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

Cory Shain

Advisor:
William Schuler

Department of Linguistics
The Ohio State University

Abstract

Psycholinguists 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

Much of the data available to psycholinguistics is generated by processes that unfold in time. Examples include behavioral measures such as eye-movement records and self-paced reading latencies as well as neural measures like electroencephalography (EEG), magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), and electrocorticography (ECoG). If left uncontrolled, temporal confounds in psycholinguistic data can be problematic for interpretation and testing of statistical models used to analyze them (Baayen et al., 2017, 2018).

This paper addresses one possible temporal confound which I will refer to as *temporal diffusion*. Temporal diffusion exists when the response of a dependent variable to some inputs evolves slowly as a function of time, with the result that a particular input observed at a particular time continues to exert an influence on the response as the rest of the process unfolds. Temporal diffusion has been carefully studied in some psychological subfields. For example, a sizeable literature on fMRI has investigated the structure of the *hemodynamic response function* (HRF), which is known to govern the relatively slow response of blood oxygenation to neural activity (Boynton et al., 1996; Friston et al., 1998; H. Glover, 1999; Ward, 2006; Lindquist and Wager, 2007; Lindquist et al., 2009). The HRF is an instantiation of the more general notion of *impulse response function* (IRF) from the field of signal processing (Madisetti, 1997), where the response $g*h$ of a dynamical system as a function of time is described as a convolution over time of an impulse g with an IRF h:

$$(g * h)(t) = \int_0^t g(\tau)h(t-\tau)d\tau$$

The process of *deconvolution* seeks to infer the structure of h (the IRF) given that the impulses g (stimuli) and responses $g * h$ (psycholinguistic response) are known.

Although particular attention has been paid to the importance of impulse responses in fMRI, there are other kinds of psycholinguistic measures in which temporal diffusion might reasonably play a role. This paper focuses on one such example: measures of reading time, specifically fixation durations in eye-tracking and response times in self-paced reading. It has been known for decades that the response to properties of words in human subjects' reading behavior may not be fully instantaneous, but may "spill over" into the reading of subsequent words (Erlich and Rayner, 1983). A standard approach for handling the possibility of temporally diffuse relationships between the word properties and reading response is to use spillover or lag regressors, where a word's properties are used to predict subsequent observations of the re-

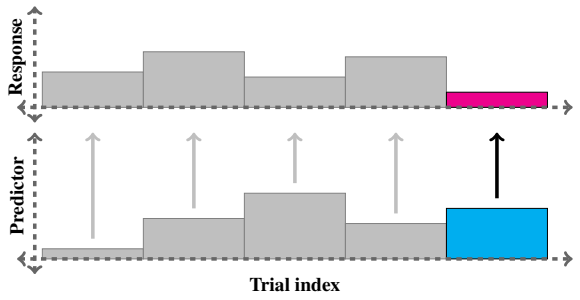


Figure 1: Effects in a linear time series model

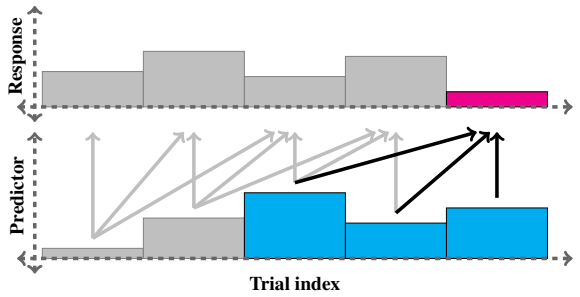


Figure 2: Linear time series model with spillover

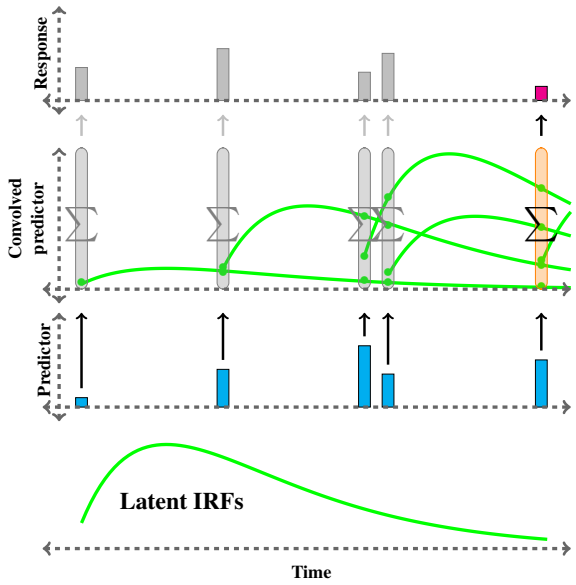


Figure 3: Effects of predictors in DTSR

sponse. But this strategy has several undesirable properties. First, the choice of spillover position(s) for a given predictor is difficult to motivate empirically. Second, since word fixations are variably long, the use of relative event indices obscures potentially important details about the actual amount of time that passed between events. 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. And fourth, if the predictors are autocorrelated, the spillover variants of each predictor will exhibit colinearities.

Deconvolutional modeling provides a way forward by supporting discovery from data of temporal diffusion in the reading response. However, major existing deconvolutional frameworks such as finite impulse response (FIR) models (Dayal and MacGregor, 1996) and vector autoregressive (VAR) models (Sims, 1980)[1] are not applicable to variably-spaced reading data because they discretize the time series, leading either (1) to severe sparsity if variable event durations are retained (few events are spaced exactly 207ms apart) or (2) distortion if they are removed (all events are treated as equally spaced).

As a solution to the problem of temporal diffusion, this paper proposes deconvolutional time series regression (DTSR), a continuous-time deconvolutional method that directly models diffusion by learning parametric IRFs of the predictors that mediate their relationship to the response variable over time. The implementation of DTSR proposed here takes advantage of the recent advent of the machine learning libraries 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 to overcome an important difficulty that otherwise holds of parametric deconvolution: the likelihood surface depends on the choice of IRF kernel, requiring re-derivation of estimators for each unique model structure. Auto-differentiation and Bayesian inference eliminate 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. DTSR mod-

[1] See Section 2 for discussion.

els are continuous-time and can therefore be optimized on naturalistic time series with variably-spaced events, including reading time data.

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 on the response as a function of time (Figure 3).

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 found 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 argued that their result constituted 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 — in this case, that the PCFG effect did not spill over — can have a dramatic impact on the conclusions supported by the statistical model. DTSR offers a way forward by building the possibility of temporal diffusion directly into the estimates, thereby avoiding the need to choose spillover positions as hyperparameters.

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). One widely-used approach to IRF discovery is finite impulse response modeling (FIR) (H. Glover, 1999; Ward, 2006). IRF models quantize the time series and use linear regression to fit estimates for each time point within some window, similarly to the spillover approach discussed above. These estimates can be unconstrained or smoothed with some form of regularization (Nikolaou and Vuthandam, 1998; Goutte et al., 2000; Pedregosa et al., 2014). Another major approach to deconvolution is vector autoregression (VAR), which discovers pairwise temporal relationships between all variables in the data (predictors and response) over some finite number of lags. VAR fits can be used to extract IRFs between pairs of variables. For both FIR and VAR models, additional post-hoc interpolation is necessary in order to obtain closed-form continuous-time IRFs. Non-parametric deconvolutional approaches like these are prone to parametric explosion and overfitting (Nikolaou and Vuthandam, 1998). Furthermore, as discussed above, their requirement of time discretization gives rise to sparsity or distortion when applied to time series with variable event duration. Finally, many psycholinguistic datasets contain data from many subjects and/or conditions, motivating the use of mixed-effects models. However, although mixed effects non-parametric deconvolutional models have been proposed (Gor-

rostieta et al., 2012), modeling random variation in the deconvolutional estimates severely increases model complexity by adding random covariates for each predictor/timepoint pair in the model.

Continuous-time mixed-effects IRF estimation for arbitrary impulse response kernels would overcome these difficulties and greatly extend the range of time series data to which deconvolutional modeling can be applied. To my knowledge, DTSR is the first mathematical formulation and software implementation of such an approach. By including a parametric impulse response as part of model design, DTSR avoids time discretization and the attendant problems with model complexity discussed above. DTSR thus expands the range of possible applications of deconvolutional modeling to include settings with variable event duration and heterogeneous sources of data.

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

A fixed-effects DTSR model therefore contains $2 + K + K \cdot R$ parameters: one intercept, K coefficients (one for each impulse), $K \cdot R$ IRF parameters (R parameters for each impulse, and one variance of the response. Mixed-effects DTSR models can also include random variation in the intercept, coefficients, and/or IRF parameters. This yields at most $1 + Z(1 + K + KR)$ estimates for a mixed effects model with Z total random grouping factor levels, although sub-maximal numbers of estimates can arise from restricting randomness to substructures of the model (e.g. to the intercept only). For fuller discussion of mixed-effects DTSR models, see the Appendix.

To define the convolution step, let g_k for $k \in \{1, 2, \ldots, K\}$ 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 $_k \in \mathbb{R}^{N \times M}$ for $k \in \{1, 2, ..., K\}$ as defined below:

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

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 convolution matrices and the design matrix $\mathbf{X}$:[3]

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

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 A note on multicolinearity

Note that the formulation in eq. 4 is simply a linear model on the convolved design matrix $\mathbf{X}'$. Therefore, the primary difference between linear and

[2] Throughout this paper I use the term *coefficients* to refer to what are often called *slopes* in linear models. This is to avoid falsely implying that the coefficients represent straight-line functions of the predictors, when in fact they are applied non-linearly to the predictors via the impulse response. Alternatively, the coefficients can be construed as slopes on the *convolved* predictors $\mathbf{X}'$, as shown in eq. 4.

[3] 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.

DTSR models is that DTSR additionally infers the parameters that generate $\mathbf{X}'$ jointly with the model intercept and coefficients.

Since DTSR depends internally on linear combination to generate its outputs, it is vulnerable to confounds from multicolinearity (correlated predictors) in much the same way that linear models are. In linear models, multicolinearity increases uncertainty about how to allocate covariation between predictors and response, since the predictors themselves covary. In the extreme case of perfect multicolinearity (i.e. one or more predictors are an exact linear combination of one or more other predictors), the model has no solution (Neter et al., 1989).

Multicolinearity in DTSR works in much the same way, with the added complexity that DTSR models also have a temporal dimension which may allow the fitting procedure to discover real characteristics of the global impulse response structure while struggling proportionally to the degree of multicolinearity to decompose that structure into predictor-wise IRFs. To understand this, note that the expected response t seconds after stimulus presentation is a weighted sum of the IRFs at t, with weights provided by the predictor values of the stimulus. When multicolinearity is low, the expected overall response can vary widely from one stimulus to another, since the IRFs are reweighted at each stimulus by roughly orthogonal predictor values. This variation in expected overall response provides clues to the system as to the magnitude, direction, and temporal shape of the individual response to each predictor. As multicolinearity increases, the expected overall response increasingly converges to a single shape which is shared across all stimuli (albeit scaled by the stimulus magnitude). In this setting, the model should still be able to correctly recover the global response characteristics, but may decompose it into predictor-wise responses that increasingly deviate from the true data generating model. In the extreme case that each predictor is identical, the expected response is identical for each stimulus, and the model will construct IRFs whose summation approximates the true global response profile but whose attribution of IRF components to predictors is random.

The above predictions are born out empirically by results presented in Section 6. While those results indicate that DTSR models are surprisingly robust to multicolinearity, models fitted to highly colinear data should be interpreted with caution, and perfectly colinear data should be avoided altogether. As in linear models, multicolinearity can be avoided by orthogonalizing predictors in advance (e.g. via principal components analysis). Empirical assessment of orthogonalization procedures in the DTSR setting is left to future work.

5 Implementation

The present implementation defines the equations from Section 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.[5] 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:

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

This is simply the probability density function of the Gamma distribution augmented with a shift parameter δ allowing the lower bound of the support of the distribution to deviate from 0. The following additional constraints are imposed: (1) δ is strictly negative, thereby allowing the model to find a non-zero instantaneous response, and (2) k is strictly greater than 1, deconfounding the shape and shift parameters. All bounded variables are constrained using the softplus bijection:

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

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).[6]

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

[5] 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.

[6] 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.

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,[7] prior means for the intercept μ and variance σ are set (respectively) to the empirical mean and variance of the response, and prior means for both fixed coefficients and random effects[8] 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`.

6 Experiment 1: Synthetic data

An initial experiment fits DTSR estimates to synthetic datasets. This experiment has two purposes: (1) to determine whether the model can recover known ground truth IRFs and (2) to assess the impact of multicolinearity in the predictors. Synthetic data were created by convolving sets of random covariates with known impulse responses in order to generate simulated response variables. Each simulation contained 20 covariates, and predictor and response streams each contained 10,000 observations spaced 100ms apart. A single set of impulse responses was randomly drawn at the outside of the experiment and shared across all synthetic datasets. To produce the impulse responses, ground truth coefficients were drawn from a uniform distribution $\mathcal{U}(-50, 50)$, and ground truth IRF parameters were drawn from the following distributions: $\alpha \sim \mathcal{U}(1, 6)$, $\beta \sim \mathcal{U}(0, 5)$, $\delta \sim \mathcal{U}(-1, 0)$.[9] The prior means for the corresponding IRF kernel parameters were placed at the centers of these ranges.

To manipulate multicolinearity in the predictors, predictor streams were drawn from multivariate normal distributions in which the variance-covariance matrix had a diagonal of 1 and all off-diagonal elements were set to the desired level of correlation. For example, predictors with correlation level $\rho = 0.5$ were drawn using the following variance-covariance matrix:

$$\begin{bmatrix} 1 & 0.5 & 0.5 & \dots & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & \dots & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & \dots & 0.5 & 0.5 & 0.5 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0.5 & 0.5 & 0.5 & \dots & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & \dots & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & \dots & 0.5 & 0.5 & 1 \end{bmatrix}$$

Four sets of predictors were generated in this way, one for each of $\rho = 0$ (uncorrelated predictors), $\rho = 0.25$, $\rho = 0.5$, $\rho = 0.75$, $\rho = 1$. Note that $\rho = 1$ is a degenerate case in which all 20 covariates are identical. As such, it is undecidable and therefore not a valid use case for DTSR. It is included in this experiment to assess the prediction outlined in Section 4 that in such a setting DTSR will faithfully recover the global response profile while decomposing and distributing it to predictors at random.

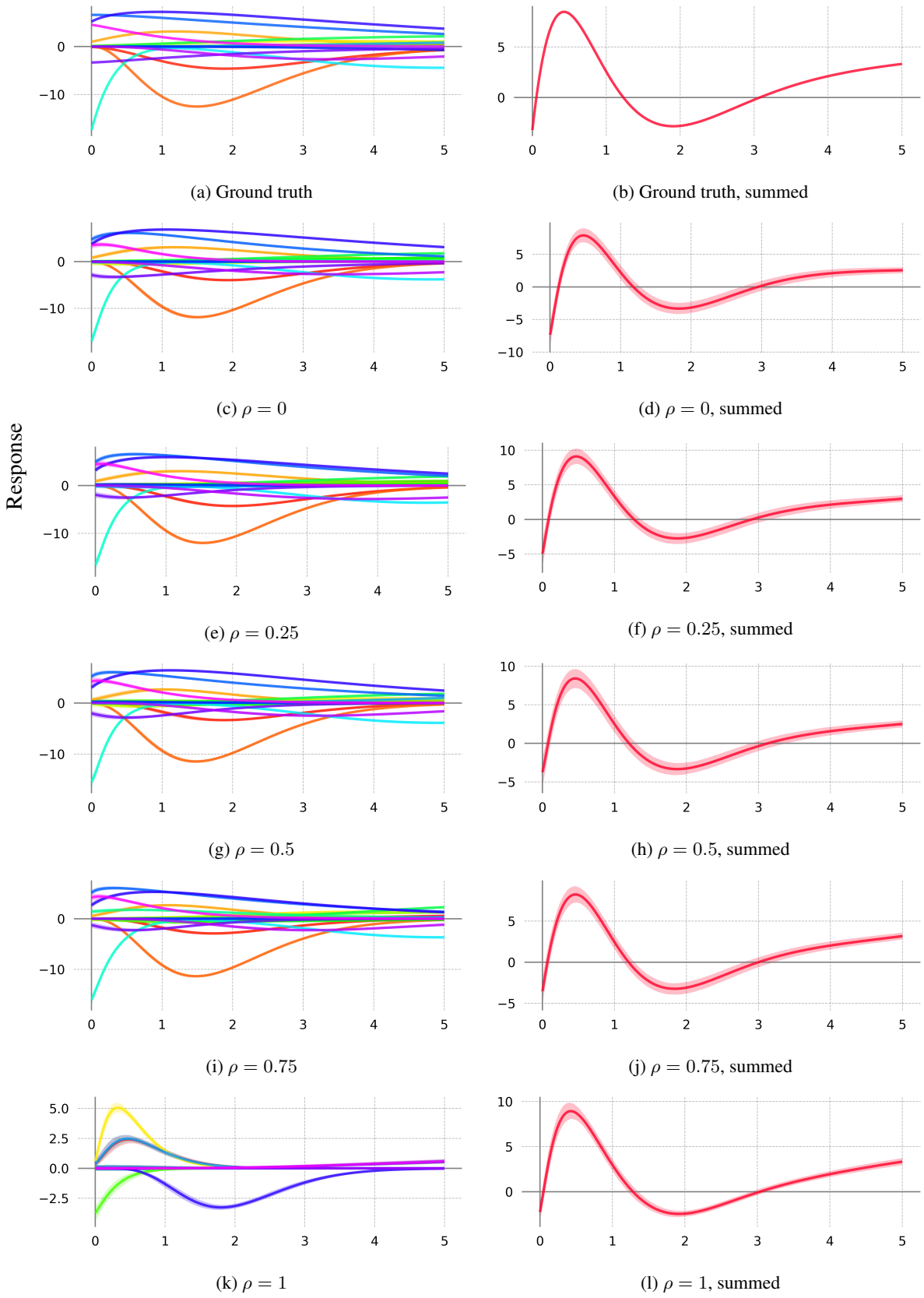
For each set of predictors, the stream of responses was generated by convolving the covariates with their corresponding IRFs and multiplying them by their coefficients. For each response vector, Gaussian noise with standard deviation 20 was drawn and added to the responses following generation.

As shown in the left column of Figure 4, the DTSR estimates for the synthetic data with $\rho \leq 0.75$ 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 even in the presence of strong multicolinearity. Nonetheless, there are small degradations in model fidelity with increasing colinearity, supporting the hypothesis that the difficulty of model identification increases

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

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

[9] These ranges generally yield IRF with peak dynamics within the system's visibility window of 12.8s. Visibility is curtailed by history truncation (128 timesteps, see above), leading to a maximum visibility of 12.8s since each trial is 0.1s long ($128 \times 0.1 = 12.8$).



(a) Ground truth

(b) Ground truth, summed

(c) $\rho = 0$

(d) $\rho = 0$, summed

(e) $\rho = 0.25$

(f) $\rho = 0.25$, summed

(g) $\rho = 0.5$

(h) $\rho = 0.5$, summed

(i) $\rho = 0.75$

(j) $\rho = 0.75$, summed

(k) $\rho = 1$

(l) $\rho = 1$, summed

Figure 4: *Synthetic data*. Predictor-wise (left) and summed IRFs in the true (top) and estimated models at varying levels of impulse multicolinearity ρ. Estimates are shown with 95% credible intervals.

with colinearity and motivating the avoidance of strongly colinear data when possible.

The right column of Figure 4 shows a sum of all the IRFs in the system. This corresponds to the expected response to a stimulus containing a unit impulse for each predictor. Since this is a deterministic function of the component IRFs, summed estimated responses should closely approximate summed true responses, which Figure 4 shows to be the case in all models. The reason the summed response is of interest here is that, as argued in Section 4, with increasing multicolinearity the synthetic response at each stimulus increasingly resembles (a multiple of) a single response profile. In these experiments where predictors have identical means and variances and are positively correlated, this response profile approaches the summed response in actual fact, and in such a setting DTSR is expected to find component IRFs that sum to the global response, even if their allocation to particular predictors is largely arbitrary. The $\rho = 1$ example (bottom row) clearly shows this to be the case. Although the model has discovered component IRFs and assigned them to predictors in ways that do not resemble the true model at all, the sum of these IRFs nonetheless captures the true summed response with very high fidelity. The summed response profile is the only characteristic of the true model that DTSR could have learned in this setting, and DTSR does successfully recover it.

7 Experiment 2: Human reading times

7.1 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 context-rich narratives that resemble fluent storytelling while nonetheless containing many grammatical constructions that rarely occur naturally in texts. The public release of the corpus contains data collected from 181 subjects who paged through the stories on a computer screen, pressing a button to reveal the next word. The amount of time spent on each word was recorded as the response variable. The stimulus set contains 10 stories with a total of 485 sentences and 10,245 word tokens, for a total 848,768 fixation events (where one event is a single subject viewing a single word token).

Dundee is an eye-tracking 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 word tokens, for a total of 260,065 fixation events (where one event is a single subject fixating a single word token for the first time from the left).

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 word tokens, for a total of 53,070 fixation events.

In all experiments, the response variable is log fixation duration (go-past duration for eye-tracking). 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),[10] *Saccade Length* (in words, eye-tracking only), *Word Length* (in characters), *Unigram Log Probability*, and *5-gram Surprisal*. *Unigram Log Probability* and *5-gram Surprisal* are computed by the KenLM toolkit (Heafield et al., 2013) trained on Gigaword 4 (Parker et al., 2009). Examples of studies using some or all of these predictors include Demberg and Keller (2008); Frank and Bod (2011); Smith and Levy (2013) and Baayen et al. (2018).

In addition, DTSR enables fitting of the impulse response to a *Rate* predictor, which is simply a vector of ones, one for each observation (Brennan et al., 2012). *Rate* can be viewed as a deconvolutional intercept, providing information about stimulus *timing* but no information about stimulus properties. The fitted response to *Rate* is an estimate of the baseline response of the system, with expected deviation from that baseline governed by the estimated responses to the other predictors, which vary from stimulus to stimulus. *Rate* estimates also provide insight into the effect of temporal *density* of stimulus presentation, since the *Rate* responses accumulate for stimuli that impinge on

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

the system before it has effectively finished responding to *Rate* from previous stimuli. 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 intercepts along with by-subject random coefficients for each of these predictors are included in all models (baseline and DTSR).[11] All predictors are rescaled by their standard deviations prior to fitting.[12] A single DTSR model was fitted to each corpus.

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 Log Probability* 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.

The predictors *Saccade Length*, *Word Length*, *Unigram Log Probability*, and *5-gram Surprisal* are all motor, perceptual, or linguistic variables to which the sentence processing system has been shown to respond upon word fixation (Demberg and Keller, 2008) and to which the response might not be perfectly instantaneous. To the extent that temporally diffuse responses to any of these predictors exist, it is desirable that the model be able to capture them. By contrast, *Trial* and *Sentence Position* merely index progress through documents and sentences respectively. They are not perceptual or linguistic properties of the experiment, and it is unclear how any diffuse impulse response attributed to them would be interpreted. Following prior work (Demberg and Keller, 2008; Baayen et al., 2018), their presence in the model is motivated by the possibility of trends in the response. For this reason, ShiftedGamma IRFs are fitted to all predictors except *Trial* and *Sentence Position*, which are assigned a Dirac delta IRF (i.e. a linear coefficient). Consequently, in plots, the *Trial* and *Sentence Position* estimates are shown as stick functions at time 0s.

Specifying the Bayesian model requires stating priors over the IRF parameters. Unfortunately, the existing literature on human reading does not provide detailed evidence as to the temporal characteristics of the response to predictors, due at least in part to the difficulty of estimating these characteristics without the aid of DTSR. Nonetheless, some relevant signposts exist. For example, studies using many spillover positions have found the influence of linguistic predictors like surprisal to decrease monotonically with spillover position (Smith and Levy, 2013). Although spillover does not have a clear continuous-time interpretation, these results support a strictly-decreasing shape for the prior on the IRF kernel. In addition, the timecourse of the sentence processing response has been very carefully studied in the domain of electroencephalography, with consistent finding of a number of distinct event-related potentials generally occurring within one second after stimulus onset (Sur and Sinha, 2009). Together, these considerations support prior bias towards a stricly-decreasing exponential-like IRF with response primarily localized to the first second after stimulus onset. In this study, the prior means used to meet this desideratum are $\alpha = 2$, $\beta = 5$, and $\delta = -0.5$. Note that these are simply choices for the prior; the model can deviate unboundedly far from them in its parameters as motivated by the data, potentially finding late-peaking responses, more rapidly decaying responses, and more slowly decaying responses. In practice, the choice of prior does not appear to be very constraining, since posterior means of fitted models often deviate quite far from the prior means.

In all reading experiments, data were partitioned into training (50%), development (25%) and test (25%) sets. Outlier filtering was also

[11] 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. By-word random intercepts, though common in psycholinguistic studies (Demberg and Keller, 2008), were also avoided because (1) estimates for *Word Length* and *Unigram Log Probability* are of interest for this study but are context-free and can therefore be wholly or partially consumed by the random intercepts and (2) early experiments suggested that by-word intercepts led to overfitting (based on development set performance) in both DTSR and baseline models.

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

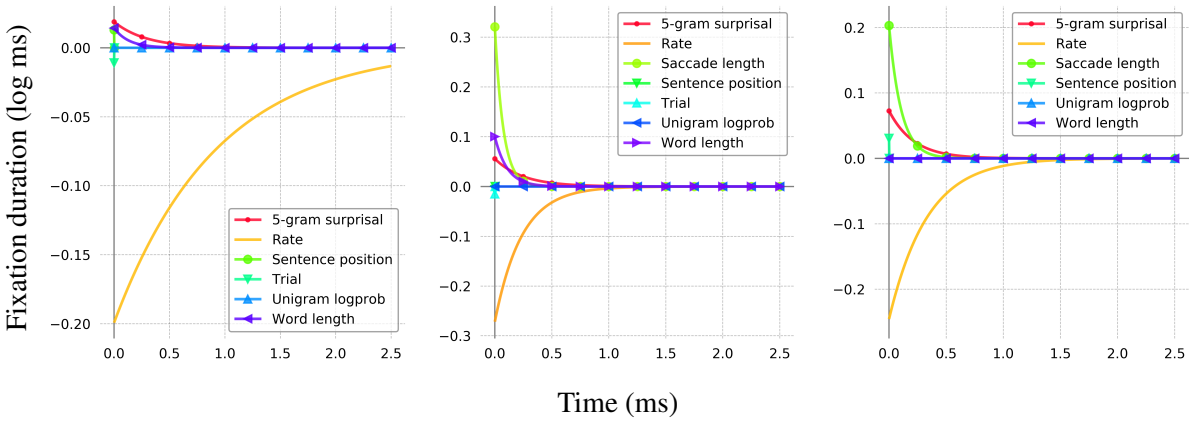


Figure 5: *Human data*. Estimated IRFs with 95% credible intervals for Natural Stories (left), Dundee (center) and UCL (right). Intervals are too tight to be seen.

performed. For Natural Stories, following Shain et al. (2016), items were 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 were 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 were excluded as well as (1) items following saccades longer than 4 words and (2) sentence starts and ends.[13] Partitioning and filtering were applied only to the response series. The entire predictor history remained 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.[14] 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*).[15]

7.2 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 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. In all estimates the response decreases monotonically in magnitude, with a peak instantaneous influence that decays with time. In general, the estimated responses for *Word Length*, *Unigram Log Probability*, and *5-gram Surprisal* are positive and concentrated in the first second after stimulus onset. There are also large-magnitude negative *Rate* estimates across all

[13] Most of these outlier filters are designed to minimize the influence of boundary effects like implicit prosody (Breen, 2014). Differences across corpora in exclusion criteria are driven by a combination of (1) differences in precedent established by studies that use these corpora (see citations), (2) differences in modality, since e.g. unfixated items and long saccades are only relevant to eye-tracking, and (3) differences in source data, since e.g. only Dundee provides information about screen, document, and line boundaries.

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

[15] 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 much longer histories than its competitors (in these experiments, 128 timepoints vs. 4), DTSR is much more constrained in its use of history in that 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

corpora, with an especially pronounced *Rate* response in Natural Stories. More detailed interpretation of these curves is provided below in Section 7.3.

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 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. 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.

Note also that even in the absence of very diffuse effects that 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.

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).[16]

[16]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. The reason DTSR outperforms GAM-S in the pooled permutation test despite underperforming it on Dundee and UCL is that it gives a large relative improvement on Natural Stories, which is itself much larger than the other two datasets.

7.3 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. If left uncontrolled, variation due to *Rate* (i.e. due to the timing structure of the stream of stimuli) might be mis-attributed to other predictors, possibly confounding 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 Log Probability* 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 Log Probability*). The biggest departure from prior expectations is the null estimate for *Word Length* in UCL. It appears that the contribution of *Word Length* in this corpus can be effectively explained

	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

by other variables.

Third, the response estimates for Dundee and UCL (both of which are eye-tracking) are 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, as shown by the existence in all corpora of a point in time after which the *5-gram Surprisal* response exceeds the *Saccade Length* and *Word Length* responses in magnitude. This is consistent with a view of sentence processing in which perceptual-motor variables have a faster response because they involve rapid bottom-up computation (e.g. visual processing or motor planning/execution), while surprisal has a slower response because it involves more expensive top-down computations about future words given context (Friederici, 2002; Connor et al., 2004; Bonte et al., 2005). 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, indicating a diffuse response that spreads to or even peaks at subsequent words, DTSR permits direct investigation of these dynamics.

8 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 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 non-parametric tests of predictive performance on in-sample or out-of-sample 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, hypothesis tests based on direct comparison of DTSR models rely on a (possibly incorrect) assumption that the models are effectively optimal. The empirical results presented in Section 6 support this assumption by showing DTSR fits that are consistently close to the true data generating model, even in the presence of strongly correlated predictors, suggesting that non-global optima may not be a severe confound in practice. The procedure of statistically comparing models fitted via non-convex optimization is identical to the one typically followed for model comparison in machine learning (Demšar, 2006), where differences between two models in performance on held-out data is subjected to statistical testing, even when the fits are obtained in a non-convex deep learning setting. The outcomes of such tests are implicitly conditional on the fitted parameters of each model. The probability of that the fitted parameters are indeed optimal for a given experiment can be increased by Monte Carlo techniques in which multiple randomly initialized models are fitted and then aggregated, an approach

to which DTSR is also amenable.

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. For example, if the average trial duration is 300ms and the estimated response to a predictor is near zero at that point, this could be used to motivate the exclusion of spillover variables for that predictor. To overcome the limitation that linear models cannot fit estimates of *Rate*, the convolved *Rate* predictor from the DTSR fit can be extracted and supplied to the linear model as a predictor. 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 (Boynton et al., 1996). The canonical HRF may be sub-optimal for the data at hand, yet likelihood maximization and model comparison are conditional on it. The same would hold of linear fits to predictors convolved using DTSR. However, unlike the canonical HRF, since DTSR is domain-general, it can be integrated into any analysis toolchain for time series.

9 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 (Boynton et al., 1996) or intracranial event-related potentials (Walter et al., 1964) in oscillatory measures like electroencephalography, suggesting a rich array of potential applications of DTSR in computational psycholinguistics.

References

Martín 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 Mané, 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 Viégas, 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.

Milene Bonte, Tiina Parviainen, Kaisa Hytönen, and Riitta Salmelin. 2005. Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. *Cerebral Cortex*, 16(1):115–123.

Geoffrey M Boynton, Stephen A Engel, Gary H Glover, and David J Heeger. 1996. Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, 16(13):4207–4221.

Mara Breen. 2014. Empirical investigations of the role of implicit prosody in sentence processing. *Language and Linguistics Compass*, 8(2):37–50.

Jonathan Brennan, Yuval Nir, Uri Hasson, Rafael Malach, David J Heeger, and Liina Pylkkänen. 2012. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, 120(2):163–173.

Charles E Connor, Howard E Egeth, and Steven Yantis. 2004. Visual attention: bottom-up versus top-down. *Current biology*, 14(19):R850–R852.

Bhupinder S Dayal and John F MacGregor. 1996. Identification of finite impulse response models: methods and robustness issues. *Industrial & engineering chemistry research*, 35(11):4078–4090.

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

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(Jan):1–30.

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 and Rens Bod. 2011. Insensitivity of the Human Sentence-Processing System to Hierarchical Structure. *Psychological Science*, 22(6):829–834.

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(4):1182–1190.

Angela D Friederici. 2002. Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, 6(2):78–84.

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, Harry J . Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2018. The Natural Stories Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

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.

Cristina Gorrostieta, Hernando Ombao, Patrick Bédard, and Jerome N Sanes. 2012. Investigating brain connectivity using mixed effects vector autoregressive models. *NeuroImage*, 59(4):3347–3355.

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. *NeuroImage*, 9:416–429.

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/1412.6.

Martin Lindquist and Tor Wager. 2007. Validity and power in hemodynamic response modeling: A comparison study and a new approach. *Human brain mapping*, 28:764–784.

Martin A Lindquist, Ji Meng Loh, Lauren Y Atlas, and Tor D Wager. 2009. Modeling the hemodynamic response function in fMRI: Efficiency, bias and mismodeling. *NeuroImage*, 45(1, Supplement 1):S187 – S198.

Vijay Madisetti. 1997. *The digital signal processing handbook*. CRC press.

Yurii E Nesterov. 1983. A method for solving the convex programming problem with convergence rate O(1/kˆ2). In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547.

John Neter, William Wasserman, and Michael H Kutner. 1989. Applied linear regression models.

Michael Nikolaou and Premkiran Vuthandam. 1998. FIR model identification: Parsimony through kernel compression with wavelets. *AIChE Journal*, 44(1):141–150.

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. *NeuroImage*, 104.

Valerie A Ramey. 2016. Macroeconomic shocks and their propagation. In John B Taylor and Harald Uhlig, editors, *Handbook of Macroeconomics*, volume 2 of *Handbook of Macroeconomics*, pages 71–162. Elsevier.

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

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.

Christopher A Sims. 1980. Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, pages 1–48.

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

Shravani Sur and V K Sinha. 2009. Event-related potential: An overview. *Industrial psychiatry journal*, 18(1):70.

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*.

W. Walter, R. Cooper, V. J. Aldridge, W. C. McCallum, and A. L. Winter. 1964. Contingent Negative Variation: An Electric Sign of Sensori-Motor Association and Expectancy in the Human Brain. *Nature*.

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, $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{d}$, $\mathbf{F}$, 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 $\in \mathbb{R}^{R \times W}$ of R random IRF kernel parameters for W random impulse vectors

Random parameters $\mathbf{o}$, $\mathbf{v}$, and are constrained to be zero-centered.

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:

$$\mathbf{q} \stackrel{\text{def}}{=} \mu + \mathbf{O}\,\mathbf{o} \tag{6}$$

Let $\in \{0,1\}^{L \times U}$ 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:

$$\mathbf{Q} \stackrel{\text{def}}{=} \mathbf{1}\,(\,\mathbf{u})^\top + \mathbf{V}'\,\text{diag}(\mathbf{v})\,\mathbf{V}^\top \tag{7}$$

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:

$$\mathbf{P}_n \stackrel{\text{def}}{=} \mathbf{A} + (\mathbf{W}'_n \odot)\,\mathbf{W}^\top \tag{8}$$

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 is performed by premultiplying the inputs $\mathbf{X}$ with L sparse matrices $_l \in \mathbb{R}^{N \times M}$ for $l \in \{1, 2, ..., L\}$:

$$(_l)_{[n,*]} \stackrel{\text{def}}{=} g_l\left(\mathbf{b}_{[n]} - \mathbf{a}^\top; (\mathbf{P}_n)_{[*,l]}\right) \odot \mathbf{F}_{[n,*]} \tag{9}$$

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.[17] The

[17] 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 $K = L$.

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, the design matrix, and the impulse indicator $\mathbf{L}$:

$$\mathbf{X}'_{[*,l]} \stackrel{\text{def}}{=}_l \mathbf{X}\,\mathbf{L}_{[*,l]} \tag{10}$$

The full model mean is the sum of (1) the intercepts and (2) the sum-product of the convolved predictors with the coefficient parameters:

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{q} + (\mathbf{X}' \odot \mathbf{Q})\,\mathbf{1}, \sigma^2\right) \tag{11}$$