Cognition 215 (2021) 104735

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Continuous-time deconvolutional regression for psycholinguistic modeling

Cory Shain *, William Schuler

The Ohio State University, Department of Linguistics, Oxley Hall, 1712 Neil Avenue, Columbus, OH 43210, United States of America

A B S T R A C T

The influence of stimuli in psycholinguistic experiments diffuses across time because the human response to language is not instantaneous. The linear models typically used to analyze psycholinguistic data are unable to account for this phenomenon due to strong temporal independence assumptions, while existing deconvolutional methods for estimating diffuse temporal structure model time discretely and therefore cannot be directly applied to natural language stimuli where events (words) have variable duration. In light of evidence that continuous-time deconvolutional regression (CDR) can address these issues (Shain & Schuler, 2018), this article motivates the use of CDR for many experimental settings, exposits some of its mathematical properties, and empirically evaluates the influence of various experimental confounds (noise, multicollinearity, and impulse response misspecification), hyperparameter settings, and response types (behavioral and fMRI). Results show that CDR (1) yields highly consistent estimates across a variety of hyperparameter configurations, (2) faithfully recovers the data-generating model on synthetic data, even under adverse training conditions, and (3) outperforms widely-used statistical approaches when applied to naturalistic reading and fMRI data. In addition, procedures for testing scientific hypotheses using CDR are defined and demonstrated, and empirically-motivated best-practices for CDR modeling are proposed. Results support the use of CDR for analyzing psycholinguistic time series, especially in a naturalistic experimental paradigm.

1. Introduction

Psycholinguistic models are evaluated by regressing their predictions against data from human subjects, but the human response to linguistic input is not strictly instantaneous; it takes time for the brain to recognize and respond to language. Consequently, measures of human cognition that are often used to test psycholinguistic hypotheses, including reaction times and neuronal activity, may capture the lingering influence of multiple preceding stimulus events, a phenomenon we will refer to as *temporal diffusion*. This article argues that temporal diffusion can be problematic for analyzing psycholinguistic data using standard tools, presents evidence that a continuous-time deconvolutional regression (CDR) technique proposed by Shain and Schuler (2018) can address these problems, and evaluates the robustness of CDR to various adverse training conditions that are likely to arise in the course of scientific modeling.[1] Results indicate that CDR is a useful technique for identifying and controlling for temporal diffusion in arbitrary time series.

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 neuronal activity (Boynton, Engel, Glover, & Heeger, 1996; Friston, Josephs, Rees, & Turner, 1998; Glover, 1999; Lindquist, Loh, Atlas, & Wager, 2009; Lindquist & Wager, 2007; Ward, 2006). 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 $h * g$ of a dynamical system as a function of time is described as a convolution over time of an impulse h with an IRF g as shown in Eq. (1), where τ is bound by the integral operation and ranges over the time interval $[0, t]$, and $h(\tau)$ is the impulse at time τ:[2]

$$(h*g)(t) = \int_0^t h(\tau)g(t-\tau)d\tau \tag{1}$$

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

* Corresponding author.
E-mail addresses: shain.3@osu.edu (C. Shain), schuler.77@osu.edu (W. Schuler).

[1] Shain and Schuler (2018) used the term *deconvolutional time series regression* (DTSR) to refer to what we are calling *continuous-time deconvolutional regression*. We have altered the name in order to stress the contribution of the approach (its continuous-time structure), since discrete-time deconvolutional regression models like finite impulse response models and vector autoregression have existed for some time.

[2] Throughout this paper, vectors and matrices are notated in **bold** lowercase and uppercase, respectively (e.g. $\mathbf{u}$, $\mathbf{U}$). Objects with indexed names are designated using subscripts (e.g. $\mathbf{v}_r$). Vector and matrix indexing operations are notated using subscript square brackets, and slice operations are notated using * (e.g. $\mathbf{X}_{[*,k]}$ denotes the kth column of matrix $\mathbf{X}$). Hadamard (pointwise) products are notated using $\odot$. The notations $\mathbf{0}$ and $\mathbf{1}$ designate conformable column vectors of 0's and 1's, respectively.

https://doi.org/10.1016/j.cognition.2021.104735
Received 20 August 2019; Received in revised form 1 April 2021; Accepted 11 April 2021

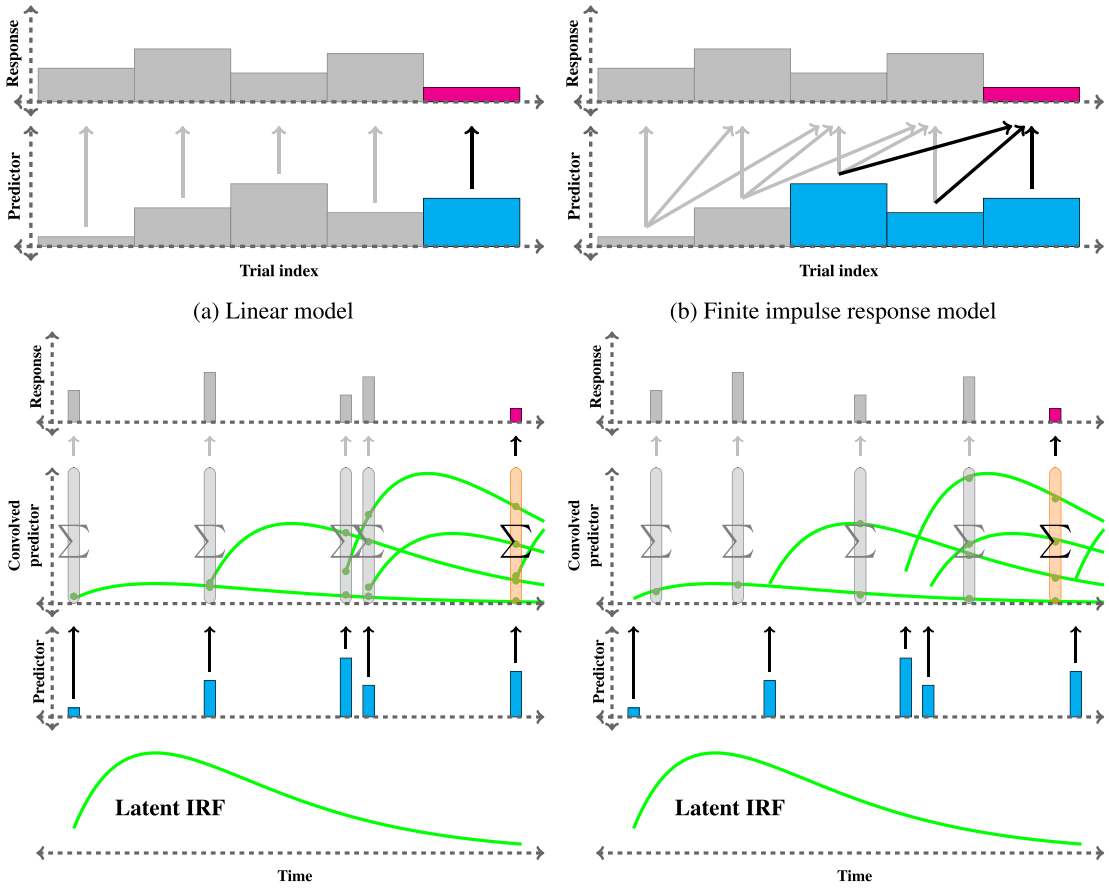


Fig. 1. Visual comparison of time series models. In linear models (a), the response (*y*-axis) is independent of previous events, while in FIR models (b), previous events are assumed to be equidistant in time (*x*-axis). In CDR models (c and d), the response is a weighted sum of all previous events, with weights provided by the IRF as a function of continuous time. Because the IRF is continuous, the response can be queried at any point, permitting direct application to both synchronous (c) and asynchronous (d) stimulus and response measures.

This article describes and evaluates a technique for *continuous-time deconvolutional regression* (CDR), which recasts the sequences of stimuli (predictors) and responses as convolutionally-related signals whose temporal relationship is mediated by one or more continuous-time IRFs with shape estimated from data (Shain & Schuler, 2018). By convolving predictors with their estimated IRFs, as in Eq. (1), the model can condition its predictions on the entire history of stimuli encountered up to a given point in an experiment, rather than e.g. on the properties of the current word alone. And by estimating the shape of the IRF from data, the model can reveal fine-grained patterns of temporal structure that are otherwise difficult to obtain.

Established techniques for IRF identification, including *finite impulse response* (FIR, also known as *distributed lag*, or DL) models (Griliches, 1967; Koyck, 1954; Neuvo, Dong, & Mitra, 1984; Robinson, 1975; Samaeki, Mitra, & Kaiser, 1993; Sims, 1971) and *vector autoregressive* (VAR) models (Sims, 1980), implicitly assume that the time series is sampled at a fixed frequency. This assumption is often ill-suited to language research because words in natural language have variable duration, whether spoken or read. The number of parameters in discrete-time deconvolutional models is also linear (or super-linear) on the length of the history window, which can easily lead to over-parameterization. These objections equally apply to the common technique in psycholinguistics of injecting "spillover" regressors into linear models (i.e. adding coefficients for predictors associated with preceding events, e.g. Erlich & Rayner, 1983; Mitchell, 1984), which turns out to be FIR/DL by a different name (§3).

By contrast, CDR defines IRFs as parametric functions of continuous time and applies the same continuous IRF to all events in the history, yielding a model that can be applied to non-uniform time series (such as language) without distorting the temporal or featural structure of the stimulus sequence, with constant parametric complexity on the length of the history window. The continuous-time nature of CDR also allows the estimated response to be queried at any timepoint without reliance on post-hoc techniques for interpolation or extrapolation.

A visual comparison of CDR and FIR models is given in Fig. 1. An ordinary least-squares model (Fig. 1a) considers the response to be independent of all preceding stimulus events:[3]

$$\mathbf{y} \sim \mathcal{N}(\mathbf{Xb}, \sigma^2) \tag{2}$$

FIR models relax this independence assumption at the expense of model complexity by including additional weights for a fixed number of preceding events (Fig. 1b).[4] Note that FIR assumes that events are equidistant in time, or, equivalently, that variation in temporal spacing is inconsequential (an assumption implicitly made by spillover models in psycholinguistics). By contrast, CDR retains real-valued timestamps of

[3] $\mathbf{y}$ is the response vector, $\mathbf{X}$ is the design matrix of predictors, $\mathbf{b}$ is the vector of coefficients, and σ^2 is the variance of the error distribution.

[4] FIR is a special case of the linear model given in eq. 2, with a structured design matrix $\mathbf{X}$ containing O values for each predictor representing a history of O timesteps into the past (see §3).

the stimulus events and fits continuous IRFs that govern the influence of each predictor on the response as a function of time (Fig. 1c and d). Unlike linear models, CDR is agnostic as to whether the stimuli and responses are measured at the same time (Fig. 1c) or at different times (Fig. 1d), a useful property for applications like fMRI modeling in which stimuli have variable duration but responses are measured at a fixed frequency (see §8 for discussion).

Shain and Schuler (2018) provide a formulation and cursory evaluation of CDR for human reading. This article builds on that foundation by empirically evaluating the influence of factors such as noise, multicollinearity, IRF misspecification, and hyperparameter selection on the final estimates, and additionally applies CDR for hemodynamic response discovery from fMRI data. The latter application is an appealing use case for continuous-time deconvolution because of the asynchrony between stimuli (words) and responses (brain scans) in naturalistic fMRI studies of language (§8).

The structure of this article is as follows. It first motivates the use of deconvolutional modeling (§2) and the use of CDR over available deconvolutional alternatives (§3) within the context of psycholinguistics. It also exposits some of CDR's mathematical properties (§4), describes a documented open-source Python implementation (§5), evaluates the model on synthetic data implementing a variety of plausible degeneracies (§6), explores the impact of various hyperparameter settings on estimates of reading (§7) and fMRI (§8) data, evaluates procedures for scientific hypothesis testing in a CDR framework (§9), and proposes empirically-motivated best practices for future CDR applications (§10). Results reveal successful identification of ground truth models in synthetic evaluations and plausible, fine-grained estimates of temporal structure in psycholinguistic evaluations that are highly consistent across a variety of hyperparameters, with improved generalization quality across domains compared to widely-used statistical approaches.

2. The importance of effect timecourses in psycholinguistics

Pyscholinguists have long recognized that the "critical region" for observing an effect may lag behind the stimulus that triggers it because of latency in human perception and information processing (Bouma & De Voogd, 1974; Erlich & Rayner, 1983; Mitchell, 1984; Morton, 1964; Rayner, 1977; Rayner, 1998; Smith & Levy, 2013; Vasishth & Lewis, 2006). Indeed, theoretical importance has been attached to questions of language processing timecourses, such as whether or not the human language comprehension architecture contains a buffer allowing information processing to lag behind perception (Bouma & De Voogd, 1974; Ehrlich & Rayner, 1981; Just & Carpenter, 1980; Mollica & Piantadosi, 2017). For this reason, statistical analyses of psycholinguistic data often include "spillover" regressors to preceding stimuli, whether in the context of ordinary least squares (OLS) models (Grodner & Gibson, 2005), linear mixed-effects (LME) models (Demberg & Keller, 2008), or generalized additive (GAM) models (Smith & Levy, 2013).[5]

While the psycholinguistic community is aware of the possibility of temporal diffusion and has directly investigated it in many studies, there may in general be insufficient concern over the severity of the potential consequences of temporal diffusion for statistical analysis of human-generated time series. For example, while many studies in psycholinguistics (such as those cited above) include spillover effects in analyses, there are also many that do not, even in naturalistic settings where the rate of presentation is not controlled and diffusion might be especially pronounced (e.g. Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Fossum & Levy, 2012; Frank & Bod, 2011; Roark, Bachrach, Cardenas, & Pallier, 2009; Smith & Levy, 2008; van Schijndel & Schuler, 2015). In general, such studies are not directly concerned with effect timecourses, and omission of spillover regressors embodies an implicit belief that an *in situ* model with no controls for diffusion is "good enough" to permit discovery of the relevant patterns. However, controlling for temporal diffusion is important regardless of whether the central research question directly concerns timecourses because failing to do so can lead to false negatives (e.g. failing to detect an effect because it occurred later than expected) or misattribution of variance due to temporally diffuse effects.

For example, using LME to evaluate the sensitivity of self-paced reading in the Natural Stories corpus (Futrell et al., 2021) to theory-driven predictors of memory retrieval cost shows significant main effects of syntactic constituent wrap-up (p= 2.33e-14) and syntactic dependency length (p= 4.87e-10; Shain, van Schijndel, Futrell, Gibson, & Schuler, 2016). However, spilling over one control variable (probabilistic context-free grammar surprisal) one position (from *in situ* to spillover-1) causes the effects of interest to vanish (p= 0.816 for constituent wrap-up and p= 0.370 for dependency length; Shain & Schuler, 2018).

The contrast between these two sets of results comes down to different assumptions about timecourse that are both reasonable *a priori* (i.e. whether the locus of surprisal effects should fall on the current word or spill over into the following one). For this particular dataset and model definition, it turns out that spilling over surprisal produces a stronger baseline that ultimately casts doubt on results obtained using a baseline inspired by preceding work. Such an outcome can be difficult to anticipate in advance. The possibility of such discrepant results based solely on assumptions about timecourse should motivate increased attention to diffusion of effects in psycholinguistic modeling.

The importance of controlling for temporal diffusion is of course dependent on experimental design. For example, there may be little impact from diffusion in a lexical decision task based on words presented in isolation with long intervals in between, while there is almost certainly a large influence of diffusion in fMRI scans of subjects listening to running speech. Psycholinguists and cognitive scientists are increasingly using naturalistic experiments in order to improve ecological validity and minimize task artifacts from artificially constructed designs (Campbell & Tyler, 2018; Demberg & Keller, 2008; Hasson & Honey, 2012). As suggested by the discussion of Shain et al. (2016) above, controlling for temporal diffusion may be of particular importance in such a setting, since measurements are taken from subjects carrying out rapid incremental sentence comprehension and multiple word fixations may take place within a short span of time (Kolers, 1976; Morton, 1964). It is nonetheless possible that even experiments with carefully constructed stimuli might benefit from improved control of temporal diffusion. For example, even holding prefixes of linguistic stimuli fixed up to a critical region cannot entirely control the influence of temporal diffusion; the same prefix can be fixated differently in different presentations, both within and across subjects, potentially leading to variation in patterns of diffuse processing that may affect the response in the critical region.

This concern about temporal diffusion in psycholinguistic data complements recent psycholinguistic interest in other kinds of temporal confounds, especially auto-correlation and non-stationarity (Baayen, van Rij, de Cat, & Wood, 2018; Baayen, Vasishth, Kliegl, & Bates, 2017).[6] Baayen et al. (2018) demonstrate the utility of including a first-order auto-regressive term in a GAM model to control for auto-correlated error, while Baayen et al. (2017) relax the assumption of

[5] GAM models (Hastie & Tibshirani, 1986) relax the linearity requirement of linear models, allowing the predictors to be related to the response via arbitrary smooth functions. A GAM with a Gaussian linking function has the following form:

$$\mathbf{y} \sim \mathcal{N}\left(b_0 + f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + \ldots + f_k(\mathbf{x}_k), \sigma^2\right) \quad (3)$$

[6] For related findings at high temporal resolution, see Cho, Brown-Schmidt, and Lee (2018).

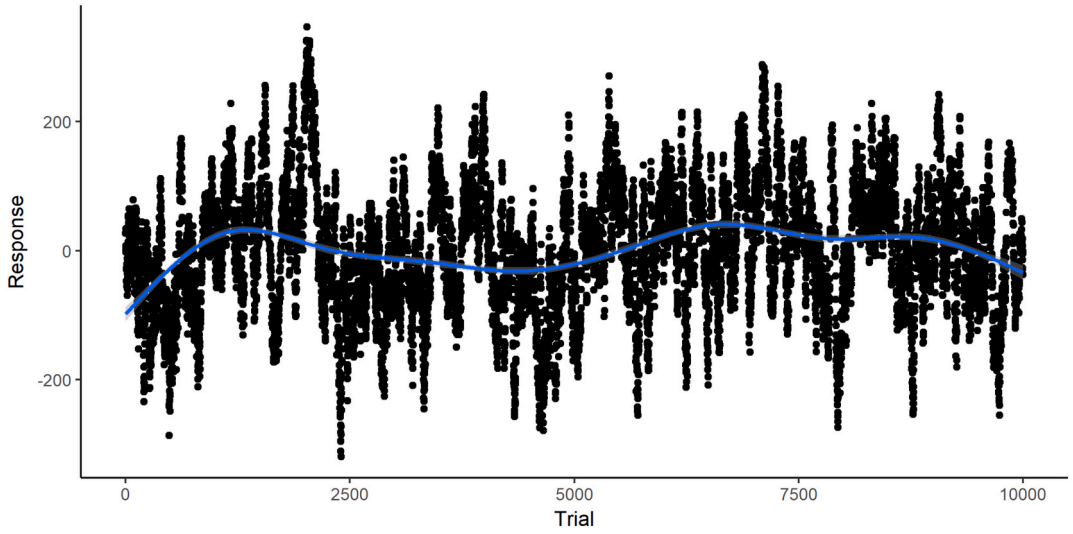


Fig. 2. Synthetic responses produced by convolving *i.i.d.* normal independent variables with stationary convolution kernels. Note the undulating GAM smooth in blue, suggesting non-stationarity that is not in fact present in the underlying generative process.

stationarity by augmenting GAM models with an independent variable $C \in [0, 1]$ representing the proportion of the series completed at the current timestep. When a spline is fitted directly to C, the obtained curve can be interpreted as an estimate of fluctuation in the base response rate over time. And when a spline is fitted to the interaction between C and a predictor, the obtained curve can be interpreted as an estimate of fluctuation in the influence of the predictor over time. Baayen et al. (2017) show that directly modeling non-stationarity can have important impacts on both effect estimates and significance testing when applied to time series generated by human subjects.

While auto-regressive/non-stationary GAM models can capture temporal effects, the crucial point of divergence from this work is that they do not capture temporal diffusion. They allow the influence of an independent variable to fluctuate with time, but continue to assume independence of the response from preceding observations of the predictor(s). In order to handle temporal diffusion, auto-regressive/non-stationary GAM models must still make use of the problematic spillover technique discussed above.

CDR and non-stationary GAM models can therefore be seen to address distinct potential confounds in time series data: temporal diffusion of effects (CDR) vs. auto-correlation and non-stationarity (GAM). All three confounds can be addressed relatively straightforwardly by deploying CDR as a pre-process to GAM fitting, resulting in a two step analysis in which the data are first convolved with CDR and then analyzed using GAM (see §4.6 for further elaboration on this general idea). However, we note in passing that convolutional structure can in some cases explain apparent autocorrelation and non-stationarity. For example, the plot in Fig. 2 shows a time series of synthetic responses generated by convolving *i.i.d.* normal independent variables with gamma-shaped convolution kernels, as described in §6. The overall base response rate in Fig. 2 appears to fluctuate with time. This is supported by the undulating GAM spline (shown in blue), suggesting non-stationarity. Responses are also clearly auto-correlated, as shown by the higher frequency oscillations evident in the plot. However, the data were in fact generated by a strictly stationary convolutional process and are *i.i.d.* normal conditional on the convolution. Apparent autocorrelation and non-stationarity are artifactual. Thus, one possible source of apparent auto-correlation and non-stationarity in time series data may be latent convolutional structure, and, in these cases, diffusion is the core temporal confound that must be brought under statistical control.

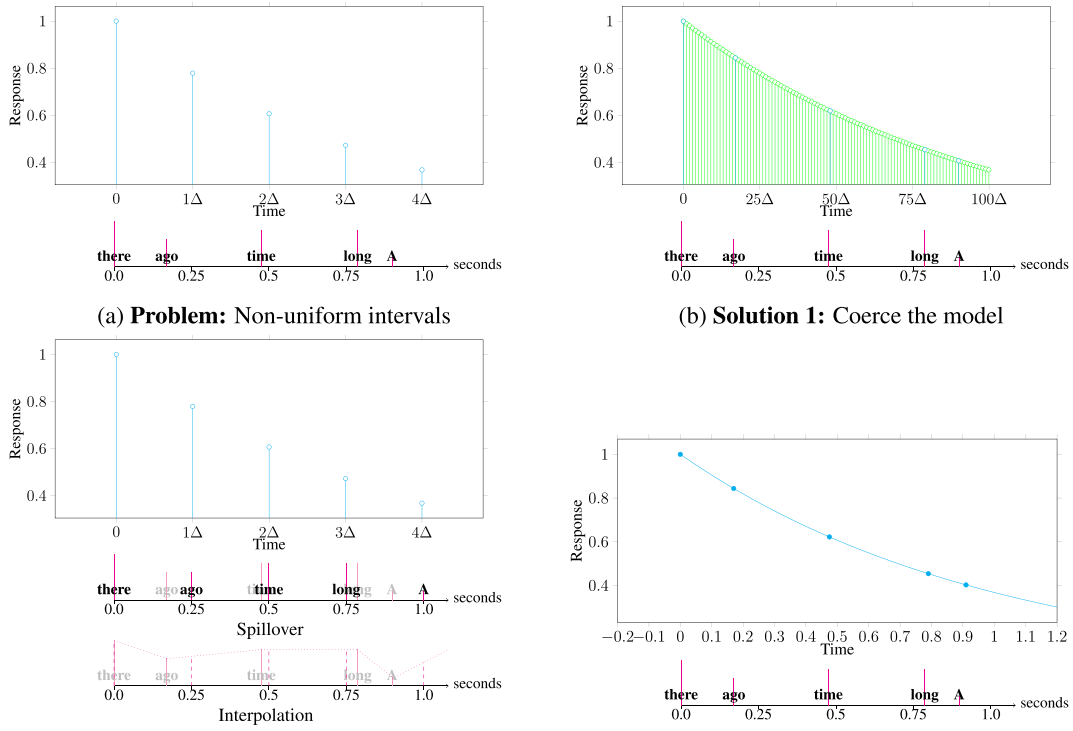
3. Existing deconvolutional models

In order to infer the structure of an IRF g in Eq. (1) from data, it is first necessary to construct a solution space over which to perform inference. One way of doing so is through **discrete-time deconvolution**, a class of methods which recast deconvolution as a special case of linear regression by discretizing time into a finite number of equidistant steps and then estimating timestep-specific parameters. Continuous-time IRFs can be inferred from these discrete estimates if desired using various post-hoc smoothing techniques. One example of this general approach, known as *finite impulse response* (FIR) models in the signal processing literature (Neuvo et al., 1984; Saramaeki et al., 1993) and *distributed lag* (DL) models in the time series literature (Griliches, 1967; Koyck, 1954), consists of including regressors from previous timesteps. For simplicity, we will henceforth refer to such models as FIR. A fixed-effects FIR model of order O with K predictors is a linear model of $\mathbf{y} \in \mathbb{R}^N$ with design matrix $\mathbf{X} \in \mathbb{R}^{N \times K}$ and parameters $\mathbf{b} \in \mathbb{R}^{1+O \cdot K}$, i.e. an intercept, plus one coefficient for each of K predictors for each of O timesteps into the past:

$$\mathbf{y}_{\text{FIR}_O} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{1}^N & \mathbf{X} & \mathbf{0}^{1\times K} & \cdots & \mathbf{0}^{(O-1)\times K} \\ & & \mathbf{X}_{[1\cdots N-1,*]} & \cdots & \mathbf{X}_{[1\cdots N-O+1,*]} \end{bmatrix}\mathbf{b}, \sigma^2\right) \quad (4)$$

The sequence of O coefficients for a given predictor defines a discrete-time IRF, and the linear combination of predictors with $\mathbf{b}$ defines a temporal convolution operation (i.e. a weighted sum along the time dimension).

Another prominent example of discrete time deconvolutional approaches is vector autoregressive (VAR) modeling (Sims, 1980). VAR generalizes FIR to predict the next timepoint (row) of $\mathbf{X}$ rather than a distinguished response $\mathbf{y}$. VAR thus estimates parameters $\mathbf{B} \in \mathbb{R}^{(1+O\cdot K)\times K}$, and generates predictions $\mathbf{Y} \in \mathbb{R}^{N\times K}$ through linear transformation:



(a) **Problem:** Non-uniform intervals

(b) **Solution 1:** Coerce the model

(c) **Solution 2:** Coerce the data

(d) **Solution 3:** CDR

Fig. 3. The problem of variably spaced events for discrete-time convolution in a hypothetical univariate FIR model on predictor *word length* (characters) with step duration Δ. Plots show hypothetical IRFs, with predictor values for each word shown in magenta in timelines below each plot. The example sentence is typeset in reverse because the impulse response describes the changing influence of words as they recede into the past. An order 5 FIR model cannot be used directly on a variably-spaced word sequence because there is no value of Δ that aligns the FIR coefficients with the stimulus events (3a). One solution (3b) is to use a high-resolution IRF of order sp, where s is the length of the history window in seconds and p is the inverse of the precision of the temporal measurement. Although this model can be directly applied to variably-spaced events, it overparameterized to the point of being unidentifiable. In this hypothetical training example where $s = 1$ and $p = 100$, only 5/100 parameters have data. Another solution is to coerce the data into a format that fits the assumptions of FIR (3c). Temporal variation can be deleted by "snapping" words to coefficients in one-to-one alignment under the assumption of a fixed but unknown value for Δ (Spillover). This technique is distortionary if the stimuli are variably spaced and their underlying contribution is a function of clock time rather than relative event index. Alternatively, the predictor can be continuously interpolated between events, and the interpolated signal is resampled at points (vertical dashed lines) that align with the discrete IRF coefficients (Interpolation). This technique is distortionary for event-based predictors that are not underlyingly continuous. CDR (3d) avoids both sparsity and distortion by replacing the discrete IRF with a parametric continuous function of time (in this example, $f(x;\beta) = e^{-\beta x}$). A continuous IRF can be queried exactly at any point, has a parametric complexity that is independent of the temporal span or resolution of the response kernel, can be applied directly and without distortion to variably-spaced time series, and is agnostic to temporal alignment between stimuli and response.

$$\mathbf{Y}_{\text{VAR}_O} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{1}^N & \mathbf{0}^{1\times K} & \cdots & \mathbf{0}^{O\times K} \\ & \mathbf{X}_{[1\cdots N-1,*]} & \cdots & \mathbf{X}_{[1\cdots N-O,*]}\end{bmatrix}\mathbf{B}, \sigma^2\right) \tag{5}$$

Thus, unlike FIR, which predicts a distinguished response variable via convolution of the history of predictor values, VAR predicts all K variables at the current timestep through summed linear transformations of the O preceding timesteps. VAR fits can be used to extract IRFs between any pair of variables in the model. Both of these techniques are widely used in the fMRI literature (Friston et al., 1994; Harrison, Penny, & Friston, 2003), and both can be augmented with random effects (Beckmann, Jenkinson, & Smith, 2003; Gorrostieta, Ombao, Bédard, & Sanes, 2012). Since VAR estimates IRFs between all pairs of variables in the data, it is perhaps unnecessarily powerful for typical psycholinguistic studies that seek to model a distinguished response variable. We therefore focus on FIR for the remainder of our discussion of discrete-time deconvolution.

As discussed in §1, in psycholinguistics, temporal diffusion is often addressed by adding "spillover" predictors encoding predictor values from preceding events. From Eq. (4) above, it follows that this approach reduces to FIR modeling: a linear model containing O spillover positions of K predictors is an FIR model of order O on those K predictors, and the set of O coefficients for each predictor defines a discrete-time IRF. In practice, the term *FIR* tends to be used in signal processing settings where the temporal distance between samples is *fixed*, while the term *spillover* tends to be used in experimental settings where the temporal distance between samples is *ignored*. While this distinction matters for the interpretation of the discovered IRFs, since discrete-time IRFs only have a clock-time interpretation when the offset between timesteps is fixed and known, it is immaterial for the definition of the statistical model itself. Even the GAM models with spillover used in e.g. Smith and Levy (2013) can be thought of as a variant of FIR with an IRF kernel whose shape depends on the values of the impulses at each timestep.

Because of the identity (in the case of LME) or close relationship (in the case of GAM) of spillover models to FIR models, the remainder of this article will no longer distinguish the two, since the same objections apply regardless of terminological choice.

Since additional spillover positions contribute additional parameters, rich spillover controls can easily create such heavily parameterized models that realistically sized datasets cannot support them, especially when used in the context of mixed-effects or spline regressions that fit random effects or multi-dimensional smooths for each spillover position of each predictor. For example, linear mixed effects models with by-subject random slopes fitted using lme4 (Bates, Mächler, Bolker, & Walker, 2015) on the Dundee corpus (Kennedy, Pynte, & Hill, 2003) fail to converge with two or more spillover regressors.[7] These models only contained spillover variants of four predictors, and the problem of overparameterization can be even more severe in models that contain more control variables (e.g. Demberg & Keller, 2008, where LME models contained up to thirteen predictors). In addition to concerns about overparameterization, spillover regressors can also introduce spurious multicollinearity to the extent that predictors are autocorrelated, which can be problematic for model identification and interpretation (Kutner, Nachtsheim, & Neter, 2003).

As a consequence, analysts are forced in practice to trade off the richness of the timecourse model with other sources of complexity. For example, Smith and Levy (2013) use GAM models containing rich spillover structures (up to three timesteps) but relatively poor random effects (by-subject random intercept), while van Schijndel and Schuler (2015) use LME models with rich random effects (by-subject and by-word random intercepts along with by-subject random slopes for every predictor) but no spillover regressors.

However, perhaps more fundamental for language research than the aforementioned computational problems is the inability of discrete-time deconvolutional models to represent variably spaced events. Fig. 3a visually exemplifies this problem, and Fig. 3b and c exemplify possible solutions to it within a discrete-time framework, each of which has undesirable properties. As shown in Fig. 3a, an FIR model assumes a single fixed interval Δ between coefficients, and thus the discrete-time IRF cannot directly convolve the properties of variably-spaced words because no such interval exists. This problem is visualized by the lack of temporal alignment between the words in the example and the FIR coefficients, and it can be addressed by coercing the model to match the data or the data to match the model. To coerce the model to match the data, the interval Δ can be reduced to the level of precision of the temporal measurement (e.g. 1 millisecond) along with a compensatory increase in the number of FIR coefficients per unit time (Fig. 3b), ensuring alignment to an FIR coefficient of all past events within some finite window. As visualized in the figure, such an approach exaggerates the problem of overparameterization and data sparsity to such a degree that we are unaware of any psycholinguistic studies that attempt to use this technique (how many events in a psycholinguistic experiment are spaced exactly 142 ms apart?). To coerce the data to match the model, the stimuli can be (1) forced into one-to-one alignment with the FIR coefficients under the simplifying assumption that Δ is fixed but unknown or (2) interpolated and resampled at points that align with the FIR coefficients (Fig. 3c). The forced alignment approach is equivalent to spillover, and it is distortionary for variably spaced events to the extent that the underlying contribution of those events is a function of clock time rather than relative event index. The interpolation approach has been used e.g. in fMRI modeling (Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016), and it is distortionary to the extent that the stimuli represent transient events rather than samples from a continuously evolving feature space.[8]

CDR avoids both problems (overparameterization and distortion) by defining the IRF as a continuous-time parametric kernel (Fig. 3d). Because a continuous IRF has infinite precision, it can be queried exactly at any point, thereby avoiding the trade-off faced by discrete-time models between parsimony on the one hand and the temporal span and resolution of the response kernel on the other. An additional advantage of continuous-time deconvolution is the ability to model asynchronously measured data (see §1). Because continuous-time deconvolution is parsimonious, faithful to the underlying temporal structure in the stimulus, and agnostic to temporal alignment between stimulus and response, it is more appropriate than FIR (spillover) approaches for analyzing many kinds of psycholinguistic data. Despite these conceptual advantages, continuous-time deconvolution is not currently used in psycholinguistics[9] and is little used in cognitive science more generally (aside from some previous neuroimaging studies that optimize the parameters of gamma-shaped hemodynamic response functions, e.g. Kruggel & Yves von Cramon, 1999; Kruggel, Wiggins, Herrmann, & von Cramon, 2000; Miezin, Maccotta, Ollinger, Petersen, & Buckner, 2000; Lindquist & Wager, 2007; Lindquist et al., 2009).

In addition to the discrete-time frameworks discussed above, related continuous-time regression models have also been proposed. Prior work has defined continuous-time extensions of distributed lag models (Bergstrom, 1984; Robinson, 1975, 1976; Sims, 1971). However, these approaches rely on Fourier analysis of the discretized covariate vector in order to model the continuous IRF. Consequently, they impose two problematic restrictions: (1) the covariates must be underlyingly continuous, and (2) discrete samples from the covariates must be taken at a uniform time interval (Robinson, 1975). They are therefore even less applicable to non-uniform discrete time series than their discrete-time analogs, which do not impose continuity constraints, while also being subject to the same critiques of the uniformity requirement.

Mathematically, the most closely related existing model to CDR is the the Hawkes process model, also known as a self-exciting counting process model (Hawkes, 1971). Hawkes process models are used to analyze stochastic point processes in which the occurrence of an event locally increases the instantaneous probability of other events occurring, and thus the intensity function of the process is *self-exciting* in continuous time. Formally, a Hawkes process generates the intensity $\lambda(t)$ given (possibly non-stationary) base intensity $\mu(t)$ and event times T via convolution with the triggering function $g(t)$ (analagous to an IRF):

$$\lambda(t) = \mu(t) + \sum_{\tau \in T, \tau < t} g(t - \tau) \tag{6}$$

Commonly, g is chosen to be the two parameter exponential function $g(t) = \alpha e^{\beta t}$, $\alpha, \beta > 0$, enforcing exponential decay of the self-excitation function with amplitude α and decay rate β,[10] and the intensity $\lambda(t)$ is taken to be the parameter of a Poisson distribution describing the instantaneous concentration of events given the history, or equivalently, the rate parameter of an exponential distribution describing the

[7] Models used log go-past durations as the dependent variable and included the predictors *word length* (in characters), *saccade length* (in words), *unigram log probability*, and *5-gram surprisal*, with probabilities computed by KenLM language models (Heafield et al., 2013) trained on the Gigaword 3 corpus (Graff & Cieri, 2003). Spillover positions from 0 (in situ) to n for each predictor were included for six models, one for each of $n \in \{0,1,2,3,4,5\}$, along with a random intercept by word and random slopes by subject for each predictor (and each spillover position of each predictor). Outlier filtering was performed following van Schijndel and Schuler (2015), yielding a total of 193,309 data points.

[8] As discussed in §4, interpolation of variably-spaced samples from predictors that *are* underlyingly continuous over time (e.g. ambient noise level) is appropriate and in fact necessary to avoid distortion in CDR's event-based convolution procedure.

[9] Aside from our own recent studies that apply CDR to psycholinguistic data (Shain, 2019; Shain, Blank, van Schijndel, Schuler, & Fedorenko, 2020; Shain & Schuler, 2018).

[10] Other kernel types, including power law kernels $h(t) = \frac{\alpha}{(t+\beta)^{\gamma+1}}$ (Lapham, 2014), non-parametric basis kernels (Zhou, Zha, & Song, 2013), and self-regulating neural network kernels (Mei & Eisner, 2017), are also widely used.

expected waiting time until the next event given the history (Cooper, 2005). Parameters μ, α, and β are usually estimated from data using non-linear numerical optimization (Ozaki, 1979). This framework has been generalized in many ways, including extension to multivariate event data (i.e. simultaneous modeling of multiple event streams; Embrechts, Liniger, & Lu, 2011), extension to *marked* processes that contain regressors in addition to timestamps (Lapham, 2014), and the use of recurrent neural network intensity functions (Mei & Eisner, 2017).

Although both CDR and Hawkes processes involve a continuous parametric convolution over the time dimension, a fundamental difference between them is that CDR seeks to model a designated response variable while Hawkes processes seek to model the future temporal realization of the sequence of events. To our knowledge, no existing formulation of Hawkes process models can be used to address the temporal diffusion problem targeted in this study.

4. Mathematical definition

The mathematical definition of CDR was previously proposed in Shain and Schuler (2018). For convenience, we reproduce it in §4.1, revised for clarity. The remainder of this section (§4.2, §4.3, §4.4, and §4.5) exposits additional mathematical properties of CDR.

4.1. CDR (fixed effects) model

For clarity, here we define a fixed-effects-only variant of CDR. The full model with random effects is defined in Appendix A.1. The CDR model assumes the following quantities as input:

- $X \in \mathbb{N}$: Number of predictor observations
- $Y \in \mathbb{N}$: Number of response observations
- $K \in \mathbb{N}$: Number of predictors
- $R \in \mathbb{N}$: Number of impulse response parameters
- $J \in \mathbb{N}$: Number of unique time series[11]
- $\mathbf{X} \in \mathbb{R}^{X \times K}$: Design matrix of X predictor observations of K dimensions each
- $\mathbf{y} \in \mathbb{R}^{Y}$: Vector of Y response observations
- $\mathbf{t} \in \mathbb{R}^{X}$: Vector of timestamps associated with each observation in $\mathbf{X}$
- $\mathbf{t}' \in \mathbb{R}^{Y}$: Vectors of timestamps associated with each observation in $\mathbf{y}$
- $\mathbf{s} \in \{1, 2, \ldots, J\}^{X}$: Vector of time series IDs associated with each observation in $\mathbf{X}$
- $\mathbf{s}' \in \{1, 2, \ldots, J\}^{Y}$: Vectors of time series IDs associated with each observation in $\mathbf{y}$
- $g_k(t; \theta) \in \mathbb{R}_+ \rightarrow \mathbb{R}$ for $k \in \{1, 2, \ldots, K\}$: Parametric IRF kernels specifying response at time t given parameters θ, one for each of K predictors

A single dataset may contain multiple time series. For example, a psycholinguistic experiment may contain data from several participants, each of whom read several texts. Each participant-text pair could be treated as a unique time series. Different time series are considered statistically independent and are indexed by unique time series IDs (represented in $\mathbf{s}$ and $\mathbf{s}'$). Note that X (number of predictor observations) and Y (number of response observations) can differ in a CDR model because $\mathbf{X}$ will be forced into a conformable dimensionality with $\mathbf{y}$ via convolution over time (see $\mathbf{X}'$; below). This property permits CDR analysis of predictor/response streams with different acquisition times.

CDR seeks to estimate the following quantities, which mediate between $\mathbf{X}$ and $\mathbf{y}$:

- a scalar intercept $\mu \in \mathbb{R}$
- a vector $\mathbf{u} \in \mathbb{R}^{K}$ of K coefficients[12]
- K vectors $\mathbf{v}_k \in \mathbb{R}^{R}$ of R IRF kernel parameters for K predictors
- a scalar variance $\sigma^2 \in \mathbb{R}_+$ of the response

To support convolution, we define a mask $\mathbf{F} \in \{0, 1\}^{Y \times X}$ that admits only those observations in $\mathbf{X}$ that precede each $\mathbf{y}_{[y]}$ in the same time series, for $1 \leq x \leq X$, $1 \leq y \leq Y$:

$$\mathbf{F}_{[y,x]} \stackrel{\text{def}}{=} \begin{cases} 1 & \left(\mathbf{s}_{[x]} = \mathbf{s}'_{[y]}\right) \text{ and } \left(\mathbf{t}_{[x]} \leq \mathbf{t}'_{[y]}\right) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

We further define K sparse convolution matrices $\mathbf{G}_k \in \mathbb{R}^{Y \times X}$ for $k \in \{1, 2, \ldots, K\}$:

$$\mathbf{G}_k \stackrel{\text{def}}{=} g_k(\mathbf{t}' \mathbf{1}^{\top} - \mathbf{1}\mathbf{t}^{\top}; \mathbf{v}_k) \odot \mathbf{F} \tag{8}$$

The convolution that yields the design matrix of convolved predictors $\mathbf{X}' \in \mathbb{R}^{Y \times K}$ is then defined using a product of the convolution matrices and the design matrix:

$$\mathbf{X}'_{[*,k]} \stackrel{\text{def}}{=} \mathbf{G}_k \mathbf{X}_{[*,k]} \tag{9}$$

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

$$\mathbf{y} \sim \mathcal{N}(\mathbf{m} + \mathbf{X}' \mathbf{u}, \sigma^2) \tag{10}$$

A summary table of all variable definitions in the full (mixed effects) model is provided in Appendix Table A1, and a step-through example of the equations in the full model is provided in Appendix A.3.

Note that this is an event-based implementation of convolution that is only exact when the predictors fully describe discrete impulse signals. Exact convolution of samples from continuous signals is generally not possible because the signal is generally not analytically integrable. For continuous signals, the CDR procedure above defines a Riemann sum approximation of the integral as long as (1) the predictor is sampled at a fixed frequency or (2) the predictor is interpolated at a fixed frequency between variably-spaced samples.

4.2. Effect estimates in (fixed effects) CDR

Many scientific applications of linear modeling involve testing a null hypothesis about the scalar-valued *effect estimate* obtained for a predictor (e.g. that it is equal to 0). Because CDR models estimate continuous functions of time rather than scalars (as in linear regression), the estimated IRF must be distilled into a scalar in order to yield a comparable notion. Here, we define CDR effect estimates by integrating the IRFs, since the integral describes the total expected influence on the response from observing a unit impulse of each predictor. In particular, the unscaled and scaled fixed effect estimates $\mathbf{g}$, $\mathbf{g}' \in \mathbb{R}^{K}$ are defined as follows, where scaling is performed using the coefficient vector $\mathbf{u}$ (§4.1):

$$\mathbf{g}_{[k]} \stackrel{\text{def}}{=} \int_0^{\infty} g_k(t; \mathbf{v}_k) dt \tag{11}$$

[11] $J \ll X, Y$ because each time series indexed by $\{1, \ldots, J\}$ contains many predictor and response observations.

[12] Throughout this paper we 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. 10.

$$\mathbf{g}' \stackrel{\text{def}}{=} \mathbf{g} \odot \mathbf{u} \tag{12}$$

Due to considerations that arise in mixed effects CDR models (Appendix A.2), we constrain the IRFs g_k to have a unit integral over the positive real line:

$$1 = \int_0^\infty g_k(t;\theta)\,dt;\ k \in \{1,2,\ldots,K\},\ \theta \in \mathbb{R}^R \tag{13}$$

An important implementational consideration for finite training data is that the model will not have empirical support over the positive infinite real line, and thus the infinite integral involves some degree of extrapolation. To ensure that effect estimates have strong empirical support, in practice we upper bound the integral in Eq. (13) to the 75th percentile of temporal offsets seen in training. We consider this to be a reasonable default that concentrates the effect estimate on empirically well-attested regions of the support of the IRF. However, particular research questions may motivate the use of other kinds of bounds (e.g. if the research domain imposes a principled constraint on the duration of interest for the IRF).

4.3. The deconvolutional intercept

In linear regression, the *intercept* is a bias term implemented by fitting a coefficient to a vector of ones, one for each data point. The intercept term estimates the base response of the system when the other predictors are equal to 0. Because CDR contains a linear model on the convolved predictors (Eq. (10)), it is just as important to include an intercept term in CDR models as in linear ones. However, in a deconvolutional setting, it is also possible that the response is partially described by the *timing* of stimuli alone, independently of their properties. This possibility can be brought under control by additionally convolving the intercept with an estimated impulse response. Analagously to a linear intercept term, this convolved intercept estimates the base response of the system when the other predictors are equal to 0, but unlike the linear intercept, the convolved intercept is sensitive to stimulus timing. The estimate for the deconvolutional intercept is therefore the expected change in the response over time from observing an event, regardless of the properties of that event. We refer to this deconvolutional intercept as *rate* (see also e.g. Brennan, Stabler, Van Wagenen, Luh, & Hale, 2016), and we consider it to be an essential control to include in CDR models. Without it, variance in the response due to event timing must be captured by other components in the model, which is potentially problematic for interpretation and hypothesis testing.

Depending on the problem definition, *rate* effects may have a theoretical interpretation. For example, as argued in §7, a negative *rate* estimate in reading data can be seen as an inertia effect, since the negative *rate* contributions of preceding words compound to suppress reading times on the current word as a function of their recency. In other words, fast reading in the recent past will engender fast reading now, a possibility which *rate* estimates allow the model to account for.

Note that *rate* can only be estimated in a continuous-time deconvolutional setting because variation over time in the rate of events is necessary to identify it. In an FIR model, the *rate* predictor is equal to 1 at every timestep, rendering it identical to the intercept. The ability to detect *rate* effects in variably-spaced time series is a major advantage of CDR.

4.4. Scale and shift in CDR models

Linear models are invariant to transformations that rescale and/or shift the design matrix, in that for a linear model f with intercept p, slopes $\mathbf{q}$, predictors $\mathbf{X}$, scale vector $\mathbf{r}$, and shift vector $\mathbf{s}$, the following identity holds:

$$\begin{aligned} f(\mathbf{X}\ diag(\mathbf{r}) + \mathbf{1}\mathbf{s}^\top; p, \mathbf{q}) &= p + (\mathbf{X}\ diag(\mathbf{r}) + \mathbf{1}\mathbf{s}^\top)\mathbf{q} \\ &= p + \mathbf{s}^\top\mathbf{q} + \mathbf{X}\ diag(\mathbf{r})\mathbf{q} \\ &= f(\mathbf{X}; p + \mathbf{s}^\top\mathbf{q}, diag(\mathbf{r})\mathbf{q}) \end{aligned} \tag{14}$$

In other words, a linear model of shifted/rescaled $\mathbf{X}$ is equivalent to a shifted/rescaled linear model of $\mathbf{X}$. This result entails that (non-interacted) predictors can be shifted and rescaled (e.g. standardized) prior to fitting without altering the solution, which can help with numerical optimization of complex linear models.

In a non-linear optimization setting such as that required for CDR, normalization can be helpful for accelerating convergence (Ba, Kiros, & Hinton, 2016; Ioffe & Szegedy, 2015; Salimans & Kingma, 2016), and we therefore explore the impact of scale and shift of the inputs to CDR models. Invariance under rescaling follows trivially from Eq. (9). It also follows from Eq. (9) that CDR is *not* invariant under additive shift $\mathbf{s}$:

$$\mathbf{G}_k\left(\mathbf{X}_{[*,k]} + \mathbf{1}\mathbf{s}_{[k]}\right) = \mathbf{G}_k\mathbf{X}_{[*,k]} + \mathbf{G}_k\mathbf{1}\mathbf{s}_{[k]} \tag{15}$$

As shown, the shift scalar $\mathbf{s}_{[k]}$ is also convolved with $\mathbf{G}_k$, resulting in an additional term $\mathbf{G}_k\,\mathbf{1}\mathbf{s}_{[k]} \in \mathbb{R}^N$ which cannot be absorbed by the intercept because its value differs for each data point (unlike $\mathbf{s}^\top\mathbf{q}$ in Eq. (14), whose value is identical for all elements of $\mathbf{X}$). However, note that by convolving a matrix of ones, $\mathbf{G}_k\,\mathbf{1}\mathbf{s}_{[k]}$ implicitly defines a deconvolutional intercept term with IRF g_k and scale $\mathbf{u}_{[k]}\mathbf{s}_{[k]}$. In other words, shifting K predictors introduces K deconvolutional intercepts, each with IRF shape and scale tied to the shape and scale estimates of the predictors themselves. Since these deconvolutional intercepts are summed together in Eq. (10) and convolution obeys the distributive property, they together define a new impulse response g_0 for the deconvolutional intercept:

$$g_0(t) = \sum_{k=1}^{K} \mathbf{u}_{[k]}\mathbf{s}_{[k]}g_k(t) \tag{16}$$

Since $g_0(t) = 0$ can only be guaranteed if $\mathbf{s} = \mathbf{0}$, the model is not invariant under shift. However, in models with an explicit *rate* predictor as recommended in §4.3, g_0 simply modulates the IRF estimate for *rate*. Under the conventional assumption that the intercept (*rate* predictor) is the first column of the design matrix $\mathbf{X}$, the implicit kernel g_1' for the deconvolutional intercept in models with shift can be computed as:

$$g_1'(t) = g_0(t) + g_1(t) \tag{17}$$

The deconvolutional intercept thus does "absorb" shift in a limited sense. It must nonetheless be kept in mind that unless identity is enforced between $g_{1,\ldots,K}$, the estimated shape of g_1' will consist of a sum of response kernels and can therefore fall outside the solution space defined by the parametric IRF kernel assigned to *rate*.

4.5. Multicollinearity

The formulation in Eq. (10) is simply a linear model on the convolved design matrix $\mathbf{X}'$. Therefore, the primary difference between linear and

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

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

Multicollinearity in CDR works in much the same way, with the added complexity that CDR 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 multicollinearity to decompose that structure into predictor-specific 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 multicollinearity 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 multicollinearity 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-specific responses that increasingly deviate from the true data generating model. In the extreme case of perfect multicollinearity, the expected response has an identical shape 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 arbitrary.

Empirical results (§6.3) indicate that CDR models are quite robust to multicollinearity. Nonetheless, models fitted to highly collinear data should be interpreted with caution, and perfectly collinear data should be avoided altogether. As in linear models, multicollinearity can be avoided by orthogonalizing predictors in advance (e.g. via principal components analysis). Empirical evaluation of orthogonalization procedures in the CDR setting is left to future work.

4.6. Hypothesis testing

The ultimate scientific purpose of most statistical models is to test a claim about nature. Hypothesis testing is challenging in a CDR context for two reasons. First, familiar hypothesis tests using e.g. standard errors in linear models are unavailable for CDR, since it lacks analytical estimators for uncertainty about its parameters. As a result, exact null hypothesis significance tests cannot be performed on the basis of a single model. Single-model tests are nonetheless also problematic in a linear regression context when predictors are collinear (Neter et al., 1989), a pervasive issue in psycholinguistics that has motivated a shift toward ablative tests based on model comparison (Frank & Bod, 2011). Second, reliance on stochastic optimization of a non-convex objective function[13] introduces estimation noise through the possibility of imperfect convergence to an optimum or convergence to a non-global optimum. As a result, training likelihood cannot be guaranteed to be maximized, which can result in degenerate outcomes for in-sample ablative tests (e.g. the ablated model can have better likelihood than the full model).

For this reason, two tests that are commonly used for linear models may be unreliable for CDR. First, tests based on credible intervals (e.g. whether the 95% credible interval for an effect estimate includes 0) may be unreliable because the credible intervals produced by (variational) Bayesian CDR reflect the local neighborhood of the discovered solution (optimum), which may not account for the existence of more distant optima.[14] Credible intervals tests in CDR are therefore anticonservative. Indeed, the analyses reported below show that CDR-estimated credible intervals tend to be very tight. Second, likelihood ratio testing (LRT) may be unreliable because the test statistic is a function of the maximum likelihood estimates, and CDR likelihood cannot be guaranteed to be maximized. Instead, we consider two types of hypothesis test for CDR models: (1) a **direct test** by bootstrap comparison of model fit to out-of-sample data, and (2) a **2-step test** in which CDR is used first to estimate a data-driven convolution $\mathbf{X}'$ of the design matrix $\mathbf{X}$ and then existing statistical models (e.g. OLS, LME, GAM) are fitted to $\mathbf{X}'$ and used to perform the test.

To perform a direct test, training and evaluation sets must be created, either by running two separate experiments or by partitioning the data from a single experiment.[15] Two CDR models are fitted to the training set, one with a fixed effect for the variable to be tested (full model), and one without one (ablated model). Out-of-sample error vectors are then generated by predicting from each model on the evaluation set, and an aggregate test statistic (e.g. absolute difference in mean squared error) is computed over the two vectors. To perform the test (a paired permutation test), an empirical distribution is created for the test statistic: for n iterations, the by-item errors from each model are randomly swapped pairwise to generate two new error vectors, and a new test statistic over the resampled errors is computed and stored. The test rejects the null hypothesis at level α if the observed test statistic is greater than $(1 - \alpha) \times 100\%$ of the resampled test statistics.

To perform a 2-step test, a single CDR model containing all fixed effects of interest is fitted to the data, and the predictors are convolved using the estimated IRFs. Standard statistical models (e.g. OLS, LME, GAM) are then fitted to the convolved predictors and used to perform any of the tests that they support (e.g. LRT). Note that to perform an ablative test like LRT in a 2-step setting, the ablation is only applied at the second step (e.g. the LME stage). If ablation is also applied at the CDR stage, then the predictors in the full and ablated models *are not necessarily the same*, invalidating the test.

Both tests potentially suffer from non-convexity, since they are both conditional on possibly sub-optimal IRF estimates. Nonetheless, we offer the following arguments in defense of using CDR for hypothesis testing. First, the synthetic experiments reported in §6 show a strong tendency for CDR to closely recover the true data-generating model, even under adverse training conditions like variably spaced events, multicollinearity, and ill-fitting IRF kernels. We thus have empirical reason to believe that convergence to a bad optimum is not a serious problem in practice. Second, the assumption that the models fall within a tolerance of the global optimum (i.e. are "good enough") also underlies ablative tests in popular linear regression libraries like *lme4*, which use numerical

[13] The *objective function* maps predictors and responses to a scalar value that the model attempts to optimize. In this study, the objective to be maximized is the (regularized) training likelihood of the response given the predictors. The objective is *non-convex* if it has multiple optima, i.e. "peaks" in parameter space where the objective is higher than it is at intermediate values of the parameters. The implementation applied in this study employs a widely used optimizer for deep neural networks (see §5), where non-convexity is a pervasive issue. Although empirically successful in many prior applications, the optimizer cannot guarantee convergence to the global (best) optimum.

[14] Maximum likelihood CDR models do not estimate uncertainty, and therefore CDR does not support (frequentist) confidence intervals.

[15] When partitioning and/or filtering outliers prior to CDR fitting, it is important to keep in mind that partitioning and outlier filtering should only be performed on the *response* vector. Partitioning/filtering the design matrix is equivalent to assuming that the removed events did not take place, which can distort the IRF estimates. The CDR software library described in §5 provides utilities for data partitioning and filtering, which automatically apply only to the response data.

Table 1
Distribution of fixed effects estimates of LME models fitted to CDR-convolved synthetic and human subjects (Experimental) data (median, 25th percentile, and 75th percentile). Estimates concentrate near 1, indicating that CDR-estimated coefficients are generally close to the global optimum given the IRF.

	Synthetic			Experimental		
Inference	Median	25%	75%	Median	25%	75%
MLE	1.000	1.000	1.001	1.002	0.997	1.011
BBVI.imp	1.000	1.000	1.000	1.002	0.997	1.014
BBVI	1.010	1.002	1.076	1.003	0.973	1.027

optimization and define a tolerance-based stopping criterion. Third, in pursuit of understanding complex non-linear phenomena like human language comprehension, CDR may permit discovery of previously unknown patterns precisely by relaxing the strict linearity and independence assumptions that support linear models' convenient statistical and mathematical properties. Optimality guarantees are of little value if the underlying generative process lies far outside the model's solution space.

The direct test potentially suffers more than the 2-step test from the issue of non-convexity, since in the 2-step test the coefficients benefit from convergence guarantees at the second step (e.g. LME) and the IRFs do not vary with ablation. However, the 2-step test potentially suffers more than the direct test from multicollinearity, since it cannot adjust IRF shapes in the ablated model that might have been influenced by multicollinear predictors in the full model. And although the 2-step procedure alone can guarantee maximum training likelihood conditional on the fitted IRF, this may not be of critical importance because the direct test does not implicitly require that training likelihood is maximized, since it is based on out-of-sample error rather than asymptotic distributional guarantees (cf. e.g. LRT, which relies on the result that the likelihood ratio test statistic asymptotically has a chi-squared distribution; Wilks, 1938). Indeed, because of the possibility of overfitting (memorizing noise in the training data), much research in statistics and machine learning has been dedicated to the study of regularization techniques and stopping criteria that avoid minimizing training error in pursuit of minimizing generalization error (Raskutti, Wainwright, & Yu, 2014; Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014; Yao, Rosasco, & Caponnetto, 2007), and out-of-sample bootstrap model comparison is one of the most widely used statistical tests for non-convex model comparison in machine learning (Demšar, 2006). Refocusing on out-of-sample rather than in-sample performance has the added benefit of building external model validity directly into the statistical test, which is potentially timely in light of growing concern over the replication crisis in psychological science (Gilmore, Diaz, Wyble, & Yarkoni, 2017; Makel & Plucker, 2014; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012; Simons, 2014; Yarkoni & Westfall, 2017). For this reason, we argue that the direct test is a defensible method for evaluating scientific hypotheses using CDR.

One possibility we also consider is a hybrid of these two approaches where 2-step fitting is used over the training set and then the second step fits are used to perform an out-of-sample non-parametric test. This approach would allow the second step fit to find globally optimal coefficients (regression weights) in the event that those discovered by CDR were sub-optimal. While this approach may be slightly more conservative than the direct test, we believe it has little practical utility because CDR is very good at finding optimal coefficients. Table 1 shows aggregated fixed-effects estimates obtained by running LME models over the CDR-convolved data $\mathbf{X}'$ from the both the synthetic and the human experimental datasets analyzed in this article. If the CDR-estimated coefficients are globally optimal given the impulse responses, then the LME estimates for all fixed effects should be 1. As shown, the LME-estimated fixed effects indeed cluster tightly around 1 for all three inference types explored here. This outcome suggests that CDR offers little room for improvement on its coefficient estimates, supporting the adequacy of the direct test for out-of-sample comparison.

In settings where the lack of optimality guarantee is unacceptable, CDR can still be a useful tool for data exploration, since it can estimate and visualize likely response profiles for predictors of interest (on exploratory data). Those estimates can then be used to construct effective and parsimonious FIR models. For example, suppose researchers have obtained a CDR-estimated IRF for predictor *A* which decays to near 0 in 300 ms, and that the mean interval between events in their experiment is 250 ms. This finding suggests they might be able to capture most of the temporally diffuse response to *A* simply by inserting one additional spillover regressor for *A* into their linear model. This kind of information would otherwise be difficult to obtain without first fitting models with multiple different spillover configurations of the same predictors, a computationally-intensive procedure. CDR can also be used to overcome the inability of discrete-time models to estimate generalized effects of stimulus timing (§4.3) by estimating a convolved *rate* predictor which can be added as an effect to standard regression models. Results in §7 indicate that this may be particularly important for some psycholinguistic response variables, especially response times in self-paced reading.

5. Methods

The experiments described in this article apply an open-source Python implementation of the CDR model described in §4 (`https://github.com/coryshain/cdr`), built using the Tensorflow (Abadi et al., 2015) and Edward (Tran et al., 2016) machine learning libraries. Eq. (10) is implemented as a Tensorflow computation graph and optimized using either MLE or variational Bayesian inference.

5.1. Initialization

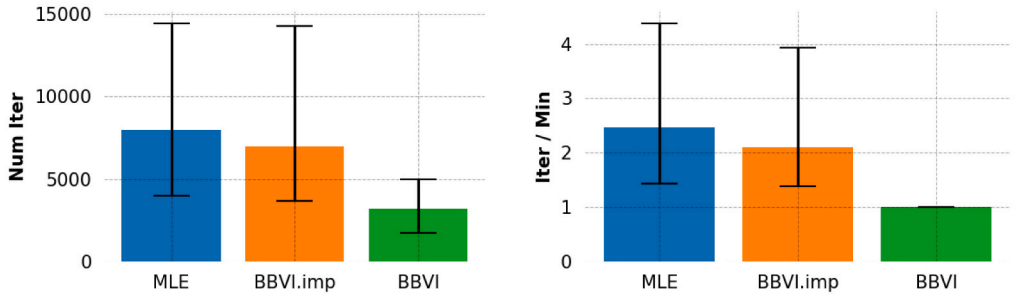
CDR fits are initially centered at the *null* model, i.e. a model in which there is no relationship between the predictors and response. In such a model, the intercept is the population mean, the variance is the population variance, all coefficients are 0 (predictors have no influence on the response), all random effects (Appendix A.1) are 0 (there is no random deviation from the population means), and the IRF shapes are inconsequential (there is no response to the predictors). Thus, μ is initialized at the mean of the response, σ^2 is initialized at the variance of the response, and all coefficients and random effects are initialized at 0.[16] Appropriate initializations for the fixed IRF parameters $\mathbf{v}_k$ are domain-specific, although kernels supported by our implementation come with overridable defaults as laid out in the documentation. The kernel initializations used in this study are described in §5.3.

5.2. Convergence

Because these CDR models use stochastic gradient optimization, it is necessary to define a convergence criterion by which the model parameters can be deemed (locally) optimal. Intuitively, the model has converged when it has ceased to improve with training. Diagnosing this condition automatically and model-independently is challenging because (1) the absolute rate of change in the loss over time depends on the scale of the data and the definition of the model and (2) the change in loss per iteration can be both noisy and non-decreasing due to stochastic optimization over a non-convex surface.

We address these challenges by retaining a history of the losses over a finite number of timesteps n and declaring convergence when the loss is uncorrelated with training time at a predetermined significance level α. Basing the convergence criterion on correlation eliminates any influence of scale on either the loss or the representation of training time, instead grounding convergence in the strength of the linear relationship

[16] The coefficients are $\mathbf{m}$ and the random effects are defined in Appendix 12.1 as $\mathbf{u}$, $\mathbf{U}$, and $\mathbf{V}_k$.



(a) Number of iterations

(b) Ratio of iterations to minimum in same experimental condition

Fig. 4. Median training time by inference type. Error bars show empirical 1st and 3rd quartiles.

between these two quantities. To reduce the influence of noise and high-frequency autocorrelation on the test statistic, we also permit a stride length m such that losses are only pushed to the history every m iterations, with intermediate values aggregated through a moving average.

In particular, given a vector of losses by iteration $\mathbf{l} \in \mathbb{R}^{\lceil n/m \rceil}$, and a vector $\mathbf{t} \in \mathbb{Z}^{\lceil n/m \rceil}$ of corresponding iteration numbers, we define correlation of loss with training time as a test statistic:

$$\rho_t \stackrel{\text{def}}{=} \text{corr}(\mathbf{l}, \mathbf{t}) \tag{18}$$

Given a significance level α, the null hypothesis $H_{0t} \stackrel{\text{def}}{=} \rho_t = 0$ can be tested by computing probability $p_{\rho t}$ using a Student's t distribution with $\lceil n/m \rceil - 2$ degrees of freedom and checking that $p_{\rho t} < \alpha$. When this test fails to reject H_{0t}, the losses are uncorrelated with training time (ρ_t is insignificantly different from 0 at level α).

This correlation-based hypothesis test defines binary criterion s:

$$s \stackrel{\text{def}}{=} \begin{cases} 1 & p_{\rho_t} > \alpha \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

In other words, $s = 1$ if and only if H_{0t} is retained. To avoid premature convergence when by chance ρ_t happens to have small magnitude, we additionally store the history of values of s from each of the past $\lceil n/m \rceil$ strides in vector $\mathbf{s} \in \{0, 1\}^{\lceil n/m \rceil}$ and compute a proportion p_s of successful convergence checks:

$$p_s \stackrel{\text{def}}{=} \frac{m\|\mathbf{s}\|_1}{n} \tag{20}$$

We declare the model converged when $p_s > \alpha$ (i.e. when at least $\alpha \times 100\%$ of the previous $\lceil n/m \rceil$ convergence checks were positive). The overall stringency of this criterion increases with both n/m (because the test is higher-powered) and α (because the tests reject at a higher threshold and a larger proportion p_s is required). Experiments reported in this study use the following convergence hyperparameters: $n = 500$, $m = 1$, and $\alpha = 0.5$. In other words, convergence is declared when H_{0t} is retained at $\alpha = 0.5$ for at least 50% of the previous 500 training iterations.

5.3. Experimental procedure: General model parameters

In all experiments reported in this article, CDR models are fitted using the Nadam optimizer (Dozat, 2016)[17] with a constant learning rate of 0.001 and minibatches of size 1024. Nadam is a gradient-based optimizer that performs well for many non-convex optimization problems, and 0.001 is a widely-used default learning rate (see Tensorflow documentation: https://www.tensorflow.org/). Minibatch sizes in powers of two improve efficiency on standard compute architectures, and we found during development that a size of 2^{10} roughly optimized iteration speed on our available hardware. These parameters were not otherwise systematically tuned.

In order to compare estimation methods, all models are fitted using maximum likelihood estimation (MLE), black-box variational inference with independent improper uniform priors and independent normal posteriors (BBVI-improper), and black-box variational inference with independent normal priors and posteriors (BBVI).[18] BBVI-improper and BBVI use the same black box estimation procedure. The only difference between them is that BBVI-improper lacks any penalties for divergence from a prior.

The analyses reported below in this article reveal dramatically faster convergence in BBVI inference mode than in MLE or BBVI-improper modes. This asymmetry is demonstrated by Fig. 4, which shows (1) the mean number of training iterations (complete passes through the training data) across all experimental conditions and (2) the mean ratio of training iterations to the minimum number of iterations used by any model within the same experimental condition. As shown, BBVI requires on average less than half as many training iterations as MLE or BBVI-improper to reach convergence (Fig. 4a) and nearly always requires fewer iterations than MLE or BBVI-improper in any given model configuration (Fig. 4b). We speculate that the priors may discourage the model from following tiny gradients, thereby accelerating convergence. Because of this computational advantage, we expect BBVI to be the most useful estimation method and therefore perform all subsequent model comparisons (both for null hypothesis significance testing and for comparison of CDR to baselines) using BBVI-estimated CDR models, even when other estimation techniques achieve better error.

In the BBVI setting, reasonable prior variances for intercepts, coefficients, and error depend on the scale of the response. For this reason, our CDR implementation uses the variance of the response in the training set as the prior variance for these parameters. Reasonable prior variances for the impulse response parameters are independent of the scale of the response. For simplicity, we use a variance of 1. In order to improve both convergence and general applicability of the hyperparameters used in this study to other kinds of data, we implicitly standardize (z-transform) the response variable prior to fitting, then invert this transform in order to compute predictions and likelihoods. As a result, all BBVI priors used in this study have unit variance and mean

[17] The Adam optimizer (Kingma & Ba, 2014) with Nesterov momentum (Nesterov, 1983)

[18] Although our software implementation includes experimental support for multivariate normal priors and variational posteriors, this results in a quadratic increase in the number of parameters, and we have not found it to provide any performance benefit.

equal to the initialization value used in MLE inference (described below for IRF parameters and in §5.1 for all other parameters).

While it is in principle sensible in a BBVI setting to use the prior as the initial value for the variational posterior, in practice we have found that doing so can lead to training divergence in complex models due to early initial sampling of poor solutions from an excessively wide distribution. We therefore initialize the variational posterior with a tighter standard deviation (one one-hundredth of the standard deviation of the prior) in all BBVI-improper and BBVI models. Note that this is merely an initialization technique — the model can adjust the width of the variational posterior throughout training as required by the data.

It is standard practice in mixed-effects modeling to penalize the random effects (Bates et al., 2015). In all psycholinguistic analyses with random effects reported below, we follow Bates et al. (2015) by penalizing all random effects (random intercepts, coefficients, and impulse response parameters, see Appendix A.1) using L2 regularization with regularization level $\lambda = 1.0$. In the BBVI setting, we implement a similar kind of constraint by imposing a tighter prior on the random effects than on the fixed effects (standard deviation of 0.1 and 1.0, respectively).

The BBVI priors used in this study were not tuned in any way, and different priors may be motivated for different datasets based on foreknowledge of the experimental domain. The size of the hyperparameter space explored in this study is so large that additional systematic exploration of the influence of different prior settings is computationally prohibitive and left to future work. That said, the present results suggest that our choices of priors do not strongly constrain the solution space, since BBVI inference finds qualitatively similar responses to the BBVI-improper and MLE inferences, which have no priors (see Appendix C).

Models reported here use some combination of *exponential*, *normal*, *shifted gamma*, and pseudo non-parametric linear combination of Gaussians (LCG) impulse response kernels. The *exponential*, *normal*, and *shifted gamma* kernels are the probability density functions associated with each type of probability distribution, with the addition of normalization terms to ensure that IRFs integrate to 1 over the positive real line (§4.2). For these parametric kernels, the normalization term is the survival function of the distribution (complement of the cumulative density function) at $x = 0$.

The probability density functions of the exponential, normal, and shifted gamma distributions are respectively:

$$f_{\text{Exp}}(x;\beta) = -\beta e^{-\beta x} \tag{21}$$

$$f_{\text{Normal}}\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{\sigma^2}} \tag{22}$$

$$f_{\text{ShiftedGamma}}(x;\alpha,\beta,\delta) = \frac{\beta^\alpha (x-\delta)^{\alpha-1} e^{-\beta(x-\delta)}}{\Gamma(\alpha)} \tag{23}$$

The exponential distribution integrates to 1 over the positive real line and requires no further normalization. For the *normal* and *shifted gamma* kernels, we require the corresponding survival functions:

$$S_{\text{Normal}}\left(x;\mu,\sigma^2\right) = 1 - \frac{1}{2}\left(1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right) \tag{24}$$

$$S_{\text{ShiftedGamma}}(x;\alpha,\beta,\delta) = 1 - \frac{1}{\Gamma(\alpha)}\gamma(\alpha,\beta(x-\delta)) \tag{25}$$

where

$$erf(x) = \frac{1}{\sqrt{\pi}}\int_{-x}^{x} e^{-t^2} dt \tag{26}$$

$$\gamma(s,x) = \int_0^x t^{s-1} e^{-t} dt \tag{27}$$

Thus, the *exponential*, *normal*, and *shifted gamma* IRF kernels are defined respectively as:

$$\text{Exp}(x;\beta) = f_{\text{Exp}}(x;\beta) \tag{28}$$

$$\text{Normal}\left(x;\mu,\sigma^2\right) = \frac{f_{\text{Normal}}(x;\mu,\sigma^2)}{S_{\text{Normal}}(0;\mu,\sigma^2)} \tag{29}$$

$$\text{ShiftedGamma}(x;\alpha,\beta,\delta) = \frac{f_{\text{ShiftedGamma}}(x;\alpha,\beta,\delta)}{S_{\text{ShiftedGamma}}(0;\alpha,\beta,\delta)} \tag{30}$$

These kernels encode increasingly flexible assumptions about the response shape. In particular, the *exponential* kernel assumes that the influence of a predictor is strongest immediately (at $t = 0$) and decreases monotonically over time. The only question is how quickly, which is determined by the rate parameter β. The *normal* kernel relaxes this monotonicity assumption. Like the *exponential* kernel, it can fit monotonically decreasing IRFs (by finding location $\mu \leq 0$), but it can also fit late-peaking (rising then falling) IRFs (by finding $\mu > 0$), allowing the peak response to occur at some delay from the stimulus (e.g. Smith & Levy, 2013, who found a peak surprisal response on the following word in a self-paced reading experiment). In the latter case, the *normal* kernel assumes a symmetric rise/fall pattern. The *shifted gamma* kernel additionally relaxes the symmetry assumption. It can find approximately symmetric late-peaking IRFs, but it can also find IRFs that e.g. rise quickly and then decay slowly. The *shifted gamma* kernel approximately subsumes the solution space of the *normal* kernel, which approximately subsumes the solution space of the *exponential* kernel.

As mentioned above, we also construct a quasi-non-parametric comparison kernel consisting of a linear combination of Gaussians (LCG; Goshtasby & O'Neill, 1994; Gimel'farb, Farag, & El-Baz, 2004). The LCG kernel contains n component *normal* kernels, each with location, scale, and amplitude parameters. Having a more flexible control allows us to assess the extent of bias introduced by parametric kernels when applied to real world data in which the ground-truth IRF is unknown (§7, §8). This LCG kernel is defined as:

$$\text{LCG}\left(x;\mu_{1,\ldots,n},\sigma^2_{1,\ldots,n},\beta_{1,\ldots,n}\right) = \frac{\sum_{i=1}^{n} f_{\text{Normal}}\left(x;\mu_i,\sigma_i^2\right)\cdot\beta_i}{\sum_{i=1}^{n} S_{\text{Normal}}(0;\mu_i,\sigma_i^2)\cdot\beta_i} \tag{31}$$

The LCG kernel is highly flexible (see §6, §7, and §8), yet computationally efficient, since it has an analytical normalization constant (the sum of the component survival functions). More expressive kernels are possible, such as spline functions or kernel smoothing, but these do not have analytical integrals and therefore require numerical integration at each optimization step in order to properly normalize them. In initial experiments, we found this to greatly slow training without clear improvement to final fit, and we therefore use LCG for our non-parametric comparison. In these experiments, all LCG kernels have 10 components. The choice of kernel type thus has a substantial influence on the size of the parameter arrays estimated by the model. For example, assuming that all predictors share an *exponential* kernel, then the IRF parameter vectors $\mathbf{v}_k$ would each be of size 1 (the rate parameter β), whereas they would be of size 3 for *shifted gamma* (rate α, scale β, and shift δ) and size 30 for LCG (location μ, scale σ, and amplitude β for each of the 10 components).

Beyond any constraints imposed by the definitions of the density functions above, the *shifted gamma* kernel additionally requires that $\alpha > 1$ and $\delta < 0$.[19] Constraints on bounded parameters are enforced using the softplus bijection:

$$\text{softplus}(x) = log(e^x + 1) \tag{32}$$

All models use the same default initializations for these kernels. For *exponential* kernels, $\beta = 1$. For *shifted gamma* kernels, $\alpha = 2$, $\beta = 1$, and $\delta = -1$. For *normal* kernels, $\mu = 0$ and $\sigma^2 = 1$. For the LCG kernels, $\mu_i = 0$, $\sigma_i^2 = i$

[19] $\alpha > 1$ helps deconfound the shape and shift parameters by ensuring that the response underlyingly has a rising-falling profile, in which case strictly falling responses can only be found by shifting the peak to the left of 0. $\delta < 0$ ensures that the instantaneous response (response at $x = 0$) is well defined.

and $\beta_i = 1$ if $i = 0$, $\beta_i = 0$ otherwise for $i \in \{1, \ldots, 10\}$. This initializes the kernel with a single non-zero component, and with incrementally wider initial components to allow the model to find later or earlier peaks. Initializing all components with $\beta = 0$ leads to numerical degeneracies because of the requirement that the LCG function integrate to 1.

Prediction from the network uses an exponential moving average of parameter iterates with a decay rate of 0.999. BBVI models are evaluated using *maximum a posteriori* estimates obtained by setting all parameters to their posterior means. This procedure is motivated by the law of large numbers: because all parameters have independent normal distributions in the variational posterior, samples from that posterior converge in probability to the posterior mean. For computational reasons, we truncate predictor histories at 256 timesteps (words) into the past.

6. Synthetic evaluation

Before applying CDR to human-generated time series, we first empirically validate it through simulations using synthetic data with known ground-truth impulse responses, since this permits direct comparison of the CDR estimates to the true data-generating model. In the process, we systematically explore the sensitivity of CDR estimates to several potential sources of influence that are likely to arise in practice: noise in the response variable, non-uniform time intervals between events, multicollinearity, and misspecification of the impulse response kernel. As shown below, CDR recovers the data-generating model in a wide range of settings and is robust to adverse training conditions like multicollinearity and IRF misspecification.

6.1. Simulation design

Several design details are common to all simulations reported here. All datasets contain twenty randomly generated predictors. In all simulations except the multicollinearity manipulations, these twenty predictors are sampled from independent standard normal distributions. In all datasets, ground-truth coefficients for each predictor are sampled from a uniform distribution $\mathscr{U}(-10, 10)$, and a ground-truth impulse response is created for each predictor by randomly sampling parameters for a given impulse response kernel. The response is then generated by convolving each predictor with its assigned impulse response, sampling the convolved signal at predetermined query points (timestamps), scaling the sampled signal by the ground-truth coefficients, and summing the scaled sample across predictors in order to generate a response vector. In all simulations except the noise manipulations, Gaussian noise with standard deviation 10 is added to the generated response vector. In all simulations except the time manipulations, time intervals between stimulus events and response samples are asynchronous and sampled from an exponential distribution with mean 100 ms. For simplicity, all synthetic datasets consist of a single timeseries containing 10,000 response samples. For all simulations, we provide (1) qualitative assessments of IRF identification by visually comparing true and estimated responses and (2) quantitative assessments of IRF identification by computing root mean squared deviation (RMSD) of the estimated response from the true response over 1000 timepoints spaced equidistantly on the 95th percentile of temporal offsets seen in training. For simplicity, only BBVI-estimated responses are shown here. Full results are given in Appendix C.1.

6.2. Simulations

6.2.1. Simulation A: Noise

Simulation A explores the sensitivity of CDR estimates to noise in the response variable. A single set of 20 independent predictors is sampled as described above, and a single set of *shifted gamma* impulse responses is sampled from the following distributions: $\alpha \sim \mathscr{U}(1, 6)$, $\beta \sim \mathscr{U}(0, 5)$, and $\delta \sim \mathscr{U}(-1, 0)$. Gaussian noise with standard deviation 0 (noise free), 1, 10, and 100 is then injected into the convolved response, and CDR models with *shifted gamma* IRF kernels are fitted separately to each level of noise. The true simulated response has a signal power (mean squared value) of 1907, and thus the signal to noise ratios of the synthetic datasets are respectively ∞, 1907, 19.07, and 0.1907.

Table 2
Number of parameters by kernel family and corpus (Simulation D, Natural Stories, and Dundee, respectively). Differences between corpora are driven by the random effects, since Natural Stories contains many more participants (181) than Dundee (10), while Simulation D contains no random effects. Note that BBVI and BBVI-improper double these figures by additionally fitting variances for each parameter in the variational posterior.

	Parameters		
Kernel	SimD	NatStor	Dundee
Exponential	42	1970	166
Normal	62	2686	232
Gaussian	82	3402	298
LCG	662	21,845	2080

6.2.2. Simulation B: Time

Simulation B explores the sensitivity of CDR estimates to different kinds of time intervals between predictor and response observations. We consider three manipulations: (1) fixed vs. variable spacing, (2) long vs. short intervals, and (3) synchronous vs. asynchronous measures of predictors and response. To evaluate these influences, we construct six conditions manipulating the length, variability, and alignment of time intervals between consecutive predictors/responses:

- **Fixed synchronous short (FSS)**: Predictors and response are aligned and placed at fixed 100 ms intervals.
- **Fixed synchronous long (FSL)**: Predictors and response are aligned and placed at fixed 500 ms intervals.
- **Random synchronous short (RSS)**: Predictors and response are temporally aligned and intervals are sampled from an exponential distribution with mean 100 ms.
- **Random synchronous long (RSL)**: Predictors and response are temporally aligned and intervals are sampled from an exponential distribution with mean 500 ms.
- **Random asynchronous short (RAS)**: Predictors and response are not temporally aligned: intervals are sampled independently for predictors on the one hand and response on the other from an exponential distribution with mean 100 ms.
- **Random asynchronous long (RAL)**: Predictors and response are not temporally aligned: intervals are sampled independently for predictors on the one hand and response on the other from an exponential distribution with mean 500 ms.

The data generating model is constructed using the same procedure as in Simulation A.

6.2.3. Simulation C: Multicollinearity

Simulation C explores the sensitivity of CDR estimates to multicollinearity in the predictors. 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:

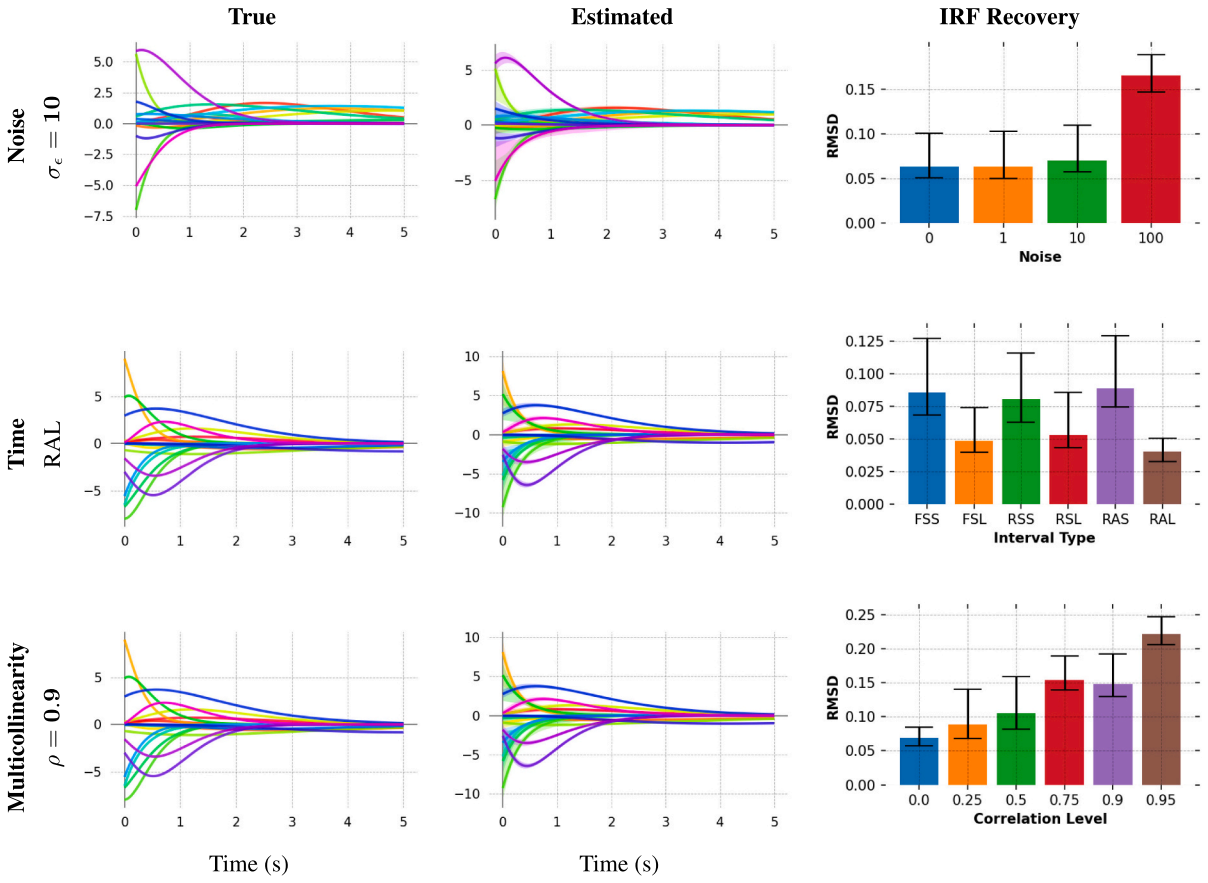


Fig. 5. Simulation A (top): *Noise*. True (left) vs. estimated (center) model under Gaussian noise with standard deviation 10. **Simulation B (middle)**: *Time*. True (left) vs. estimated (center) model under the random asynchronous long (RAL) condition. **Simulation C (bottom)**: *Multicollinearity*. True (left) vs. estimated (center) model under a pairwise correlation level of $\rho = 0.9$. The twenty IRFs (corresponding to the twenty random covariates) are represented by distinct curves in each plot. 95% credible intervals are shown. Right: Root mean squared deviation (RMSD) of BBVI-estimated from true models across simulation conditions. Error bars show 95% Monte Carlo credible intervals.

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

Six sets of predictors were generated in this way, one for each of $\rho = 0$ (uncorrelated predictors), $\rho = 0.25$, $\rho = 0.5$, $\rho = 0.75$, $\rho = 0.9$, and $\rho = 0.95$. The data-generating model was constructed using the same procedure as in Simulations A and B.

6.2.4. Simulation D: IRF misspecification

Simulation D explores the ability of CDR to find reasonable IRF estimates in the presence of mismatch between the true and modeled IRF kernels. To this end, we construct three data-generating models with different underlying response shapes: *exponential* (E), *normal* (N), and *shifted gamma* (G). For each of these datasets, we fit CDR models of each kernel type, such that two out of three modeled kernels for each dataset are not matched to the underlying model (e.g. fitting *exponential* IRFs to the output of a *normal* data-generating model). The target outcome under kernel mismatch is to discover the best available estimate given the solution space defined by the modeled kernel. We also explore the use of more flexible LCG kernels (described in §5.3), since their solution space approximately subsumes all ground truth models used in this simulation. As shown in Table 2, the LCG kernel is much more heavily parameterized than the others.

Note that these different kernels have asymmetrical patterns of compatibility. For example, the solution space of the *shifted gamma* kernel contains the *exponential* kernel, since the exponential distribution is a special case of the shifted gamma distribution (i.e. when $\alpha = 1$ and $\delta = 0$).[20] The converse does not hold: *shifted gamma* contains late-peaking responses that fall outside the strictly monotonic solution space of the *exponential* kernel. The *normal* kernel is more flexible than the *exponential* kernel and may therefore be able to better approximate *shifted gamma* responses, but it is additionally constrained by symmetry about the mean.

The predictors, coefficients, and *shifted gamma* data-generating model are constructed using the same procedure as in Simulations A and C. The impulse responses for the *exponential* data-generating model are constructed by sampling $\beta \sim \mathcal{U}(0, 5)$. The impulse responses for the *normal* data-generating model are constructed by sampling $\mu \sim \mathcal{U}(-2, 2)$, $\sigma^2 \sim \mathcal{U}(0, 2)$.

6.3. Results and discussion

Figs. 5 and 6 show that CDR accurately recovers the underlying model across all conditions explored here. For brevity, we include only

[20] Because these values lie at the parameter bounds for the *shifted gamma* kernel used here, the model cannot exactly reach them. However, it can come arbitrarily close within 32-bit floating point precision.

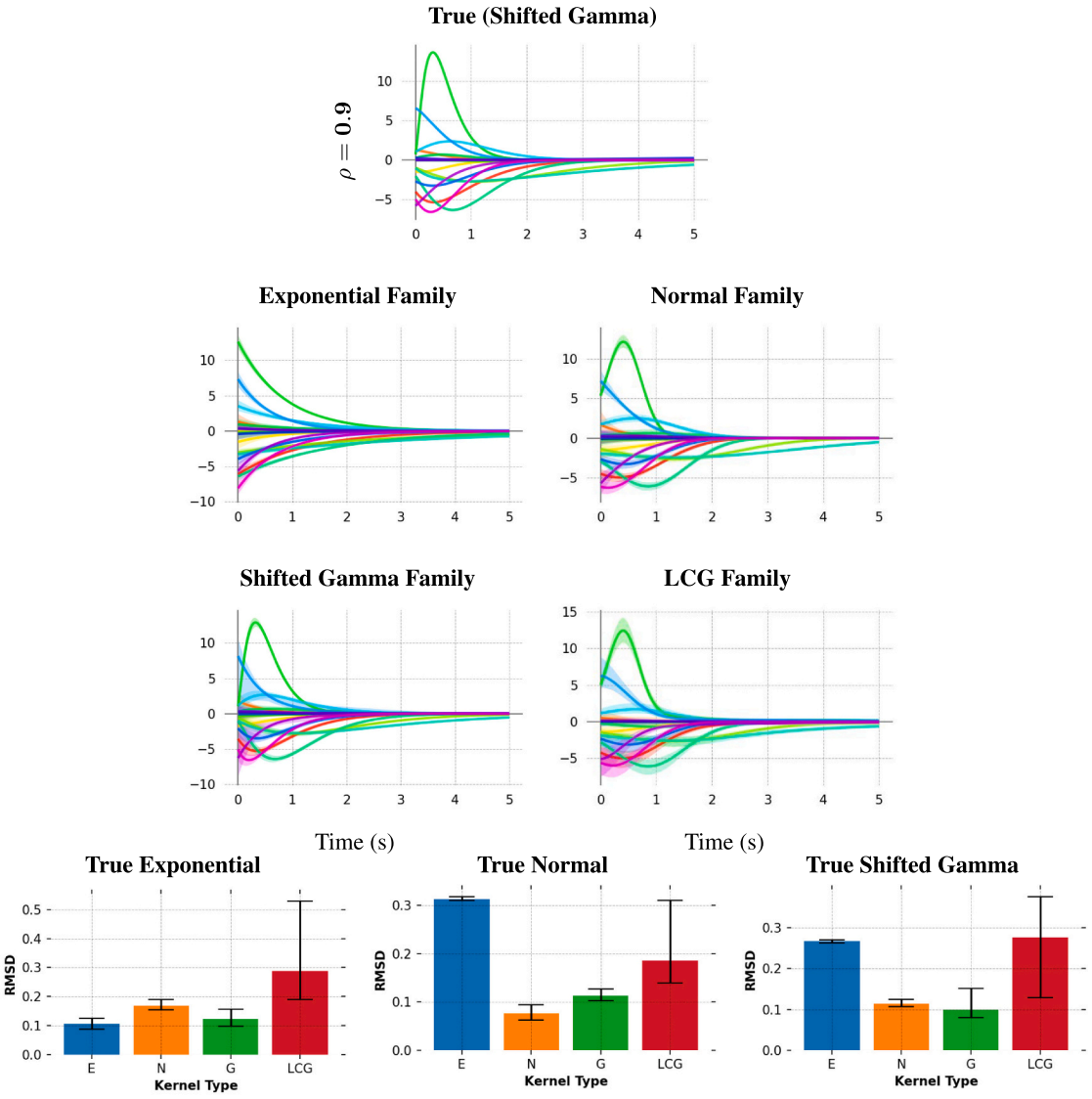


Fig. 6. Simulation D: *IRF Misspecification*. True shifted gamma model (top) vs. models estimated from four different parametric families. The twenty IRFs (corresponding to the twenty random covariates) are represented by distinct curves in each plot. 95% credible intervals are shown. Bottom: Root mean squared deviation (RMSD) of BBVI-estimated from true models across simulation conditions — exponential (E), normal (N), shifted gamma (G), and linear combination of Gaussians (LCG) kernels applied to true exponential, normal, and shifted gamma responses. Error bars show 95% Monte Carlo credible intervals.

visual comparisons of representative true and BBVI-estimated responses across simulations. Full results are given in Appendix C.1. Visually, CDR estimates closely resemble the underlying model across conditions. This impression is supported quantitatively by low root mean squared deviation (RMSD) of estimated from true curves across conditions: mean RMSDs are nearly always within about 0.3, and mostly within 0.2, both of which are a small fraction of the maximum absolute peak amplitude (~18) in these experiments. Synthetic results therefore indicate that CDR can reliably identify the underlying temporal dynamics, even in the presence of high levels of noise (Simulation A), asynchronous predictor-response alignment (Simulation B), multicollinearity (Simulation C), and IRF misspecification (Simulation D).

Figs. 5 and 6 provide some additional finer-grained insights. For example, as expected, RMSDs increase under noisier data or more severe multicollinearity (Fig. 5, top and bottom rows). In addition, RMSDs are consistently higher in Simulation B when the interval between events is short (Fig. 5, middle row), which results from poorer temporal coverage of slow response components (Appendix C.1). This outcome underlines the importance of selecting a history window with sufficient coverage to support identification of all plausible impulse response components.

Furthermore, results in Fig. 6 indicate that success does not crucially rely on foreknowledge of the true impulse response family. Although the underlying model has a shifted gamma shape, CDR estimates using *normal* or LCG kernels largely recover its underlying structure. As expected, there are limits to this generalization. For example, as discussed above, exponential kernels cannot discover late-peaking responses, resulting in large RMSDs for exponential kernels fitted to data from underlyingly normal- or shifted-gamma-shaped models. In addition, the high dimensionality of the LCG kernels makes them more prone to find wiggly estimates, leading to higher overall RMSD.

This simulation study shows that CDR consistently finds solutions near the global optimum across a range of inference methods, simulation types, and adverse training conditions. Thus, despite the theoretical vulnerability to local optima due to non-convex optimization (§4.6), simulation results (1) suggest that such potential problems often have little practical impact and therefore (2) support the use of CDR modeling to test scientific hypotheses.

7. Experiment A: Naturalistic reading

We now apply CDR to discover effect timecourses in human reading behavior from two large naturalistic datasets. As argued in §2, analyzing naturalistic reading is a key application area of interest, both because (1) diffusion of effects may be especially pronounced in the naturalistic reading paradigm and (2) existing tools like FIR/spillover models have a number of shortcomings when applied to this domain (see §3). Our analyses are inspired by those reported in Shain (2019) but focus primarily on the validation of CDR rather than on an empirical claim about human sentence processing. To this end, much like in the simulation studies reported in §6, we apply multiple estimation techniques and impulse response kernels to each corpus.

The primary interest of CDR as an explanatory model is that it yields detailed estimates of diffuse temporal structure that existing discrete-time regression techniques cannot provide. However, unlike the simulations above, the data-generating model for human reading responses is unknown, and its temporal structure is currently poorly understood. Thus, direct comparison to the ground truth is not available for model validation on human data; indeed, a potential benefit of CDR modeling is its ability to shed light on this important question. For this benefit to be realized, it is first necessary to establish that CDR provides a "good" model of human responses according to some standard; otherwise, its estimates should not be trusted. In this study, we propose that such a standard can be constructed using the predictive performance of established statistical methods. If CDR models generalize less well to unseen data than standard models, then their estimates of temporal structure should be treated with skepticism. However, if CDR predictions perform competitively with those of standard models, then this indicates that the model has tapped into generalizable properties of the response. In this study, we compare the predictive performance of CDR to that of linear mixed effects (LME) and generalized additive (GAM) models, both of which have been used extensively in psycholinguistics. In doing so, we stress that the primary advantage of CDR lies in its ability to estimate continuous diffusion over time, with performance comparisons serving as a sanity check. Below, we show that CDR predictive performance is competitive with that of all baselines in each dataset and superior to all baselines overall, supporting the reliability of its estimates of temporal structure. In addition, we show that CDR performance is stable across a range of estimation methods and IRF kernels.

7.1. Data

This reading study uses the Natural Stories (Futrell et al., 2021) and Dundee (Kennedy, Pynte, & Hill, 2003) datasets.[21] Natural Stories is a self-paced reading (SPR) corpus consisting of context-rich narratives read by 181 subjects. Stimuli are designed to resemble naturally-occurring texts while increasing the representation of rare words and syntactic constructions. Subjects 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 corpus contains a total 1,013,290 events (where one event is a single subject viewing a single word token).[22]

Dundee is an eye-tracking corpus containing newspaper editorials read by 10 subjects. The corpus contains a total of 340,840 events (where one event is a single subject entering and then exiting a single word region).[23]

In both reading experiments, data are partitioned into training (50%), exploratory (25%) and test (25%) sets. The partitioning strategy attempts to respect the non-independence of words within the same sentence, using modular arithmetic to cycle sentence IDs e into different bins of the partition with a different phase for each subject u: partition(e, u) = $(e + u)$ mod 4, assigning outputs 0 and 1 to the training set, 2 to the exploratory set, and 3 to the test set. Outlier filtering is also performed, largely following the procedures described in Shain and Schuler (2018).[24] Because CDR's convolution operation is only correct if applied to all preceding events within the history window, partitioning and filtering are applied only to the response, retaining all events in the predictor matrix.[25]

7.2. Experimental setup

The purpose of Experiment A is to evaluate CDR as a statistical model

[21] Although we have previously reported deconvolutional results (Shain, 2019; Shain & Schuler, 2018) on the UCL eye-tracking corpus (Frank, Monsalve, Thompson, & Vigliocco, 2013a), we have chosen not to include UCL in the current study because we believe it is not a reliable test case for deconvolutional modeling. While Natural Stories and Dundee contain long texts, UCL contains randomized sentences presented in isolation. Since the public version of UCL provides no information about inter-stimulus intervals, the only way to run CDR on UCL is to treat each sentence as a distinct time series, which results in insufficient amounts of elapsed time between impulses and responses to reliably identify IRFs. In the case of Natural Stories and Dundee, the 95th percentile of temporal offsets seen in training is 123s and 93s, respectively, while in UCL (treating sentences as independent time series) it is less than 2s. In general, data consisting of isolated sentences (a frequent stimulus design in psycholinguistics) are perfectly compatible with CDR modeling as long as the recording session (rather than the sentence) is treated as a time series and inter-stimulus intervals are considered when computing event timestamps. An enriched version of the UCL data that contains inter-stimulus intervals would thus permit deconvolution by allowing timestamps to be accurately computed over the full recording session. We leave this possibility to future work.

[22] This figure differs from the published count in Futrell et al. (2021) because we do not filter events from the stimulus sequence, since they are needed for accurate deconvolution.

[23] A limitation of the Dundee corpus is the number of participants (10). Although each of these participants is quite densely sampled and should therefore be able to be reliably modeled, the small number of participants may limit the degree to which results based on Dundee can be expected to generalize to the population as whole. While we acknowledge this concern, the purpose of the present study is not to test hypothesized effects in human sentence processing, but rather to evaluate the empirical properties of a new modeling approach (CDR). Dundee is one of the most extensively analyzed naturalistic eye-tracking corpora in psycholinguistics (e.g. Demberg & Keller, 2008; Frank & Bod, 2011; Fossum & Levy, 2012; Smith & Levy, 2013; van Schijndel & Schuler, 2015; Goodkind & Bicknell, 2018, inter alia) and therefore serves as a "standard" dataset for initial evaluation of CDR for eye-tracking. CDR analysis of eye-tracking corpora with larger numbers of participants (e.g. Cop, Dirix, Drieghe, & Duyck, 2017) is left to future work.

[24] For Natural Stories, following Shain et al. (2016), items were excluded if they have fixations shorter than 100 ms or longer than 3000 ms, if they start or end a sentence, or if subjects missed 4 or more subsequent comprehension questions. We additionally removed any subjects with fewer than 100 datapoints after application of the other filters, both because such subjects are likely uncooperative (missing excessive comprehension questions or paging too rapidly through the text) and because their data are likely insufficient to support estimation of random effects. 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. In addition, following common practice in psycholinguistics, we removed items whose duration included a blink (e.g. Schotter, Leinenger, & von der Malsburg, 2018). 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.

[25] In these experiments, an entire document is treated as a time series, with the result that words can continue to influence the response across sentence boundaries.

of human reading. To this end, all analyses share the following design features.

7.2.1. Response variable

In all experiments, the response variable of interest is reading latency, under the eye-mind assumption (Just & Carpenter, 1980) that longer latencies index greater processing difficulty. The definition of latency varies by experimental modality. For self-paced reading data (Natural Stories), reading latency is defined as reaction time — the interval between button presses. For eye-tracking data (Dundee), a number of latency measures are possible (Rayner, 1998). In this study, we focus on *scan path* duration, defined as the time elapsed between entering a word region (from either direction) and entering a different word region (in either direction). Under the assumption that word features (e.g. frequency and surprisal) do not accumulate in influence from consecutive saccades to the same word, we sum over all consecutive saccades to the same word region. Compared to other commonly-used duration measures (e.g. first pass and go-past durations), scan path durations follow the temporal sequence of eye movements rather than the spatial sequence of words on the screen (these two sequences can differ due to regressive — backward — eye movements). Appendix C.2 contains additional, and generally consistent, results from applying CDR to first pass and go-past durations in Dundee.[26] In all cases, latency is measured in milliseconds. Because of the non-normal distribution of reading times in psycholinguistic experiments (e.g. Frank, Monsalve, Thompson, & Vigliocco, 2013b), we also consider log-transformed variants of reading time durations, following e.g. Smith and Levy (2011).[27]

7.2.2. Predictor variables

Models use the following predictors: *sentence position* (index of word in sentence), *document position* (index of word in document), incoming *saccade length* (in words, eye-tracking only), *previous was fixated* (indicator for whether the preceding word was fixated, eye-tracking only), *word length* (in characters), *unigram surprisal*,[28] and *5-gram surprisal*.[29] *Unigram surprisal* and *5-gram surprisal* are computed by the KenLM toolkit (Heafield, Pouzyrevsky, Clark, & Koehn, 2013) trained on Gigaword 3 (Graff, Kong, Chen, & Maeda, 2007). 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). Models also include a deconvolutional intercept, referred to as *rate*, which is designed to capture any generalized response to stimuli, independently of their properties (§4.3). We exclude *rate* from all non-CDR baseline models reported below because it is identical to the intercept, and thus these baselines are unable to identify *rate* effects.

Models include a rich random effects structure to capture variation between individuals, with by-subject random intercepts, slopes, and IRF parameters.[30] By-word random intercepts, though common in psycholinguistic studies (Demberg & Keller, 2008), are avoided because (1) they can absorb context-independent effects like *word length* and *unigram log probability* and (2) early experiments suggest that by-word intercepts lead to overfitting (based on exploratory set performance) in both CDR and baseline models. All predictors are rescaled by their standard deviations prior to fitting.[31]

Prior work suggests the following a priori expectations about the effect estimates for these predictors. *Saccade length*, *word length*, and *5-gram surprisal* are expected to increase processing difficulty (Demberg & Keller, 2008). According to several preceding studies, *unigram surprisal* should also positively modulate processing difficulty (see Staub, 2015, for review), although this pattern has recently been called into question (Shain, 2019). *Previous was fixated* has been shown to have a facilitation effect (van Schijndel & Schuler, 2015). *Sentence position* and *document position* are designed to capture trends in the response over different timescales (sentences and documents). Previous work indicates that reading times decrease over the course of experiments (Baayen et al., 2017), suggesting an expected negative effect of *document position*. Previous estimates for *sentence position* have been small-magnitude and negative (Demberg & Keller, 2008; van Schijndel & Schuler, 2015). *Rate* effects in reading data have not been carefully studied, in part for lack of CDR (though see Shain & Schuler, 2018).

The predictors *saccade length*, *word length*, *unigram surprisal*, 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 & 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, *document position* 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 (Baayen et al., 2018; Demberg & Keller, 2008), their presence in the model is motivated by the possibility of trends in the response. For this reason, parametric IRFs are fitted to all predictors except *document position* and *sentence position*, which are assigned a (parameter-free) Dirac delta IRF (i.e. a linear coefficient). *Document position* and *sentence position* are thus omitted from IRF plots. Because the functional family of the underlying response is unknown, we explore the impact of the modeled response kernel by fitting *exponential* (E), *normal* (N), *shifted gamma* (G), and pseudo non-parametric linear combination of Gaussians (LCG) kernels with default initializations (§5.1), as in Simulation D (§6).

Because we are analyzing scan path durations in Dundee (which contain regressive fixations), it is plausible that the variables of interest exert different influences in the scan path record depending on whether they belong to a word that is being fixated for the first time vs. a word that is being re-fixated or fixated as part of a regression. This is especially true of e.g. surprisal — presumably a surprising word is less surprising after it has already been observed. While there are many conceivable ways of accounting for the possibility of such interactions in the model design, in the interests of parsimony we opted for a simple approach of splitting each variable into two predictors, one corresponding to fixations that are part of a regressive eye movement, and one corresponding to fixations that are not. These two variants thus partition the variable among the fixations. For example, the single vector of surprisal values by fixation is split into two vectors, one containing only those values that are associated with regressive fixations, and one containing only those values that are associated with non-regressive fixations, with zeros elsewhere. In all results, we distinguish regressive estimates with "(+reg)". In addition, we include an indicator variable for *notregression*, to account for any generalized difference in response profile for regressive vs. non-regressive fixations.

[26] The question of how to define events and measures in the reading record (scan path, first pass, go-past, regression probability, etc.) is orthogonal to the question of how to analyze the data (LME, GAM, CDR): as shown in our full ensemble of Dundee results (here and in Appendix C.2), comparable discrete-time LME/GAM models with spillover can also be constructed for each response definition, including scan paths.

[27] Appendix C.2 additionally reports on results using an alternative non-normal error distribution (sinh-arcsinh) for estimating IRFs from reading data.

[28] We represent unigrams on a surprisal scale (negative log probability) simply to facilitate comparison with 5-gram effects, but recognize that they are a degenerate (memory-less) model of surprisal and are usually included in psycholinguistic models to capture lexical retrieval rather than prediction effects (Shain, 2019; Staub, 2015).

[29] All surprisals used in this study are fully lexicalized in that the support of their underlying probability models is (a subset of) the vocabulary of English, rather than syntactic abstractions like parts of speech.

[30] See Appendix A.1 for mixed model equations and Appendix D for an empirical analysis of random effects in CDR.

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

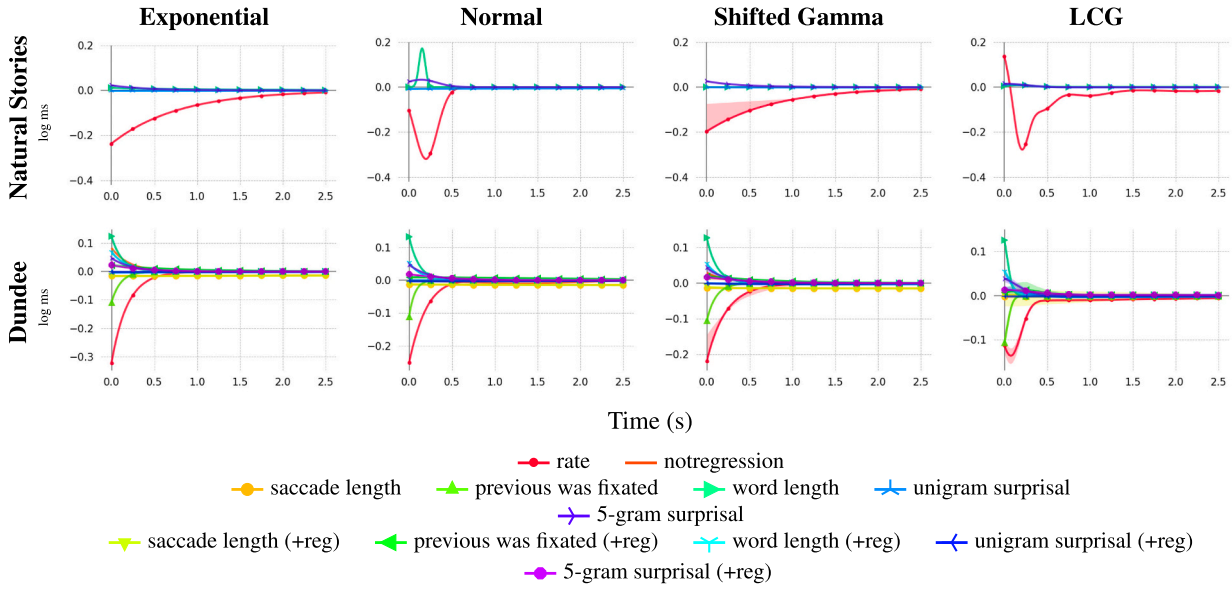


Fig. 7. BBVI-estimated IRFs by kernel for Natural Stories (**top**) and Dundee (**bottom**).

Table 3
Natural Stories. CDR vs. baselines, mean-squared error. CDR results shown using (E)xponential, (N)ormal, Shifted (G)amma, and non-parametric LCG response kernels, fitted using MLE, BBVI improper, and BBVI. Best-performing models within the sets of baseline and CDR models are shown in *italics*. Best-performing overall models are shown in **bold**. Daggers (†) indicate convergence failures.

	Natural Stories (ms)			Natural Stories (log-ms)		
Model	Train	Expl	Test	Train	Expl	Test
LME	20,179	20,624	20,369	0.0803	0.0818	0.0815
LME-S	19,980†	20,471†	20,230†	0.0789†	0.0807†	0.0803†
GAM	20,070	20,501	20,255	0.0798	0.0814	0.0810
GAM-S	*19,873*	*20,349*	*20,109*	*0.0784*	*0.0802*	*0.0799*
CDR-E-MLE	17,766	18,172	–	0.0630	0.0643	–
CDR-E-BBVI.imp	17,765	18,168	–	0.0630	0.0643	–
CDR-E-BBVI	18,106	18,361	–	0.0644	0.0651	–
CDR-N-MLE	17,487	18,060	–	0.0622	0.0641	–
CDR-N-BBVI.imp	17,500	18,058	–	0.0622	0.0640	–
CDR-N-BBVI	17,686	18,182	–	0.0630	0.0645	–
CDR-G-MLE	17,510	18,053	–	0.0620	0.0656	–
CDR-G-BBVI.imp	17,551	18,045	–	0.0623	0.0643	–
CDR-G-BBVI	18,118	18,373	**18,212**	0.0646	0.0652	**0.0654**
CDR-LCG-MLE	16,222	18,785	–	0.0569	0.0657	–
CDR-LCG-BBVI.imp	***16,153***	18,770	–	***0.0567***	0.0659	–
CDR-LCG-BBVI	17,437	***17,805***	–	0.0613	***0.0627***	–

7.2.3. Model comparison

To establish a standard of comparison for evaluating predictive performance, we also fit baseline LME and GAM models. Because the purpose of CDR is scientific modeling rather than engineering, the primary results of interest are the IRFs themselves and the insights they provide into human sentence processing. Therefore, the baseline models are used to construct a standard of reliability in predictive performance for each dataset, and comparison to them is intended to validate the CDR estimates. If CDR performs comparably to or better than baseline models in terms of generalization error, then this serves as evidence that the detailed estimates of temporal dynamics provided by CDR reliably characterize the response variable of interest. Both baseline types (LME and GAM) are fitted with and without three preceding spillover positions for each predictor (baselines with spillover are designated throughout this paper with the suffix *-S*), since a fourth-order FIR filter (spillover 0 through 3) is among the longest filters attested in previous naturalistic reading experiments (e.g. Smith & Levy, 2013).

We compare to GAM models because of their established usage in psycholinguistic data analysis. However, we acknowledge that CDR and GAM are designed to address different limitations of linear models. CDR addresses the possible existence of continuous temporal diffusion of effects in non-uniform time series, while GAM does not. GAM addresses the possible existence of arbitrary smooth non-linear functional relationships between predictors and response, while CDR (as currently defined) does not. The relative performance of CDR vs. GAM may therefore vary by dataset according to the relative importance of temporal diffusion vs. non-linear effects in describing the underlying response function. Extension of CDR to directly estimate non-linear response functions is left to future work, though see §2 for elaboration on a proposal to combine CDR and GAM models in a two-step regression framework.

In summary, the principal advantage of CDR for scientific modeling is the fact that it produces high-resolution estimates of temporal diffusion that cannot be obtained using established techniques like LME or GAM. The fact that CDR additionally outperforms those other techniques in terms of overall generalization error (see Table 5 below) primarily supports the reliability of the model's estimates.

Table 4
Dundee (scan path). CDR vs. baselines, mean-squared error. CDR results shown using (E)xponential, (N)ormal, Shifted (G)amma, and non-parametric LCG response kernels, fitted using MLE, BBVI improper, and BBVI. Best-performing models within the sets of baseline and CDR models are shown in *italics*. Best-performing overall models are shown in **bold**. Daggers (†) indicate convergence failures. LME-S performance metrics could not be obtained for scan paths because of long runtimes required for training.

	Dundee (ms)			Dundee (log-ms)		
Model	Train	Expl	Test	Train	Expl	Test
LME	14,645†	15,215†	15,230†	0.1790†	0.1807†	0.1801†
LME-S	–	–	–	–	–	–
GAM	14,476	15,055	15,064	0.1779	0.1795	0.1791
GAM-S	*14,340*	***14,942***	***14,973***	*0.1757*	***0.1776***	***0.1773***
CDR-E-MLE	14,510	*15,055*	–	0.1769	0.1784	–
CDR-E-BBVI.imp	14,508	*15,055*	–	0.1769	0.1784	–
CDR-E-BBVI	14,573	15,107	–	0.1775	0.1788	–
CDR-N-MLE	14,493	*15,055*	–	0.1765	0.1783	–
CDR-N-BBVI.imp	14,501	15,069	–	0.1764	*0.1782*	–
CDR-N-BBVI	14,545	15,081	–	0.1775	0.1788	–
CDR-G-MLE	14,501	15,063	–	0.1767	0.1785	–
CDR-G-BBVI.imp	14,508	15,058	–	0.1766	0.1784	–
CDR-G-BBVI	14,574	15,104	15,171	0.1776	0.1789	0.1789
CDR-LCG-MLE	14,109	15,204	–	***0.1711***	0.1810	–
CDR-LCG-BBVI.imp	***14,083***	15,283	–	***0.1711***	0.1806	–
CDR-LCG-BBVI	18,295	18,675	–	0.1779	0.1792	–

7.3. Results

Fig. 7 shows IRF estimates for Natural Stories (top) and Dundee (bottom). For brevity, we only present the BBVI estimates using untransformed durations, and we focus our discussion of Dundee on non-regressive estimates unless stated otherwise. The full set of estimates using all inference types and error transforms is given in Appendix C.2. In general, they are similar to the estimates presented here. *Rate* obtains a large-magnitude, negative, and slowly decaying IRF across datasets and kernel families. This is consistent with the existence of an *inertia* effect, such that quicker reading in the recent past tends to engender quicker reading at the current word as well. This effect is especially pronounced in Natural Stories, where the *rate* estimate is many times larger in magnitude than that of any other predictor, suggesting that self-paced reading may be particularly susceptible to influences from inertia (i.e. habituation to repeated button presses). Nevertheless, other predictors are also estimated to influence reading latencies over and above *rate*. *Word length* and *5-gram surprisal* are given positive estimates, consistent with prior expectations of processing costs associated with each of these variables. IRFs in Dundee decay more quickly than those in Natural Stories, suggesting a less pronounced influence of temporal diffusion in the eye-tracking modality compared to the self-paced reading modality. This suggestion is further supported by a much stronger improvement in predictive performance from using CDR for Natural Stories than Dundee (see Tables 3 and 4). Within the Dundee estimates, IRFs for *word length* generally decay more quickly than those for *5-gram surprisal* or *unigram surprisal* (in models like LCG that find a non-zero *unigram surprisal* effect), which is consistent with the hypothesis that higher-level predictive coding and/or lexical retrieval processes entail more computation and therefore engender a slower response than a low-level perceptual variable like *word length* (Shain & Schuler, 2018). This pattern does not seem to obtain in Natural Stories, suggesting an influence of modality on comprehension patterns. Estimates for *unigram surprisal* in both corpora are mostly negative or null, consistent with results reported in Shain (2019). Dundee models also generally find a large-magnitude, negative, rapidly decaying IRF for *previous was fixated*, suggesting a strong but brief facilitation effect for single-word saccades, perhaps due to parafovial processing from the preceding word. Dundee results also indicate that there are indeed substantially different estimates for variables depending the [±regression] dimension: all effects are attenuated under a regressive eye-movement.

As in the synthetic experiments reported in §6.3, the LCG estimates show wiggly dynamics but generally recapitulate the overall shape trends that emerge using parametric kernels. LCG models also generally do not achieve much better generalization error than parametric models, although improvements to training fit are sometimes quite large (Tables 3 and 4). These findings suggest that the simpler parameteric kernels (1) are not too constraining to find near-optimal response shapes and (2) are less prone to overfitting than the pseudo-non-parametric kernels.

The models generally find exponential-like kernel shapes across datasets and kernel families. This suggests that the effects of these variables on reading behavior are mostly monotonic — the properties of a word exert the biggest influence on fixations to that word, with diminishing influence on fixations to subsequent words. While this has often been assumed in reading research (e.g. Rayner, 1998), here it is an emergent finding. Two key exceptions occur in the Natural Stories estimates, where *word length* is given a large late-peaking estimate under the *normal* kernel and the *rate* response is initially strongly positive and quickly dips strongly negative under the *LCG* kernel. The late-peaking *word length* possibly merits further investigation, although several considerations cast suspicion on it. First, *word length* is generally considered to be a low-level effect of visual word recognition and is thus not expected to have a late effect *a priori*. Second, many implementationally similar CDR models do not recover the same response profile for word length (see Appendix C.2). Third, the model that found this particular effect shape for *word length* does not achieve systematically better generalization performance over those that did not (Table 3). That particular response component is thus plausibly an artifact, although follow-up study may be warranted. The initial positivity in the LCG estimate for *rate* is likely explained by two facts: (1) the *rate* estimate at time 0 is confounded with the model intercept, and (2) fixations shorter than 100 ms (approximately the length of the positivity) are filtered out by standard preprocessing in Natural Stories. There is thus little training signal for the IRF shape over the interval 0–100 ms, and the more flexible LCG kernel is capable of finding spurious estimates there.

In summary, the IRFs recovered by CDR accord with prior expectations about the reading response and are largely consistent across modalities and kernel definitions, while shedding additional light on fine-grained details of temporal structure.

To validate the CDR estimates, comparisons to baseline LME and GAM models are presented in Tables 3 and 4. LME-S performance metrics could not be obtained for Dundee scan paths because the fitting time exceeded the two week runtime allocation provided by our compute resource. CDR models outperform (i.e. achieve better exploratory and — where relevant — test set error than) all baselines on

Table 5
Reading data model comparison (Dundee scan path). Permutation tests of improvement on test set from CDR-G-BBVI over baselines (pooled across all tasks), along with binomial tests of the probability of CDR models improving over each baseline, aggregating over training and exploratory sets (those on which all CDR models are evaluated). **Note**: Dundee scan path durations are excluded from the LME-S comparison because runtime limits prevented training from terminating.

	Permutation test	Binomial test	
Baseline	*p*	Success rate	*p*
LME	**1.0e-4*****	0.96	**2.2e-16*****
LME-S	**1.0e-4*****	1.0	**2.2e-16*****
GAM	**1.0e-4*****	0.74	**1.4e-6*****
GAM-S	**1.0e-4*****	0.54	0.24

Natural Stories, and they generally outperform the LME-based models on Dundee.[32] GAM without spillover mostly outperforms CDR on Dundee without log transformation and mostly underperforms CDR with log transformation, although the errors are similar in magnitude throughout. GAM with spillover systematically outperforms CDR in Dundee, though again the errors are relatively close, especially under log transformation. The limited performance of CDR on Dundee relative to GAM with spillover is likely due to some combination of (1) the relatively constrained degree of temporal diffusion in Dundee, as revealed by the rapidly decaying response estimates in the bottom panels of Fig. 7,[33] (2) the requirement of CDR to generate responses through linear combination of the convolved predictors, while GAM estimates non-linear relationships between predictors and response, and (3) the greater concision of CDR models, which contain substantially fewer parameters than GAM models with spillover (GAM fits multidimensional smooth functions for each spillover position of each predictor, in addition to the random effects model). Nonetheless, the ensemble of exploratory set comparisons support the reliability of CDR estimates, since they yield similar or improved generalization performance across the board.

We statistically evaluate CDR against the baseline models in two ways: (1) by comparing test-set performance against each baseline and (2) by comparing overall *success rates* of CDR models against each baseline on the training and exploratory sets. For (1), to avoid multiple comparisons, we select only CDR-G-BBVI as the representative CDR model across reading datasets and compare its generalization performance on the test set to that of all baseline models.[34] We use a paired permutation test (Demšar, 2006) with 10,000 resampling iterations, pooling error vectors from all tasks (both linear and log responses from both Natural Stories and Dundee) into a single comparison.[35] For (2), we test the empirical probability of a CDR model outperforming each baseline model on the training or exploratory set[36] (a *success*) against a null hypothesis of chance probability (0.5), using a binomial test. This test assesses the general robustness of CDR compared to the baselines, aggregating over CDR hyperparameters by testing whether the probability of improvement from using an arbitrarily chosen CDR model type differs from chance. Results are given in Table 5. As shown, CDR-G-BBVI significantly outperforms all baselines in terms of test-set error, indicating that CDR models achieve better overall out-of-sample error than any of the baselines. In addition, the CDR success rate over all baselines but GAM-S is significantly greater than chance, indicating that an arbitrarily chosen CDR configuration is likely to outperform these baselines, even without careful tuning on an exploratory set.[37]

7.4. Discussion

Our application of CDR to the study of human reading latencies shows consistent estimates across response kernels that largely align with prior expectations but additionally provide high-resolution insights into underlying temporal dynamics that are difficult to obtain using standard statistical models. Results additionally show a significant overall improvement in generalization error from CDR over baseline models, supporting the trustworthiness of estimated response functions.

8. Experiment B: Naturalistic fMRI

We now use CDR to infer the shape of the hemodynamic response function (HRF) from fMRI measures of brain responses to variably-spaced naturalistic stimuli. There is already an extensive literature on HRF discovery using discrete-time deconvolutional methods (Josephs, Turner, & Friston, 1997; Friston et al., 1998; Miezin et al., 2000; Gitelman, Penny, Ashburner, & Friston, 2003; Lindquist et al., 2009; Pedregosa, Eickenberg, Ciuciu, Gramfort, & Thirion, 2014, inter alia). These approaches rely on stimulus designs in which events are regularly spaced and aligned with the fMRI scan times. For fMRI researchers seeking the benefits to ecological validity afforded by the naturalistic experimental paradigm (Campbell & Tyler, 2018; Hasson, Egidi, Marelli, & Willems, 2018; Hasson & Honey, 2012; Hasson, Malach, & Heeger, 2010), this requirement poses a problem, since many naturalistic stimuli (including language) do not consist of regularly spaced events occurring at integer multiples of the fMRI scanner's acquisition rate. Naturalistic fMRI experiments are therefore an important target application of continuous-time deconvolution, since CDR imposes no such requirement on the stimulus design. Here we evaluate CDR models trained on a naturalistic fMRI dataset and compare their performance to that of existing methods for modeling fMRI measures of neural responses to naturalistic language stimuli.

8.1. Data

We use the same fMRI dataset as Shain, Blank, van Schijndel,

[32] Although we were unable to achieve convergence in the more expressive LME-S baseline, comparisons using first pass and go-past durations in Dundee (Appendix C.2) show that it systematically underperforms CDR.

[33] This outcome nevertheless highlights an advantage of CDR: it can discover the extent of temporal diffusion empirically, which can be of scientific interest even in datasets where the latent processes are not very diffuse and do no afford large gains in fit from CDR modeling.

[34] The choice of BBVI for this analysis was motivated in §5.3. Of the BBVI models, we focus on *shifted gamma* because it is the most flexible of the parametric kernels and therefore likely to be of interest to future applications of CDR to reading times. Differences in generalization performance between kernels tend to be small.

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

[36] I.e., those sets for which performance data are available for all models.

[37] As shown in Appendix C, the reported pattern of permutation testing results is unchanged whether first pass, go-past, or scan path durations in Dundee are used, or indeed whether all Dundee durations are considered simultaneously, despite the fact that this greatly overrepresents Dundee, where the relative performance of CDR is considerably worse than in Natural Stories. Binomial tests using the other duration types in Dundee (first pass and go-past) are significant even over GAM-S, though the success rates are lower than against the other baselines. This again suggests that the relative importance of controlling for temporal diffusion (CDR) vs. non-linear effects (GAM) may be less great in the Dundee corpus, where diffusion appears to be relatively constrained. Nonetheless, a major advantage of the CDR approach is its ability to estimate the extent of diffusion in general, and thus to reveal circumstances in which diffusion plays a more or less pronounced role.

Schuler, and Fedorenko (2020).[38] Data were collected from 78 participants (30 males) exposed to auditory presentation of texts from the Natural Stories corpus (Futrell et al., 2021) read by one of two speakers (1 male, 1 female). Left-hemisphere fronto-temporal language regions are functionally localized on a participant-specific basis using a separate localizer task (Fedorenko, Hsieh, Nieto-Castañón, Whitfield-Gabrieli, & Kanwisher, 2010; Braze et al., 2011; Vagharchakian, Dehaene-Lambertz, Pallier, & Dehaene, 2012; Blank, Balewski, Mahowald, & Fedorenko, 2016; Scott, Gallée, & Fedorenko, 2017, inter alia). The response variable consists of average blood oxygen level dependent (BOLD) contrast imaging signal within the voxels of six functionally defined regions of interest (fROIs) constituting the left-hemisphere fronto-temporal language network: inferior frontal gyrus (IFG) and its orbital part (IFGorb), middle frontal gyrus (MFG), anterior temporal cortex (ATL), posterior temporal cortex (PTL), and angular gyrus (AngG). For full details of the fMRI data acquisition, preprocessing, and functional localization methods, see Shain, Blank et al (2020).

Similarly to Experiment A, training (50%), exploratory (25%), and test (25%) sets are created using modular arithmetic. Following Shain, Blank et al (2020), we use a slower partitioning cycle (15 TRs or 30s) compared to Experiment A, which is motivated by a desire to reduce correlation between the elements of the partition in light of strong prior evidence that the BOLD signal is highly auto-correlated. In particular, we cycle TR numbers e into different bins of the partition with a different phase for each subject u: $\text{partition}(e,u) = \frac{\lfloor e+u \rfloor}{15} mod\ 4$, again assigning outputs 0 and 1 to the training set, 2 to the exploratory set, and 3 to the evaluation (test) set.

8.2. Predictor and response variables

Following Shain, Blank et al (2020), the dependent variable is average BOLD response within each fROI, with all fROIs combined into a single model. Unlike reading latencies, BOLD measurements are not strictly positive and generally not heavily skewed. Therefore, in contrast to the reading experiments above, we do not explore any normalizing (e.g. logarithmic) transformations on the fMRI response. In addition to *rate*, *unigram surprisal*, and *5-gram surprisal*, all implemented identically to Experiment A, we also include a predictor for *sound power* (e.g. Brennan et al., 2016), estimated as frame-by-frame root mean squared energy (RMSE) of the audio stimuli computed using the Librosa software library (McFee et al., 2015). Because *sound power* is a continuous rather than event-based predictor, we implement it by taking regular RMSE samples every 100 ms. *Sound power* thus uses RMSE sample times as timestamps rather than word onsets, and CDR's event-based deconvolutional procedure thus implicitly uses a Riemann sum approximation of the continuous *sound power* convolution integral. To implement the assumption of a fixed-shape hemodynamic response in a given cortical region, we tie the parameters of the IRF kernel across all predictors within each region, while giving each predictor its own coefficient in order to estimate different response amplitudes. We also add a linear predictor for repetition time number (*TR number*, the sample's index within the current story), designed to capture any linear trends in the overall response. Models contain by-fROI random intercepts, slopes, and HRF parameters for each of these predictors, along with by-subject random intercepts.[39]

8.3. Hemodynamic response kernels

We consider IRF kernels based on the double-gamma hemodynamic response function (Boynton et al., 1996):

[38] Data are available at https://osf.io/eyp8q/.

[39] As discussed in Shain, Blank et al (2020), this dataset does not appear to support identification of richer by-subject random effects models, which generalize poorly to the out-of-sample sets.

Table 6
Number of parameters by kernel in the fMRI experiment. Note that BBVI and BBVI-improper double these figures by additionally fitting variances for each parameter in the variational posterior, and that sinh-arcsinh models additionally include parameters for the skewness and tailweight of the response.

Kernel	Parameters
HRF1	128
HRF2	135
HRF3	142
HRF4	149
HRF5	156
LCG	331

$$f(x;\alpha_1,\beta_1,c,\alpha_2,\beta_2) = \frac{\beta_1^{\alpha_1} x^{\alpha_1-1} e^{-\beta_1 x}}{\Gamma(\alpha_1)} - c\frac{\beta_2^{\alpha_2} x^{\alpha_2-1} e^{-\beta_2 x}}{\Gamma(\alpha_2)} \tag{33}$$

The normalization constant for this kernel is simply $\frac{1}{1-c}$, since it consists of a sum of two scaled gamma probability densities whose integrals over the positive reals are 1 and $-c$, respectively. We therefore define a 5-parameter HRF5 kernel as:

$$\text{HRF5}(x;\alpha_1,\beta_1,c,\alpha_2,\beta_2) \stackrel{\text{def}}{=} f(x;\alpha_1,\beta_1,c,\alpha_2,\beta_2) \ / \ (1-c) \tag{34}$$

Because the double-gamma HRF is fairly heavily parameterized (5 parameters), we explore the impact of reparameterizations that reduce the flexibility of the kernel through parameter tying, constraining the kernel toward the canonical HRF, where $\alpha_1 = 6$, $\beta_1 = 1$, $c = \frac{1}{6}$, $\alpha_2 = 16$, and $\beta_2 = 1$ (Lindquist et al., 2009). Thus, in addition to the 5-parameter kernel shown in Eq. (34) (HRF5), we also consider the following kernel variants:

- A 4-parameter variant (HRF4) with tied rate parameter β:

$$\text{HRF4}(x;\alpha_1,\beta,c,\alpha_2) \stackrel{\text{def}}{=} \left(\frac{\beta^{\alpha_1} x^{\alpha_1-1} e^{-\beta x}}{\Gamma(\alpha_1)} - c\frac{\beta^{\alpha_2} x^{\alpha_2-1} e^{-\beta x}}{\Gamma(\alpha_2)}\right) \Big/ \ (1-c) \tag{35}$$

- A 3-parameter variant (HRF3) which additionally ties the shape parameters α_1 and α_2 to have a constant offset of 10, as used by SPM's canonical HRF:

$$\text{HRF3}(x;\alpha,\beta,c) \stackrel{\text{def}}{=} \left(\frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} - c\frac{\beta^{\alpha+10} x^{\alpha+9} e^{-\beta x}}{\Gamma(\alpha+10)}\right) \Big/ \ (1-c) \tag{36}$$

- A 2-parameter variant (HRF2) which additionally fixes the undershoot constant c at SPM's default value of $\frac{1}{6}$:

$$\text{HRF2}(x;\alpha,\beta) \stackrel{\text{def}}{=} \left(\frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} - \frac{\beta^{\alpha+10} x^{\alpha+9} e^{-\beta x}}{6\Gamma(\alpha+10)}\right) \Big/ \ \left(\frac{5}{6}\right) \tag{37}$$

- A 1-parameter variant which fixes the shape parameter α at SPM's default value of 6, implementing a "stretchable" canonical HRF:

$$\text{HRF1}(x;\beta) \stackrel{\text{def}}{=} \left(\frac{\beta^{6} x^{5} e^{-\beta x}}{\Gamma(6)} - \frac{\beta^{16} x^{15} e^{-\beta x}}{6\Gamma(16)}\right) \Big/ \ \left(\frac{5}{6}\right) \tag{38}$$

In all of these kernels, we initialize the parameters at the SPM defaults for the canonical HRF presented above.

As in earlier experiments, we also consider LCG kernels. As discussed above, the LCG models are more heavily parameterized than those with parametric kernels (Table 6, see Table 2 for reading models).

8.4. Model comparison

To validate the CDR estimates of the HRF, we compare CDR fits to those produced by four existing approaches for modeling naturalistic fMRI experiments. First, we pre-convolve the stimuli using the canonical

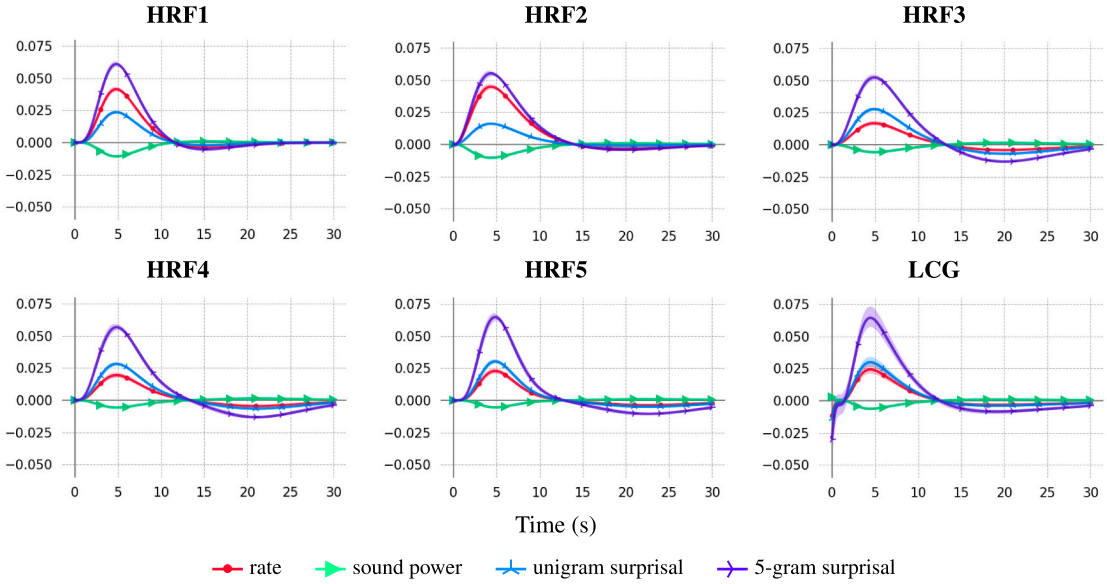


Fig. 8. Natural Stories: BBVI-estimated IRFs by kernel.

HRF (Canonical HRF), as is done in many naturalistic studies (Brennan et al., 2012; Willems, Frank, Nijhof, Hagoort, & den Bosch, 2015; Henderson, Choi, Luke, & Desai, 2015; Henderson, Choi, Lowder, & Ferreira, 2016; Lopopolo, Frank, den Bosch, & Willems, 2017, inter alia). This approach has advantages for parsimony, since it avoids the need to fit coefficients at multiple time offsets. But it assumes a fixed, universal hemodynamic response that may not accurately describe the response profile in a given brain region (Handwerker, Ollinger, & D'Esposito, 2004). Second, we use piecewise linear interpolation to resample the predictors at timepoints that align with the fMRI scan times (Interpolated). This approach distorts the predictor series by treating it as a sequence of samples from an underlyingly continuous signal (see §3 for discussion). Third, we use the fMRI scan times to define discrete temporal bins and then average the predictor values within each bin (Averaged). This approach has been used, for example, by Wehbe et al. (2021). Fourth, we follow Huth et al. (2016) in downsampling the predictor series to the temporal resolution of the fMRI signal by convolving it with a low-pass Lanczos filter with 3 lobes and a cutoff frequency of 0.25, the Nyquist frequency of the fMRI scanner (Lanczos).[40] This method essentially implements a "soft" variant of the averaging approach by taking a weighted sum of the stimuli in the neighborhood of an fMRI sample, weighted by a function (the Lanczos kernel) of the temporal distance between the stimulus and the sample. Mixed models with the same random effects structure as the CDR models described above were fitted using the lme4 package. For the Canonical HRF baseline, we applied a single fixed and by-fROI random slope per predictor, since the temporal modeling is implemented by the convolutional preprocess. For the Interpolated, Averaged, and Lanczos baselines, we applied fourth-order FIR models with fixed and by-fROI random slopes for the four TR's preceding an fMRI sample. These FIR kernels implement a discretized version of the HRF and estimate its (non-parametric) shape from data.

8.5. Results

Fig. 8 shows plots of the IRF estimates from each kernel. For brevity, we only present the BBVI estimates. The full set of estimates using all inference types is given in Appendix C.3. As in previous experiments, results are highly consistent across these dimensions.

Models estimate *5-gram surprisal* to have the largest influence on the language network's response, generally followed by *rate* and then by *unigram surprisal*. The *sound power* predictor tends to be assigned a small-magnitude negative response.

Fig. 8 shows that all models find estimates that closely resemble the canonical HRF, and that these estimates do not change dramatically in

Table 7
Natural Stories fMRI. CDR vs. baselines, mean-squared error. CDR results shown using 1-, 2-, 3-, 4-, and 5-parameter double-gamma hemodynamic response kernels, along with non-parametric LCG response kernels, fitted using MLE, BBVI improper, and BBVI. Best-performing models within the sets of baseline and CDR models are shown in *italics*. Best-performing overall models are shown in **bold**. Daggers (†) indicate convergence failures.

	Natural Stories fMRI		
Model	Train	Expl	Test
Canonical HRF	11.3548†	*11.8263*†	*11.5661*†
Interpolated	11.4236†	11.9888†	11.6654†
Averaged	*11.3478*†	11.9280†	11.6090†
Lanczos	11.3536†	11.9059†	11.5871†
CDR-HRF1-MLE	11.3442	11.8729	–
CDR-HRF1-BBVI.imp	11.3442	11.8732	–
CDR-HRF1-BBVI	11.3469	11.8600	–
CDR-HRF2-MLE	11.3365	11.8551	–
CDR-HRF2-BBVI.imp	11.3365	11.8550	–
CDR-HRF2-BBVI	11.3386	11.8410	–
CDR-HRF3-MLE	11.2810	11.7131	–
CDR-HRF3-BBVI.imp	11.2809	11.7126	–
CDR-HRF3-BBVI	11.2840	11.7058	–
CDR-HRF4-MLE	11.2758	11.7033	–
CDR-HRF4-BBVI.imp	11.2757	11.7034	–
CDR-HRF4-BBVI	11.2808	11.7002	–
CDR-HRF5-MLE	11.2730	11.6956	–
CDR-HRF5-BBVI.imp	11.2730	11.6956	–
CDR-HRF5-BBVI	11.2774	11.6928	**11.5369**
CDR-LCG-MLE	***11.2585***	***11.6819***	–
CDR-LCG-BBVI.imp	11.2607	11.6861	–
CDR-LCG-BBVI	11.2762	11.7023	–

[40] Source code for this technique is available at https://github.com/HuthLab/speechmodeltutorial.

Table 8
fMRI data model comparison. Permutation tests of improvement from CDR-G-BBVI over baselines, along with binomial tests of the probability of CDR models improving over each baseline, aggregating over training and exploratory sets (those on which all CDR models are evaluated).

	Permutation test	Binomial test	
Baseline	*p*	Success rate	*p*
Canonical HRF	**4.0e-4*****	0.83	**3.5e-5*****
Interpolated	**1.0e-4*****	1.0	**1.5e-11*****
Averaged	**1.0e-4*****	1.0	**1.5e-11*****
Lanczos	**1.0e-4*****	1.0	**1.5e-11*****

Table 9
Ablative testing results against null hypothesis of no effect for *5-gram surprisal*, using three different CDR-appropriate testing procedures. Mean and 95% credible intervals for the CDR effect estimate g' for *5-gram surprisal* are shown in the g' columns. Rejections of the null are shown in **bold**. Daggers (†) indicate convergence failures in one or both models. Plusses (+) indicate conceptual reproductions of tests in Shain (2019).

		g'			*p*-value		
Dataset	Response	Mean	2.5%	97.5%	Direct PT	2-Step LRT	2-Step PT
Natural Stories	ms	0.729	0.706	0.753	1.0	**2.2e-16*****†	1.0†
+Natural Stories	log-ms	0.011	0.011	0.011	**3.0e-4*****	**2.2e-16*****	1.0
Dundee (FP)	ms	3.52	3.32	3.71	**1.0e-4*****	**2.6e-14*****†	0.074†
Dundee (FP)	log-ms	0.012	0.011	0.013	**2.0e-4*****	**2.6e-14*****	**0.016***
fMRI	BOLD	0.180	0.175	0.184	**1.0e-4*****	**1.0e-9*****	**1.0e-4*****

the simplified reparameterizations of the HRF. At the same time, some models deviate from the canonical HRF in noteworthy ways. First, more parameterized models (HRF3+, which allow tuning of the undershoot amplitude *c*) tend to find a larger-magnitude undershoot component (the negative dip at the tail of the response kernel) than that of the canonical HRF. The ability to tune *c* and thus find deeper undershoots also corresponds to a striking improvement in both training and generalization performance (see the drop in error from HRF2 to HRF3 shown in Table 7). Second, the LCG model finds an early negative response consistent with prior evidence of an "initial dip" in the HRF (Yacoub et al., 2001; Röther et al., 2002; Hu & Yacoub, 2012, inter alia). Such a dip is outside the solution space of the parametric kernels used here and could only be discovered by LCG.

These findings have several implications. First, they reassuringly show that CDR models discover patterns that resemble the canonical HRF and are thus consistent with decades of prior research on the hemodynamic response, even using kernels with highly unconstrained solution spaces (LCG). Second they bear on prior concerns that the hemodynamic response is neither strictly stationary (Logothetis, 2003) nor strictly additive (Friston, Josephs, et al., 1998, Friston, Mechelli, Turner, & Price, 2000), possibly undermining the usefulness of fMRI measures from long-running naturalistic exposures where confounds from e.g. response saturation might be more pronounced (Lindquist et al., 2009). Our results are reassuring for naturalistic fMRI modeling since they directly support the hypothesis that the double-gamma shape continues to characterize the HRF not only in short constructed sensory experiments where it is typically studied, but also in long-running naturalistic experiments using indicators of high-level cognitive processes like language comprehension. Third, the deeper undershoot components and better fits obtained using more heavily parameterized models suggest that the canonical HRF may underestimate the size of the undershoot in the functional language network during naturalistic sentence comprehension, although further experiments would be needed to bear this out more convincingly.

Table 7 compares the performance of CDR against that of the baselines described in §8.4. As shown, among the baseline models, pre-convolution with the canonical HRF performs quite well in terms of both in-sample and out-of-sample error (achieving the best generalization performance of any baseline model), despite its reduced number of parameters and its inability to adapt the HRF to the data. However, CDR generally outperforms all baselines on all datasets. This indicates that CDR constitutes a substantial improvement over existing methods for modeling fMRI data in naturalistic experiments.

We select CDR-HRF5-BBVI as the representative CDR model for the fMRI dataset because it is the most flexible and best-performing of the BBVI parametric models explored here. We compare CDR-HRF5-BBVI generalization performance on the test set to that of all baseline models.[41] As shown in Table 7, CDR-HRF5-BBVI significantly outperforms all baselines on the test set, and as shown in the *Permutation test* column of Table 8. As in Experiment A, we also compare the combined training and development set results from all CDR models against each baseline, using a binomial test of the success rate (rate at which any CDR model outperforms a baseline model). As shown in Table 8, the CDR success rate over each baseline is significantly greater than chance, indicating that CDR models generally achieve better in-sample and out-of-sample error than any of the baselines, even without careful tuning on an exploratory set.

9. Hypothesis testing: Surprisal effects

This section compares null hypothesis significance testing (NHST) methods using CDR models on a familiar result from psycholinguistics: *surprisal* effects in human sentence processing, which have been argued to support the existence of a predictive coding component of the language comprehension architecture that generates expectations about upcoming words (Demberg & Keller, 2008; Frank & Bod, 2011; Smith & Levy, 2013). In particular, we are interested in statistical tests of CDR estimates of *5-gram surprisal* against the null hypothesis of no effect. One possible approach is a credible intervals test, checking whether Monte-Carlo-estimated 95% credible intervals of the effect size (§4.2) in CDR for *5-gram surprisal* include zero, which implements NHST at a 0.05 level of significance. This test rejects the null hypothesis for all comparisons, as shown in Table 9, where the upper and lower bounds on the 95% credible interval for the effect estimate g' are always positive. However, as argued in §4.6, such a test is anticonservative in a CDR setting because of non-convexity, since the credible interval estimates only consider the local neighborhood of the mode to which the model has converged. Furthermore, such single-model tests are viewed with increasing skepticism in psycholinguistics because they are influenced by multicollinearity, leading many researchers to favor ablative model

[41] The choice of BBVI for this analysis is motivated in §5.3.

comparisons against a baseline in which the fixed effect of interest is removed (Frank & Bod, 2011). Finally, in-sample tests such as a credible intervals test or a likelihood ratio test evaluate on the training data and are therefore unable to directly diagnose overfitting. This limitation can be addressed by the use of non-parametric out-of-sample tests, such as the paired permutation test.

For the purposes of this hypothesis testing demonstration, we apply three testing paradigms (discussed in §4.6) against the null hypothesis of no effect for *5-gram surprisal* in each of the datasets explored above:

1. **Direct PT**: Ablative held-out paired permutation test (PT) of CDR models with and without a fixed effect for *5-gram surprisal*.[42]
2. **2-step LRT**: Ablative likelihood ratio test (LRT) of LME models with and without a fixed effect for *5-gram surprisal*, with models fitted to the training set using predictors convolved using the full CDR model.
3. **2-step PT**: Ablative held-out paired permutation test of LME models with and without a fixed effect for *5-gram surprisal*, with models fitted to the training set using predictors convolved using the full CDR model. Tests are based on out-of-sample mean squared error.

The key difference between the direct and 2-step approaches is that the 2-step tests use LME to estimate globally optimal intercepts and linear coefficients on the convolved data. The key difference between the LRT and PT approaches to 2-step testing is that PT is a non-parametric evaluation on out-of-sample data, while LRT is a parametric evaluation on in-sample data under asymptotic guarantees about the distribution of the likelihood ratio statistic (Wilks, 1938). In order to avoid evaluating multiple CDR models on the test set, tests use the same BBVI-estimated CDR models that were selected in the previously reported baseline comparisons (BBVI inference, *shifted gamma* kernels for reading data and *HRF5* kernels for fMRI data). LME models in 2-step LRT tests are fitted to data convolved using the full CDR model as a preprocess (see §4.6 for details). To facilitate convergence, we simplify the LME structure by using uncorrelated random intercepts and slopes (Bates et al., 2015). Since these analyses are primarily for demonstration purposes and we seek uniformity within testing procedures across datasets, we take no further steps to address LME convergence problems in 2-step tests, although in practice it is recommended to simplify LME models until convergence is obtained before using them in scientific tests (Barr, Levy, Scheepers, & Tily, 2013). Out-of-sample permutation tests use likelihood difference (direct PT) or mean squared error difference (2-step PT) as the test statistic. We join the exploratory and test sets in each corpus to create the PT evaluation set.

Results are shown in Table 9.[43] The p-values of 1.0 observed in some cells indicate that the ablated model outperformed the full model on the evaluation set. Despite the use of simpler LME models with uncorrelated random intercepts and slopes, convergence failures affect the 2-step results for the linear response (ms) models of Natural Stories and Dundee.

The 2-step LRT test is the least conservative, rejecting the null (and supporting the existence of surprisal effects) in all models. This is unsurprising both because the likelihood ratio test is maximally powerful (Neyman & Pearson, 1933) and because it is in-sample and therefore unable to directly account for external validity, unlike PT, which is based on generalization quality. Generalization-based tests like PT are arguably more likely to favor replicable findings than in-sample tests like LRT, since an effect that is significant by LRT on the training data but does not generalize to a different sample is of limited scientific interest. The direct PT test rejects the null for all models except Natural Stories (ms), where ablated models outperform the full model. We leave further exploration of this exception to future research. Nonetheless, direct PT results overall support the existence of surprisal effects on all three kinds of experimental measures considered here. The 2-step PT test appears to be the most conservative of the testing procedures evaluated here, only rejecting the null for two out of five comparisons. These results suggest that LME models fitted to CDR-convolved data generalize less well than the underlying CDR model itself. In light of this finding and the evidence from §4.6 that CDR-estimated coefficients are near globally optimal, the added complexity of the 2-step approach may be of limited value.

In sum, using multiple testing procedures, CDR models generally reveal evidence for surprisal effects across all datasets considered here. Although all three procedures described here are appropriate for testing scientific hypotheses, we recommend the direct test because of its simplicity and in light of the combined evidence that (1) CDR-estimated coefficients are near-optimal (§4.6), (2) LME models in 2-step tests are prone to convergence problems, and (3) LME estimates in 2-step PT tend to generalize less well than CDR estimates.

10. General recommendations

The foregoing results suggest certain empirically-motivated best practices for future CDR analyses of psycholinguistic time series. These best practices are used as defaults in the CDR implementation proposed here, although they can easily be overridden on a model-by-model basis as motivated by the experimental design.

10.1. Inference type: BBVI

Of the three inference types examined here (MLE, BBVI-improper, and BBVI), results show that BBVI tends to converge more quickly (§5.3). As shown in Appendix C, they also tend to yield more conservative estimates of uncertainty (§6). For these reasons, we suggest BBVI inference as a general default. MLE and BBVI-improper inference modes are still useful for sanity checking and sometimes obtain better error.

10.2. Kernel type: parametric

Results show a strong tendency for low-dimensional parametric response kernels to perform at least as well as high-dimensional LCG kernels in terms of synthetic IRF recovery and generalization error. Furthermore, depending on compute architecture, LCG kernels can be many times slower per iteration than parametric ones, in part because they contain many more parameters (Table 2). At the same time, over-constrained kernels can lead to high model bias (see e.g. *exponential* kernels fitted to *shifted gamma* responses in Simulation D). For this reason, we recommend the use of parametric kernels for research purposes whenever possible (i.e. whenever domain knowledge suggests a parametric kernel that covers the space of plausible solutions), although sanity checking the results against LCG estimates can be a useful step for datasets rich enough to support discovery of LCG models.

10.3. Convergence criterion: Time-loss correlation

We have proposed and applied a CDR convergence criterion based on statistical tests for non-decreasing loss. All parametric CDR models in this study met this criterion within a reasonable number of training iterations, suggesting that it is robust and scale-independent, as argued in

[42] Tests are based on out-of-sample likelihood rather than mean-squared error to enable consistent application to models with asymmetric error distributions explored in Appendices B and C.4, since the latter do not optimize mean squared error.

[43] For full hypothesis testing results, including all error distributions and duration definitions considered in this study, see Appendix C.4.

§5.2.[44] The use of a reliable automatic stopping criterion reduces the number of experimenter degrees of freedom by eliminating the need for researchers to decide when model training has completed.

11. Conclusion

This article motivated, defined, implemented, and evaluated a framework for continuous-time deconvolutional regression (CDR) of arbitrary time series data, building on a proposal from Shain and Schuler (2018). Synthetic evaluations explored the influence of several plausible confounds, showing that CDR recovers the data-generating model under various kinds of noise, temporal structure in the predictor and response signals, multicollinearity, and impulse response function (IRF) misspecification, using multiple forms of statistical inference (maximum likelihood, black-box variational Bayes, and black-box variational Bayes with improper uniform priors). Real-world evaluations on reading latencies and fMRI measures from human subjects explored the influence of different IRF kernel types, inference types, and error transforms, showing highly consistent IRF estimates across these dimensions. Statistical comparisons to standard baselines (linear mixed-effects and generalized additive models, with and without finite impulse response filters) on human subjects data showed that CDR improves generalization performance in multiple domains (self-paced reading, eye-tracking during reading, and fMRI). This article also proposed multiple procedures for testing scientific hypotheses in a CDR framework and evaluated their application to tests of surprisal effects in human sentence processing. These findings support the use of CDR for analyzing many classes of time series data, since it provides fine-grained estimates of temporal structure and directly controls for diffusion of effects, especially in settings (e.g. naturalistic language processing) where discrete-time deconvolutional methods are difficult to apply.

Acknowledgments

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. The authors would additionally like to thank Evelina Fedorenko for generously allowing access to the fMRI data discussed in this article.

Source code for the CDR software is available at https://github.com/coryshain/cdr. For reading data access, see the citations in §7. Data preprocessing was performed using the ModelBlocks repository (van Schijndel & Schuler, 2013): https://github.com/modelblocks/modelblocks-release.

Appendix A. CDR (mixed effects) model

Table A1
Summary of variables in CDR model definition.

	Name	Type	Description
Dimensions	X	$\mathbb{N}$	Number of predictor observations
	Y	$\mathbb{N}$	Number of response observations
	Z	$\mathbb{N}$	Number of random grouping factor levels
	K	$\mathbb{N}$	Number of predictors
	R	$\mathbb{N}$	Number of impulse response parameters
	J	$\mathbb{N}$	Number of unique time series
Data	$\mathbf{X}$	$\mathbb{R}^{X\times K}$	X predictor observations
	$\mathbf{y}$	$\mathbb{R}^{Y}$	Y response observations
	$\mathbf{Z}$	$\{0, 1\}^{Y\times Z}$	Random effects indicator
	$\mathbf{t}$	$\mathbb{R}^{X}$	Timestamps of observations in $\mathbf{X}$
	$\mathbf{t}'$	$\mathbb{R}^{Y}$	Timestamps of observations in $\mathbf{y}$
	$\mathbf{s}$	$\{1, 2, \ldots, J\}^{X}$	Time series IDs of observations in $\mathbf{X}$
	$\mathbf{s}'$	$\{1, 2, \ldots, J\}^{Y}$	Time series IDs of observations in $\mathbf{y}$
Parameters	μ	$\mathbb{R}$	Fixed intercept
	$\mathbf{m}$	$\mathbb{R}^{Z}$	Random intercepts
	$\mathbf{u}$	$\mathbb{R}^{K}$	Fixed coefficients
	$\mathbf{U}$	$\mathbb{R}^{Z\times K}$	Random coefficients
	$\mathbf{v}_k$	$\mathbb{R}^{R}$	Fixed IRF parameters for the k^{th} predictor
	$\mathbf{V}_k$	$\mathbb{R}^{Z\times R}$	Random IRF parameters for the k^{th} predictor
	σ^2	$\mathbb{R}_+$	Variance
Model	$g_k(t;\theta)$	$\mathbb{R}_+ \rightarrow \mathbb{R}$	IRF kernel for the k^{th} predictor, function of time t given parameters θ
	$\mathbf{m}'$	$\mathbb{R}^{Y}$	Fixed + random intercepts
	$\mathbf{U}'$	$\mathbb{R}^{Y\times K}$	Fixed + random coefficients
	$\mathbf{V}_k'$	$\mathbb{R}^{Y\times R}$	Fixed + random IRF parameters for predictor k
	$\mathbf{F}$	$\{0, 1\}^{Y\times X}$	Convolution mask
	$\mathbf{G}_k$	$\mathbb{R}^{Y\times X}$	Convolution matrix for the k^{th} predictor
	$\mathbf{X}'$	$\mathbb{R}^{Y\times X}$	Convolved design matrix

A.1. Mixed effects definition

In addition to the variable definitions assumed in §4.1, the mixed-effects CDR assumes the following quantities:

- $Z \in \mathbb{N}$: Number of random grouping factor levels[45]

[44] While the criterion is robust to the scale of the loss, it is sensitive to low-level training parameters like the learning rate, optimizer, and batch size. For example, the 500-iteration window used for convergence diagnosis in these experiments may lead to unnecessarily long training times at smaller batch sizes or learning rates. Users who manipulate these optimization settings should also revisit the convergence parameters in order to ensure that they are still appropriate.

[45] The sum total of all levels of each random grouping factor in the model, e.g. the number of subjects plus the number of items.

- $\mathbf{Z} \in \{0, 1\}^{Y \times Z}$: Boolean matrix indicating random grouping factor levels associated with each response observation

Following e.g. `lme4` (Bates et al., 2015), we use *random grouping factor* to refer to variables that capture categorical random variation in a model (e.g. *participant* or *item*) and *random grouping factor level* to refer to individual values of a random grouping factor (e.g. the value `participant A` of the random grouping factor *participant*).

In addition to the fixed parameters defined in §4.1, mixed-effects CDR seeks to estimate the following quantities:

- a vector $\mathbf{m} \in \mathbb{R}^Z$ of Z random intercepts
- a matrix $\mathbf{U} \in \mathbb{R}^{Z \times K}$ of ZK random coefficients, i.e. random estimates for each of K predictors for each of Z random effects levels
- K matrices $\mathbf{V}_k \in \mathbb{R}^{Z \times R}$ of ZR random IRF kernel parameters, i.e. random estimates for each of K predictors for each of R IRF parameters for each of Z random effects levels

Random parameters $\mathbf{m}$, $\mathbf{U}$, and $\mathbf{V}_k$ are constrained to be zero-centered within each random grouping factor.[46]

A fixed-effects CDR model therefore contains $2 + K + KR$ parameters: one intercept, K coefficients (one for each predictor), KR IRF parameters (R parameters for each predictor), and one variance of the response. Mixed-effects CDR models can also include random variation in the intercept, coefficients, and/or IRF parameters. This yields at most $1 + (Z + 1)(1 + K + KR)$ estimates for a mixed effects model with Z total random grouping factor levels (for example, the number of subjects plus items). Sub-maximal numbers of estimates can arise by forcing random effects components to zero. For example, zeroing out $\mathbf{v}_k$ and $\mathbf{V}_k$ eliminates random coefficients and IRF parameters for the k th predictor.

To support mixed modeling, the fixed and random effects must first be combined by adding fixed effects with their random offsets using the indicator matrix $\mathbf{Z}$, resulting in intercept vector $\mathbf{m}' \in \mathbb{R}^Y$, coefficient matrix $\mathbf{U}' \in \mathbb{R}^{Y \times K}$ and IRF parameter matrices $\mathbf{V}_k' \in \mathbb{R}^{Y \times R}$ for $k \in \{1, 2, \ldots, K\}$:

$$\mathbf{m}' \stackrel{\text{def}}{=} \mu + \mathbf{Z}\mathbf{m} \tag{39}$$

$$\mathbf{U}' \stackrel{\text{def}}{=} \mathbf{1}\mathbf{u}^\top + \mathbf{Z}\mathbf{U} \tag{40}$$

$$\mathbf{V}_k' \stackrel{\text{def}}{=} \mathbf{1}\mathbf{v}_k^\top + \mathbf{Z}\mathbf{V}_k \tag{41}$$

In a mixed model, we must redefine the convolution procedure to use the summed fixed+random effects above:

$$\mathbf{G}_k \stackrel{\text{def}}{=} g_k\left(\mathbf{t}'\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top; \mathbf{V}_k'\right) \odot \mathbf{F} \tag{42}$$

$$\mathbf{X}'_{[*,k]} \stackrel{\text{def}}{=} \mathbf{G}_k \mathbf{X}_{[*,k]} \tag{43}$$

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{m}' + (\mathbf{X}' \odot \mathbf{U}')\mathbf{1}, \sigma^2\right) \tag{44}$$

A summary table of the variable definitions above is provided in Appendix Table A1, and a step-through example of the CDR equations is provided in Appendix A.3.

A.2. Effect estimates in (Mixed Effects) CDR

Similarly to the unscaled and scaled fixed effects estimates $\mathbf{g}$, $\mathbf{g}'$, unscaled and scaled random effect estimates $\mathbf{H}$, $\mathbf{H}' \in \mathbb{R}^{Z \times K}$ are computed by integrating the IRF, in this case for each level of each random grouping factor:

$$\mathbf{H}_{[z,k]} \stackrel{\text{def}}{=} \int_0^\infty g_k\left(t; \mathbf{V}'_{k[z,*]}\right) dt \tag{45}$$

$$\mathbf{H}' \stackrel{\text{def}}{=} \mathbf{H} \odot \mathbf{U} - \mathbf{1}\mathbf{g}'^\top \tag{46}$$

The fixed effect estimate $\mathbf{g}'$ is subtracted out to ensure that $\mathbf{H}'$ denotes the deviation at each level from the fixed effect size.

In mixed-effects CDR models with random impulse response parameters, the IRF shape — and therefore the integral of the IRF — can vary between levels of the random grouping factor. As a result, zero-centering the random coefficients $\mathbf{U}$ within each grouping factor is insufficient to guarantee zero-centered random effect estimates. For example, in a 2-level mixed univariate CDR model with fixed effect sizes $\mathbf{g}' = [1 \quad 1]^\top$, random coefficients $\mathbf{U} = [-1 \quad 1]^\top$, and unscaled random effect estimates (IRF integrals) $\mathbf{H} = [1 \quad 10]^\top$, $\mathbf{U}$ has mean 0, but the random effect estimate vector $\mathbf{H}' = \mathbf{H} \odot \mathbf{U} - \mathbf{1}\mathbf{g}'^\top = [-2 \quad 9]^\top$ has mean 3.5, yielding a biased population level effect estimate.

To overcome this, we constrain the IRFs g_k to have a unit integral over the positive real line, as discussed in §4.2. Under this constraint, zero-centered coefficients are guaranteed to yield zero-centered effect estimates, and the population-level (fixed) effect estimates are unbiased.[47]

A.3. CDR model: A worked example

Consider a model containing two predictors p_1 and p_2 and two random effects levels s_1 and s_2. Assume the following 6 rows for each of $\mathbf{X}$ (predictors), $\mathbf{t}$ (predictor timestamps), and $\mathbf{s}$ (predictor series IDs):

[46] In practice, we also assume normally distributed random effects, either implicitly (via L_2 regularization) or explicitly (via variational priors and posteriors). See §5.3 for details.

[47] In models without random IRF parameters, the IRF integrals are identical across levels of the random grouping factor, and effect estimates are thus unbiased with or without normalization. Normalizing can still have numerical advantages since it factors effect size and shape, so we apply normalization to all models reported here, regardless of whether they contain random IRF parameters.

$$\mathbf{X} = \begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} p_1 & p_2 \end{matrix} \begin{bmatrix} 5 & 0 \\ -1 & 3 \\ 6 & 1 \\ 2 & 2 \\ 0 & 1 \\ -2 & -1 \end{bmatrix}, \mathbf{t} = \begin{bmatrix} 0 \\ 2 \\ 3 \\ 0 \\ 1 \\ 4 \end{bmatrix}, \mathbf{s} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{bmatrix}$$

Assume the following four rows for each of **Z** (random effects indicator), $\mathbf{t}'$ (response timestamps), and $\mathbf{s}'$ (response series IDs):

$$\mathbf{Z} = \begin{matrix} \\ \\ \\ \\ \end{matrix} \begin{matrix} s_1 & s_2 \end{matrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \mathbf{t}' = \begin{bmatrix} 1 \\ 4 \\ 2 \\ 3 \end{bmatrix}, \mathbf{s}' = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \end{bmatrix},$$

Assume a Gaussian IRF kernel with scalar location and scale parameters w_1, w_2:

$$g_1(x; w_1, w_2) = g_2(x; w_1, w_2) = e^{\frac{-(x-w_1)^2}{w_2}}$$

Assume the following model parameters μ, **m**, **u**, **U**, $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{V}_1$, $\mathbf{V}_2$, and σ_2. Note that random effects **m**, **U**, $\mathbf{V}_1$, and $\mathbf{V}_2$ are zero-centered:

$$\mu = 1.2, \mathbf{m} = \begin{matrix} s_1 \\ s_2 \end{matrix} \begin{bmatrix} -0.2 \\ 0.2 \end{bmatrix}, \mathbf{u} = \begin{matrix} p_1 \\ p_2 \end{matrix} \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix}, \mathbf{U} = \begin{matrix} s_1 \\ s_2 \end{matrix} \begin{matrix} p_1 & p_2 \end{matrix} \begin{bmatrix} -0.3 & 0.7 \\ 0.3 & -0.7 \end{bmatrix}$$

$$\mathbf{v}_1 = \begin{matrix} w_1 \\ w_2 \end{matrix} \begin{matrix} p_1 \end{matrix} \begin{bmatrix} 0.8 \\ 1.7 \end{bmatrix}, \mathbf{v}_2 = \begin{matrix} w_1 \\ w_2 \end{matrix} \begin{matrix} p_2 \end{matrix} \begin{bmatrix} 1.1 \\ 0.5 \end{bmatrix}, \mathbf{V}_1 = \begin{matrix} s_1 \\ s_2 \end{matrix} \begin{matrix} w_1 & w_2 \end{matrix} \begin{bmatrix} -0.2 & 0.1 \\ 0.2 & -0.1 \end{bmatrix}, \mathbf{V}_2 = \begin{matrix} s_1 \\ s_2 \end{matrix} \begin{matrix} w_1 & w_2 \end{matrix} \begin{bmatrix} 0.3 & 0.4 \\ -0.3 & -0.4 \end{bmatrix}$$

$$\sigma^2 = 1.3$$

We use the CDR equations to generate estimates for the four elements of **y** using the inputs and parameters. The vector $\mathbf{m}' = \mu + \mathbf{Z}\,\mathbf{m}$ contains intercepts for each element of **y** and is computed as follows:

$$\begin{aligned} \mathbf{m}' &= \overbrace{1.3}^{\mu} + \overbrace{\begin{matrix} s_1 & s_2 \end{matrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}}^{\mathbf{Z}} \overbrace{\begin{bmatrix} -0.2 \\ 0.2 \end{bmatrix}}^{\mathbf{m}} \\ &= \overbrace{1.3}^{\mu} + \overbrace{\begin{bmatrix} -0.2 \\ -0.2 \\ 0.2 \\ 0.2 \end{bmatrix}}^{\mathbf{Zm}} \\ &= \overbrace{\begin{bmatrix} 1.1 \\ 1.1 \\ 1.5 \\ 1.5 \end{bmatrix}}^{\mu+\mathbf{Zm}} \end{aligned}$$

$$\mathbf{m}' = \overbrace{1.3}^{\mu} + \overbrace{\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}}^{\mathbf{Z}} \overbrace{\begin{bmatrix} -0.2 \\ 0.2 \end{bmatrix}}^{\mathbf{m}}$$

$$= \overbrace{1.3}^{\mu} + \overbrace{\begin{bmatrix} -0.2 \\ -0.2 \\ 0.2 \\ 0.2 \end{bmatrix}}^{\mathbf{Zm}}$$

$$= \overbrace{\begin{bmatrix} 1.1 \\ 1.1 \\ 1.5 \\ 1.5 \end{bmatrix}}^{\mu+\mathbf{Zm}}$$

The matrix $\mathbf{U}' = \mathbf{1}\mathbf{u}^\top + \mathbf{Z}\,\mathbf{U}$ contains coefficients for both predictors for each element of $\mathbf{y}$ and is computed as follows:

$$\mathbf{V}_1 = \overbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}}^{\mathbf{1}} \overbrace{\begin{bmatrix} 0.8 & 1.7 \end{bmatrix}}^{\mathbf{v}_1^\top} + \overbrace{\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}}^{\mathbf{Z}} \overbrace{\begin{bmatrix} -0.2 & 0.1 \\ 0.2 & -0.1 \end{bmatrix}}^{\mathbf{V}_1}$$

$$= \overbrace{\begin{bmatrix} 0.8 & 1.7 \\ 0.8 & 1.7 \\ 0.8 & 1.7 \\ 0.8 & 1.7 \end{bmatrix}}^{\mathbf{1}\mathbf{v}_1^\top} + \overbrace{\begin{bmatrix} -0.2 & 0.1 \\ -0.2 & 0.1 \\ 0.2 & -0.1 \\ 0.2 & -0.1 \end{bmatrix}}^{\mathbf{z}\mathbf{V}_1}$$

$$= \overbrace{\begin{bmatrix} 0.6 & 1.8 \\ 0.6 & 1.8 \\ 1.0 & 1.6 \\ 1.0 & 1.6 \end{bmatrix}}^{\mathbf{1}\mathbf{v}_1^\top + \mathbf{z}\mathbf{V}_1}$$

The matrices $\mathbf{V}_1' = \mathbf{1}\mathbf{v}_1^\top + \mathbf{Z}\,\mathbf{V}_1$ and $\mathbf{V}_2' = \mathbf{1}\mathbf{v}_2^\top + \mathbf{Z}\,\mathbf{V}_2$ contain IRF parameters for responses to p_1 and p_2, respectively, for each of the four elements of $\mathbf{y}$. They are computed as follows:

$$\mathbf{V}_2' = \overbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}}^{\mathbf{1}} \overbrace{\begin{bmatrix} 1.1 & 0.5 \end{bmatrix}}^{\mathbf{v}_2^\top} + \overbrace{\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}}^{\mathbf{Z}} \overbrace{\begin{bmatrix} 0.3 & 0.4 \\ -0.3 & -0.4 \end{bmatrix}}^{\mathbf{V}_2}$$

$$= \overbrace{\begin{bmatrix} 1.1 & 0.5 \\ 1.1 & 0.5 \\ 1.1 & 0.5 \\ 1.1 & 0.5 \end{bmatrix}}^{\mathbf{1}\mathbf{v}_2^\top} + \overbrace{\begin{bmatrix} 0.3 & 0.4 \\ 0.3 & 0.4 \\ -0.3 & -0.4 \\ -0.3 & -0.4 \end{bmatrix}}^{\mathbf{z}\mathbf{V}_2}$$

$$= \overbrace{\begin{bmatrix} 1.4 & 0.9 \\ 1.4 & 0.9 \\ 0.8 & 0.1 \\ 0.8 & 0.1 \end{bmatrix}}^{\mathbf{1}\mathbf{v}_2^\top + \mathbf{z}\mathbf{V}_2}$$

The mask matrix $\mathbf{F}_{[y,x]} \stackrel{\text{def}}{=} \begin{cases} 1 & \left(\mathbf{s}_{[x]} = \mathbf{s}'_{[y]}\right) \text{ and } \left(\mathbf{t}_{[x]} \leq \mathbf{t}'_{[y]}\right) \\ 0 & \text{otherwise} \end{cases}$ indicates predictor observations that precede each element of $\mathbf{y}$ in the same time series. Timestamp vectors $\mathbf{t}$, $\mathbf{t}'$ and series ID vectors $\mathbf{s}$, $\mathbf{s}'$ are shown on the top and left axes for expository purposes.

$$\mathbf{F} = \begin{array}{cc} & \begin{matrix} 0 & 2 & 3 & 0 & 1 & 4 \end{matrix} \}\mathbf{t} \\ & \begin{matrix} 1 & 1 & 1 & 2 & 2 & 2 \end{matrix} \}\mathbf{s} \\ \begin{matrix} 1 \\ 4 \\ 2 \\ 3 \end{matrix} \begin{matrix} 1 \\ 1 \\ 2 \\ 2 \end{matrix} & \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \end{array}$$
$$\begin{matrix} \underbrace{}_{\mathbf{t}'} & \underbrace{}_{\mathbf{s}'} \end{matrix}$$

To compute convolution matrices $\mathbf{G}_1$, $\mathbf{G}_2$, we first compute an array $\mathbf{t}'\mathbf{1}^\top - \mathbf{1}\mathbf{t}$ containing distance in time of predictors from responses:

$$\mathbf{t}'\mathbf{1}^\top - \mathbf{1}\mathbf{t} = \overbrace{\begin{bmatrix} 1 \\ 4 \\ 2 \\ 3 \end{bmatrix}}^{\mathbf{t}'} \overbrace{\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}}^{\mathbf{1}^\top} - \overbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}}^{\mathbf{1}} \overbrace{\begin{bmatrix} 0 & 2 & 3 & 0 & 1 & 4 \end{bmatrix}}^{\mathbf{t}^\top}$$
$$= \overbrace{\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 4 & 4 & 4 & 4 & 4 & 4 \\ 2 & 2 & 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 & 3 & 3 \end{bmatrix}}^{\mathbf{t}'\mathbf{1}^\top} - \overbrace{\begin{bmatrix} 0 & 2 & 3 & 0 & 1 & 4 \\ 0 & 2 & 3 & 0 & 1 & 4 \\ 0 & 2 & 3 & 0 & 1 & 4 \\ 0 & 2 & 3 & 0 & 1 & 4 \end{bmatrix}}^{\mathbf{1}\mathbf{t}^\top}$$
$$= \overbrace{\begin{bmatrix} 1 & -1 & -2 & 1 & 0 & -3 \\ 4 & 2 & 1 & 4 & 3 & 0 \\ 2 & 0 & -1 & 2 & 1 & -2 \\ 3 & 1 & 0 & 3 & 2 & -1 \end{bmatrix}}^{\mathbf{t}'\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top}$$

These distances are supplied as inputs to the impulse response functions g_1, g_2, and irrelevant cells (i.e. cells from the future or cells from other time series) are masked using $\mathbf{F}$. The resulting convolution matrices $\mathbf{G}_1 = g_1(\mathbf{t}'\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top; \mathbf{V}_1') \odot \mathbf{F}$, $\mathbf{G}_2 = g_2(\mathbf{t}'\ \mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top; \mathbf{V}_2') \odot \mathbf{F}$ contain the convolution weights to apply to the elements of $\mathbf{X}$ in order to generate $\mathbf{y}$. They are computed as follows, where g_k is parameterized row-wise by $\mathbf{V}_k'$ and applied elementwise to $\mathbf{t}'\ \mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top$:

$$\mathbf{G}_1 = g_1\left(\overbrace{\begin{bmatrix} 1 & -1 & -2 & 1 & 0 & -3 \\ 4 & 2 & 1 & 4 & 3 & 0 \\ 2 & 0 & -1 & 2 & 1 & -2 \\ 3 & 1 & 0 & 3 & 2 & -1 \end{bmatrix}}^{\mathbf{t}'\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top} ; \overbrace{\begin{bmatrix} 0.6 & 1.8 \\ 0.6 & 1.8 \\ 1.0 & 1.6 \\ 1.1 & 1.6 \end{bmatrix}}^{\mathbf{V}'\mathbf{1}} \right) \odot \overbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}}^{\mathbf{F}}$$
$$= \overbrace{\begin{bmatrix} 0.91 & 0.24 & 0.02 & 0.91 & 0.82 & 0.00 \\ 0.00 & 0.34 & 0.91 & 0.00 & 0.04 & 0.00 \\ 0.54 & 0.54 & 0.08 & 0.54 & 1.00 & 0.00 \\ 0.08 & 1.00 & 0.54 & 0.08 & 0.54 & 0.08 \end{bmatrix}}^{g_1(\mathbf{t}'\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top; \mathbf{V}_1')} \odot \overbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}}^{\mathbf{F}}$$
$$= \overbrace{\begin{bmatrix} 0.91 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.34 & 0.91 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.54 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.08 & 0.54 & 0.00 \end{bmatrix}}^{g_1(\mathbf{t}'\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top; \mathbf{V}_1') \odot \mathbf{F}}$$

$$\mathbf{G}_2 = g_2\left(\overbrace{\begin{bmatrix} 1 & -1 & -2 & 1 & 0 & -3 \\ 4 & 2 & 1 & 4 & 3 & 0 \\ 2 & 0 & -1 & 2 & 1 & -2 \\ 3 & 1 & 0 & 3 & 2 & -1 \end{bmatrix}}^{\mathbf{t}'\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top}; \overbrace{\begin{bmatrix} 1.4 & 0.9 \\ 1.4 & 0.9 \\ 0.8 & 0.1 \\ 0.8 & 0.1 \end{bmatrix}}^{\mathbf{V}'\mathbf{1}}\right) \odot \overbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}}^{\mathbf{F}}$$

$$= \overbrace{\begin{bmatrix} 0.84 & 0.00 & 0.00 & 0.84 & 0.11 & 0.00 \\ 0.00 & 0.67 & 0.84 & 0.00 & 0.06 & 0.11 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.67 & 0.00 \\ 0.00 & 0.67 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}}^{g_2\left(\mathbf{t}'\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top; \mathbf{V}_2'\right)} \odot \overbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}}^{\mathbf{F}}$$

$$= \overbrace{\begin{bmatrix} 0.84 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.67 & 0.84 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.67 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}}^{g_2\left(\mathbf{t}'\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top; \mathbf{V}_2'\right) \odot \mathbf{F}}$$

The two columns of the convolved predictor matrix $\mathbf{X}' = \mathbf{G_2}\,\mathbf{X}_{[*,2]}$ are computed by pre-multiplying each column with its corresponding convolution matrix:

$$\mathbf{X}'_{[*,1]} = \overbrace{\begin{bmatrix} 0.91 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.34 & 0.91 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.54 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.08 & 0.54 & 0.00 \end{bmatrix}}^{\mathbf{G}_1} \overbrace{\begin{bmatrix} 5 \\ -1 \\ 6 \\ 2 \\ 0 \\ -2 \end{bmatrix}}^{\mathbf{X}_{[*,1]}}$$

$$= \overbrace{\begin{bmatrix} 4.55 \\ 5.15 \\ 1.08 \\ 0.16 \end{bmatrix}}^{\mathbf{G}_1\mathbf{X}_{[*,1]}}$$

$$\mathbf{X}'_{[*,2]} = \overbrace{\begin{bmatrix} 0.84 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.67 & 0.84 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.67 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}}^{\mathbf{G}_2} \overbrace{\begin{bmatrix} 0 \\ 3 \\ 1 \\ 2 \\ 1 \\ -1 \end{bmatrix}}^{\mathbf{X}_{[*,2]}}$$

$$= \overbrace{\begin{bmatrix} 0.00 \\ 2.85 \\ 0.67 \\ 0.00 \end{bmatrix}}^{\mathbf{G}_2\mathbf{X}_{[*,2]}}$$

$$\mathbf{X}' = \begin{bmatrix} 4.55 & 0.00 \\ 5.15 & 2.85 \\ 1.08 & 0.67 \\ 0.16 & 0.00 \end{bmatrix}$$

The expected response $\widehat{\mathbf{y}} = \mathbf{m}' + (\mathbf{X}' \odot \mathbf{U}')\mathbf{1}$ is computed by rescaling $\mathbf{X}'$ by the coefficient matrix $\mathbf{U}'$, summing across predictors, and shifting by the intercept $\mathbf{m}'$, as shown:

$$\begin{aligned}\widehat{\mathbf{y}} &= \overbrace{\begin{bmatrix}1.1\\1.1\\1.5\\1.5\end{bmatrix}}^{\mathbf{m}'} + \left(\overbrace{\begin{bmatrix}4.55 & 0.00\\5.15 & 2.85\\1.08 & 0.67\\0.16 & 0.00\end{bmatrix}}^{\mathbf{X}'} \odot \overbrace{\begin{bmatrix}-0.5 & 1.2\\-0.2 & 1.2\\0.4 & -0.2\\0.4 & -0.2\end{bmatrix}}^{\mathbf{U}'}\right)\overbrace{\begin{bmatrix}1\\1\end{bmatrix}}^{\mathbf{1}}\\ &= \overbrace{\begin{bmatrix}1.1\\1.1\\1.5\\1.5\end{bmatrix}}^{\mathbf{m}'} + \overbrace{\begin{bmatrix}-0.91 & 0.00\\-1.03 & 3.42\\0.43 & -0.13\\0.06 & 0.00\end{bmatrix}}^{\mathbf{X}' \odot \mathbf{U}'}\overbrace{\begin{bmatrix}1\\1\end{bmatrix}}^{\mathbf{1}}\\ &= \overbrace{\begin{bmatrix}1.1\\1.1\\1.5\\1.5\end{bmatrix}}^{\mathbf{m}'} + \overbrace{\begin{bmatrix}-0.91\\2.39\\0.30\\0.06\end{bmatrix}}^{(\mathbf{X}' \odot \mathbf{U}')\mathbf{1}}\\ &= \overbrace{\begin{bmatrix}0.19\\3.49\\1.80\\1.56\end{bmatrix}}^{\mathbf{m}' + (\mathbf{X}' \odot \mathbf{U}')\mathbf{1}}\end{aligned}$$

Appendix B. Addressing non-normally distributed error: Log-normal and sinh-arcsinh transforms

Like ordinary linear regression, the CDR model defined in §4 assumes a Gaussian error distribution. However, it is often necessary to analyze data that violate this assumption. A common solution to non-normal errors is to apply a normalizing transform, such as a log transform or a power transform (Box & Cox, 1964). Transforms can complicate model interpretation by changing the linking function. For example, log transforming the response creates a log-linear rather than linear model on the convolved data, and model estimates thus describe a multiplicative rather than additive change in the response as a function of the predictors.

In this study, we therefore also consider an alternative approach, made possible by our use of stochastic gradient optimization directly on the likelihood surface. Instead of defining the error distribution as Gaussian, we can define it as a sinh-arcsinh transform on the Gaussian distribution.[48] The sinh-arcsinh transformed Gaussian is a generalization of the Gaussian distribution that additionally contains skewness and tailweight parameters $\epsilon \in \mathbb{R}$ and $\delta \in \mathbb{R}_+$ (Jones & Pewsey, 2009). When $\epsilon = 0$ and $\delta = 1$, the distribution is Gaussian. When $\varepsilon < 0$, the distribution has negative skew, and when $\epsilon > 0$, the distribution has positive skew. Tail thickness increases with δ. Both ϵ and δ are estimated from data, along with all other model parameters. The advantage of using sinh-arcsinh error over normalizing transforms is that it can flexibly adapt to asymmetrically distributed data without transforming it, thus preserving the original scale of the response as well as the additive interpretation of model estimates while also relaxing normality assumptions. Normalizing transforms and sinh-arcsinh error distributions are explored in §14 for the reading and fMRI experiments. As shown below, sinh-arcsinh improves goodness of fit over Gaussian error across all model designs, supporting its adoption for CDR modeling. However, we stress that sinh-arcsinh error is not appropriate for settings in which estimates will ultimately be used in ways that assume normally-distributed error. For example, researchers may wish to evaluate CDR models with respect to squared error or percent variance explained. Such evaluations assume a Gaussian likelihood and are therefore not appropriate for asymmetric error distributions like sinh-arcsinh.

Appendix C. Full results

Here we present the full set of results from all analysis conducted in this study, including synthetic datasets, self-paced-reading and eye-tracking during reading datasets, and the fMRI dataset. As discussed in Appendix B, in the reading and fMRI analyses we additionally explore the effect of using a sinh-arcsinh transform of a normal error distribution.

C.1. Simulation

Figs. A1, A2, A3, A4, A5 and A6 show impulse response estimates for all inference types and (in Simulation D) response kernels used in all simulation studies. As shown, estimates are highly consistent across inference types and response kernels and are robust to noise, non-uniform temporal distribution of events, multicollinearity, and impulse response misspecification, as long as the underlying response falls within the model's solution space.

[48] The sinh-arcsinh transform on the standard normal distribution with skewness ϵ and tailweight δ yields probability density $f_{\epsilon,\delta}$:

$$f_{\epsilon,\delta}(x) \stackrel{\text{def}}{=} \{2\pi(1+x^2)\}^{-1/2}\delta C_{\epsilon,\delta}(x) exp\left\{-S^2_{\epsilon,\delta}(x)\Big/2\right\} \tag{47}$$

$$S_{\epsilon,\delta}(x) \stackrel{\text{def}}{=} sinh\{\delta sinh^{-1}(x) - \epsilon\} \tag{48}$$

$$C_{\epsilon,\delta}(x) \stackrel{\text{def}}{=} \left\{1 + S^2_{\epsilon,\delta}(x)\right\}^{1/2x} \tag{49}$$

In practice, location and scale can also be parameterized. For additional details, see Jones and Pewsey (2009).

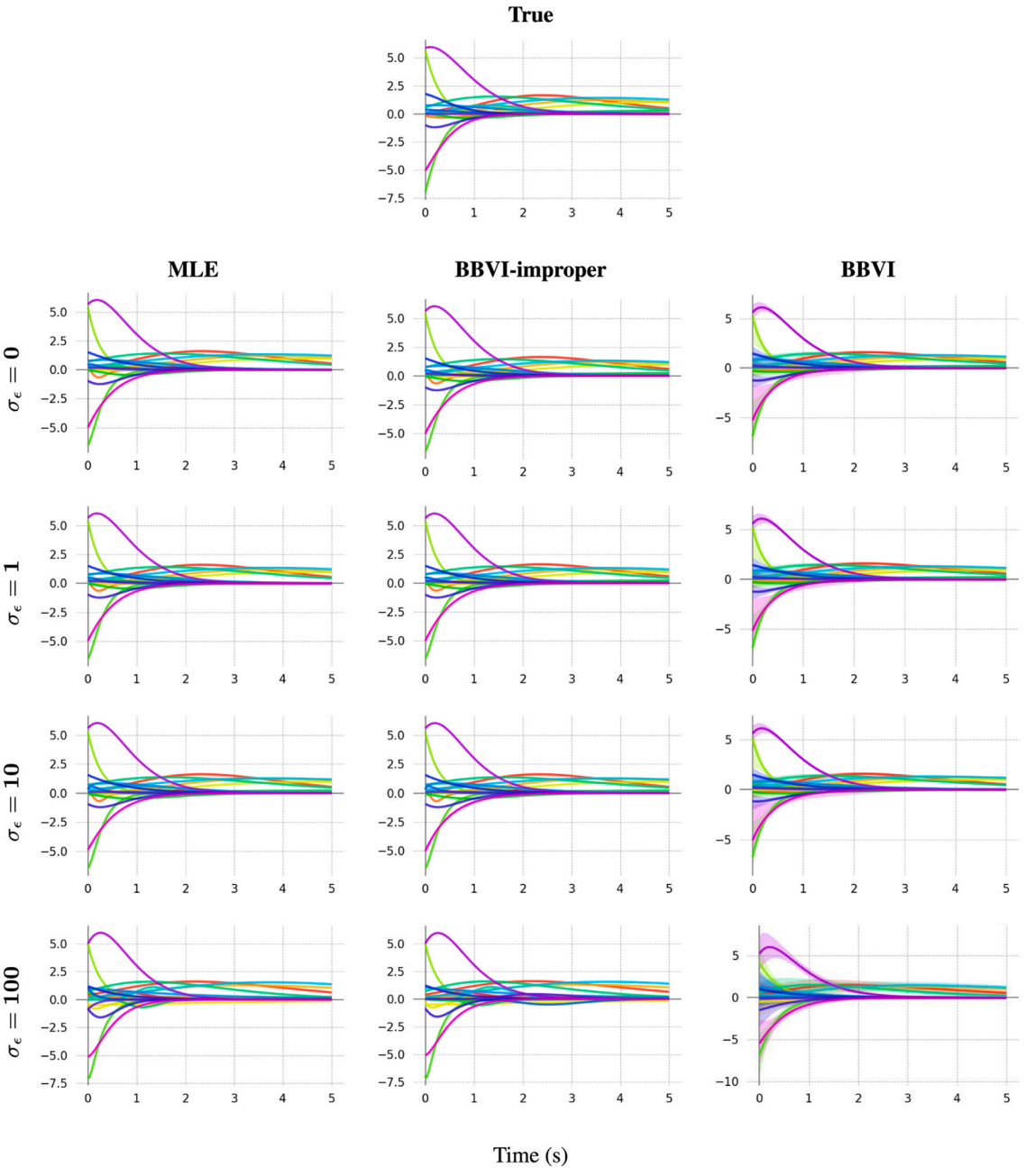


Fig. A1. Simulation A: *Noise*. True synthetic model vs. CDR-estimated models with increasingly large standard deviation σ_ϵ of the Gaussian noise distribution. The twenty IRFs (corresponding to the twenty random covariates) are represented by distinct curves in each plot. BBVI and BBVI-improper are plotted with 95% credible intervals.

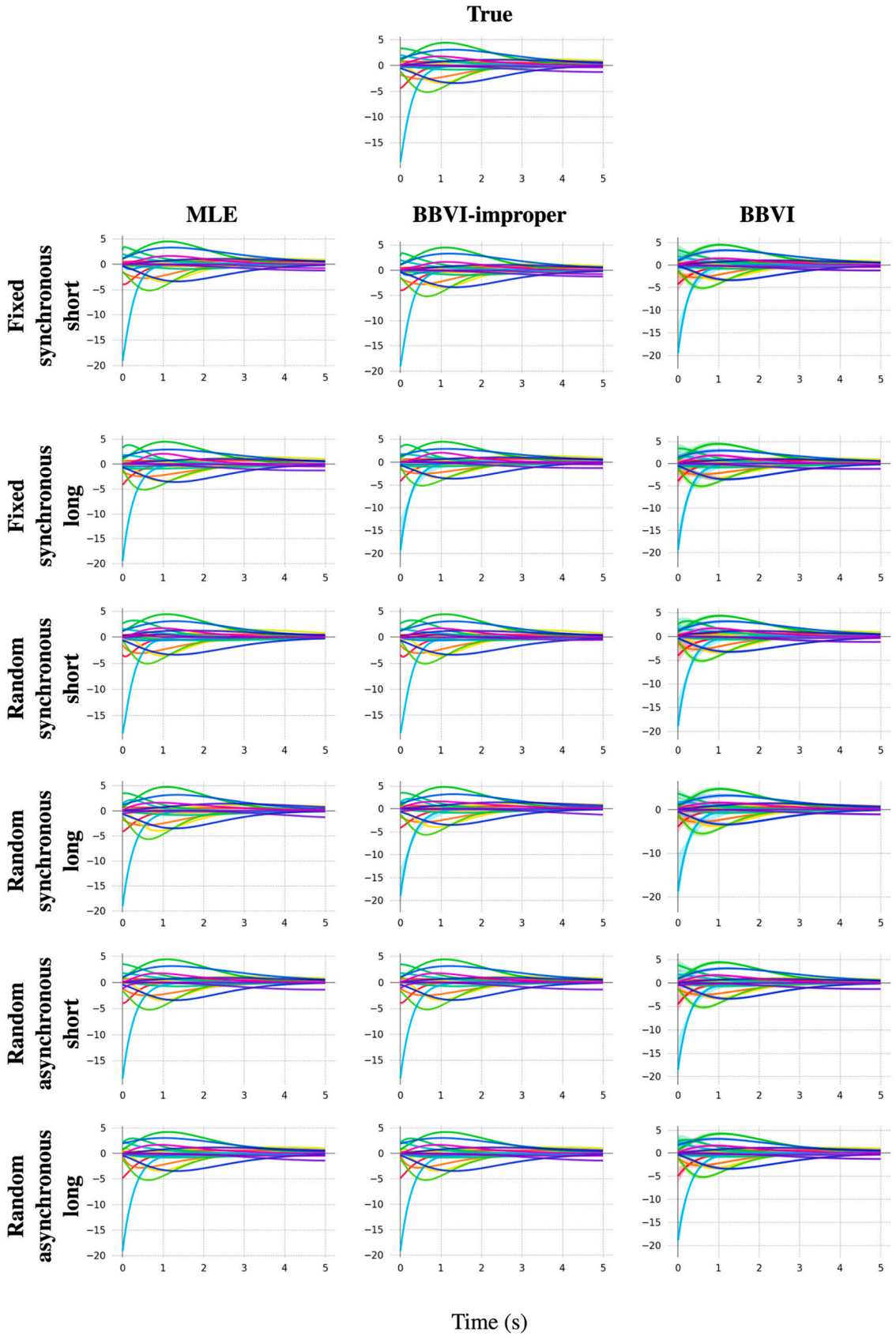


Fig. A2. Simulation B: *Time*. True synthetic model vs. CDR-estimated models with varying types of time interval. The twenty IRFs (corresponding to the twenty random covariates) are represented by distinct curves in each plot.

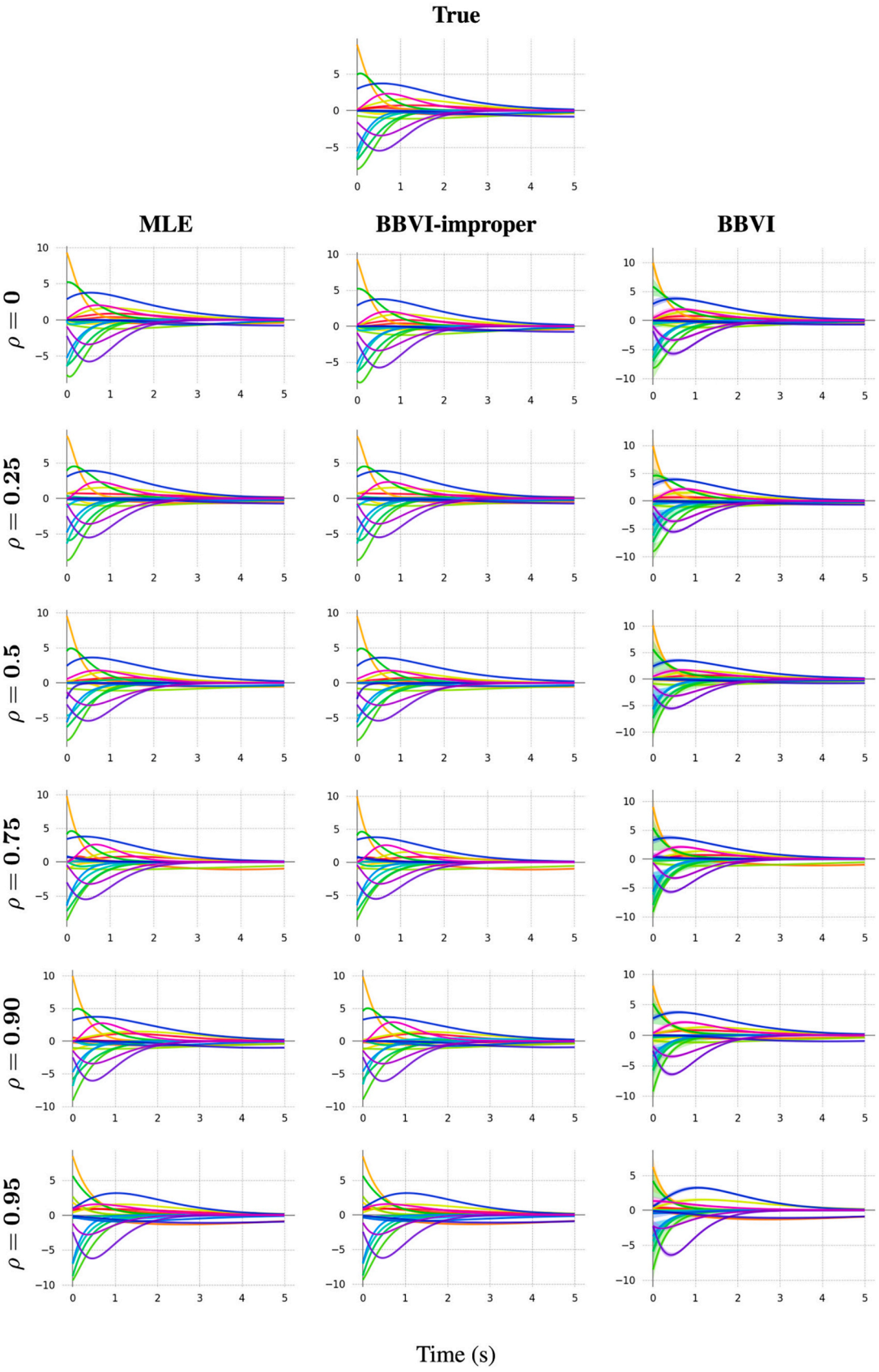


Fig. A3. Simulation C: *Multicollinearity*. True synthetic model vs. CDR-estimated models with increasingly multicollinear predictors. The twenty IRFs (corresponding to the twenty random covariates) are represented by distinct curves in each plot.

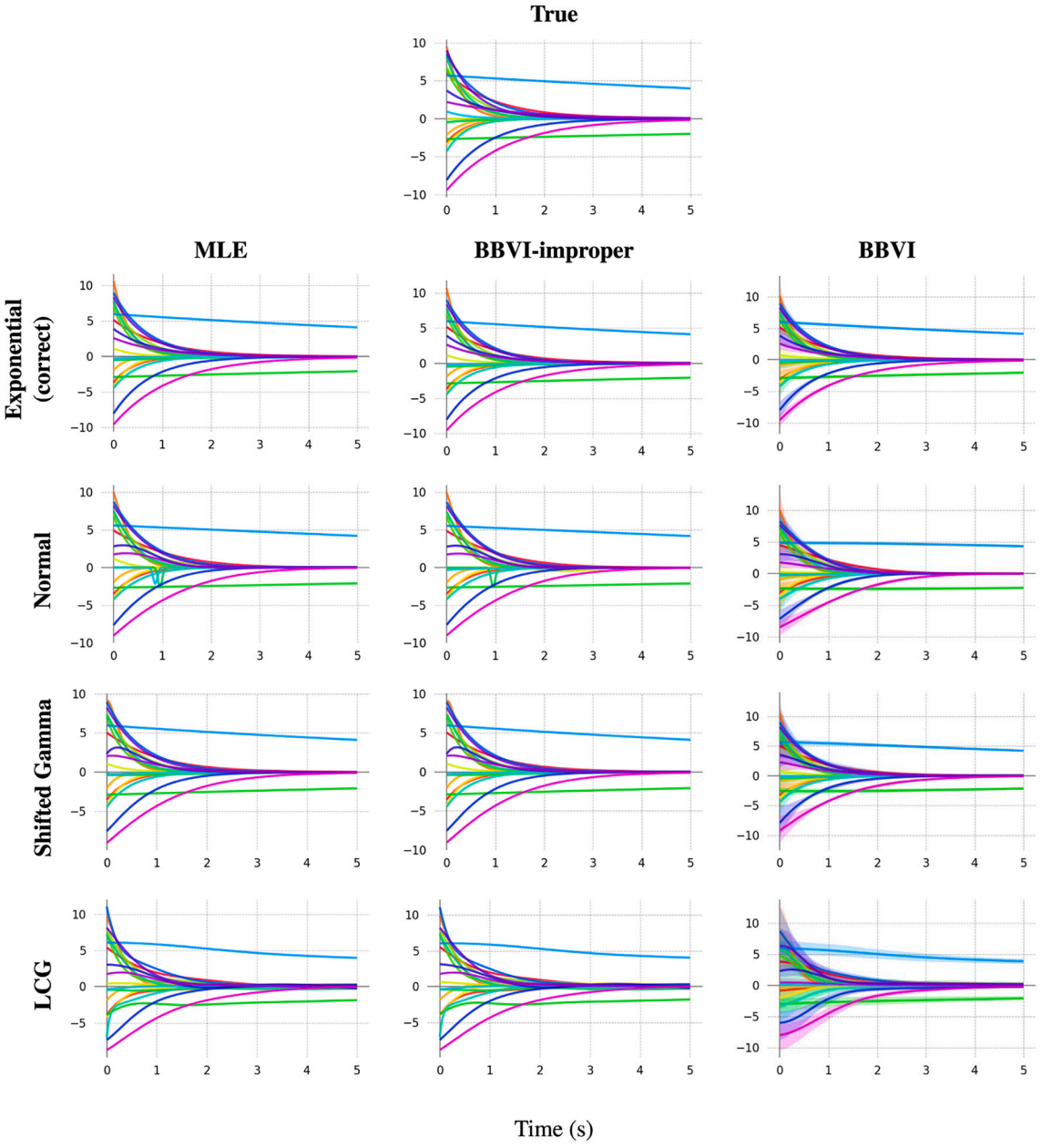


Fig. A4. Simulation D: *Misspecification (exponential ground truth)*. True exponential model vs. CDR-estimated models using various IRF kernels. The twenty IRFs (corresponding to the twenty random covariates) are represented by distinct curves in each plot.

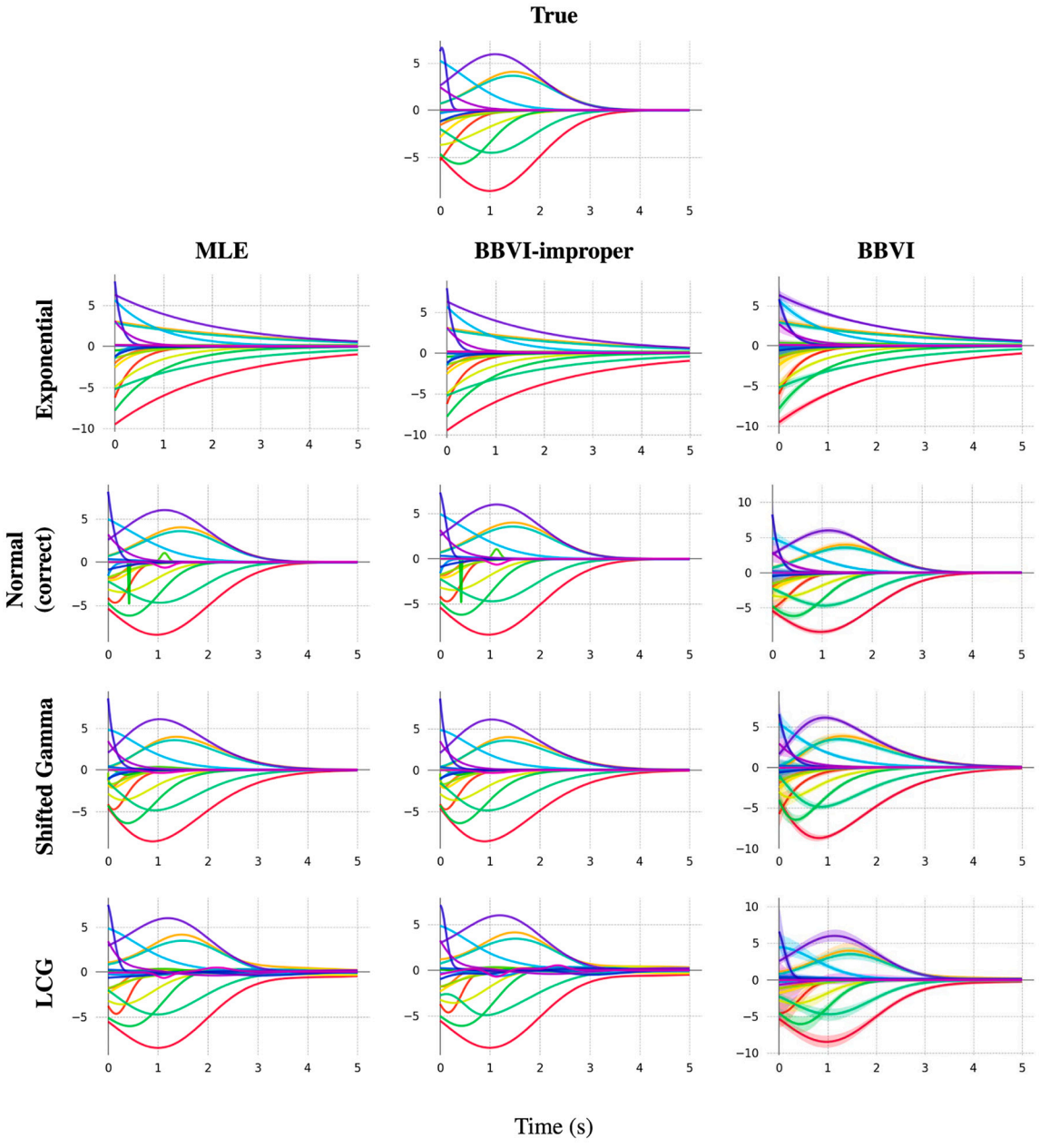


Fig. A5. Simulation D: *Misspecification* (*normal ground truth*). True normal model vs. CDR-estimated models using various IRF kernels. The twenty IRFs (corresponding to the twenty random covariates) are represented by distinct curves in each plot.

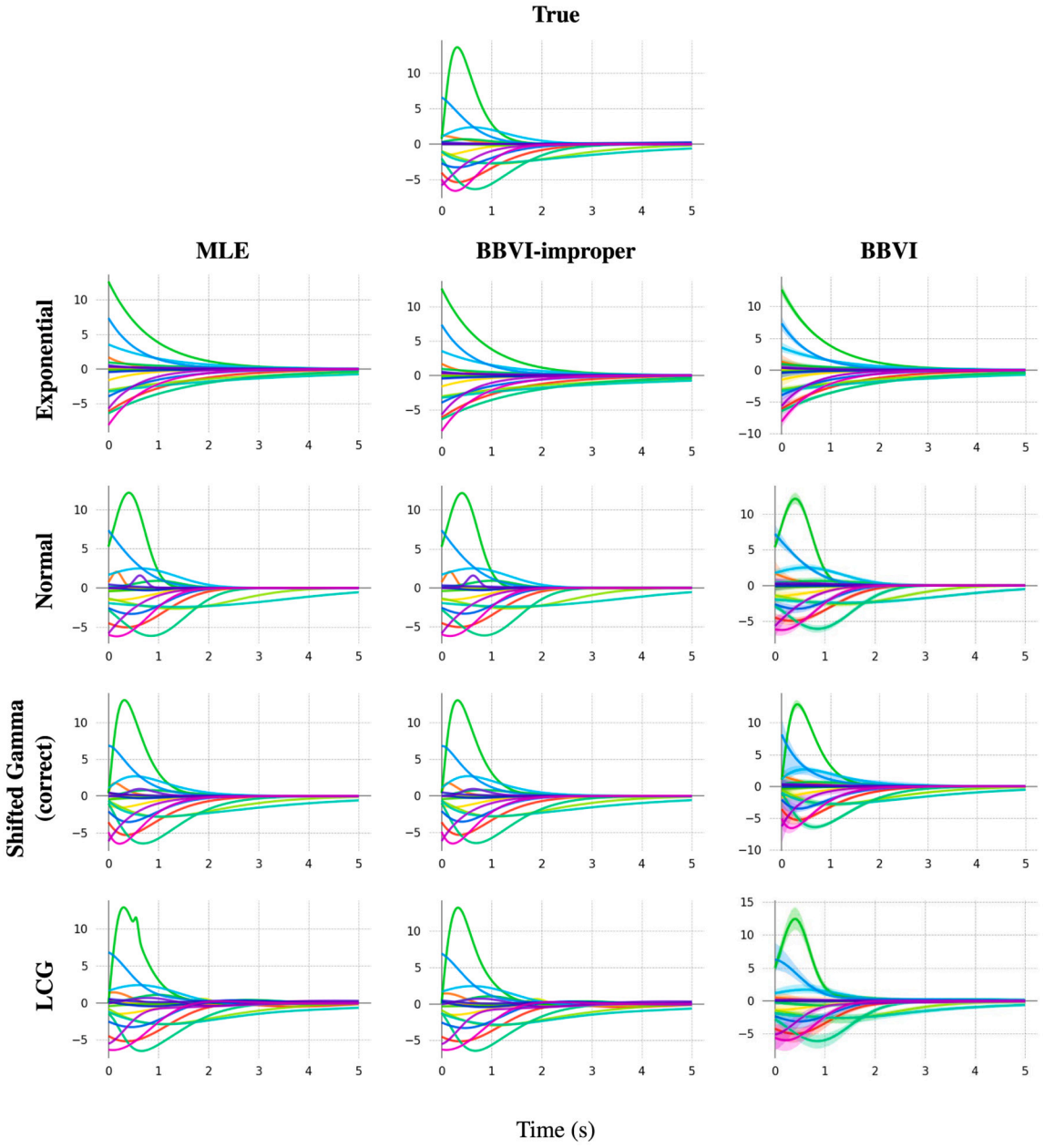


Fig. A6. Simulation D: *Misspecification (shifted gamma ground truth)*. True shifted gamma model vs. CDR-estimated models using various IRF kernels. The twenty IRFs (corresponding to the twenty random covariates) are represented by distinct curves in each plot.

C.2. Reading

We apply *exponential*, *normal*, *shifted gamma*, and LCG kernels to Gaussian and sinh-arcsinh error distributions of linear and log-transformed reading durations in Natural Stories (self-paced reading) and Dundee (eye-tracking), using MLE, BBVI improper, and BBVI inference. For Dundee, we consider scan pass durations (discussed in the main article) along with first pass and go-past durations (presented here). First pass duration is defined as the time elapsed between entering a word region from the left and entering a different word region to its left or right. Go-past duration is defined as the time elapsed between entering a word region from the left and entering a word region to its right (including all intervening regressive fixations).

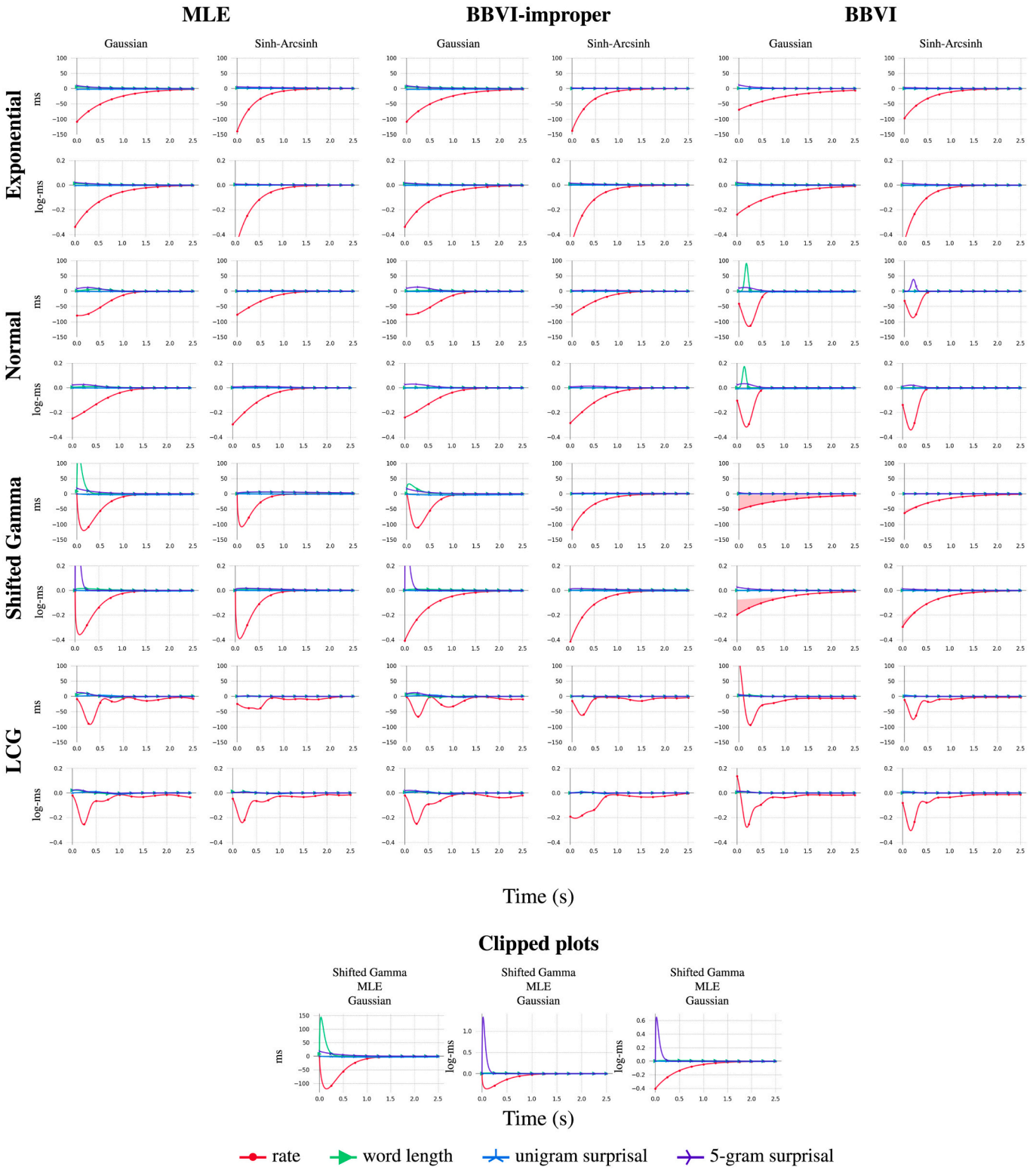


Fig. A7. Natural Stories: IRF estimates using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels.

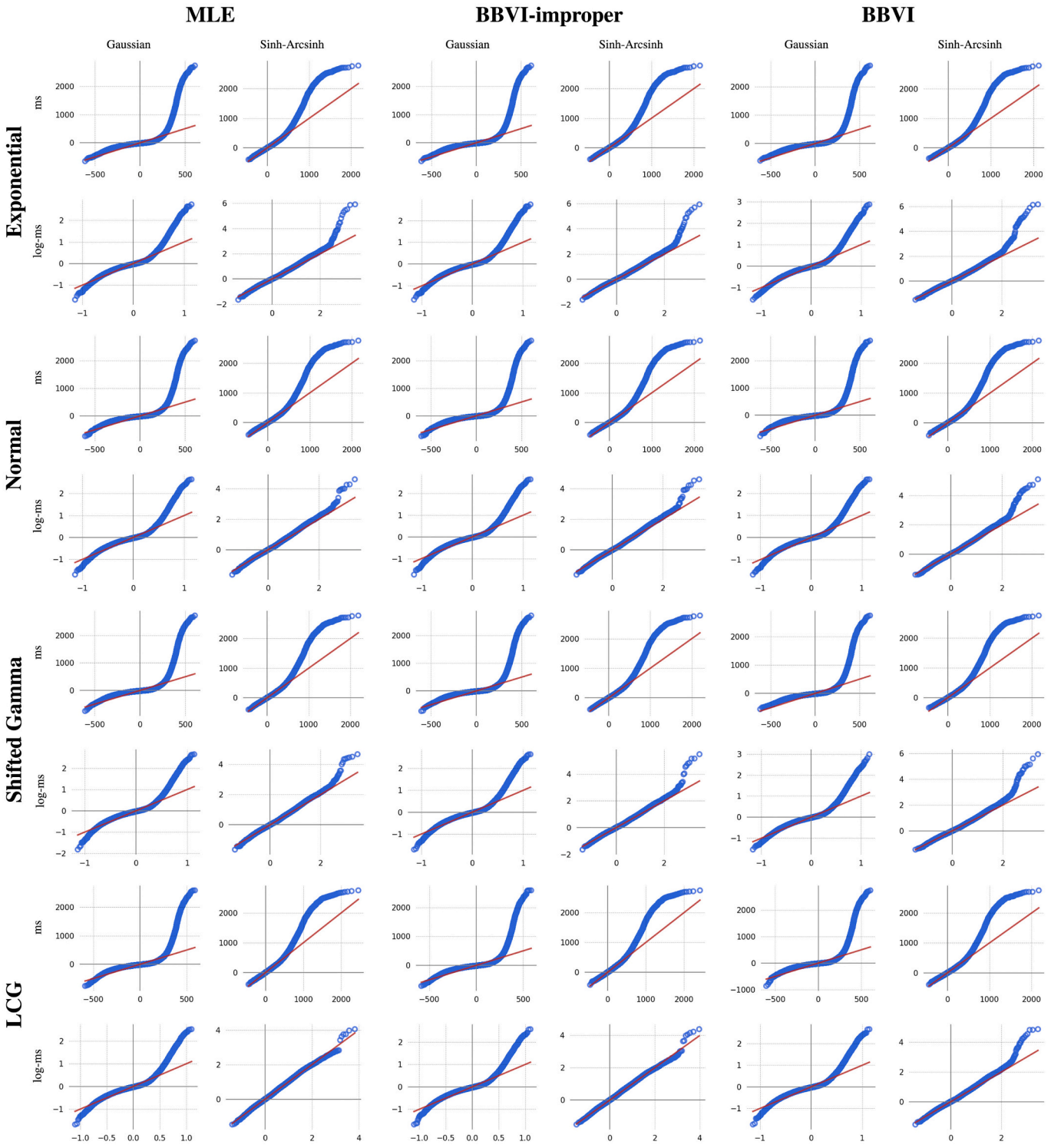


Fig. A8. Natural Stories: Quantile-quantile plots of fitted (*x*-axis) to true (*y*-axis) error distributions on the training set, using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels. The theoretical best-fit line is plotted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

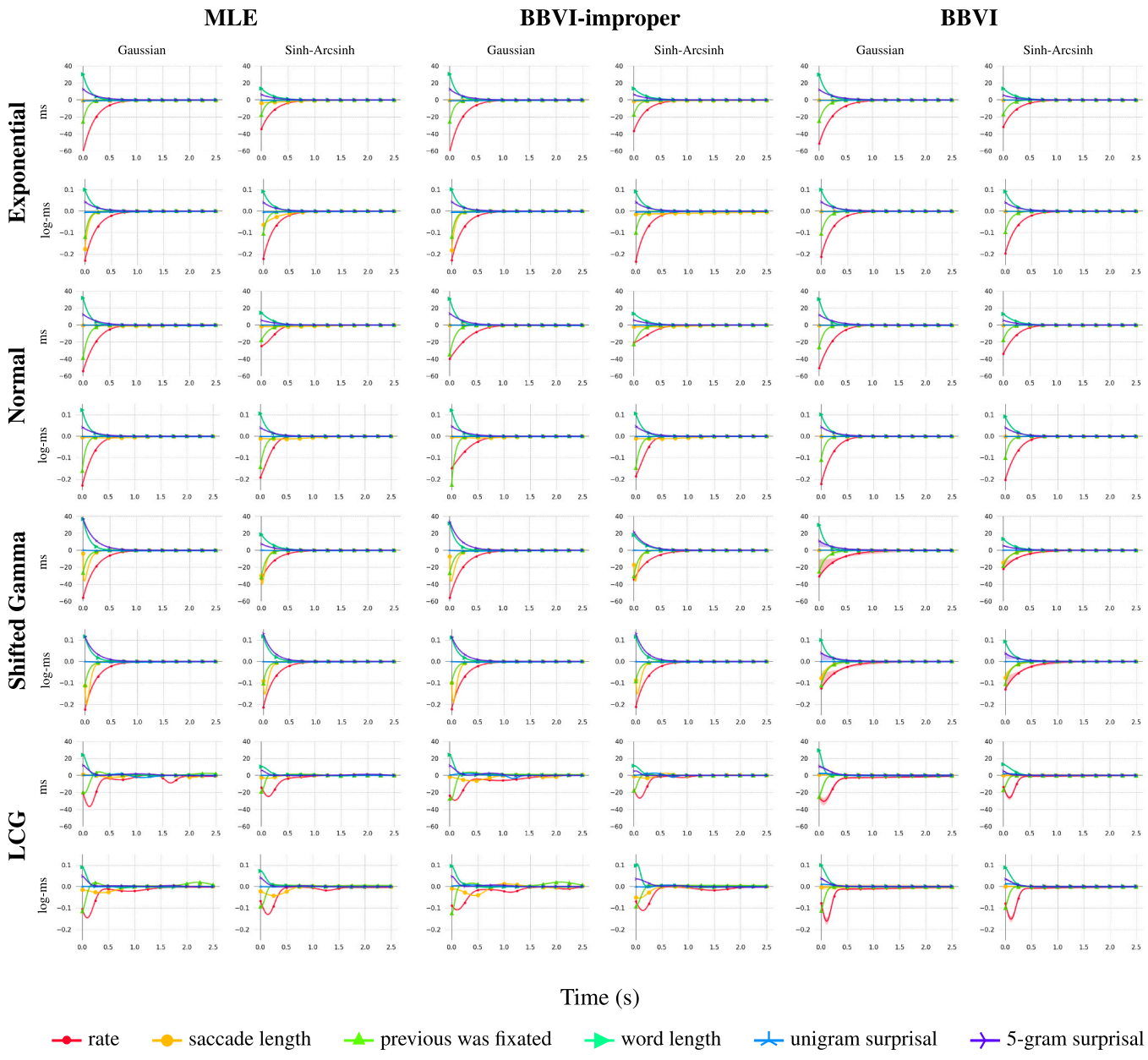


Fig. A9. Dundee (first past): IRF estimates using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels.

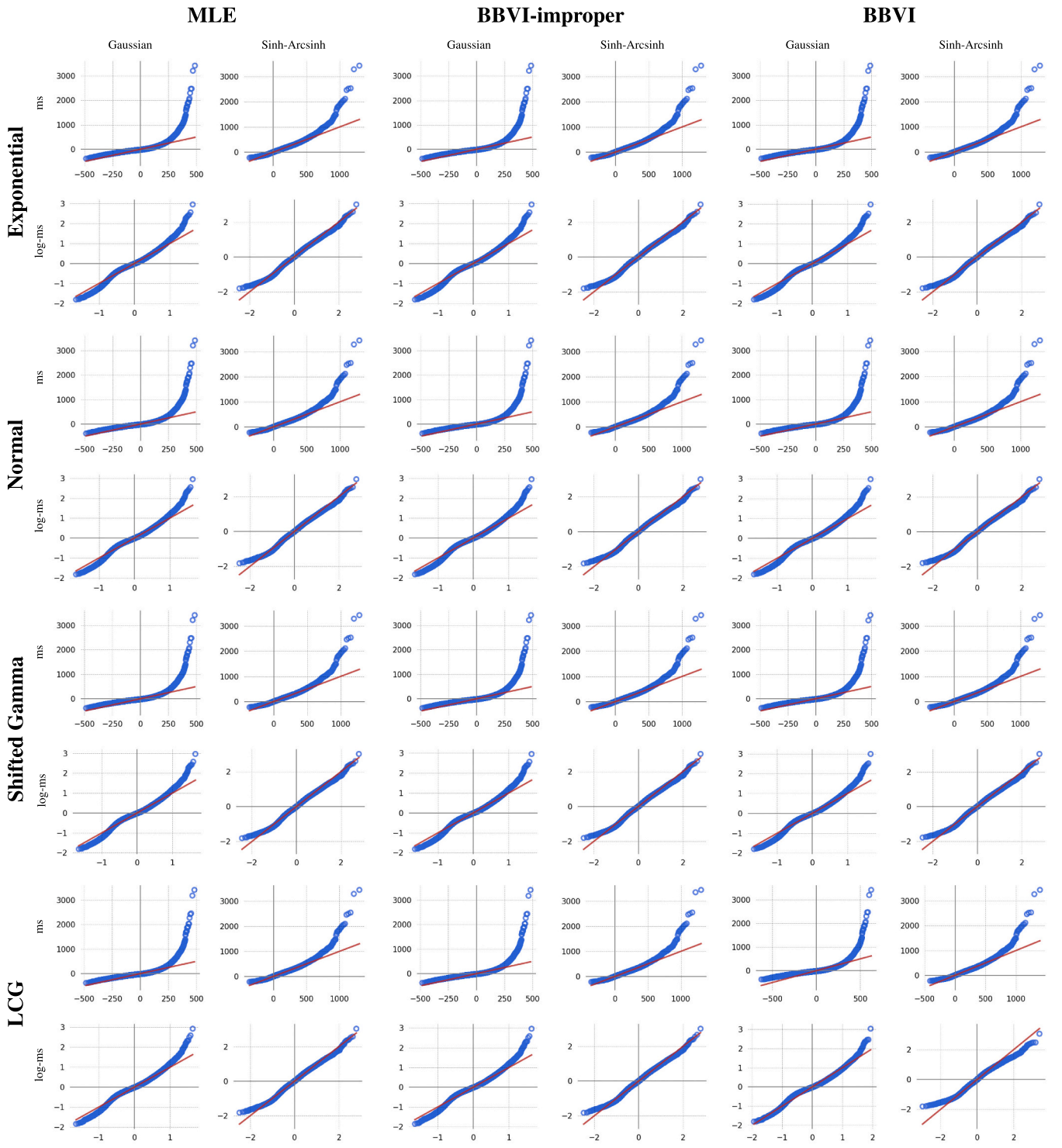


Fig. A10. Dundee (first pass): Quantile-quantile plots of fitted (*x*-axis) to true (*y*-axis) error distributions on the training set, using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels. The theoretical best-fit line is plotted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

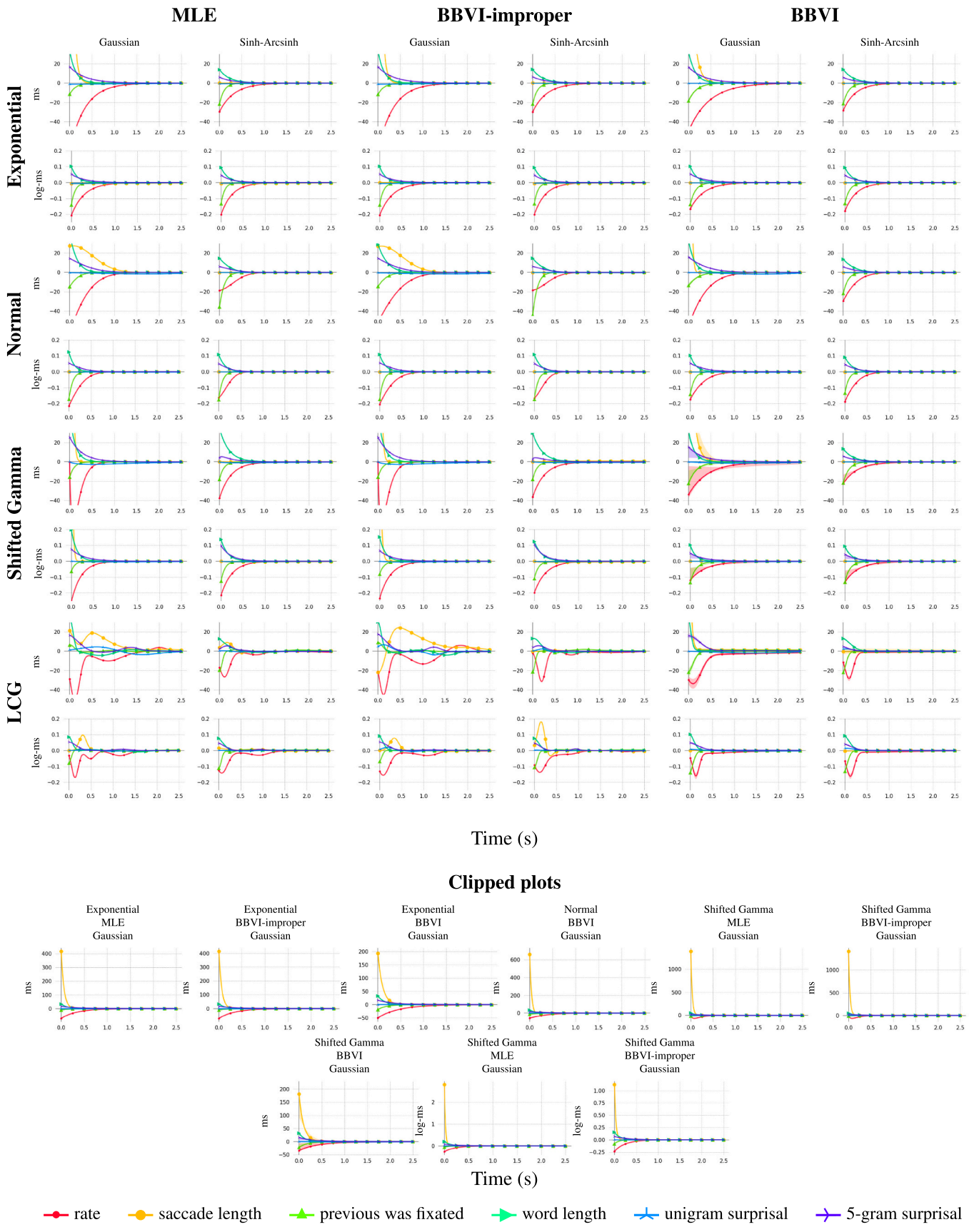


Fig. A11. Dundee (go-past): IRF estimates using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels.

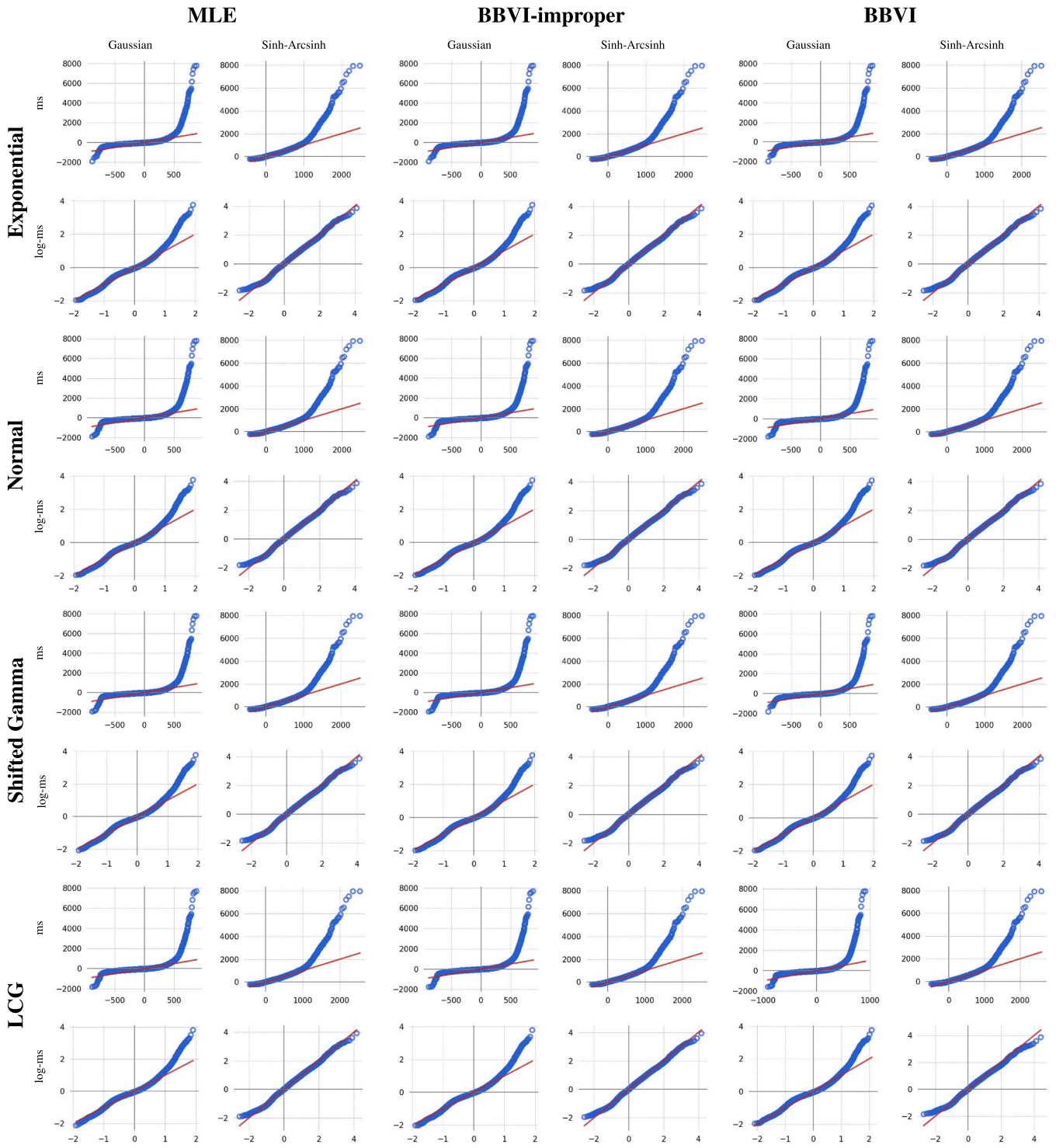


Fig. A12. Dundee (go-past): Quantile-quantile plots of fitted (*x*-axis) to true (*y*-axis) error distributions on the training set, using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels. The theoretical best-fit line is plotted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

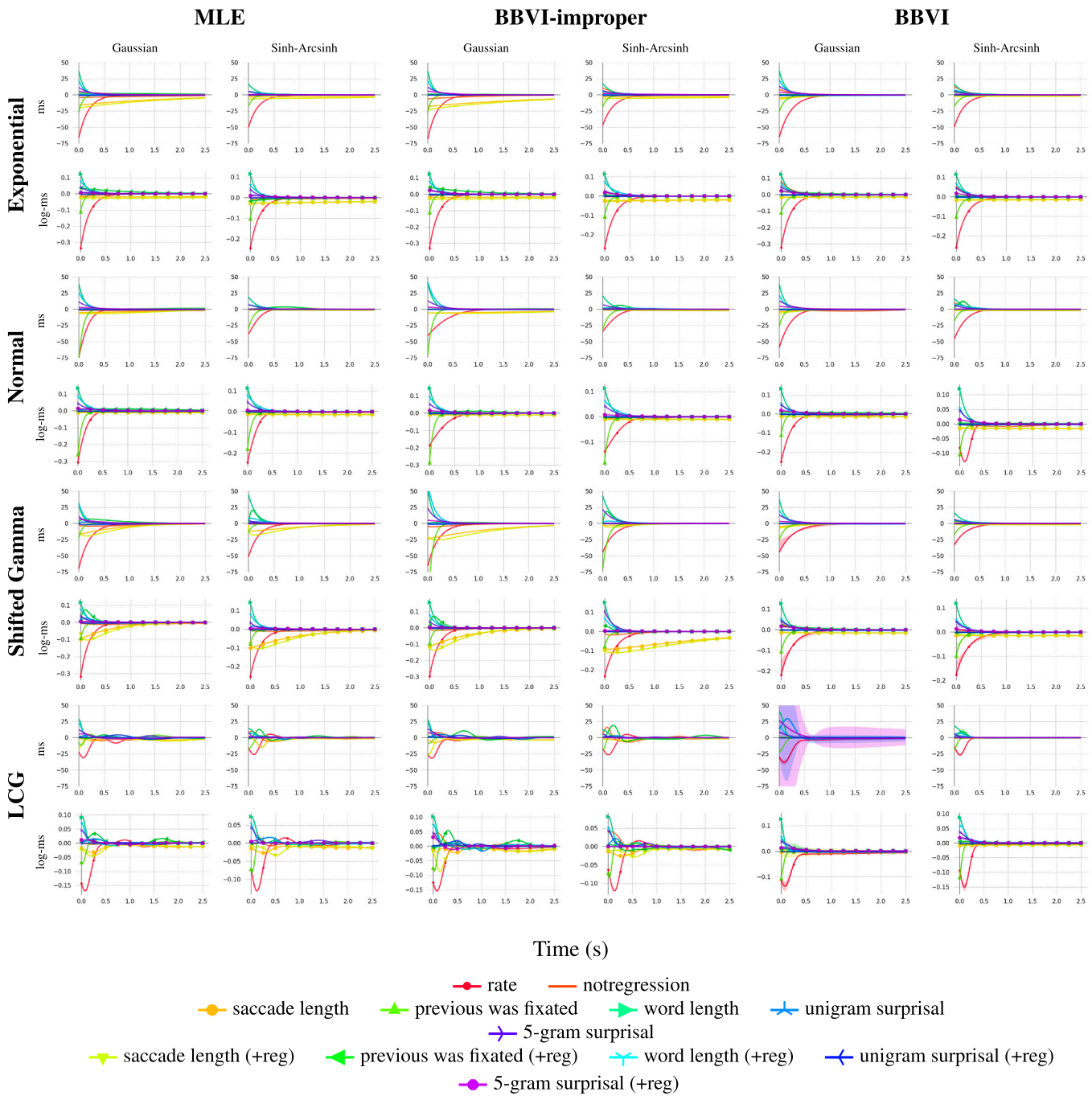


Fig. A13. Dundee (scan path): IRF estimates using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels. IRFs are distinguished for predictors under regressive eye movements are distinguished by (+reg).

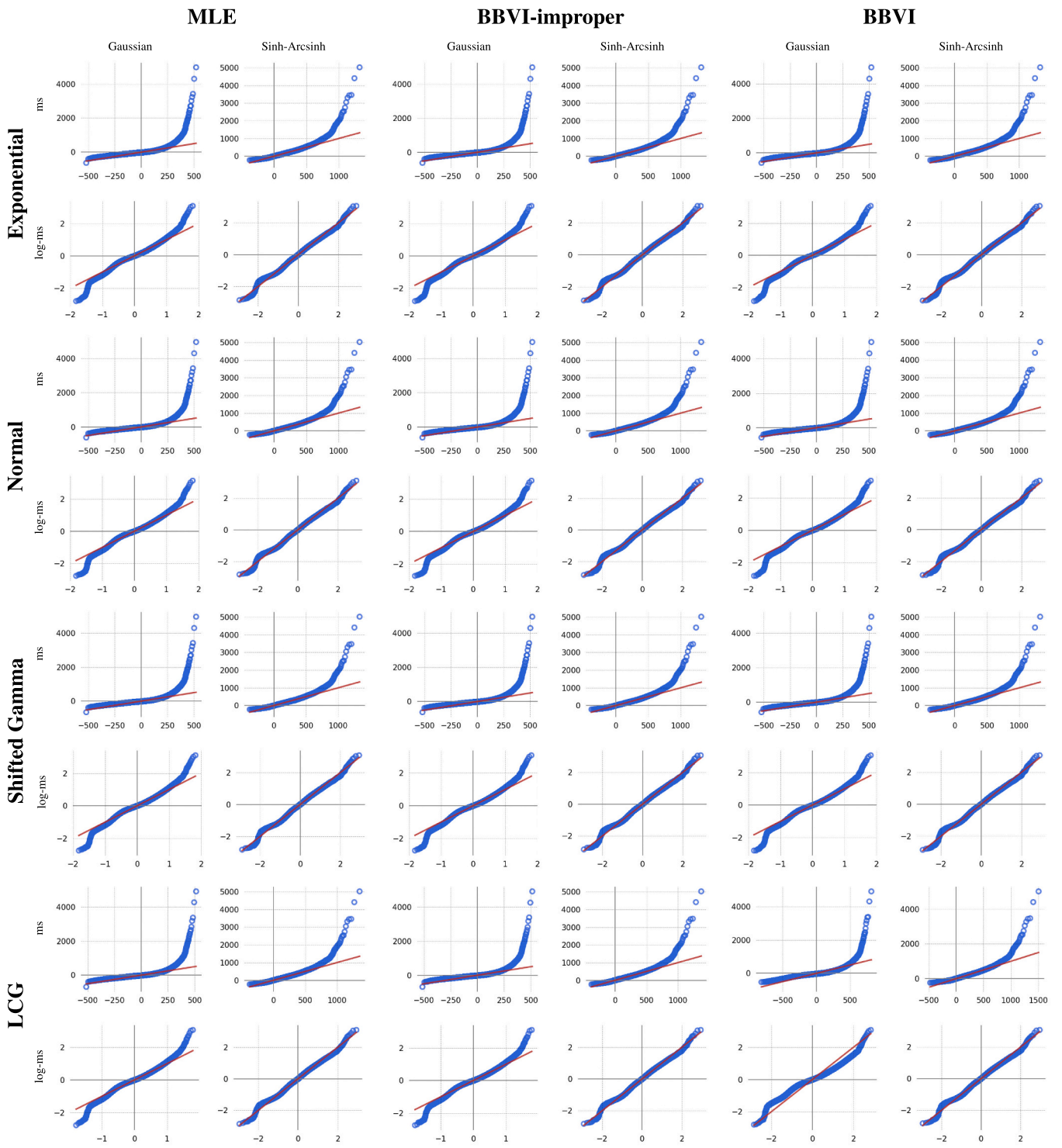


Fig. A14. Dundee (scan path): Quantile-quantile plots of fitted (*x*-axis) to true (*y*-axis) error distributions on the training set, using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels. The theoretical best-fit line is plotted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Figs. A7, A9, A11, and A13 show the full response estimates for Natural Stories, Dundee first pass, Dundee go-past, and Dundee scan path, respectively. For ease of comparison, *x*- and *y*-axis dimensions are shared within each response definition per dataset. However, some estimates (Natural Stories and Dundee go-past) are so extreme that including them would compromise visual clarity. In these cases, we clip the estimates in the main plots, then plot the full responses with adjusted *y*-axes under the heading "Clipped plots".

Models show similar estimates of temporal dynamics across response definitions, error definitions, and IRF kernels, and generally conform to prior expectations about effect sizes and direction, as discussed in §7, including across different duration definitions in Dundee. There are some key exceptions to this pattern. First, some models of go-past duration in Dundee find very large early saccade length effects, in some cases orders of magnitude larger than any other effect in the model. This outcome is relatively uncommon. Exceptions are primarily found in Gaussian error models of

untransformed Dundee go-past durations, although two such examples also occur in Gaussian error models of log-transformed go-past durations (see "Clipped plots" in Fig. A11). The source of these apparent outlier estimates using go-past durations is left to future research. Second, the Gaussian error BBVI-estimated LCG model of Dundee scan path durations is a clear outlier compared to the rest of the Dundee scan path models: it finds very large credible intervals for some estimates, and its predictive performance is very poor, both in-sample and out-of-sample (Table 4). This result suggests divergent training and motivates caution when using heavily parameterized CDR models, at least under BBVI estimation, where initial random sampling of parameter estimates can yield poor-quality samples and large gradients. If this particular model were the basis of a scientific investigation, such evidence of training divergence would motivate the use of different hyperparameters (e.g. a lower learning rate). However, it is reassuring that this outcome is quite rare in practice, occurring in only one of the hundreds of models fitted for this study.

Goodness of fit results are visualized as quantile-quantile plots in Figs. A8, A10, A12, and A14. As shown, both the normalizing (log) transform and the asymmetric sinh-arcsinh error distribution improve goodness of fit across model types, with the best fit consistently occurring in sinh-arcsinh models of log-transformed fixation durations. The true error distribution for raw fixation durations tends to have a heavier right tail than that of the fitted distribution, even under sinh-arcsinh. This suggests that sinh-arcsinh on its own may not be sufficiently expressive to account for very skewed data. However, sinh-arcsinh consistently eliminates the heavy *left* tail visible in the Gaussian models, and it tends to better account for the right tail, thus improving fit across the board compared to Gaussian models in matched experimental conditions. These results indicate that sinh-arcsinh error is beneficial across designs, even if it cannot completely eliminate poor fit on its own.

Table A2
Dundee (first pass). CDR vs. baselines, mean-squared error. CDR results shown using (E)xponential, (N)ormal, Shifted (G)amma, and non-parametric LCG response kernels, fitted using MLE, BBVI improper, and BBVI. Best-performing models within the sets of baseline and CDR models are shown in *italics*. Best-performing overall models are shown in **bold**. Daggers (†) indicate convergence failures.

	Dundee (ms)			Dundee (log-ms)		
Model	Train	Expl	Test	Train	Expl	Test
LME	13,152†	14,204†	14,026†	0.1516	0.1542	0.1531
LME-S	13,112†	14,162†	14,024†	0.1507†	0.1532†	0.1526†
GAM	13,007	14,065	13,871	0.1510	0.1536	0.1525
GAM-S	*12,882*	***13,948***	***13,771***	*0.1491*	***0.1518***	***0.1508***
CDR-E-MLE	13,040	14,077	–	0.1500	0.1529	–
CDR-E-BBVI.imp	13,040	14,079	–	0.1500	0.1530	–
CDR-E-BBVI	13,071	14,103	–	0.1505	0.1529	–
CDR-N-MLE	13,030	14,163	–	0.1498	0.1542	–
CDR-N-BBVI.imp	13,037	*14,068*	–	0.1498	0.1543	–
CDR-N-BBVI	13,063	14,077	–	0.1504	*0.1526*	–
CDR-G-MLE	13,034	14,069	–	0.1499	0.1534	–
CDR-G-BBVI.imp	13,037	14,072	–	0.1499	0.1534	–
CDR-G-BBVI	13,073	14,106	13,960	0.1505	0.1539	0.1520
CDR-LCG-MLE	***12,769***	14,189	–	***0.1465***	0.1551	–
CDR-LCG-BBVI.imp	12,788	14,109	–	0.1466	0.1547	–
CDR-LCG-BBVI	13,069	14,092	–	0.1501	*0.1526*	–

Table A3
Dundee (go-past). CDR vs. baselines, mean-squared error. CDR results shown using (E)xponential, (N)ormal, Shifted (G)amma, and non-parametric LCG response kernels, fitted using MLE, BBVI improper, and BBVI. Best-performing models within the sets of baseline and CDR models are shown in *italics*. Best-performing overall models are shown in **bold**. Daggers (†) indicate convergence failures.

	Dundee (ms)			Dundee (log-ms)		
Model	Train	Expl	Test	Train	Expl	Test
LME	44,184	39,523	42,948	0.2073†	0.2072†	0.2070†
LME-S	44,097†	39,476†	43,014†	0.2057†	0.2057†	0.2058†
GAM	43,976	***39,289***	42,704	0.2063	0.2061	0.2060
GAM-S	*43,476*	39,483	*42,180*	*0.2034*	***0.2036***	***0.2035***
CDR-E-MLE	43,094	41,322	–	0.2049	0.2046	–
CDR-E-BBVI.imp	43,094	41,338	–	0.2049	0.2046	–
CDR-E-BBVI	43,178	41,926	–	0.2054	0.2051	–
CDR-N-MLE	42,935	41,449	–	0.2045	0.2043	–
CDR-N-BBVI.imp	42,944	41,351	–	0.2048	0.2042	–
CDR-N-BBVI	42,975	40,776	–	0.2053	0.2047	–
CDR-G-MLE	***42,864***	39,844	–	0.2039	*0.2036*	–
CDR-G-BBVI.imp	***42,864***	39,876	–	0.2041	0.2037	–
CDR-G-BBVI	43,222	40,970	**40,018**	0.2054	0.2051	0.2052
CDR-LCG-MLE	42,329	41,798	–	***0.1998***	0.2061	–
CDR-LCG-BBVI.imp	42,336	41,678	–	0.2002	0.2052	–
CDR-LCG-BBVI	42,995	39,540	–	0.2047	0.2044	–

Table A4
Reading data model comparison (Dundee first pass). Permutation tests of improvement on test set from CDR-G-BBVI over baselines (pooled across all tasks), along with binomial tests of the probability of CDR models improving over each baseline, aggregating over training and exploratory sets (those on which all CDR models are evaluated).

	Permutation test	Binomial test	
Baseline	*p*	Success rate	*p*
LME	**1.0e-4*****	0.96	**2.2e-16*****
LME-S	**1.0e-4*****	0.91	**2.2e-16*****
GAM	**1.0e-4*****	0.72	**1.1e-5*****
GAM-S	**1.0e-4*****	0.59	**0.02***

Table A5
Reading data model comparison (Dundee go-past). Permutation tests of improvement on test set from CDR-G-BBVI over baselines (pooled across all tasks), along with binomial tests of the probability of CDR models improving over each baseline, aggregating over training and exploratory sets (those on which all CDR models are evaluated).

	Permutation test	Binomial test	
Baseline	*p*	Success rate	*p*
LME	**1.0e-4*****	0.88	**9.2e-16*****
LME-S	**1.0e-4*****	0.86	**6.0e-14*****
GAM	**1.0e-4*****	0.86	**6.0e-14*****
GAM-S	**1.0e-4*****	0.65	**0.003****

Table A6
Reading data model comparison (Dundee, all duration types). Permutation tests of improvement on test set from CDR-G-BBVI over baselines (pooled across all tasks), along with binomial tests of the probability of CDR models improving over each baseline, aggregating over training and exploratory sets (those on which all CDR models are evaluated). **Note**: Dundee scan path durations are excluded from the LME-S comparison because runtime limits prevented training from terminating.

	Permutation test	Binomial test	
Baseline	*p*	Success rate	*p*
LME	**1.0e-4*****	0.90	**2.2e-16*****
LME-S	**1.0e-4*****	0.87	**2.2e-16*****
GAM	**1.0e-4*****	0.67	**2.2e-6*****
GAM-S	**1.0e-4*****	0.36	0.36

Performance comparisons for first pass and go-past models of Dundee are given in Tables A2 and A3 (see §7 for results in Natural Stories and Dundee scan path durations). CDR mostly outperforms the LME baselines in exploratory set error on first pass durations and on log-transformed go-past durations. Exploratory set performance on linear go-past durations is generally below that of the baselines, although test set performance is substantially better than that of all baselines. GAM generally achieves similar generalization performance to CDR, while GAM-S generally improves upon CDR's generalization performance.

The outcome of our primary model comparison (permutation testing of test-set error from CDR-G-BBVI vs. baselines) is unchanged from Table 5: CDR significantly outperforms all baselines on aggregated Natural Stories and Dundee reading data, whether considering first pass durations (Table A4), go-past durations (Table A5), scan path durations (Table 5), or all three duration types combined (Table A6). Thus, results still support the claim of overall generalization improvement from using CDR over baselines.

The binomial test for training and exploratory set improvement from arbitrarily chosen CDR models reveals a significant rate of improvement from CDR over LME, LME-S, and GAM, no matter which duration type(s) are used for Dundee (Tables A4, A5, 5, A6). In addition, CDR outperforms GAM-S at a significant rate when scan paths are not considered (Tables A4 and A5). However, CDR fails to outperform GAM-S at a significant rate when scan paths are considered (Tables 5 and A6). The strong performance of GAM-S vs. CDR in Dundee is possibly due to partially non-overlapping strengths and weaknesses of the two models (see §7.3 for discussion). While an arbitrarily chosen hyperparameterization for CDR for Dundee scan paths does not provide an expected performance improvement over GAM-S, CDR performance is still strong in this domain (e.g. relative to the other baselines), justifying its application when temporal diffusion plausibly affects the measured response.

C.3. fMRI

Full results for the Natural Stories fMRI dataset (§8) are shown in Fig. A15. Results are highly consistent across kernels, error distributions, and estimation methods. As in the reading data, sinh-arcsinh substantially improves goodness of fit of modeled error distribution (Fig. A16).

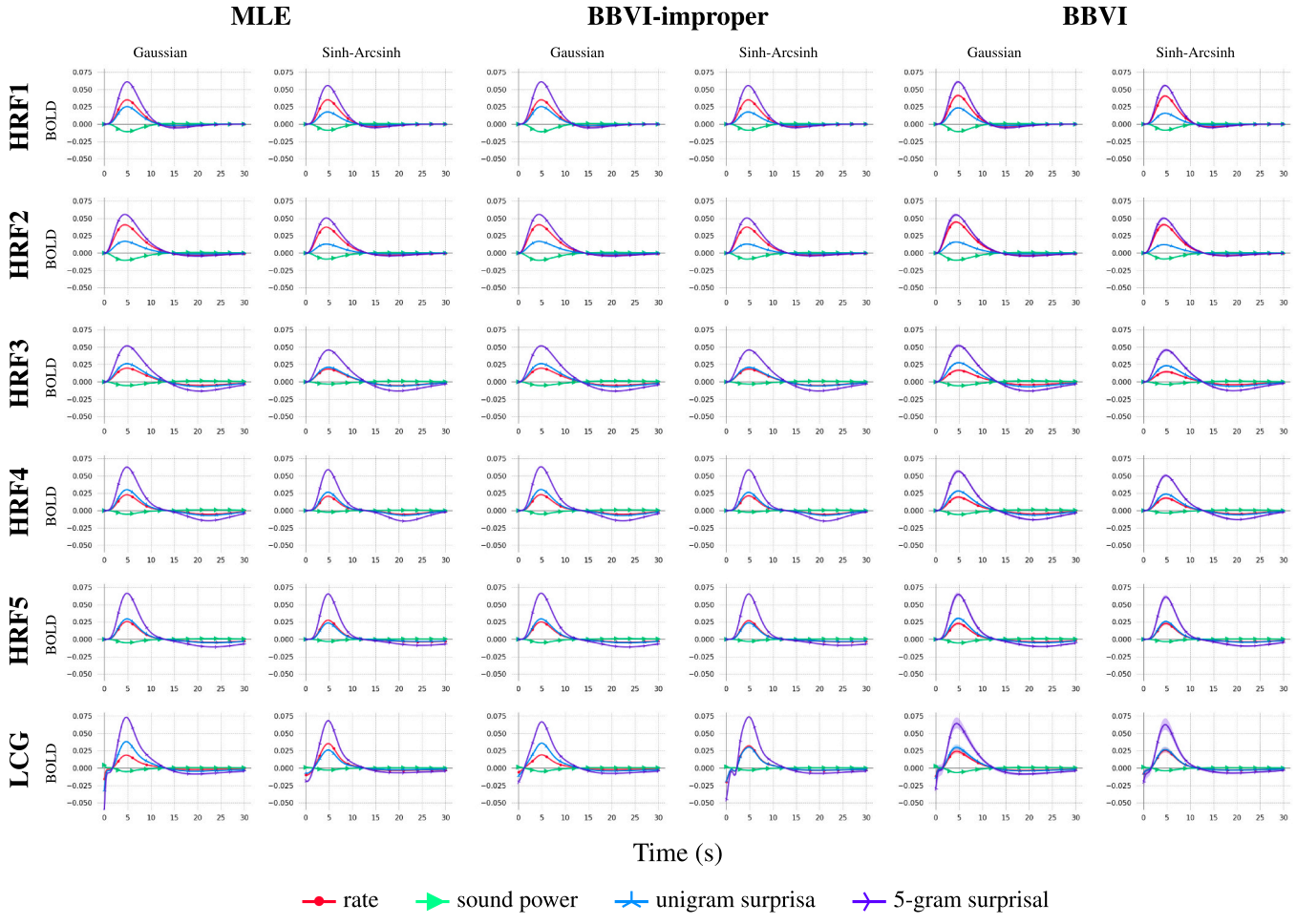


Fig. A15. Natural Stories fMRI: HRF estimates using Gaussian vs. sinh-arcsinh error distributions and various hemodynamic response kernels.

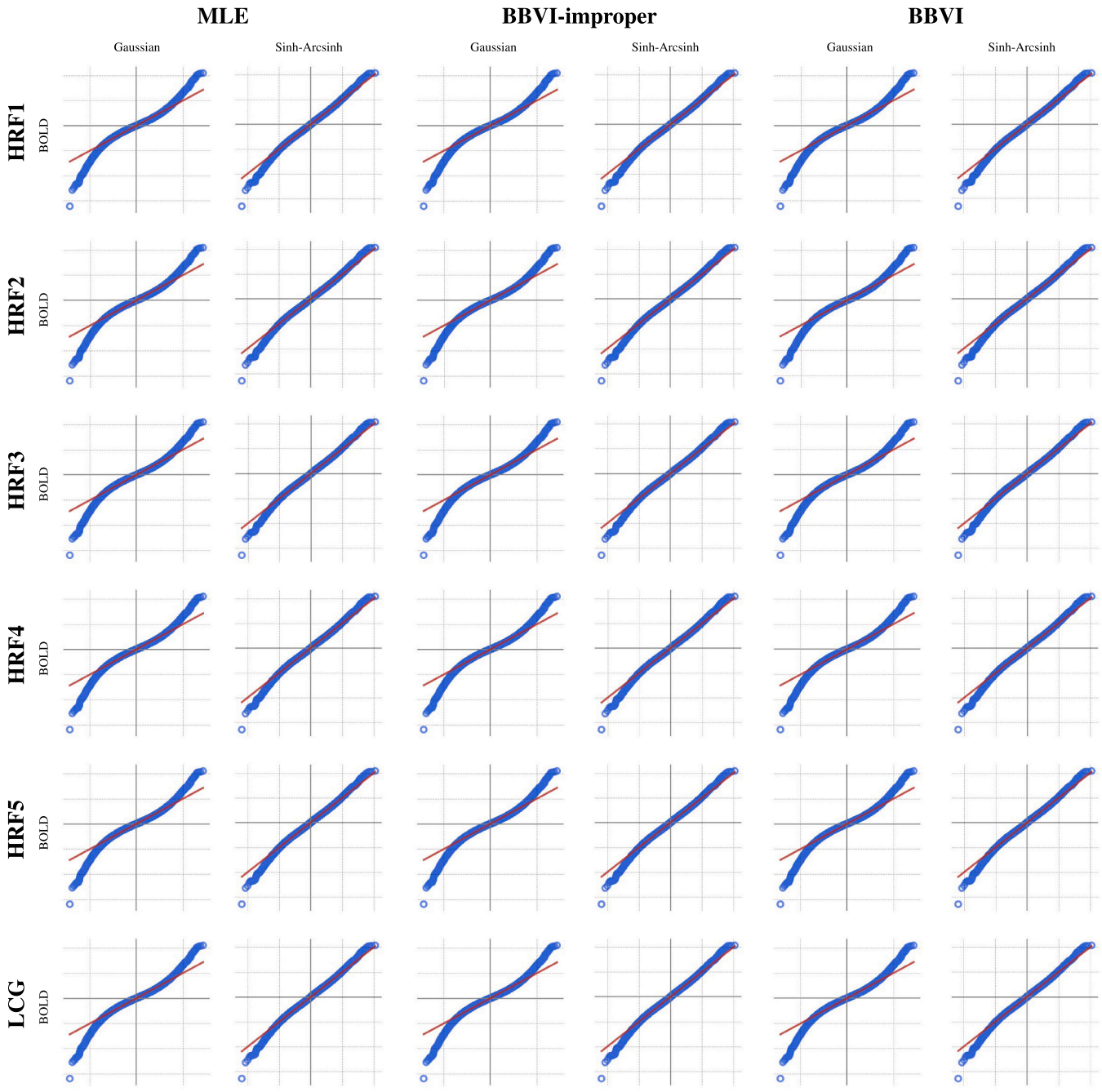


Fig. A16. Natural Stories fMRI: Quantile-quantile plots of fitted (*x*-axis) to true (*y*-axis) error distributions on the training set, using Gaussian vs. sinh-arcsinh error distributions and various hemodynamic response kernels. The theoretical best-fit line is plotted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table A7
Ablative testing results against null hypothesis of no effect for *5-gram surprisal*, using three different CDR-appropriate testing procedures. Mean and 95% credible intervals for the CDR effect estimate g' for *5-gram surprisal* are shown in the g' columns. Rejections of the null are shown in **bold**. Daggers (†) indicate convergence failures in one or both models. Plusses (+) indicate conceptual reproductions of tests in Shain (2019).

			g'			*p*-value		
Dataset	Response	Error	Mean	2.5%	97.5%	Direct PT	2-Step LRT	2-Step PT
Natural Stories	ms	Gaussian	0.729	0.706	0.753	1.0	**2.2e-16*****†	1.0†
Natural Stories	ms	sinh-arcsinh	2.47	2.42	2.53	1.0	**2.2e-16*****†	0.44†
+Natural Stories	log-ms	Gaussian	0.011	0.011	0.011	**3.0e-4*****	**2.2e-16*****	1.0
Natural Stories	log-ms	sinh-arcsinh	0.009	0.009	0.009	**1/0e-4*****	**2.2e-16*****	1.0
Dundee (FP)	ms	Gaussian	3.52	3.32	3.71	**1.0e-4*****	**2.6e-14*****†	0.074†
Dundee (FP)	ms	sinh-arcsinh	1.75	1.69	1.80	**1.0e-4*****	**3.1e-8*****†	1.0†
Dundee (FP)	log-ms	Gaussian	0.012	0.011	0.013	**2.0e-4*****	**2.6e-14*****	**0.016***
Dundee (FP)	log-ms	sinh-arcsinh	0.011	0.010	0.011	**1.0e-4*****	**2.9e-12*****	0.50
Dundee (GP)	ms	Gaussian	5.75	5.42	6.05	**1.0e-4*****	**4.0e-13*****†	0.99†
Dundee (GP)	ms	sinh-arcsinh	1.93	1.88	1.98	**1.0e-4*****	**8.7e-8*****†	0.39†
+Dundee (GP)	log-ms	Gaussian	0.018	0.017	0.019	**2.0e-4*****	**6.9e-15*****	0.52
Dundee (GP)	log-ms	sinh-arcsinh	0.014	0.013	0.015	**1.0e-4*****	**9.7e-11*****	0.58
Dundee (SP)	ms	Gaussian	3.88	3.78	3.98	**1.0e-4*****	**2.2e-16**†	0.76†
Dundee (SP)	ms	sinh-arcsinh	0.703	0.592	0.816	**1.0e-4*****	1.0†	0.42†
Dundee (SP)	log-ms	Gaussian	0.009	0.008	0.009	**1.0e-4*****	**1.7e-13*****	0.38
Dundee (SP)	log-ms	sinh-arcsinh	0.006	0.006	0.007	**1.0e-4*****	**9.0e-12*****	0.41
fMRI	BOLD	Gaussian	0.180	0.175	0.184	**1.0e-4*****	**1.0e-9*****	**1.0e-4*****
fMRI	BOLD	sinh-arcsinh	0.140	0.137	0.144	**1.0e-4*****	**2.0e-8*****	**5.0e-4*****

C.4. Hypothesis testing

Table A7 shows the full set of hypothesis testing results, including sinh-arcsinh error distributions and all duration definitions applied to the Dundee corpus.

Appendix D. Random effects in CDR: A case study

Random effects estimation is an essential feature of statistical modeling approaches in behavioral psycholinguistics and cognitive neuroscience, and it is therefore critical that CDR support it. Beyond this, the use of random effects in our models of reading and fMRI data is not central to any of our core claims, and detailed empirical evaluation of the contribution of random effects to all models considered here is beyond the scope of this study. Nonetheless, as a sanity check, in this section we provide a case study of random effects in one of our model configurations: the 5-parameter BBVI-estimated HRF model of fMRI data (HRF5), selected because it was the model used in our critical fMRI evaluation.

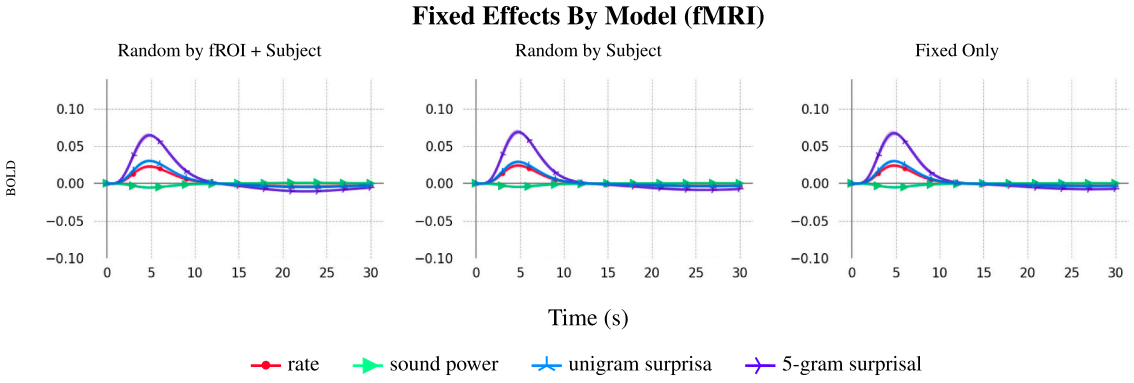


Fig. A17. Fixed HRF estimates across model configurations in decreasing order of random effects complexity. Estimates are highly stable across designs.

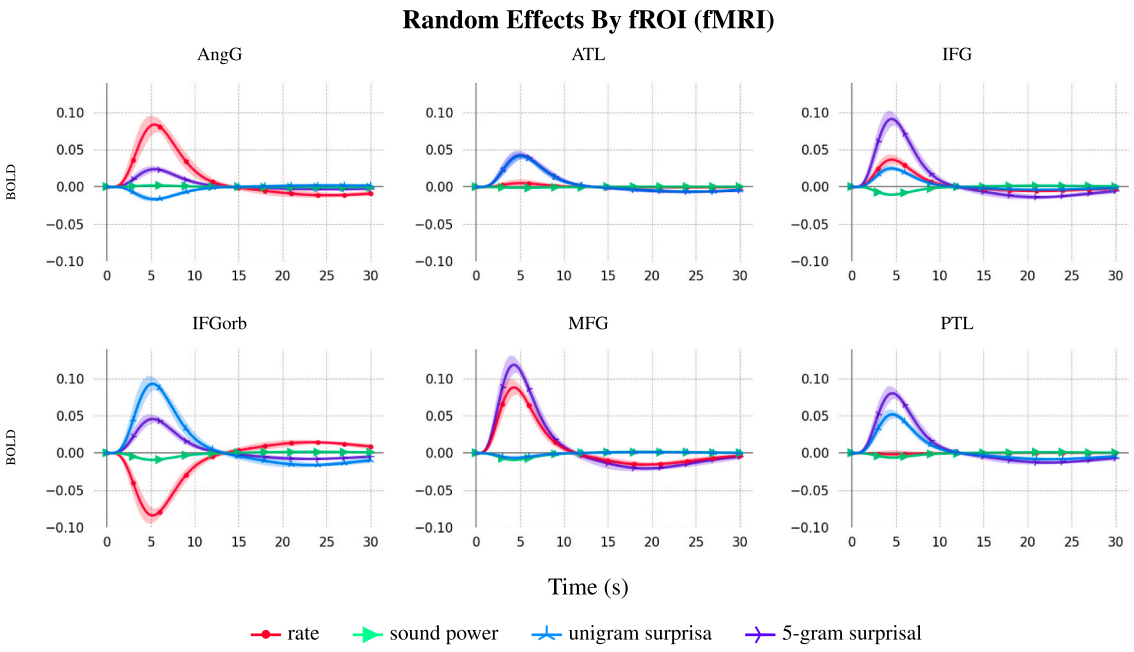


Fig. A18. HRF estimates by fROI in the full (fROI + subjects) model. Results show substantial variability between estimates for angular gyrus (AngG), anterior temporal lobe (ATL), inferior frontal gyrus (IFG), the orbital part of IFG (IFGorb), middle frontal gyrus (MFG), and posterior temporal lobe (PTL).

As discussed in §8, the HRF5 model contains random IRFs by fROI and random intercepts by subject. We explore the influence of these random terms by sequentially ablating them, first removing the random effects by fROI, then removing the random effects by subject (leaving a fixed-effects-only model).[49] As shown in Fig. A17, all models find similar estimates, indicating that our main results do not depend critically on our particular random effects design. Random HRF estimates from the full model (Fig. A18) show that this is not due to over-constraining priors collapsing the random effects: effect sizes — and, to a lesser extent, shapes — vary between regions.

The influence of random effects design on model performance is given in Table A8, which shows the in-sample (Train) and out-of-sample (Expl) error from the CDR models in comparison to the canonical HRF baseline (the best performing baseline from §8.5) in which the random effects have been comparably manipulated. CDR achieves better in-sample and out-of-sample error over each baseline with matched random effects. Both models show evidence that random effects can drive overfitting in this domain: in both model types, training set error decreases with additional random effects terms, indicating better fit to the training data, but exploratory set does not show the same pattern. The best-generalizing design is the fixed-effects-only model for both CDR and LME, which in both cases has the highest in-sample error but lowest out-of-sample error, suggesting that the other models have overfit to the training data. Nevertheless, in both model types, adding random effects by fROI improves over a variant with by-subject random effects only. For both CDR and LME, paired permutation tests show a statistically significant degradation in out-of-sample performance from the fixed-effects model to the by-fROI model and a statistically significant out-of-sample performance improvement from the by-fROI model to the by-fROI, by-subjects model ($p = 0.0001$*** in all comparisons). In addition, performance changes as a function of random effects design to a similar degree in both CDR and Canonical HRF models (the latter of which are implemented via LME). Together, these results suggest a qualitatively similar influence of random effects design in both CDR and LME.

Table A8
Model performance (mean squared error) on fMRI data as a function of random effects design.

Model	By-fROI	By-Subject	Natural Stories (fMRI)	
			Train	Expl
Canonical HRF			11.4269	*11.7526*
Canonical HRF		+	11.3857	11.8451
Canonical HRF	+	+	*11.3548*	11.8263
CDR-HRF5-BBVI			11.3571	***11.6546***
CDR-HRF5-BBVI		+	11.3185	11.7239
CDR-HRF5-BBVI	+	+	***11.2774***	11.6928

In summary, our analysis shows that (1) fixed effects estimates are highly stable across different random effects definitions, (2) effect sizes and shapes vary substantially between random effects levels (fROIs), (3) the influence of random effects on predictive performance in CDR is similar to their influence in a linear mixed-effects comparison case, and (4) random effects can harm generalization performance (i.e. overfit to the training data), both in CDR and linear mixed models. Systematic investigation of the influence of different model configurations and datasets on fixed vs. random effects estimates is left to future work.

[49] This is of course not exhaustive – we could have ablated by-subject effects and kept by-fROI effects, simplified by-fROI effects, etc. Our purpose here is to explore an illustrative subset of possible random effects configurations.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2015). *TensorFlow: large-scale machine learning on heterogeneous distributed systems*.

Ba, J. L., Kiros, J. R., & Hinton, G. E.. *Layer normalization. arXiv preprint.* (2016). *arXiv:1607.06450*.

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language, 94*(Supplement C), 206–234.

Baayen, R. H., van Rij, J., de Cat, C., & Wood, S. (2018). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. In D. Speelman, K. Heylen, & D. Geeraerts (Eds.), *Mixed effects regression models in linguistics*. Berlin: Springer.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *Neuroimage, 20*(2), 1052–1063.

Bergstrom, A. R. (1984). Continuous time stochastic models and issues of aggregation over time. In Griliches, & Intriligator (Eds.), *vol. 2. Handbook of econometrics* (pp. 1145–1212). Elsevier.

Blank, I., Balewski, Z., Mahowald, K., & Fedorenko, E. (2016). Syntactic processing is distributed across the language system. *Neuroimage, 127*, 307–323.

Boston, M. F., Hale, J. T., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research, 2*(1), 1–12.

Bouma, H., & De Voogd, A. H. (1974). On the control of eye saccades in reading. *Vision Research, 14*(4), 273–284.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B: Methodological, 26*(2), 211–243.

Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience, 16*(13), 4207–4221.

Braze, D., Mencl, W. E., Tabor, W., Pugh, K. R., Constable, R. T., Fulbright, R. K., … Shankweiler, D. P. (2011). Unification of sentence processing via ear and eye: An fMRI study. *cortex, 47*(4), 416–431.

Breen, M. (2014). Empirical investigations of the role of implicit prosody in sentence processing. *Lang & Ling Compass, 8*(2), 37–50.

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language, 120*(2), 163–173.

Brennan, J., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language, 157*, 81–94.

Campbell, K. L., & Tyler, L. K. (2018). Language-related domain-specific and domain-general systems in the human brain. *Current Opinion in Behavioral Sciences, 21*, 132–137.

Cho, S.-J., Brown-Schmidt, S., & Lee, W.-y. (2018). Autoregressive generalized linear mixed effect models with crossed random effects: An application to intensive binary time series eye-tracking data. *Psychometrika, 83*(3), 751–771.

Cooper, J. C. B. (2005). The poisson and exponential distributions. *Mathematical Spectrum, 37*(3), 123–125.

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods, 49*(2), 602–615.

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

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

Dozat, T. (2016). Incorporating Nesterov momentum into Adam. In *ICLR workshop*.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20*(6), 641–655.

Embrechts, P., Liniger, T., & Lu, L. (2011). Multivariate Hawkes processes: an application to financial data. *Journal of Applied Probability, 48A*, 367–378.

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

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology, 104*(2), 1177–1194.

Fossum, V., & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics*. Association for Computational Linguistics.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science, 22*(6), 829–834.

Frank, S. L., Monsalve, I. F., Thompson, R. L., & Vigliocco, G. (2013a). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods, 45*(4), 1182–1190.

Frank, S. L., Monsalve, I. F., Thompson, R. L., & Vigliocco, G. (2013b). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods, 45*(4), 1182–1190.

Friston, K. J., Fletcher, P., Oliver, J., Holmes, A., Rugg, M. D., & Turner, R. (1998). Event-related fMRI: Characterizing differential responses. *Neuroimage, 7*(1), 30–40.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping, 2*(4), 189–210.

Friston, K. J., Josephs, O., Rees, G., & Turner, R. (1998). Nonlinear event-related responses in fMRI. *Magnetic Resonance in Medicine*, 41–52.

Friston, K. J., Mechelli, A., Turner, R., & Price, C. J. (2000). Nonlinear responses in fMRI: The balloon model, volterra kernels, and other hemodynamics. *NeuroImage, 12*(4), 466–477.

Futrell, Richard, Gibson, Edward, Tily, Harry J., Blank, Idan, Vishnevetsky, Anastasia, Piantadosi, Steven T., & Fedorenko, Evelina (2021). The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation, 55*(1), 63–77.

Gilmore, R. O., Diaz, M. T., Wyble, B. A., & Yarkoni, T. (2017). Progress toward openness, transparency, and reproducibility in cognitive neuroscience. *Annals of the New York Academy of Sciences, 1396*(1), 5–18.

Gimel'farb, G., Farag, A. A., & El-Baz, A. (2004). Expectation-Maximization for a linear combination of Gaussians. In , *vol. 3. Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004* (pp. 422–425). IEEE.

Gitelman, D. R., Penny, W. D., Ashburner, J., & Friston, K. J. (2003). Modeling regional and psychophysiologic interactions in fMRI: The importance of hemodynamic deconvolution. *Neuroimage, 19*(1), 200–207.

Glover, H. G. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage, 9*, 416–429.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on cognitive modeling and computational linguistics (CMCL 2018)* (pp. 10–18).

Gorrostieta, C., Ombao, H., Bédard, P., & Sanes, J. N. (2012). Investigating brain connectivity using mixed effects vector autoregressive models. *NeuroImage, 59*(4), 3347–3355.

Goshtasby, A., & O'Neill, W. D. (1994). Curve fitting by a sum of Gaussians. *CVGIP: Graphical Models and Image Processing, 56*(4), 281–288.

Graff, D., & Cieri, C. (2003). *English Gigaword LDC2003T05*.

Graff, D., Kong, J., Chen, K., & Maeda, K. (2007). *English Gigaword third edition LDC2007T07*.

Griliches, Z. (1967). Distributed lags: A survey. *Econometrica: Journal of the Econometric Society*, 16–49.

Grodner, D. J., & Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science, 29*, 261–291.

Handwerker, D. A., Ollinger, J. M., & D'Esposito, M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage, 21*(4), 1639–1651.

Harrison, L., Penny, W. D., & Friston, K. (2003). Multivariate autoregressive modeling of fMRI time series. *Neuroimage, 19*(4), 1477–1491.

Hasson, U., Egidi, G., Marelli, M., & Willems, R. M. (2018). Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition, 180*, 135–157.

Hasson, U., & Honey, C. J. (2012). Future trends in neuroimaging: Neural processes as expressed within real-life contexts. *NeuroImage, 62*(2), 1272–1278.

Hasson, U., Malach, R., & Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences, 14*(1), 40–48.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science, 1*(3), 297–310.

Hawkes, A. G. (1971). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B: Methodological, 33*(3), 438–443.

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 690–696). Bulgaria: Sofia.

Henderson, J. M., Choi, W., Lowder, M. W., & Ferreira, F. (2016). Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *Neuroimage, 132*, 293–300.

Henderson, J. M., Choi, W., Luke, S. G., & Desai, R. H. (2015). Neural correlates of fixation duration in natural reading: Evidence from fixation-related fMRI. *NeuroImage, 119*, 390–397.

Hu, X., & Yacoub, E. (2012). The story of the initial dip in fMRI. *Neuroimage, 62*(2), 1103–1108.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature, 532*(7600), 453–458.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *In International conference on machine learning* (pp. 448–456).

Jones, M. C., & Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika, 96*(4), 761–780.

Josephs, O., Turner, R., & Friston, K. (1997). Event-related fMRI. *Human Brain Mapping, 5*(4), 243–248.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329–354.

Kennedy, A., Pynte, J., & Hill, R. (2003). The Dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization. CoRR, abs/1412.6*.

Kolers, P. A. (1976). Buswell's discoveries. *Eye Movements and Psychological Processes*, 371–395.

Koyck, L. M. (1954). *Distributed lags and investment analysis* (vol. 4). North-Holland Publishing Company.

Kruggel, F., Wiggins, C. J., Herrmann, C. S., & von Cramon, D. Y. (2000). Recording of the event-related potentials during functional MRI at 3.0 tesla field strength. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 44*(2), 277–282.
Kruggel, F., & Yves von Cramon, D. (1999). Temporal properties of the hemodynamic response in functional MRI. *Human Brain Mapping, 8*(4), 259–271.
Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2003). *Applied linear regression models*. Boston: McGraw-Hill Higher Education.
Lapham, B. M. (2014). *Hawkes processes and some financial applications. Ph.D. thesis*. University of Cape Town.
Lindquist, M. A., Loh, J. M., Atlas, L. Y., & Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling. *NeuroImage, 45*(1, Supplement 1), S187–S198.
Lindquist, M., & Wager, T. (2007). Validity and power in hemodynamic response modeling: A comparison study and a new approach. *Human Brain Mapping, 28*, 764–784.
Logothetis, N. K. (2003). The underpinnings of the BOLD functional magnetic resonance imaging signal. *Journal of Neuroscience, 23*(10), 3963–3971.
Lopopolo, A., Frank, S. L., den Bosch, A., & Willems, R. M. (2017). Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PLoS One, 12*(5), Article e0177794.
Madisetti, V. (1997). *The digital signal processing handbook*. CRC press.
Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43*(6), 304–316.
McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference* (pp. 18–25).
Mei, H., & Eisner, J. (2017). The neural Hawkes process: A neurally self-modulating multivariate point process. In *Proceedings of the 31st international conference on neural information processing systems, NIPS'17* (pp. 6757–6767). Red Hook, NY, USA: Curran Associates Inc.
Miezin, F. M., Maccotta, L., Ollinger, J. M., Petersen, S. E., & Buckner, R. L. (2000). Characterizing the hemodynamic response: Effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *Neuroimage, 11*(6), 735–759.
Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. *New Methods in Reading Comprehension Research*, 69–89.
Mollica, F., & Piantadosi, S. (2017). An incremental information-theoretic buffer supports sentence processing. In *Proceedings of the 39th annual cognitive science society meeting*.
Morton, J. (1964). The effects of context upon speed of reading, eye movements and eye-voice span. *Quarterly Journal of Experimental Psychology, 16*(4), 340–354.
Nesterov, Y. E. (1983). A method for solving the convex programming problem with convergence rate O(1/k2). In *, vol. 269. Dokl. Akad. Nauk SSSR* (pp. 543–547).
Neter, J., Wasserman, W., & Kutner, M. H. (1989). *Applied linear regression models*. Boston: McGraw-Hill/Irwin.
Neuvo, Y., Dong, C.-Y., & Mitra, S. (1984). Interpolated finite impulse response filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 32*(3), 563–570.
Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*(694–706), 289–337.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716.
Ozaki, T. (1979). Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics, 31*(1), 145–155.
Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528–530.
Pedregosa, F., Eickenberg, M., Ciuciu, P., Gramfort, A., & Thirion, B. (2014). Data-driven HRF estimation for encoding and decoding models. *NeuroImage, 104*.
Raskutti, G., Wainwright, M. J., & Yu, B. (2014). Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *The Journal of Machine Learning Research, 15*(1), 335–366.
Rayner, K. (1977). Visual attention in reading: Eye movements reflect cognitive processes. *Memory & Cognition, 5*(4), 443–448.
Rayner, K. (1998). Eye movements in Reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422.
Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 conference on empirical methods in natural langauge processing* (pp. 324–333).
Robinson, P. M. (1975). Continuous time regressions with discrete data. *The Annals of Statistics*, 688–697.
Robinson, P. (1976). Fourier estimation of continuous time models. In *Statistical inference in continuous time economic models* (pp. 215–266).
Röther, J., Knab, R., Hamzei, F., Fiehler, J., Reichenbach, J. R., Büchel, C., & Weiller, C. (2002). Negative dip in BOLD fMRI is caused by blood flow—Oxygen consumption uncoupling in humans. *Neuroimage, 15*(1), 98–102.
Salimans, T., & Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in neural information processing systems* (pp. 901–909).
Saramaeki, T., Mitra, S. K., & Kaiser, J. F. (1993). Finite impulse response filter design. *Handbook for Digital Signal Processing, 4*, 155–277.
van Schijndel, M., & Schuler, W. (2013). An analysis of frequency- and memory-based processing costs. In *Proceedings of human language technologies: The 2013 annual conference of the North American chapter of the ACL*.
van Schijndel, M., & Schuler, W. (2015). Hierarchic syntax improves reading time prediction. In *Proceedings of NAACL-HLT 2015*. Association for Computational Linguistics.
Schotter, E. R., Leinenger, M., & von der Malsburg, T. (2018). When your mind skips what your eyes fixate: How forced fixations lead to comprehension illusions in reading. *Psychonomic Bulletin & Review, 25*(5), 1884–1890.
Scott, T. L., Gallée, J., & Fedorenko, E. (2017). A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive Neuroscience, 8*(3), 167–176.
Shain, C. (2019). A large-scale study of the effects of word frequency and predictability in naturalistic reading. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long and Short Papers)* (pp. 4086–4094).
Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia, 138*.
Shain, C., & Schuler, W. (2018). Deconvolutional time series regression: A technique for modeling temporally diffuse effects. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2679–2689).
Shain, C., van Schijndel, M., Futrell, R., Gibson, E., & Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the computational linguistics for linguistic complexity workshop* (pp. 49–58). Association for Computational Linguistics.
Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76–80.
Sims, C. A. (1971). Discrete approximations to continuous time distributed lags in econometrics. *Econometrica: Journal of the Econometric Society*, 545–563.
Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, 1–48.
Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In *, vol. 30. Proceedings of the annual meeting of the cognitive science society*.
Smith, N. J., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the 33rd CogSci conference*.
Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*, 302–319.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.
Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Lang & Ling Compass, 9*(8), 311–327.
Tran, D., Kucukelbir, A., Dieng, A. B., Rudolph, M., Liang, D., & Blei, D. M.. *Edward: A library for probabilistic modeling, inference, and criticism*. (2016). *arXiv preprint arXiv:1610.09787*.
Vagharchakian, L., Dehaene-Lambertz, G., Pallier, C., & Dehaene, S. (2012). A temporal bottleneck in the language comprehension network. *Journal of Neuroscience, 32*(26), 9089–9102.
Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language, 82*(4), 767–794.
Ward, B. D. (2006). *Deconvolution analysis of fMRI time series data*.
Wehbe, L., Blank, I., Shain, C., Futrell, R., Levy, R., von der Malsburg, T., … Fedorenko, E. (2021). Neural activity in the fronto-temporal language network is predicted by incremental language comprehension difficulty. *Cerebral Cortex*. https://academic.oup.com/cercor/advance-article/doi/10.1093/cercor/bhab065/6249820?guestAccessKey=1f6090c3-6fab-4326-8cba-7ea7ca2bae02.
Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics, 9*(1), 60–62.
Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & den Bosch, A. (2015). Prediction during natural language comprehension. *Cerebral Cortex, 26*(6), 2506–2516.
Yacoub, E., Shmuel, A., Pfeuffer, J., Van De Moortele, P.-F., Adriany, G., Ugurbil, K., & Xiaoping, H. (2001). Investigation of the initial dip in fMRI at 7 tesla. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo, 14*(7–8), 408–412.
Yao, Y., Rosasco, L., & Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation, 26*(2), 289–315.
Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122.
Zhou, K., Zha, H., & Song, L. (2013). Learning triggering kernels for multi-dimensional Hawkes processes. In *30th international conference on machine learning, ICML 2013* (pp. 2338–2346).