutation testing reveals a significant improvement in MSE on test data of DTSR over each baseline system ($p = 0.0001$ for all comparisons).[14]

6.3 Discussion

Some key generalizations emerge from the DTSR estimates shown in Figure 5. The first is the pronounced facilitative role of *Rate* in all three models, but especially in Natural Stories. This means that fast reading in the recent past engenders fast reading in the present, because (1) observing a stimulus exerts a large-magnitude, diffuse, and negative (facilitative) influence on the subsequent response, and (2) the *Rate* contributions of the stimuli are additive. This result demonstrates an important pre-linguistic influence of *inertia* — a tendency toward slow overall change in base response rate. This effect is especially large-magnitude and diffuse in Natural Stories, which is self-paced reading and therefore differs in modality from the other datasets (which are eye-tracking). This suggests that SPR participants strongly habituate to repeated button pressing and stresses the importance of deconvolutional regression for bringing this low-level confound under control in analyzing SPR data, since it appears to have a large influence on the response and might otherwise confound model interpretation.

Second, effects are generally consistent with expectations: positive effects for *Saccade Length*, *Word Length*, and *5-gram Surprisal*, and a negative effect of *Trial*. The null influence of *Unigram Logprob* is likely due to the presence in the model of both *5-gram Surprisal* (which interpolates unigram probabilities) and *Word Length* (which is inversely correlated with *Unigram Logprob*). The biggest departure from prior expectations is the null estimate for *Word Length* in UCL. It appears

[13] Note that GAM-S is more heavily parameterized than DTSR in that it fits multidimensional spline functions of each spillover position of each predictor. This makes it difficult to generalize information about effect timecourses from GAM fits, motivating the use of DTSR for studies in which timecourses are a quantity of interest.

[14] To ensure comparability across corpora with different error variances, per-datum errors were first scaled by their standard deviations within each corpus. Standard deviations were computed over the joint set of error values in each pair of DTSR and baseline models.

	Natural Stories			Dundee			UCL		
System	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
LME	0.0803	0.0818	0.0815	0.2135	0.2133	0.2128	0.2613	0.2776	0.2561
LME-S	0.0789†	0.0807†	0.0804†	0.2099†	0.2103†	0.2095†	0.2509†	0.2754†	0.2557†
GAM	0.0798	0.0814	0.081	0.212	0.2116	0.2111	0.2576	0.2741	0.2538
GAM-S	0.0784	0.0802	0.0799	**0.2083**	**0.2085**	**0.2078**	**0.2440**	**0.2661**	**0.2457**
DTSR	**0.0648**	**0.0655**	**0.0650**	0.2100	0.2094	0.2088	0.2590	0.2752	0.2543

Table 2: Mean squared prediction error by system (daggers indicate convergence warnings)

Baseline	DTSR improvement (z-units)	*p*-value
LME	0.059	0.0001***
LME-S	0.054	0.0001***
GAM	0.057	0.0001***
GAM-S	0.051	0.0001***

Table 3: Overall pairwise significance of prediction improvement from DTSR vs. baselines

that the contribution of *Word Length* in this corpus can be effectively explained by other variables.

Third, the response estimates for Dundee and UCL (both of which are eye-tracking) are very similar, which suggests that DTSR is discovering replicable population-level features of the temporal profile for eye-tracking data.

Fourth, there is a general asymmetry in degree of diffusion between low-level perceptual-motor variables like *Saccade Length* and *Word Length*, whose responses tend to decay quickly, and the high-level *5-gram Surprisal* variable, whose response tends to decay more slowly. This is consistent with expectations from the sentence processing literature. Perceptual-motor variables involve rapid bottom-up computation (e.g. visual processing or motor planning/execution) and are therefore not expected to have a diffuse response, while surprisal involves top-down computation of future words given context, which might be more computationally expensive and therefore engender a slower response. While this outcome is suggested e.g. by the aforementioned finding that spillover 1 winds up being a stronger position for a surprisal predictor in the Shain et al. (2016) models, DTSR permits direct investigation of these dynamics.

7 A note on hypothesis testing

As a Bayesian model, DTSR supports hypothesis testing by querying the variational posterior. For example, as shown in Table 1, the credible interval (CI) for *5-gram Surprisal* in Natural Stories does not include zero (rejecting the null hypothesis of no effect), while the CI for *Unigram logprob* does (failing to reject). To control for effects of multicolinearity, one could perform ablative tests of fitted null and alternative models using (1) likelihood comparison or (2) predictive performance on unseen data.

However, DTSR estimates are obtained through non-convex stochastic optimization, which complicates hypothesis testing because of possible *estimation noise* due to (1) convergence to a local but not global optimum, (2) imperfect convergence to the local optimum, and/or (3) Monte Carlo estimation of the test statistic via posterior sampling. It cannot therefore be guaranteed that hypothesis testing results are due to differences in model structure rather than differences in relative amounts of estimation noise introduced by the fitting procedure. Thus, *p*-values (and, consequently, hypothesis tests) based on direct comparison of DTSR models should be considered approximate.

However, even in situations where such uncertainty in hypothesis testing is not acceptable, DTSR is appropriate for certain important use cases. First, DTSR can be used for *exploratory data analysis* in order to empirically motivate the spillover structure of the linear model. Spillover variables can be excluded or included based on the degree of temporal diffusion revealed by DTSR, permitting construction of linear models that are both parsimonious and effective for controlling temporal diffusion. Second, DTSR can be used to fit a *data transform* which is then applied to the data prior to statistical analysis. This approach is identical in spirit to e.g. the use of the canonical HRF to convolve predictors in fMRI models prior to linear regression. However, since DTSR is domain-general, it can be a valuable component in any analysis toolchain for time series.

8 Conclusion

This paper presented a variational Bayesian deconvolutional time series regression method as a solution to the problem of temporal diffusion in psycholinguistic time series data and applied it to both synthetic and human responses in order to better understand and control for latent temporal dynamics. Results showed that DTSR can yield a plausible, replicable, parsimonious, insightful, and predictive model of a complex dynamical system like the human sentence processing response and therefore support the use of DTSR for psycholinguistic time series modeling. While the present study explored the use of DTSR to understand human reading times, DTSR can in principle also be used to deconvolve other kinds of response variables, such as the HRF in fMRI modeling or the power/coherence response in oscillatory measures like electroencephalography, suggesting a rich array of potential applications of DTSR in computational psycholinguistics.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments. This work was supported by National Science Foundation grants #1551313 and #1816891. All views expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Martn Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Man, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Vigas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. Tensorflow: Large-scale machine learning on heterogeneous distributed systems.

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

R. Harald Baayen, Jacolien van Rij, Cecile de Cat, and Simon Wood. 2018. 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.

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

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

Timothy Dozat. 2016. Incorporating Nesterov momentum into Adam. In *ICLR Workshop*.

Kate Erlich and Keith Rayner. 1983. Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning & Verbal Behavior*, 22:75–87.

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

Karl J. Friston, Oliver Josephs, Geraint Rees, and Robert Turner. 1998. Nonlinear event-related responses in fMRI. *Magn. Reson. Med*, pages 41–52.

Richard Futrell, Edward Gibson, Hal Tily, Anastasia Vishnevetsky, Steve Piantadosi, and Evelina Fedorenko. 2018. The Natural Stories corpus. In *LREC 2018*.

Edward Goldstein, Benjamin J. Cowling, Allison E. Aiello, Saki Takahashi, Gary King, Ying Lu, and Marc Lipsitch. 2011. Estimating incidence curves of several infections using symptom surveillance data. *PLOS ONE*, 6(8):1–8.

C. Goutte, F. A. Nielsen, and K. H. Hansen. 2000. Modeling the hemodynamic response in fMRI using smooth FIR filters. *IEEE Transactions on Medical Imaging*, 19(12):1188–1201.

Gary H. Glover. 1999. Deconvolution of impulse response in event-related BOLD fMRI. 9:416–29.

Trevor Hastie and Robert Tibshirani. 1986. Generalized additive models. *Statist. Sci.*, 1(3):297–310.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

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

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Martin Lindquist and Tor Wager. 2007. Validity and power in hemodynamic response modeling: A comparison study and a new approach. 28:764–84.

Martin A. Lindquist, Ji Meng Loh, Lauren Y. Atlas, and Tor D. Wager. 2009. Modeling the hemodynamic response function in fmri: Efficiency, bias and mis-modeling. *NeuroImage*, 45(1, Supplement 1):S187 – S198. Mathematics in Brain Imaging.

Yurii E Nesterov. 1983. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. *English Gigaword LDC2009T13*.

Fabian Pedregosa, Michael Eickenberg, Philippe Ciuciu, Alexandre Gramfort, and Bertrand Thirion. 2014. Data-driven hrf estimation for encoding and decoding models. 104.

V.A. Ramey. 2016. Macroeconomic shocks and their propagation. volume 2 of *Handbook of Macroeconomics*, pages 71 – 162. Elsevier.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. 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.

Dustin Tran, Alp Kucukelbir, Adji B. Dieng, Maja Rudolph, Dawen Liang, and David M. Blei. 2016. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*.

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

B. Douglas Ward. 2006. Deconvolution analysis of fmri time series data.

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

A Definition of mixed effects DTSR

For expository purposes, in Section 3 the DTSR model was defined only for fixed effects. However, DTSR is compatible with mixed modeling and the implementation used here supports random effects in the model intercepts, coefficients, and IRF parameters. The full mixed-effects DTSR equations are presented below.

The definitions of $\mathbf{X}$, $\mathbf{y}$, μ, σ^2, $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{d}$, $\mathbf{F}$, M, N, K, and R presented in Section 3 are retained for the mixed model definition. The remaining variables and equations must be redefined to some extent. Mixed-effects DTSR models additionally contain the following parameters:

- a vector $\mathbf{o} \in \mathbb{R}^O$ of O random intercepts
- a vector $\mathbf{u} \in \mathbb{R}^U$ of U fixed coefficients
- a vector $\mathbf{v} \in \mathbb{R}^V$ of V random coefficients
- a matrix $\mathbf{A} \in \mathbb{R}^{R \times L}$ of R fixed IRF kernel parameters for L fixed impulse vectors
- a matrix $\mathbf{B} \in \mathbb{R}^{R \times W}$ of R random IRF kernel parameters for W random impulse vectors

Random parameters $\mathbf{o}$, $\mathbf{v}$, and $\mathbf{B}$ are constrained to be zero-centered within each random grouping factor.

To support mixed modeling, the fixed and random effects must first be combined using additional utility matrices. Let $\mathbf{O} \in \{0,1\}^{N \times O}$ be a mask matrix for random intercepts. A vector $\mathbf{q} \in \mathbb{R}^N$ of intercepts is:

$$\mathbf{q} \stackrel{\text{def}}{=} \mu + \mathbf{O}\,\mathbf{o} \tag{7}$$

Let $\mathbf{U} \in \{0,1\}^{L \times U}$ be an indicator matrix for fixed coefficients, $\mathbf{V} \in \{0,1\}^{L \times V}$ be an indicator matrix for random coefficients, and $\mathbf{V}' \in \{0,1\}^{N \times V}$ be a mask matrix for random coefficients. A matrix $\mathbf{Q} \in \mathbb{R}^{N \times L}$ of coefficients is:

$$\mathbf{Q} \stackrel{\text{def}}{=} \mathbf{1}\,(\mathbf{U}\,\mathbf{u})^\top + \mathbf{V}'\,\mathrm{diag}(\mathbf{v})\,\mathbf{V}^\top \tag{8}$$

Let $\mathbf{W} \in \{0,1\}^{L \times W}$ be an indicator matrix for random IRF parameters and $\mathbf{W}'_1, \ldots, \mathbf{W}'_n \in \{0,1\}^{R \times W}$ be mask matrices for random IRF parameters. Then matrices $\mathbf{P}_n \in \mathbb{R}^{R \times L}$ for $n \in \{1, 2, \ldots, N\}$ are:

$$\mathbf{P}_n \stackrel{\text{def}}{=} \mathbf{A} + (\mathbf{W}'_n \odot \mathbf{B})\,\mathbf{W}^\top \tag{9}$$

In each equation above, the random effects parameters are masked using the random effects filter associated with each data point. $\mathbf{Q}$ and $\mathbf{P}_n$ are then transformed into the impulse vector space using the indicator matrices $\mathbf{V}$ and $\mathbf{W}$, respectively. This procedure sums the random effects associated with each data point and adds them to the population-level parameters.

To define the convolution step, let g_l for $l \in \{1, 2, \ldots, L\}$ be parametric IRF kernels, one for each impulse. Convolution of $\mathbf{X}$ with each IRF kernel is performed by premultiplying the inputs

$\mathbf{X}$ with sparse matrix $\mathbf{G}_l \in \mathbb{R}^{N \times M}$ for $l \in \{1, 2, ..., L\}$:

$$(\mathbf{G}_l)_{[n,*]} \stackrel{\text{def}}{=} g_l \left(\mathbf{b}_{[n]} - \mathbf{a}^\top ; (\mathbf{P}_n)_{[*,l]} \right) \odot \mathbf{F}_{[n,*]} \tag{10}$$

Finally, let $\mathbf{L} \in \{0, 1\}^{K \times L}$ be an indicator matrix mapping the K predictors of $\mathbf{X}$ to the corresponding L impulse vectors of the model.[15] The convolution that yields the design matrix of convolved predictors $\mathbf{X}' \in \mathbb{R}^{N \times L}$ is then defined using a product of the convolution matrices $\mathbf{G}$, the design matrix $\mathbf{X}$, and the impulse indicator $\mathbf{L}$:

$$\mathbf{X}'_{[*,l]} \stackrel{\text{def}}{=} \mathbf{G}_l \, \mathbf{X} \, \mathbf{L}_{[*,l]} \tag{11}$$

The full model mean is the sum of (1) the intercepts and (2) the sum-product of the convolved predictors $\mathbf{X}'$ with the coefficient parameters $\mathbf{Q}$:

$$\mathbf{y} \sim \mathcal{N} \left(\mathbf{q} + (\mathbf{X}' \odot \mathbf{Q}) \, \mathbf{1}, \sigma^2 \right) \tag{12}$$

[15] Predictors and impulse vectors are distinguished because in principle multiple IRFs can be applied to the same predictor. In the usual case where this distinction is not needed, $\mathbf{L}$ is identity and $K = L$.