# Gaussian Processes for time-marked time-series data

**John P. Cunningham** **Zoubin Ghahramani** **Carl E. Rasmussen**
Department of Engineering, University of Cambridge, Cambridge, UK

## Abstract

In many settings, data is collected as multiple time series, where each recorded time series is an observation of some underlying dynamical process of interest. These observations are often time-marked with known event times, and one desires to do a range of standard analyses. When there is only one time marker, one simply aligns the observations temporally on that marker. When multiple time-markers are present and are at different times on different time series observations, these analyses are more difficult. We describe a Gaussian Process model for analyzing multiple time series with multiple time markings, and we test it on a variety of data.

## 1 Introduction

In many settings, data is collected as multiple time series, such as repeated measurements of a variable that fluctuates daily, or an experimental trial in some scientific application. For all such trials $n \in \{1, ..., N\}$, each recording $y^{(n)}(t)$ is a time series observation of some underlying process of interest. These observations are often time-marked with known event times $m_k^{(n)}$ for events $k \in \{1, ..., K\}$. As a concrete example which we will use in this work, consider an experiment where the velocity $y(t)$ of an arm is recorded during a reaching experiment, where reaches are repeatedly made from a central target out to a peripheral target, and back again. The events that occur during each experimental trial include markers $m_k^{(n)}$ such as the start time of trial $n$ and the times when the subject's movement begins in one direction or another. The process of interest is the underlying dynamical response to these events. Here we seek to effectively model this and other settings.

Appearing in Proceedings of the $15^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. 

When there is only one time marker (such as trial start or beginning of a day), analysis can be simply done by aligning the observations temporally. Multiple time markers at different times on different trials complicate matters. In the reaching example, if trial $n$ begins at $m_1^{(n)} = t_{\text{start}}$, the subject begins to move outwardly at some random time $m_2^{(n)} = t_{\text{move}}$ (these times vary due to reaction time and imposed randomness in experimental cues), and then after another variable reaction time the subject moves back to the central target at $m_3^{(n)} = t_{\text{return}}$, how can we describe an average response in this experiment? Certainly $t_{\text{move}}$ correlates to $t_{\text{start}}$ via the subject's reaction time, and also $t_{\text{move}}$ will influence the subject's readiness to make an inbound movement at $t_{\text{return}}$. Figure 1 shows four observations ($N = 4$) of a time-marked data process with three markers ($K = 3$). These observations are behavioral data from the experiment described above, but the occurrence of different markers at different times makes it difficult to analyze this data by just considering one notion of time (*i.e.*, time with respect to only one marker). For example, it appears that aligning the observations to the first or third marker (the squares in the left panel or diamonds in the right panel) describe different parts of the data best, but in general this can not be determined. Further, we desire a method that can learn these relationships automatically from data and is not restricted to one simple alignment.

Typically, efforts to analyze this "time-marked data" involve clipping the data in a time window around each marker $m_k^{(n)}$ and analyzing within those windows across $n$, thereby treating each time series observation as a set of independent time series $y_k^{(n)}(t)$ recorded serially, and losing any shared statistical power across temporal epochs (this is very common in experimental sciences; see for example Sugrue et al. (2004); Churchland et al. (2006)). Here we describe a model for time series with multiple time markings. We treat each time series not as a univariate time series with the single input of absolute time $t$, but rather as a multidimensional time-series where each input dimension is time with respect to a particular marker, which corresponds to augmenting the definition of

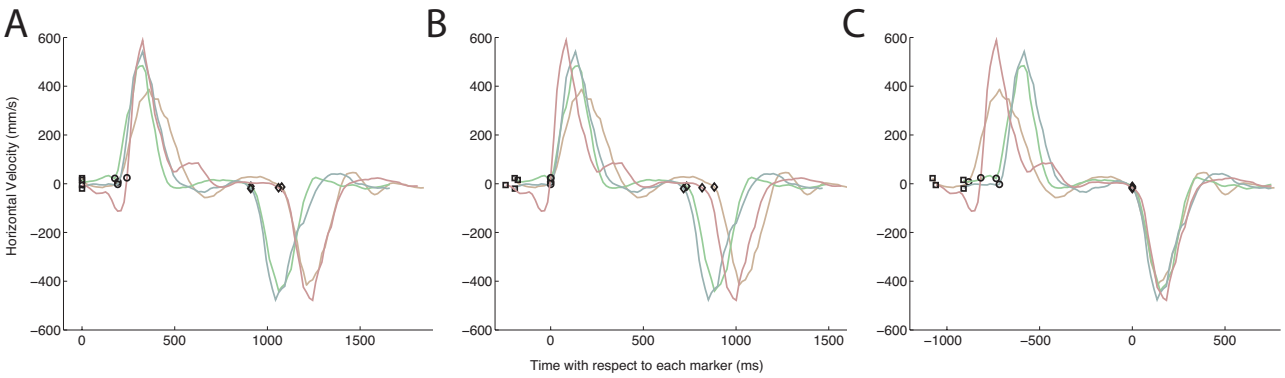


Figure 1: Four observations of time-marked data ($N = 4$). The squares, circles, and diamonds correspond to each of the three ($K = 3$) markers. In the left panel, all four time series observations are aligned to the first square marker $m_1$. In the middle and right panels, the same observations are plotted with respect to the occurrence of the second and third markers, respectively. These four observations have been shared with us from a larger set of behavioral experiments (Cunningham et al., 2011) and are part of the first data set that will be analyzed in this work.

$y(t)$ to a $K$-dimensional input $y^{(n)}(t, \mathbf{m}) = y^{(n)}(\mathbf{t}) = y^{(n)}([t - m_1^{(n)}, ..., t - m_K^{(n)}])$. Gaussian Processes (see Rasmussen and Williams (2006) for a thorough background) allow a convenient and natural framework for regression of a signal $y(\cdot)$ against time in this multidimensional extension, and we model time-marked time series within this GP framework.

While this approach will address the general problem, to accurately model this sort of data, more consideration is warranted. Specifically, a GP (and our basic multidimensional time-marked GP) is acausal, in that the signal $y^{(n)}(t)$ is influenced by the times of future events. While appropriate in some scenarios, in many settings it is desirable to enforce causality. In a *causal* multidimensional model, the observation could only be influenced by the occurrence of a particular marker after that marker occurs. We develop a causal GP model. While other non-parametric Gaussian Processes have been studied (Sampson and Guttorp, 1992; Nott and Dunsmuir, 2002; Schmidt and O'Hagan, 2003; Murray-Smith and Pearlmutter, 2003; Paciorek and Schervish, 2006, 2003), we consider in depth the particular case of causal GPs.

Of course, GP are not the only possible framework for analyzing such a model. Other statistical frameworks such as splines could be brought to bear on the problem of time-marked data and for causality. However, GP offer a principled non-parametric approach that incorporates causality and these multiple time-markers in a natural way. Furthermore, other attractive features of Bayesian non-parametrics are inherited by our choice of GP: for example, the automatic relevance determination (ARD) properties of learning lengthscale hyperparameters of particular covariance kernels can lend insight about the relevance of particular time markers to the measured behavioral response.

Our construction for causal GP models of time-marked data is simple, analytically tractable, and computationally efficient, and it recovers classical smoothing procedures as a special case. We demonstrate the model and test it on different types of data. This approach outperforms conventional methods. Thus, consideration of temporal event markers may have impact across a range of time series analyses.

## 2 Models for Time-Marked Data

We model time-marked time series as data over a multidimensional input space, where each input dimension is defined as time with respect to each given time marker or event time. As such, this GP is a conditional model, conditioned on both the event-markers and input times. Specifically, we say a collection of $N$ time series $\{y^{(n)}(t)\}_{n=1,...,N}$ with given time markers $\{m_k^{(n)}\}_{\substack{n=1,...,N \\ k=1,...,K}}$ are drawn from a GP with time-marked covariance $k_{\mathrm{TM}}$:

$$k_{\mathrm{TM}}\left(t_i^{(p)}, t_j^{(q)}\right) = k\left(t_i^{(p)} - \begin{bmatrix} m_1^{(p)} \\ \vdots \\ m_K^{(p)} \end{bmatrix}, t_j^{(q)} - \begin{bmatrix} m_1^{(q)} \\ \vdots \\ m_K^{(q)} \end{bmatrix}\right) \tag{1}$$

where the superscripts $p$ and $q$ indicate that the data points in question come from two possibly different

time series observations. We have used the subscript TM to highlight the fact that a time-marked GP is simply a particular choice of covariance. For example, we can choose a squared exponential kernel with lengthscales $l_k$, and then:

$$k_{\mathrm{TM}}\left(t_i^{(p)}, t_j^{(q)}\right) = \sigma^2 \exp\left\{-\sum_{k=1}^{K} \frac{1}{2l_k^2}\left(\left(t_i^{(p)} - m_k^{(p)}\right) - \left(t_j^{(q)} - m_k^{(q)}\right)\right)^2\right\}. \quad (2)$$

### 2.1 Causal Gaussian Processes

As noted above, modeling causality is often a natural desire in these settings. Formally, we define a causal GP $y(\mathbf{t}) \sim \mathcal{CGP}(0,k)$ where $\mathbf{t} \in I\!R^K$ and $k(\cdot,\cdot)$ is a stationary, positive semi-definite kernel function. Unlike a typical stationary GP, a causal GP has the nonstationarity property that, for a given $\omega \in \Omega$ (a single outcome from the sample space), $y(\mathbf{t}_1,\omega) = y(\mathbf{t}_2,\omega) \;\; \forall \;\; \mathbf{t}_1,\mathbf{t}_2 : (\mathbf{t}_1)_+ = (\mathbf{t}_2)_+$, where $(\cdot)_+$ is the positive part of the vector $\mathbf{t}$. In words, this means that the signal is flat in the direction of each dimension in the negative halfplane of that dimension. For simplicity, in this definition we have defined causality with respect to the origin $\mathbf{t} = \mathbf{0}$, but this can be generalized to any point by replacing $(\cdot)_+$ with $(\cdot)_{>\gamma}$, or to enforce anticausality with $(\cdot)_-$. It is also worth noting that this definition of causality is equivalent to enforcing that a draw $y(\mathbf{t})$ will have $\frac{\partial y(\mathbf{t})}{\partial t_k} = 0 \; \forall \; t_k < 0$ (or $t_k < m_k$ in the non-origin case). In one dimension, this definition reduces to the familiar notion of a signal $y(t)$ that is constant until it causally responds to an event $m_k$ (or the origin).

To model a causal GP, one might first consider conditioning on fictitious observations of a fixed value in the past, effectively constraining that part of the draw. Despite closure under conditioning being a convenient and frequently used property of the Gaussian, in the context of a causal GP, conditioning is theoretically and practically inappropriate (see for example Popov (2008)), and thus we will not consider this possible model further.

Instead, one might choose to warp the input space to enforce causality, namely $y(t) = x(h(t))$, where $h(t) = t\,\mathcal{I}(t > 0)$, and $x(t)$ is a stationary GP. These time warpings, also sometimes called spatial deformations (Sampson and Guttorp, 1992; Schmidt and O'Hagan, 2003), preserve the GP by definition. These other works focus on inferring nonstationary warping/deformation mappings using splines and multidimensional scaling (Sampson and Guttorp, 1992) or with a GP mapping true space $t$ to deformed space $h(t)$ (Schmidt and O'Hagan, 2003). The focus of those works is thus very different from this work, where we presume a fixed, known warping. Furthermore, the existence of a fixed, simple warping simplifies inference and learning considerably.

One might also consider a nonstationary linear filter $y(t) = \int x(u)g(t,u)du$, for a suitably defined filter $g(t,u)$. This filter can be any function, but to enforce causality as defined above we use:

$$g(t,u) = \begin{cases} \delta(u) & t < 0 \\ \delta(t-u) & t \geq 0 \end{cases} \quad (3)$$

where $\delta(\cdot)$ is the Dirac delta. If $x(t)$ is a GP with covariance $k_x$, then $y(t)$ is a causal GP $y(t) \sim \mathcal{CGP}(0,k_x)$. Extending either this filtering definition or the time warping definition to higher input dimensions is trivial - either definition is just repeated on all input dimensions.

Linear filtering is a convenient, simple, and theoretically sound way to construct a causal GP. In fact, it is just a generalization of the time warping example above, which can be simply seen by using a specific filter $g(\mathbf{t},\mathbf{u}) = \delta(h(\mathbf{t})-\mathbf{u})$, from which we see that $y(\mathbf{t}) = \int g(\mathbf{t},\mathbf{u})x(\mathbf{u})d\mathbf{u} = \int \delta(h(\mathbf{t})-\mathbf{u})x(\mathbf{u})d\mathbf{u} = x(h(\mathbf{t}))$. The filter $g(t,u)$ can also express more complex relationships, such as AR or MA properties, which are an important consideration outside the scope of this work. However, the instantaneous relationship of the type $g(\mathbf{t},\mathbf{u}) = \delta(h(\mathbf{t}) - \mathbf{u})$ suffices for the causal GP setting of interest here. Thus, as far as causal GP are concerned, these two models are equivalent. Hereafter we claim this model is the simple and appropriate choice for causal GP.

Conveniently, with the time-marked model and the causal GP, we can now easily make a causal time-marked GP: $\{y^{(n)}(t)\}_{n=1,\ldots,N} \sim \mathcal{CGP}(0,k_{\mathrm{TM}})$, which amounts to replacing $(t_i - m_k^{(p)})$ with $h(t_i - m_k^{(p)})$ above in Equation 2. By construction, the causal or acausal time-marked data model is a set of mappings of the input space of a standard GP, which allows this model to inherit all conventional GP machinery. Thus we defer learning and inference to standard literature such as Rasmussen and Williams (2006). Draws from this process and the corresponding time series observations (one dimensional slices) are shown in Figure 2.

GP models typically include terms for observation noise, and that can be easily done here also by adding independent noise covariance to each observation. Because our observations themselves are time-series, it often makes sense to add temporally colored noise rather than (or in addition to) the more standard white Gaussian noise at every data point. The intuition here is that each time-series measurement may have its own

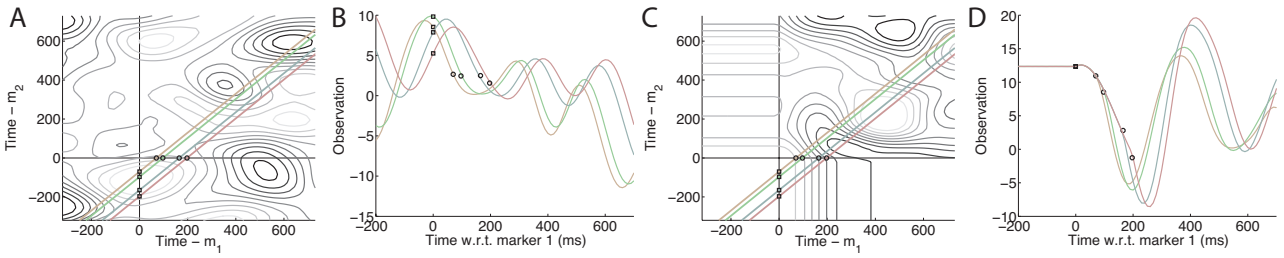


Figure 2: GP time-marked data model. Panel A shows an acausal model with lines which show the progress of time in each of the $N = 4$ trials. The projection of the surface onto these lines - namely the time series observation itself - is shown in Panel B. Panels C and D show the same features for a causal model.

temporally-varying signal that varies trial-to-trial, in addition to the common underlying dynamical process. In our experiments, we add colored noise terms modeled independently across $N$ trials with a standard squared exponential kernel. To be concrete, our final causal time-marked data covariance for our experiments will be:

$$k_{\mathrm{TM}}\left(t_i^{(p)}, t_j^{(q)}\right) = \sigma^2 \exp\left\{-\sum_{k=1}^{K} \frac{1}{2l_k^2}\left(h\left(t_i^{(p)} - m_k^{(p)}\right) - h\left(t_j^{(q)} - m_k^{(q)}\right)\right)^2\right\} + \delta(p-q)\sigma_r^2 \exp\left\{-\frac{1}{2l_r^2}\left(t_i^{(p)} - t_j^{(q)}\right)^2\right\} + \delta(p-q)\delta(t_i^{(p)} - t_j^{(q)})\sigma_v^2, \quad (4)$$

where the second term is colored noise that is 'private' to each time series observation and does not involve time markings, and the third term is conventional additive noise at every data point ($\sigma_v$, which we set to be small and do not learn in our experiments). The hyperparameters of this covariance - $\{\sigma, l_1, ..., l_K\}$ for the time-marked component and $\{\sigma_r, l_r\}$ for the noise term - can be readily learned from data jointly according to standard GP procedure. Also, this model inherits the computational complexity of other GP, but can also benefit from the literature in fast GP, such as Snelson and Ghahramani (2006).

### 2.2 Alternative models for time-marked data

We now briefly discuss alternative ways in which such data has been treated, which will serve as our comparison in the following experiments. First, the simplest algorithm for analyzing time-marked data is to simply ignore the presence of time markers altogether. We can treat the data as conventional time series and use GP regression accordingly. This choice is a special case of a time-marked data model where $K = 1$. One can also do basic averaging of the time series observations instead of GP regression, since we have all datasets at a fine enough resolution to average all training data directly.

Second, when time-markers are recognized as important features in a data set, often it is handled by 'clipping' the data: ad hoc temporal windows are chosen around each marker, and data is treated as a set of $K$ independent time series, with each trial aligned to the $k$th marker in the $k$th window. On each of these clipped windows, GP regression or simple averaging can be used interchangeably. Clipping models have often been used in experimental science (for example Sugrue et al. (2004); Churchland et al. (2006)). Despite its simple description, this model has a number of critical drawbacks. First, there is no notion of causality, and in addition the time of other events can not effect the analysis of the data belonging to a particular window. Second, this clipping model suffers from an inability to share statistical power across time epochs. Third and perhaps most importantly, the actual size of the windows for analysis can not be chosen without guessing. Finally, also critically, any choice of windows will necessary over-count or under-count particular pieces of data, which confounds analysis considerably.

## 3 Experiments and Results

We have introduced causal GP and their application to analyzing time-marked time series data. To validate the utility of the model, we now show regression on a range of time-marked data sets. First, we describe the experimental conditions and sources of these data, and then we show results for our GP methods and for many other possible methods. In all cases, the time-marked causal GP has lower error in regression on held out test data, indicating the utility of this model in experimental analysis.

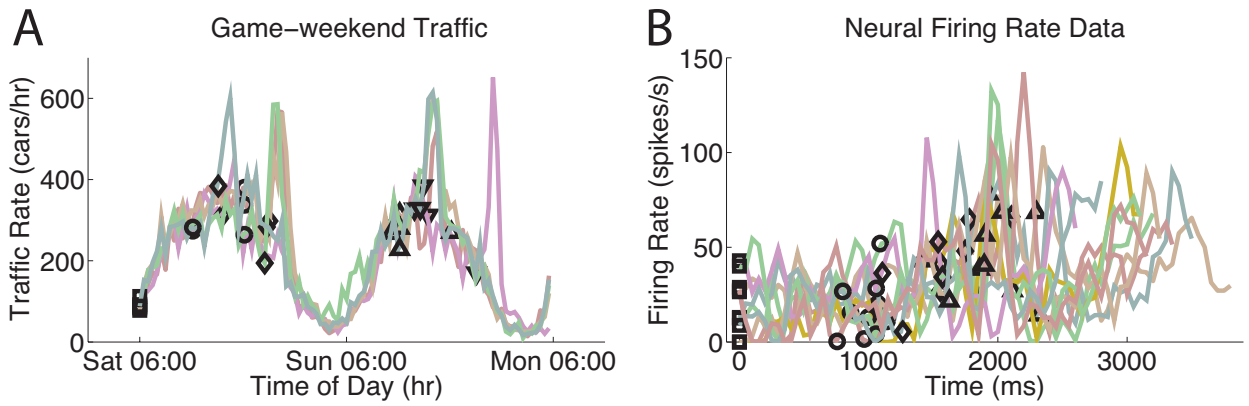


Figure 3: Panel A: five observations from the game-weekend traffic dataset of Section 3.1.2. This data has five markers: square - start of the weekend; circle - 1hr before the start of the Saturday game; diamond - 1hr before the end of the Saturday game; upward triangle - 1hr before the start of the Sunday game; downward triangle - 1hr before the end of the Sunday game. These examples show the superposition of standard daily oscillations and game-related traffic spikes that occur at variable times. Panel B: Ten observations from the neural firing rate dataset of Section 3.1.3. This data has four markers: square - start of the experimental trial; circle - time of the 'target on' cue; diamond - time of the 'go' cue; upward triangle - 150ms before the movement onset.

### 3.1 Data sets

Here we introduce the data sets that we will use to test the performance of time-marked data models.

#### 3.1.1 Arm Reach Data

The first data set involves human subjects making point-to-point arm movements between two on-screen targets - a central target in the middle of the screen and a peripheral target 80mm to the right of the central target. Each time series observation $y^{(n)}(t)$ (for $n \in \{1, ..., 13\}$ in this case) begins when the subject's hand is at rest at the central target. After 20ms, the central target is turned off, the peripheral target is turned on, and the subject reacts and moves to that target, coming to rest when touching the peripheral target. After the subject remains at the peripheral target for 500ms, the central target is again turned on, and the subject reacts and moves back to the central target. The data of interest is the horizontal velocity of the hand, shared with us directly from the experiments in Cunningham et al. (2011). These data are a larger data set ($N = 13$) from which we drew the four representative observations shown in Figure 1. This data is time marked with three markers (shown in Figure 1): the beginning of the experimental trial, the movement onset of the outbound reach, and the movement onset of the inbound reach.

#### 3.1.2 Gameday Traffic Data

Time-marked data appears in broader contexts than experimental science, such as economics and finance (where for example equity volatility may depend both on calendar events and company-specific events), and population dynamics. Here we investigate one such case of automobile traffic on a Los Angeles highway: this data measures the number of cars on the Glendale offramp of US-101 North during weekends in 2005 (this data was previously used in Ihler et al. (2006)). The data consists of the number of cars on the offramp every five minutes, which we have summed to be a data point every half hour. Five observations from this dataset are shown in Figure 3A. The particular choice of the Glendale offramp is interesting due to its proximity to the LA Dodgers baseball stadium, where weekend games see attendance of 40-55 thousand people. Figure 3A shows the expected day/night increase/decrease in traffic (time-marked with conventional "time of day"). In addition, traffic fluctuates significantly in response to the end of a game (but less so the start, given the highway configuration). Our dataset has eleven ($N = 11$) 48-hr weekend periods with both Saturday and Sunday games at Dodger Stadium. The games begin and end at variable times, and thus this "game-weekend traffic data" is time-marked with five events ($K = 5$): start of the weekend, 1hr before the start of the Saturday game, 1hr before the end of the Saturday game, 1hr before the start of the Sunday game, and 1hr before the end of the Sunday game. An hour before these events is used because that is a sensible time to consider the effect of a base-

ball game on traffic, as people tend to arrive and leave early. With this dataset we also analyze just the Saturday "gameday traffic data" to add another set of results. Here we use just the first 24hrs of data, and only the events from Saturday games ($K = 3$).

#### 3.1.3 Neural Firing Rate Data

The last data set is electrophysiological recordings of a neuron's spiking activity. These data, shown in Figure 3B, were recorded from a subject's motor cortex while the subject did $N = 10$ identical curved reaches. Spiking data was shared with us from and recorded as described in Churchland et al. (2010) (neuron 184, condition 24), and firing rates were calculated by smoothing the point process data with a Gaussian kernel of $\sigma = 26$ms, which is conventional for neural firing rate analysis. The data we analyze here is those smoothed data points binned to give one data point every 50ms. Each trial proceeded with the following events: the trial began at time $t_{\text{start}}$; the reach was instructed, but the subject was not allowed to move, at random time $t_{\text{target}}$; the cue to reach was given randomly at $t_{\text{go}}$; and fourthly $t_{\text{move}-150\text{ms}}$ is marked as 150ms before the beginning of the reaching movement (the time by when movement related neural activity should be largely present).

### 3.2 Evaluation methods and metrics

To test performance in all data sets, we used leave-one-out cross validation (LOOCV). Specifically, for each of the $N$ time series observations, we held that observation out as the test set, and we trained the model of Equation 4 on the remaining training set. This learned model was then used to infer the held out test data. The predicted test data is then compared to the true test data, and the root mean squared error (RMSE) between the inferred and the true is calculated across all $N$ LOOCV test observations. RMSE is a more appropriate choice here than likelihood, for two reasons: first, we compare our GP method to non-probabilistic averaging methods; and second, in many settings researchers are principally concerned with RMSE, so this metric has practical value. Examples of true data and prediction from LOOCV data are seen in Figure 4 for the neural data, where we show results on a single trial for all model choices, which are as follows:

i. **Conventional GP regression** Here we consider only the absolute time of the observations from the beginning of the observation. This case is equivalent to ignoring the subsequent time-markers and doing standard GP regression. As seen in Figure 4A, this case demonstrates the failure of ignoring known heterogeneity in the temporal structure of the observations.

ii. **Conventional Averaging**. This is the same as above, except we use simple averaging across the training series instead of GP regression. Because averaging is often used in experimental contexts, it is important to compare to these simple non-GP options. The point here in Figure 4A is to show that the GP is not significantly reducing the error over averaging - indeed, the results are quite similar. The reduction in error will come from considering time-marked data appropriately, which the GP does naturally.

iii. **Clipped GP regression**. As seen in blue in Figure 4B, time-marked data is often handled by clipping. Here we do so, performing GP regression in each epoch. Each epoch is chosen uniformly over a range of experimentally reasonable possibilities (for example, the neural data epochs start 200-400ms before a marker and end 400-600ms after).

iv. **Clipped Averaging** (Figure 4B in red). Here we do the same clipping procedure as above, but within epochs we simply average the training data. This method is the conventional procedure in experimental literature and thus is an important comparison point.

v. **Acausal Time-marked GP**. Figure 4C shows the results in blue of GP inference with the acausal time-marked GP model. This model does a superior job modeling the occurrence of the large peak in the data, and indeed the error is accordingly reduced.

vi. **Causal Time-marked GP**. Finally, Figure 4D shows the results in blue of GP inference with the causal time-marked GP model. By inspection this model is the best at modeling all features of the data, which will be seen also in the overall RMSE.

Figure 4 and the above description of the six comparison cases give an indication that appropriately modeling time-marked time series data can have significant effect on the quality of regression. For brevity we do not show an example of the results on the reaching or traffic data, but those examples tell a similar story.

The neural data provides interesting results for discussing the features of the time-marked data model, as shown in Figure 4. First, by inspection of the true data (green trace) and the remainder of the data set shown in Figure 3, it is unclear how strong the underlying dynamical process is, as the data are highly noisy and hold no obvious features like the reach and traffic data. The conventional GP and averaging models in Figure 4A reflect this noisiness - this panel shows

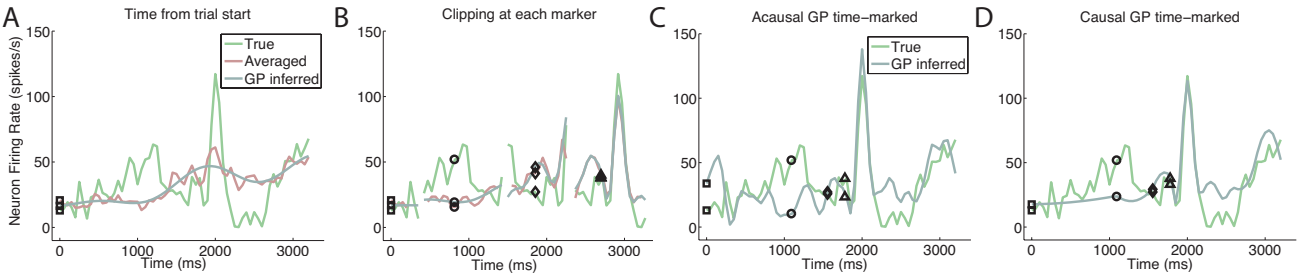


Figure 4: Example results from different model choices in the neural firing rate data. In all panels, the green line is the same true observed data (a representative sample). Each panel demonstrates a different model choice. Panel A shows conventional regression. The blue line shows GP inference (posterior mean based on the learned hyperparameters and LOOCV training set), and the red line shows the result of simple averaging. Panel B shows the same results for the clipping method, with GP regression (blue) and averaging (red). Panels C and D show the results of the time-marked GP model, for an acausal GP and a causal GP.

a lowly increasing function but assigns most of the data variance to noise. The clipping model of Figure 4B shows a spike around the last marker but is largely uninstructive and counfounding. On the other hand, the time-marked models in Figure 4C and D tell a different story: both predict a fairly inactive signal until the last marker, when a large spike in activity is then followed by a ramp for the last 800ms. The time-marked data model is able to model these shared features without data snooping or any heuristic realignments of the data.

Finally, the example in Figure 4C and D also nicely demonstrates the difference between the causal and acausal models. Both models infer the latter half of the time series very similarly. However, because the acausal model has learned a higher variance and shorter lengthscale to model the latter half of the data, it is also reflected in the first half, where the model is seemingly fitting noise. On the other hand, the causal model better assigns early variability to noise, inferring a flat response and correctly assigning the causal influence of the last time-marker.

This benefit of the causal model also raises an important caveat: while the acausal time-marked model is invariant to shifts in the time-markers, the causal model is not. Here for example, the last marker is 150ms before the subject's movement begins, which is a sensible time given delays between neural activity and corresponding movement. If instead we inappropriately chose 150ms after the movement, the causal model would produce a different result (and our tests with this indicate a higher error). Thus, one should take care to choose the event times in a data-appropriate manner when using the causal model.

### 3.3 Summary of Results

Thus far we have only seen example results. We now show the overall LOOCV errors for all datasets, summarized in Table 1. There are a few key takeaways. First, in all cases we see that a time-marked model outperforms - often significantly - competing models, either those that ignore time markings (the leftmost two columns of Table 1) or those that clip the data (middle two columns). As a reminder, the time-marked GP model includes the special case of one time marker ($K = 1$), and thus "ignoring time markings" is a choice contained in our more general time-marked model class. Furthermore, the clipping method is problematic in that a heuristic choice of window interval (the amount of time before and after each marker to consider) is required, and also that all window choices require either double counting some data or not including it at all. Thus, given the inferior performance and practical complications of clipping models, we claim that time-marked GP models are a superior choice for modeling this sort of data.

The second key takeaway is that the causal model always outperforms the acausal model, sometimes considerably. The largest performance difference comes in the neural data, where the design of the data suggests that serious improvements should be seen by causal modeling, as the causal model prevents fitting noise in early parts of the trials. In other data sets, small performance gains are still seen even where we would not expect a big difference from enforcing causality. This may seem surprising since causal models are a special case of acausal models. However, the changes in these data should be in fact causal, as the occurrence of events causes a change in the observation in a causal fashion (for example, the end of a baseball game causes traffic to spike). Thus this is another example of an

| | LOOCV test error (RMSE) | | | | | |
|---|---|---|---|---|---|---|
| | Time from start | | Clipping | | Time-marked GP | |
| | GP | Averaging | GP | Averaging | Acausal | Causal |
| Arm Reach Velocity (mm/s) | 96.2 | 96.5 | 70.8 | 69.8 | 61.2 | **56.2** |
| Game-weekend Traffic (cars/hr) | 58.6 | 58.2 | 73.4 | 71.0 | 53.9 | **52.6** |
| Gameday Traffic (cars/hr) | 59.6 | 59.4 | 66.0 | 65.6 | 44.5 | **43.2** |
| Neural Firing Rates (spikes/s) | 20.8 | 21.4 | 18.4 | 18.3 | 19.4 | **15.7** |

Table 1: Performance of Algorithms on all data sets. All results shown are RMSE between the true time-series observation and the inferred observation based on the remaining data used as a training set (standard LOOCV). To give average performance, error for clipping results is the RMSE over ten reasonable choices for the time windows over which the data was clipped. Furthermore, the GP results are RMSE for the posterior mean of the learned model, *i.e.*, after hyperparameter optimization on the LOOCV training set. We show the average error over ten random initializations of the hyperparameters, though all the learned models and their errors were very similar. In all cases, the time-marked GP model outperforms competing models. Furthermore, the causal GP always improves performance, though in some cases only to a small degree compared with the acausal time-marked GP model.

appropriately constrained prior model class improving generalization to test data.

We noted previously that RMSE is the natural metric here as it allows comparison of all these methods, and because it is commonly used in these settings. For the probabilistic models, the marginal likelihood of the learned model is also a potential choice. Using this metric, the performance hierarchy remains the same: the time-marked model always significantly outperforms the GP regression that ignores time-markings, and further the causal model outperforms the acausal model. Interestingly, when comparing the causal and acausal models, the marginal likelihood reflects a larger benefit to the causal model in the traffic data sets, but no performance difference in the neural data. The benefit in the arm reach data is of similar magnitude in both likelihood and RMSE. Nonetheless the story is unchanged: the time-marked data models have significantly better performance, and the causal model outperforms the acausal model.

## 4 Discussion

We have introduced a model for analyzing time-marked time-series data, data which occur frequently in experimental science and beyond. Our GP model significantly outperforms alternatives and does so in a sensible probabilistic framework. We also introduced causal GP and showed that this constrained model class outperforms further on all datasets we have analyzed.

The models for causal GP and for time-marked data can both be boiled down to simple mappings of the input space and the given data markers. As such, this model can be seen as a particular choice of covariance function. Conveniently, this structure allows us to compose these methods with many existing GP technologies, such as different observation models (Kuss and Rasmussen, 2005) or sparsification methods (Snelson and Ghahramani, 2006). This structure also allows learning and inference to be entirely standard, and thus our model can be easily added to existing GP software packages.

In this work we stipulate the knowledge of time markers, and in the causal case these markers induce a change in the observation. Thus this work can be seen as complementary to work in change-point detection, where models have been developed to learn where such events occur (Saatci et al., 2010). We might also treat the time markers as additional latent parameters and learn those alongside the rest of the model. Though outside the scope of this work and irrelevant for many of the applied contexts of time-marked data models, exploring the utility of this model for change point detection is an interesting direction of future work.

**Acknowledgements**

We thank Paul Nuyujukian and Krishna Shenoy for the arm reach data, and Matt Kaufman and Mark Churchland for the neural data. The traffic data was obtained from the Freeway Performance Measurement System (PeMS) via the UCI machine learning database. Funded by EPSRC EP/H019472/1.

## References


Churchland, M., Yu, B., Ryu, S., Santhanam, G., and Shenoy, K. (2006). Neural variability in premotor cortex provides a signature of motor preparation. *J Neurosci*, 26:3697–3712.

Churchland, M. M., Cunningham, J. P., Kaufman, M., Ryu, S., and Shenoy, K. (2010). Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron*, 68:387–400.

Cunningham, J. P., Nuyujukian, P., Gilja, V., Chestek, C. A., Ryu, S. I., and Shenoy, K. V. (2011). A closed-loop human simulator for investigating the role of feedback control in brain-machine interfaces. *J. Neurophysiology*, 105:1932–1949.

Ihler, A., Hutchins, J., and Smyth, P. (2006). Adaptive event detection with time-varying Poisson processes. In *Proceedings of the 12th ACM SIGKDD Conference (KDD-06)*.

Kuss, M. and Rasmussen, C. (2005). Assessing approximate inference for binary gaussian process classification. *Journal of Machine Learning Res.*, 6:1679–1704.

Murray-Smith, R. and Pearlmutter, B. A. (2003). Transformations of Gaussian Process priors. *Technical Report TR-2003-149, Glasgow University*.

Nott, D. J. and Dunsmuir, W. T. M. (2002). Estimation of nonstationary spatial covariance structure. *Biometrika*, 89:819–829.

Paciorek, C. and Schervish, M. (2003). Nonstationary covariance functions for gaussian process regression. *Advances in NIPS*, 15.

Paciorek, C. and Schervish, M. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17:483–506.

Popov, A. A. (2008). Information characteristics and properties of a random signal considered as a sub algebra of a generalized algebra with a measure. *Radioelectronics and Communications Systems*, 51(11):615–621.

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge.

Saatci, Y., Turner, R., and Rasmussen, C. (2010). Gaussian Process change point models. *Proceedings of the 27th annual ICML (ICML-2010)*.

Sampson, P. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of Amer. Stat. Assoc.*, 87(419):108–119.

Schmidt, A. M. and O'Hagan, A. (2003). Bayesian Inference for non-stationary spatial covariance structure via spatial deformations. *J. Royal Statistical Society. Series B*.

Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. *Advances in NIPS*, 18.

Sugrue, L., Corrado, G., and Newsome, W. (2004). Matching behavior and the representation of value in the parietal cortex. *Science*, 304:1782–1786.