Reliable Decision Support using Counterfactual Models

Peter Schulam
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21211
pschulam@cs.jhu.edu

Suchi Saria
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21211
ssaria@cs.jhu.edu

Abstract

Making a good decision involves considering the likely outcomes under each possible action. For example, would drug A or drug B lead to a better outcome for this patient? Ideally, we answer these questions using an experiment, but this is not always possible (e.g., it may be unethical). As an alternative, we can use non-experimental data to learn models that make counterfactual predictions of what we would observe had we run an experiment. To learn such models for decision-making problems, we propose the use of *counterfactual objectives* in lieu of classical supervised learning objectives. We implement this idea in a challenging and frequently occurring context, and propose the counterfactual GP (CGP), a counterfactual model of continuous-time trajectories (time series) under sequences of actions taken in continuous-time. We develop our model within the potential outcomes framework of Neyman [1923] and Rubin [1978]. The counterfactual GP is trained using a joint maximum likelihood objective that adjusts for dependencies between observed actions and outcomes in the training data. We report two sets of experimental results. First, we show that the CGP's predictions are reliable; they are stable to changes in certain characteristics of the training data that are not relevant to the decision-making problem. Predictive models trained using classical supervised learning objectives, however, are not stable to such perturbations. In the second experiment, we use data from a real intensive care unit (ICU) and qualitatively demonstrate how the CGP's ability to answer "What if?" questions offers medical decision-makers a powerful new tool for planning treatment.

1 Introduction

Making a good decision involves considering the likely outcomes under each possible action. Would changing the color or text of an ad be more effective for increasing a firm's revenue? Would drug A or drug B lead to a better outcome for this patient? Ideally, we would run an *experiment*: clone the patient and try each action on a different clone, compare the outcomes across scenarios, and choose the one with the best result. Experiments, however, are not always feasible. An alternative is to learn models from non-experimental data (i.e. data where we do not control actions) that can make *counterfactual predictions* of the outcomes we would have observed had we run an experiment (see e.g., Pearl 2009). The key challenge when learning counterfactual models from non-experimental data is that it is difficult to distinguish between statistical dependence and *causal* relationships. For instance, consider a drug that is often given to sicker patients who are also more likely to die. Without accounting for this bias in the treatment policy, a statistical model would predict that the drug kills patients even if it is actually beneficial.

This challenge is commonly addressed using the *potential outcomes* framework [Neyman, 1923, 1990, Rubin, 1978], which introduces a collection of *counterfactual* random variables $\{Y[a] : a \in \mathcal{C}\}$ for an outcome Y and each action a from a set of choices $\mathcal{C}$. The counterfactuals may be interpreted
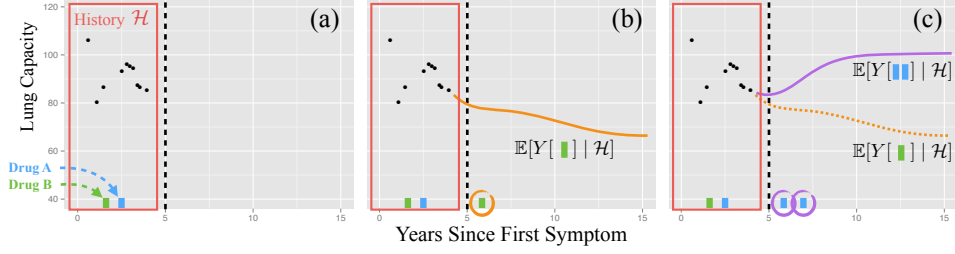
Figure 1: Best viewed in color. An illustration of the counterfactual GP applied to health care. The red box in (a) shows previous lung capacity measurements (black dots) and treatments (the history). Panels (b) and (c) show the type of predictions we would like to make. We use $Y[a]$ to represent the potential outcome under action a.

as probabilistic models of outcomes obtained by running experiments. Using the potential outcomes framework, we can clearly state assumptions under which the distribution of the counterfactuals can be learned from non-experimental data. When this is possible, learned counterfactual models can make "What if?" predictions to guide decisions.

In this paper, we use the potential outcomes framework to develop the counterfactual Gaussian process (CGP), an approach to modeling the effects of sequences of actions on continuous-time trajectories (i.e. time series). Figure 1 illustrates this idea. We show an individual with a lung disease, and would like to predict how her lung capacity (y-axis) will progress in response to different treatment plans. Panel (a) shows the *history* in the red box, which contains previous lung capacity measurements (black dots) and previous treatments (green and blue bars). Panels (b) and (c) illustrate the type of predictions we would like to make: given the history, what is the likely future trajectory of lung capacity if we prescribe Drug B (green)? What if we prescribe two doses of drug A (blue)? A physician might use these "What if?" queries to decide that two doses of A is best.

Many authors have studied counterfactual models for *discrete* time series data. For instance, in health care and epidemiology, Robins [1997] develops counterfactual models of a single outcome measured after a sequence of actions in discrete time. Brodersen et al. [2015] build counterfactual models to estimate the impact that a single, dicrete event has on a discrete time series of interest (e.g. daily sales after a product launch). Others have modeled the effect of actions taken in continuous time on a single outcome (e.g. Arjas and Parner 2004, Lok 2008). The CGP is unique in that it allows us to predict how future trajectories in continuous-time will change in response to sequences of interventions.

Contributions. For problems in which learned predictive models are used to guide decision-makers in choosing actions, this paper proposes the use of *counterfactual objectives* in lieu of classical supervised learning objectives. We implement this idea in a challenging and frequently occurring context; problems where outcomes are measured and actions are taken at discrete points in continuous-time, and may be freely interwoven. Our key methodological contribution is an adjusted maximum likelihood objective for Gaussian processes that allows us to learn counterfactual models of continuous-time trajectories from *observational traces*: irregularly sampled sequences of actions and outcomes denoted using $\mathcal{D} = \{\{(y_{ij}, a_{ij}, t_{ij})\}_{j=1}^{n_i}\}_{i=1}^{m}$, where $y_{ij} \in \mathbb{R} \cup \{\varnothing\}$, $a_{ij} \in \mathcal{C} \cup \{\varnothing\}$, and $t_{ij} \in [0, \tau]$.[1] We derive the technique by jointly modeling actions and outcomes using a marked point process (MPP; see e.g., Daley and Vere-Jones 2007), where the GP models the conditional distribution of the marks. When using potential outcomes, several assumptions are typically required to show that the learned statistical model estimates the target counterfactuals. We describe one set of assumptions sufficient for recovering the counterfactual GP from non-experimental data.

We report two sets of experimental results. First, we show that the CGP's predictions are reliable; they are stable to changes in certain characteristics of the training data that are not relevant to the decision-making problem. Predictive models trained using classical supervised learning objectives, however, are not stable to such perturbations. In this experiment, we use simulated data so that (1) we can control characteristics of the training data, and (2) we have access to the ground truth consequences of our actions on test data for evaluation. In our second experiment, we use data from a real intensive care unit (ICU) to learn the CGP, and qualitatively demonstrate how its ability to answer "What if?" questions offers medical decision-makers a powerful new tool for planning treatment.

---

[1] y_{ij} and a_{ij} may be the null variable $\varnothing$ to allow for the possibility that an action is taken but no outcome is observed and vice versa. $[0, \tau]$ denotes a fixed period of time over which the trajectories are observed.

1.1 Related Work

The difference between counterfactual predictions of an outcome if an action had been taken and if it had not been taken is defined as the *causal effect* of the action in the causal inference community (see e.g., Pearl 2009 or Morgan and Winship 2014). *Potential outcomes* are commonly used to formalize counterfactual predictions and obtain causal effect estimates [Neyman, 1923, 1990, Rubin, 1978]; we will review them shortly. Potential outcomes are often applied to cross-sectional data (see examples in Morgan and Winship 2014; recent examples from machine learning are Bottou et al. 2013, Johansson et al. 2016), but have also been used to estimate the causal effect of a sequence of actions in discrete time on a final outcome (e.g. Robins 1997, Taubman et al. 2009). Conversely, Brodersen et al. [2015] estimate the effect that a *single discrete intervention* has on a *discrete* time series.

Recent work on optimal dynamic treatment regimes uses the sequential potential outcomes framework proposed by Robins [1997] to learn lists of discrete-time treatment rules that optimize a scalar outcome. Algorithms for learning these rules often use value-function approximations (Q-learning; e.g., Nahum-Shani et al. 2012). Alternatively, A-learning directly learns the relative differences between competing policies [Murphy, 2003].

Others have extended the potential outcomes framework in Robins [1997] to learn causal effects of actions taken in continuous-time on a single final outcome using observational data. Lok [2008] proposes an estimator based on structural nested models [Robins, 1992] that learns the instantaneous effect of administering a single type of treatment. Arjas and Parner [2004] develop an alternative framework for causal inference using Bayesian posterior predictive distributions to estimate the effects of actions in continuous time on a final outcome. Xu et al. [2016] also learn effects of actions in continuous-time on outcomes measured in continuous-time, but make one-step-ahead scalar predictions instead of trajectory-valued predictions. Both Lok [2008] and Arjas and Parner [2004] use marked point processes to formalize assumptions that make it possible to learn causal effects from continuous-time observational data. We build on these ideas to learn causal effects of actions on continuous-time *trajectories* instead of a single outcome. Cunningham et al. [2012] introduce the Causal Gaussian Process, but their use of the term "causal" is different from ours, and refers to a constraint that holds for all samples drawn from the GP.

Causal effects in continuous-time have also been studied using differential equations. Mooij et al. [2013] formalize an analog of Pearl's "do" operation for deterministic ordinary differential equations. Sokol and Hansen [2014] make similar contributions for stochastic differential equations by studying limits of discrete-time non-parametric structural equation models [Pearl, 2009].

Reinforcement learning (RL) algorithms learn from data where actions and observations are interleaved in discrete time (see e.g., Sutton and Barto 1998). In RL, however, the focus is on learning a *policy* (a map from states to actions) that optimizes the expected reward, rather than a model that predicts the effects of the agent's actions on future observations. In model-based RL, a model of an action's effect on the subsequent state is produced as a by-product either offline before optimizing the policy (e.g., Ng et al. 2006) or incrementally as the agent interacts with its environment. In most RL problems, however, learning algorithms rely on active experimentation to collect samples. This is not always possible, for example in health care we cannot actively experiment on patients, and so we must rely on retrospective observational data.

In RL, a related problem known as off-policy evaluation also uses retrospective observational data (see e.g., Dudík et al. 2011, Jiang and Li 2016, Păduraru et al. 2012). The goal is to use state-action-reward sequences generated by an agent operating under an unknown policy to estimate the expected reward of a target policy. Off-policy algorithms typically use value-function approximations, importance reweighting, or doubly robust combinations of the two to estimate the expected reward.

2 Counterfactual Models from Observational Traces

Counterfactual GPs build on ideas from potential outcomes [Neyman, 1923, 1990, Rubin, 1978], Gaussian processes [Rasmussen and Williams, 2006], and marked point processes [Daley and Vere-Jones, 2007]. In the interest of space, we review potential outcomes and marked point processes, but refer the interested reader to Rasmussen and Williams [2006] for background on GPs.

Background: Potential Outcomes. To formalize counterfactuals, we adopt the potential outcomes framework [Neyman, 1923, 1990, Rubin, 1978], which uses a collection of random variables $\{Y[a] : a \in \mathcal{C}\}$ to model the distribution over outcomes under each action a from a set of choices $\mathcal{C}$. To make

counterfactual predictions, we must learn the distribution $P(Y[a])$ for each action $a \in \mathcal{C}$. If we can freely experiment by repeatedly taking actions and recording the effects, then it is straightforward to learn a probabilistic model for each potential outcome $Y[a]$. Conducting experiments, however, may not be possible. Alternatively, we can use observational data, where we have example actions A and outcomes Y, but do not know how actions were chosen. Note the difference between the action a and the random variable A that models the *observed actions* in our data. The notation $Y[a]$ serves to distinguish between the observed distribution $P(Y \mid A = a)$ and the target distribution $P(Y[a])$.

In general, we can use observational data to estimate $P(Y \mid A = a)$. Under two assumptions, however, we can show that this conditional distribution is equivalent to the counterfactual model $P(Y[a])$. The first is known as the Consistency Assumption.

Assumption 1 (Consistency). *Let Y be the observed outcome, $A \in \mathcal{C}$ be the observed action, and $Y[a]$ be the potential outcome for action $a \in \mathcal{C}$, then:* $(Y \triangleq Y[a]) \mid A = a$.

Under consistency, we have that $P(Y \mid A = a) = P(Y[a] \mid A = a)$. Now, the potential outcome $Y[a]$ may depend on the action A, so in general $P(Y[a] \mid A = a) \neq P(Y[a])$. The next assumption posits that we have additional observed variables X known as confounders [Morgan and Winship, 2014] that are sufficient to d-separate $Y[a]$ and A.

Assumption 2 (No Unmeasured Confounders (NUC)). *Let Y be the observed outcome, $A \in \mathcal{C}$ be the observed action, $X = x$ be a vector containing potential confounders, and $Y[a]$ be the potential outcome under action $a \in \mathcal{C}$, then:* $(Y[a] \perp A) \mid X = x$.

Under Assumptions 1 and 2, $P(Y \mid A = a, X = x) = P(Y[a] \mid X = x)$. By marginalizing with respect to $P(X)$ we can estimate $P(Y[a])$. An extension of Assumption 2 introduced by Robins [1997] known as *sequential NUC* allows us to estimate the effect of a sequence of actions in discrete time on a single outcome. In continuous-time settings, where both the type and *timing* of actions may be statistically dependent on the potential outcomes, Assumption 2 (and sequential NUC) cannot be applied as-is. We will describe an alternative that serves a similar role for CGPs.

Background: Marked Point Processes. Point processes are distributions over sequences of timestamps $\{T_i\}_{i=1}^{N}$, which we call points, and a marked point process (MPP) is a point process where each point is annotated with an additional random variable X_i, called its mark. For example, a point T might represent the arrival time of a customer, and X the amount that she spent at the store. We emphasize that both the annotated points (T_i, X_i) and the number of points N are random variables.

A point process can be characterized as a counting process $\{N_t : t \geq 0\}$ that counts the number of points that occured up to and including time t: $N_t = \sum_{i=1}^{N} \mathbb{I}_{(T_i \leq t)}$. By definition, this processes can only take integer values, and $N_t \geq N_s$ if $t \geq s$. In addition, it is commonly assumed that $N_0 = 0$ and that $\Delta N_t = \lim_{\delta \to 0^+} N_t - N_{t-\delta} \in \{0, 1\}$. We can parameterize a point process using a probabilistic model of ΔN_t given the history of the process $\mathcal{H}_{t^-}$ up to but not including time t (we use t^- to denote the left limit of t). Using the Doob-Meyer decomposition [Daley and Vere-Jones, 2007], we can write $\Delta N_t = \Delta M_t + \Delta \Lambda_t$, where M_t is a martingale, Λ_t is a cumulative intensity function, and

$$P(\Delta N_t = 1 \mid \mathcal{H}_{t^-}) = \mathbb{E}[\Delta N_t \mid \mathcal{H}_{t^-}] = \mathbb{E}[\Delta M_t \mid \mathcal{H}_{t^-}] + \Delta \Lambda_t(\mathcal{H}_{t^-}) = 0 + \Delta \Lambda_t(\mathcal{H}_{t^-}),$$

which shows that we can parameterize the point process using the conditional intensity function $\lambda^*(t) \, dt \triangleq \Delta \Lambda_t(\mathcal{H}_{t^-})$ The star superscript on the intensity function serves as a reminder that it depends on the history $\mathcal{H}_{t^-}$. For example, in non-homogeneous Poisson processes $\lambda^*(t)$ is a function of time that does not depend on the history. On the other hand, a Hawkes process is an example of a point process where $\lambda^*(t)$ *does* depend on the history [Hawkes, 1971]. MPPs are defined by an intensity that is a function of both the time t and the mark x: $\lambda^*(t, x) = \lambda^*(t)p^*(x \mid t)$. We have written the joint intensity in a factored form, where $\lambda^*(t)$ is the intensity of *any* point occuring (that is, the mark is unspecified), and $p^*(x \mid t)$ is the pdf of the observed mark given the point's time. For an MPP, the history $\mathcal{H}_t$ contains each prior point's time and mark.

2.1 Counterfactual Gaussian Processes

Let $\{Y_t : t \in [0, \tau]\}$ denote a continuous-time stochastic process, where $Y_t \in \mathbb{R}$, and $[0, \tau]$ defines the interval over which the process is defined. We will assume that the process is observed at a discrete set of irregular and random times $\{(y_j, t_j)\}_{j=1}^{n}$. We use $\mathcal{C}$ to denote the set of possible *action types*, $a \in \mathcal{C}$ to denote the elements of the set, and define an action to be a 2-tuple (a, t) specifying

an action type $a \in \mathcal{C}$ and a time $t \in [0, \tau]$ at which it is taken. To refer to multiple actions, we use $\mathbf{a} = [(a_1, t_1), \ldots, (a_n, t_n)]$. Finally, we define the history $\mathcal{H}_t$ at a time $t \in [0, \tau]$ to be a list of all previous observations of the process and all previous actions. Our goal is to model the counterfactual:

$$P(\{Y_s[\mathbf{a}] : s > t\} \mid \mathcal{H}_t), \text{ where } \mathbf{a} = \{(a_j, t_j) : t_j > t\}_{j=1}^m. \tag{1}$$

To learn the counterfactual model, we will use *traces* $\mathcal{D} \triangleq \{\mathbf{h}_i = \{(t_{ij}, y_{ij}, a_{ij})\}_{j=1}^{n_i}\}_{i=1}^m$, where $y_{ij} \in \mathbb{R} \cup \{\varnothing\}$, $a_{ij} \in \mathcal{C} \cup \{\varnothing\}$, and $t_{ij} \in [0, \tau]$. Our approach is to model $\mathcal{D}$ using a marked point process (MPP), which we learn using the traces. Using Assumption 1 and two additional assumptions defined below, the estimated MPP recovers the counterfactual model in Equation 1.

We define the MPP mark space as the Cartesian product of the outcome space $\mathbb{R}$ and the set of action types $\mathcal{C}$. To allow either the outcome or the action (but not both) to be the null variable $\varnothing$, we introduce binary random variables $z_y \in \{0, 1\}$ and $z_a \in \{0, 1\}$ to indicate when the outcome y and action a are not $\varnothing$. Formally, the mark space is $\mathcal{X} = (\mathbb{R} \cup \{\varnothing\}) \times (\mathcal{C} \cup \{\varnothing\}) \times \{0, 1\} \times \{0, 1\}$. We can then write the MPP intensity as

$$\lambda^*(t, y, a, z_y, z_a) = \underbrace{\lambda^*(t)p^*(z_y, z_a \mid t)}_{\text{[A] Event model}} \underbrace{p^*(y \mid t, z_y)}_{\text{[B] Outcome model (GP)}} \underbrace{p^*(a \mid y, t, z_a)}_{\text{[C] Action model}}, \tag{2}$$

where we have again used the $*$ superscript as a reminder that the hazard function and densities above are implicitly conditioned on the history $\mathcal{H}_{t^-}$. The parameterization of the event and action models can be chosen to reflect domain knowledge about how the timing of events and choice of action depend on the history. The outcome model is parameterized using a GP (or any elaboration such as a hierarchical GP, or mixture of GPs), and can be simply designed as a regression model that predicts how the future trajectory will progress given the previous actions and outcome observations.

Learning. To learn the CGP, we maximize the likelihood of observational traces over a fixed interval $[0, \tau]$. Let $\boldsymbol{\theta}$ denote the model parameters, then the likelihood for a single trace is

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^n \log p_\theta^*(y_j \mid t_j, z_{yj}) + \sum_{j=1}^n \log \lambda_\theta^*(t)p_\theta^*(a_j, z_{yj}, z_{aj} \mid t_j, y_j) - \int_0^\tau \lambda_\theta^*(s)\, ds. \tag{3}$$

We assume that traces are independent, and so can learn from multiple traces by maximizing the sum of the individual-trace log likelihoods with respect to $\boldsymbol{\theta}$. We refer to Equation 3 as the adjusted maximum likelihood objective. We see that the first term fits the GP to the outcome data, and the second term acts as an adjustment to account for dependencies between future outcomes and the timing and types of actions that were observed in the training data.

Connection to target counterfactual. By maximizing Equation 3, we obtain a statistical model of the observational traces $\mathcal{D}$. In general, the statistical model may not recover the target counterfactual model (Equation 1). To connect the CGP to Equation 1, we describe two additional assumptions. The first assumption is an alternative to Assumption 2.

Assumption 3 (Continuous-Time NUC). *For all times t and all histories $\mathcal{H}_{t^-}$, the densities $\lambda^*(t)$, $p^*(z_y, z_a \mid t)$, and $p^*(a \mid y, t, z_a)$ do not depend on $Y_s[\mathbf{a}]$ for all times $s > t$ and all actions $\mathbf{a}$.*

The key implication of this assumption is that the policy used to choose actions in the observational data did not depend on any unobserved information that is predictive of the future potential outcomes.

Assumption 4 (Non-Informative Measurement Times). *For all times t and any history $\mathcal{H}_{t^-}$, the following holds: $p^*(y \mid t, z_y = 1)\, dy = P(Y_t \in dy \mid \mathcal{H}_{t^-})$.*

Under Assumptions 1, 3, and 4, we can show that Equation 1 is equivalent to the GP used to model $p^*(y \mid t, z_y = 1)$. In the interest of space, the argument for this equivalence is in Section A of the supplement. Note that these assumptions are not statistically testable (see e.g., Pearl 2009).

3 Experiments

Counterfactual models that answer "What if?" questions are more reliable decision support tools than predictive models trained using classical supervised learning objectives because counterfactual models are, by construction, invariant to the policy used to choose actions in the training data. In our first experiments, we use simulated data so that (1) we can control characteristics of the training data, and (2) we have access to the ground truth consequences of our actions on test data for evaluation. In our second experiment, we use data from a real ICU to learn the CGP, and qualitatively demonstrate how its ability to answer "What if?" questions has the potential to give medical decision-makers a powerful new tool for planning treatment.

5

	Regime A		Regime B		Regime C	
	Baseline GP	CGP	Baseline GP	CGP	Baseline GP	CGP
Risk Score Δ from A	0.000	0.000	0.083	0.001	0.162	0.128
Kendall's τ from A	1.000	1.000	0.857	0.998	0.640	0.562
AUC	0.853	0.872	0.832	0.872	0.806	0.829

Table 1: Results measuring reliability for simulated data experiments. See Section 3.1 for details.

3.1 Reliable Decision-making with CGPs

We focus on a decision-making problem where the goal is to decide whether or not to treat a patient. The decision-maker (a clinician), should treat a patient if the value of a severity marker is likely to fall below a threshold in the future. A common approach to solving this problem is to build a predictive model of the future outcome, and use the prediction to make the decision (i.e. if the predicted value is below the threshold, then treat). We will use this approach as our baseline. If we could look into the future and see what the marker value would be under different actions, then the best decision would be to treat if the severity marker will fall below the threshold without treatment. It is precisely this type of reasoning that we can perform with the CGP.

We simulate the value of a severity marker recorded over a period of 24 hours in the hospital; high values indicate that the patient is healthy. For the baseline approach, we learn a GP that predicts the future trajectory given clinical history up until time t; i.e. $P(\{Y_s : s > t\} \mid \mathcal{H}_t)$. Using the CGP, we model the counterfactual "What if we do not treat this patient?"; i.e. $P(\{Y_s[\varnothing] : s > t\} \mid \mathcal{H}_t)$. For all experiments, we consider a single decision time $t = 12$hrs. We create a risk score using the negative of each model's predicted value at the end of 24 hours, normalized to lie in $[0, 1]$.

Data. We simulate data from three regimes. In regimes A and B, we simulate severity marker trajectories that are treated by policies π_A and π_B respectively, that are both unknown to the baseline model and CGP at train time. Both π_A and π_B are designed to satisfy Assumptions 1, 3, and 4. In regime C, we use a policy that *does not* satisfy these assumptions. This regime will demonstrate the importance of verifying whether the assumptions hold when applying the CGP. We train both the baseline model and CGP on data simulated from all three regimes. In all regimes, we test decisions on a common set of trajectories treated up until $t = 12$hrs with policy π_A.

Simulator. For each patient, we randomly sample outcome measurement times from a homogeneous Poisson process with with constant intensity λ over the 24 hour period. Given the measurement times, outcomes are sampled from a mixture of three GPs. The covariance function is shared between all classes, and is defined using a Matérn $3/2$ kernel (variance 0.2^2, lengthscale 8.0) and independent Gaussian noise (scale 0.1) added to each observation. Each class has a distinct mean function parameterized using a 5-dimensional, order-3 B-spline. The first class has a declining mean trajectory, the second has a trajectory that declines then stabilizes, and the third has a stable trajectory.[2] All classes are equally likely *a priori*. At each measurement time, the treatment policy π determines a probability p of treatment administration (we use only a single treatment type). The treatments increase the severity marker by a constant amount for 2 hours. If two or more actions occur within 2 hours of one another, the effects do not add up (i.e. it is as though only one treatment is active). Additional details about the simulator and policies can be found in the supplement.

Model. For both the baseline predictive model and CGP outcome model, we use a mixture of three GPs (as was used to simulate the data). We assume that the mean function coefficients, the covariance parameters, and the treatment effect size are unknown and must be learned. We emphasize that both the CGP and RGP have identical forms, but are trained using different objectives; the RGP marginalizes over future actions, inducing an implicit dependence on the treatment policy in the training data, while the CGP explicitly controls for them while learning. For both the baseline model and CGP, we analytically sum over the mixture component likelihoods to obtain a closed form expression for the likelihood, which we optimize using BFGS [Nocedal and Wright, 2006].[3] Predictions for both models are made using the posterior predictive mean given data and interventions up until 12 hours.

Results. When trained on data where actions were taken according to different policies, the baseline model produces different risk scores *for the same patient*. In Table 1, the first row shows the average

---

[2]The exact B-spline coefficients can be found in the simulation code included in the supplement.

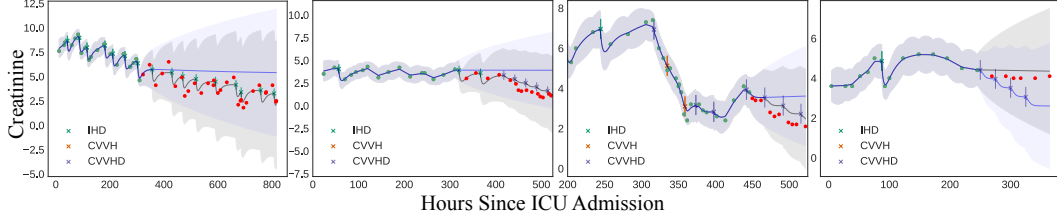[3]Additional details can be found in the supplement.

Figure 2: Example factual (grey) and counterfactual (blue) predictions on real ICU data using the CGP.

difference in risk scores (calibrated to lie in $[0, 1]$) produced by the models trained in each regime and produced by the models trained in regime A. In row 1, column B we see that the baseline GP's risk scores differ for the same person on average by around eight points ($\Delta = 0.083$). From the perspective of a decision-maker, this behavior could make the system appear less reliable. Intuitively, the risk for a given patient should not depend on the policy used to determine treatments in retrospective data. On the other hand, the CGP's scores change very little when trained on different regimes ($\Delta = 0.001$), as long as Assumptions 1, 3, and 4 are satisfied. Note, however, that the scores *do* change for the CGP in row 1, column C where the policy π_C does not satisfy these assumptions ($\Delta = 0.128$). Although we illustrate stability of the CGP comapred to the baseline GP using two regimes, this property is not specific to the choice of policies used in regimes A and B. Rather, the issue persists as we generate different training data by varying the distribution over the action choices.

A cynical reader might ask: even if the risk scores are unstable, perhaps it has no consequences on the downstream decision-making task? In the second row of Table 1, we report Kendall's τ computed between each regime and regime A using the risk scores to rank the patient's in the test data according to severity (i.e. scores closer to 1 are more severe). In the third row, we report the AUC for both models trained in each regime on the common test set. We label a patient as "at risk" if the last marker value in the untreated trajectory is below zero, and "not at risk" otherwise.

In row 2, column B we see that the CGP has a high rank correlation ($\tau = 0.998$) between the two regimes where the policies satisfy our key assumptions. The baseline GP model trained on regime B, however, has a lower rank correlation of $\tau = 0.857$ with the risk scores produced by the same model trained on regime A. Similarly, in row three, columns A and B, we see that the CGP's AUC is unchanged (AUC $= 0.872$). The baseline GP, however, is unstable and creates a risk score with poorer discrimination in regime B (AUC $= 0.832$) than in regime A (AUC $= 0.853$). Finally, we see that in column C, where the policy π_C *does not* satisfy our key assumptions, the CGP's rank correlation degrades ($\tau = 0.562$), and the AUC degrades to 0.829 (note that the baseline GP's rank correlation and AUC also degrade). This further emphasizes the importance of verifying Assumptions 1, 3, and 4 when using the CGP.

These results have broad consequences for the practice of building predictive models from observational data for decision support. Classical supervised learning is commonly used to build predictive models for assessing risk from non-experimental data. Our experiments show that the predictions these models make and the decisions that they may lead to are highly dependent on action policies in the training data. Intuitively, this is troubling because we do not expect that past action policies should have any impact on the assessment of risk. As predictive models are becoming more widely used in domains like health care (e.g., Li-wei et al. 2015, Schulam and Saria 2015, Alaa et al. 2016, Wiens et al. 2016, Cheng et al. 2017) where safety is critical, the framework proposed here is increasingly pertinent. Others have noted this issue and studied the impact that action policies in the training data have on models fit using supervised learning objectives (e.g., Dyagilev and Saria 2016). Counterfactual GPs (and counterfactual models more broadly) make predictions that are independent of policies in the training data, and offer a new more reliable way to train predictive models for decision support.

3.2 CGPs for Medical Decision Support

CGPs offer a powerful new tool for decision-makers to evaluate different actions using data-driven models. In health care, for instance, we can move closer towards the vision of evidence-based medicine by allowing clinicians to answer "What if?" questions at the point of care with predictions tailored to the patient's clinical history. When treatments are expensive or have potent side effects, the CGP can estimate the effects, which can be weighed against the cost of its administration.

7

To demonstrate how the CGP can be used as a decision support tool, we extract observational creatinine traces from the publicly available MIMIC-II database [Saeed et al., 2011]. Creatinine is a compound produced as a by-product of the chemical reaction in the body that breaks down creatine to fuel muscles. Healthy kidneys normally filter creatinine out of the body, which can otherwise be toxic in large concentrations. During kidney failure, however, creatinine levels rise and the compound must be extracted using a medical procedure called dialysis.

We extract patients in the database who tested positive for abnormal creatinine levels, which is a sign of kidney failure. We also extract the times at which three different types of dialysis were given to each individual: intermittent hemodialysis (IHD), continuous veno-venous hemofiltration (CVVH), and continuous veno-venous hemodialysis (CVVHD). The data set includes a total of 428 individuals, with an average of 34 (± 12) creatinine observations each. We shuffle the data and use 300 traces for training, 50 for validation and model selection, and 78 for testing.

Model. We parameterize the outcome model of the CGP using a mixture of GPs. We always condition on the initial creatinine measurement and model the deviation from that initial value. The mean for each class is zero (i.e. there is no deviation from the initial value on average). We parameterize the covariance function using the sum of two non-stationary kernel functions. Let $\phi : t \to [1, t, t^2]^\top \in \mathbb{R}^3$ denote the quadratic polynomial basis, then the first kernel is $k_1(t_1, t_2) = \phi^\top(t_1)\Sigma\phi(t_2)$, where $\Sigma \in \mathbb{R}^{3\times 3}$ is a positive-definite symmetric matrix parameterizing the kernel. The second kernel is the covariance function of the integrated Ornstein-Uhlenbeck (IOU) process (see e.g., Taylor et al. 1994), which is parameterized by two scalars α and ν and defined as

$$k_{\text{IOU}}(t_1, t_2) = \frac{\nu^2}{2\alpha^3}\left(2\alpha\min(t_1, t_2) + e^{-\alpha t_1} + e^{-\alpha t_2} - 1 - e^{-\alpha|t_1 - t_2|}\right).$$

The IOU covariance corresponds to the random trajectory of a particle whose velocity drifts according to an OU process. We assume that each creatinine measurement is observed with independent Gaussian noise with scale σ. Each class in the mixture has a unique set of covariance parameters. To model the treatment effects in the outcome model, we define a short-term function and long-term response function. If an action is taken at time t_0, the outcome $\delta = t - t_0$ hours later will be additively affected by the response function $g(\delta; h_1, a, b, h_2, r) = g_s(\delta; h_1, a, b) + g_\ell(\delta; h_2, r)$, where $h_1, h_2 \in \mathbb{R}$ and $a, b, r \in \mathbb{R}^+$. The short-term and long-term response functions are defined as $g_s(\delta; h_1, a, b) = \frac{h_1 a}{a - b}\left(e^{-b \cdot t} - e^{-a \cdot t}\right)$, and $g_\ell(\delta : h_2, r) = h_2 \cdot (1.0 - e^{-r \cdot t})$. The two response functions are included in the mean function of the GP, and each class in the mixture has a unique set of response function parameters. We assume that Assumptions 1, 3, and 4 hold, and that the event and action models have separate parameters, so can remain unspecified when estimating the outcome model. We fit the CGP outcome model using Equation 3, and select the number of classes in the mixture using fit on the validation data (we choose three components).

Results. Figure 2 demonstrates how the CGP can be used for medical decision support. Each panel in the figure shows data for an individual drawn from the test set. The green points show measurements on which we condition to obtain a posterior distribution over mixture class membership and the individual's latent trajectory under each class. The red points are unobserved, future measurements. In grey, we show predictions under the *factual* sequence of actions extracted from the MIMIC-II database. Treatment times are shown using vertical bars marked with an "x" (color indicates which type of treatment was given). In blue, we show the CGP's *counterfactual* predictions under an alternative sequence of actions. The posterior predictive trajectory is shown for the MAP mixture class (mean is shown by a solid grey/blue line, 95% credible intervals are shaded).

We qualitatively discuss the CGP's counterfactual predictions, but cannot quantitatively evaluate them without prospective experimental data from the ICU. We can, however, measure fit on the factual data and compare to baselines to evaluate our modeling decisions. Our CGP's outcome model allows for heterogeneity in the covariance parameters and the response functions. We compare this choice to two alternatives. The first is a mixture of three GPs that *does not* model treatment effects. The second is a single GP that *does* model treatment effects. Over a 24-hour horizon, the CGP's mean absolute error (MAE) is 0.39 (95% CI: 0.38-0.40),[4] and for predictions between 24 and 48 hours in the future the MAE is 0.62 (95% CI: 0.60-0.64). The pairwise mean difference between the first baseline's absolute errors and the CGP's is 0.07 (0.06, 0.08) for 24 hours, and 0.09 (0.08, 0.10) for 24-48 hours. The mean difference between the second baseline's absolute errors and the CGP's is 0.04 (0.04, 0.05) for 24 hours and 0.03 (0.02, 0.04) for 24-48 hours. The improvements over the

---

[4] 95% confidence intervals computed using the pivotal bootstrap are shown in parentheses

baselines suggest that modeling treatments and heterogeneity with a mixture of GPs for the outcome model are useful for this problem.

Figure 2 shows factual and counterfactual predictions made by the CGP. In the first (left-most) panel, the patient is factually administered IHD about once a day, and is responsive to the treatment (creatinine steadily improves). We query the CGP to estimate how the individual *would have* responded had the IHD treatment been stopped early. The model reasonably predicts that we would have seen no further improvement in creatinine. In the third panel, an individual with erratic creatinine levels receives CVVHD for the last 100 hours and is responsive to the treatment. As before, the CGP counterfactually predicts that she would not have improved had CVVHD not been given. Interestingly, panel four shows the opposite situation: the individual did not receive treatment and did not improve for the last 100 hours, but the CGP counterfactually predicts an improvement in creatinine similar to that in panel 3 if daily CVVHD had been administered.

4 Discussion

Our key message is that predictive models used in decision-making problems should be trained using counterfactual objectives (like the one shown in Equation 3). One reason this approach should be preferred is because the models produced are *stable* to information in the training data that is irrelevant to the downstream decision-making task (the action policies in the training data). We studied this idea in the context of problems where outcomes are measured and actions are taken at discrete points in continuous-time, and proposed the *counterfactual Gaussian process* (CGP) to model the effects of sequences of actions taken on continuous-time trajectories (time series). The CGP builds on previous ideas in continuous-time causal inference (e.g., Robins 1997, Arjas and Parner 2004, Lok 2008), but is unique in that it can predict the full counterfactual *trajectory*; we combined marked point processes (MPP) with GPs to model *observational traces*, and described three assumptions that are sufficient to connect the statistical model to the target counterfactuals.

We presented two sets of experimental results. In the first, we used simulated data to show risk score models fit using classical supervised learning objectives are sensitive to the action policies used in the training data; information that is irrelevant to the downstream decision problem. We showed that this sensitivity can alter risk assessments for *the same individual*, change relative risk assessments across individuals, and can cause differences in AUC. On the other hand, the CGP is not sensitive to the action policies in the training data, as long as they satisfy Assumptions 1, 3, and 4. In the second set of experiments, we demonstrated how the CGP offers a powerful new tool for medical decision support by learning the effects of dialysis on creatinine trajectories from real ICU data and demonstrating the types of "What if?" questions that it can be used to answer about patient prognosis under various treatment plans.

These results suggest a number of new questions and directions for future work. First, the validity of the CGP is conditioned upon a set of assumptions (this is true for all counterfactual models). In general, these assumptions are not testable. The reliability of approaches using counterfactual models therefore critically depends on the plausibility of those assumptions in light of domain knowledge. Formal procedures, such as sensitivity analyses (e.g., Robins et al. 2000, Scharfstein et al. 2014), that can identify when causal assumptions conflict with a data set will help to make these methods more easily applied in practice. In addition, there may be other sets of structural assumptions beyond those presented that allow us to learn counterfactual GPs from non-experimental data. For instance, the back door and front door criteria are two separate sets of structural assumptions discussed by Pearl [2009] in the context of estimating parameters of causal Bayesian networks from observational data.

More broadly, there are implications for recent pushes to introduce transparency, interpretability, and accountability into machine learning systems embedded in decision-making processes. We have characterized a notion of model stability relative to information in the training data that is not relevant to its downstream application, and showed that models fit using counterfactual objectives achieve stability. The framework can be further extended to incorporate recent ideas on model interpretability and accountability in the context of supervised learning objectives (e.g., Caruana et al. 2015, Ribeiro et al. 2016).

References

A.M. Alaa, J. Yoon, S. Hu, and M. van der Schaar. Personalized Risk Scoring for Critical Care Patients using Mixtures of Gaussian Process Experts. In *ICML Workshop on Computational Frameworks for Personalization*, 2016.

E. Arjas and J. Parner. Causal reasoning from longitudinal data. *Scandinavian Journal of Statistics*, 31(2):171–187, 2004.

L. Bottou, J. Peters, J.Q. Candela, D.X. Charles, M. Chickering, E. Portugaly, D. Ray, P.Y. Simard, and E. Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research (JMLR)*, 14(1):3207–3260, 2013.

K.H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S.L. Scott. Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1):247–274, 2015.

R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1721–1730. ACM, 2015.

L.F. Cheng, G. Darnell, C. Chivers, M.E. Draugelis, K. Li, and B.E. Engelhardt. Sparse multi-output Gaussian processes for medical time series prediction. *arXiv preprint arXiv:1703.09112*, 2017.

J. Cunningham, Z. Ghahramani, and C.E. Rasmussen. Gaussian processes for time-marked time-series data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 255–263, 2012.

D.J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes*. Springer Science & Business Media, 2007.

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning (ICML)*, 2011.

K. Dyagilev and S. Saria. Learning (predictive) risk scores in the presence of censoring due to interventions. *Machine Learning*, 102(3):323–348, 2016.

A.G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, pages 83–90, 1971.

N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 652–661, 2016.

F.D. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning (ICML)*, 2016.

H.L Li-wei, R.P. Adams, L. Mayaud, G.B. Moody, A. Malhotra, R.G. Mark, and S. Nemati. A physiological time series dynamics-based approach to patient monitoring and outcome prediction. *IEEE Journal of Biomedical and Health Informatics*, 19(3):1068–1076, 2015.

J.J. Lok. Statistical modeling of causal effects in continuous time. *The Annals of Statistics*, pages 1464–1507, 2008.

J.M. Mooij, D. Janzing, and B. Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. 2013.

S.L. Morgan and C. Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2014.

S.A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

I. Nahum-Shani, M. Qian, D. Almirall, W.E. Pelham, B. Gnagy, G.A. Fabiano, J.G. Waxmonsky, J. Yu, and S.A. Murphy. Q-learning: A data analysis method for constructing adaptive interventions. *Psychological Methods*, 17(4):478, 2012.

J. Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

J. Neyman. On the application of probability theory to agricultural experiments. *Statistical Science*, 5(4):465–472, 1990.

A.Y. Ng, A. Coates, M. Diel, V. Ganapathi, J. Schulte, B. Tse, E. Berger, and E. Liang. Autonomous inverted helicopter flight via reinforcement learning. In *Experimental Robotics IX*, pages 363–372. Springer, 2006.

J. Nocedal and S.J. Wright. Numerical optimization 2nd, 2006.

C. Păduraru, D. Precup, J. Pineau, and G. Comănici. An empirical analysis of off-policy learning in discrete mdps. In *Workshop on Reinforcement Learning*, page 89, 2012.

J. Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2009.

C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning*. the MIT Press, 2006.

M.T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144. ACM, 2016.

J.M. Robins. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, 79(2):321–334, 1992.

J.M. Robins. Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pages 69–117. Springer, 1997.

J.M. Robins, A. Rotnitzky, and D.O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer, 2000.

D.B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L.W. Lehman, G. Moody, T. Heldt, T.H. Kyaw, B. Moody, and R.G. Mark. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. *Critical Care Medicine*, 39(5):952, 2011.

D. Scharfstein, A. McDermott, W. Olson, and F. Wiegand. Global sensitivity analysis for repeated measures studies with informative dropout: A fully parametric approach. *Statistics in Biopharmaceutical Research*, 6(4):338–348, 2014.

P. Schulam and S. Saria. A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. In *Advances in Neural Information Processing Systems (NIPS)*, pages 748–756, 2015.

A. Sokol and N.R. Hansen. Causal interpretation of stochastic differential equations. *Electronic Journal of Probability*, 19(100):1–24, 2014.

R.S. Sutton and A.G. Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

S.L. Taubman, J.M. Robins, M.A. Mittleman, and M.A. Hernán. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *International Journal of Epidemiology*, 38(6):1599–1611, 2009.

J. Taylor, W. Cumberland, and J. Sy. A stochastic model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association*, 89(427):727–736, 1994.

J. Wiens, J. Guttag, and E. Horvitz. Patient risk stratification with time-varying parameters: a multitask learning approach. *Journal of Machine Learning Research (JMLR)*, 17(209):1–23, 2016.

Y. Xu, Y. Xu, and S. Saria. A Bayesian nonparametric approach for estimating individualized treatment-response curves. In *Machine Learning for Healthcare Conference (MLHC)*, pages 282–300, 2016.