

Errors-in-variables Modeling of Personalized Treatment-Response Trajectories

Guangyi Zhang¹, Reza Ashrafi¹, Anne Juuti², Kirsi Pietiläinen², and Pekka Marttinen¹

¹ Aalto University, Finland

{guangyi.zhang,reza.ashrafi,pekka.marttinen}@aalto.fi

² University of Helsinki, Finland

anne.juuti@hus.fi,kirsi.pietilainen@helsinki.fi

Abstract. Estimating the effect of a treatment on a given outcome, conditioned on a vector of covariates, is central in many applications. However, learning the impact of a treatment on a continuous temporal response, when the covariates suffer extensively from measurement error and even the timing of the treatments is uncertain, has not been addressed. We introduce a novel data-driven method that can estimate treatment-response trajectories in this challenging scenario. We model personalized treatment-response curves as a combination of parametric response functions, hierarchically sharing information across individuals, and a sparse Gaussian process for the baseline trend. Importantly, our model considers measurement error not only in treatment covariates, but also in treatment times, a problem which arises in practice for example when treatment information is based on self-reporting. In a challenging and timely problem of estimating the impact of diet on continuous blood glucose measurements, our model leads to significant improvements in estimation accuracy and prediction.

Keywords: Treatment-response trajectory · Bayesian methods · Errors-in-variables · Hierarchical modeling · Gaussian process · Wearable self-monitoring devices

1 Introduction

Increasing popularity of electronic health records (EHRs) and smart healthcare services has led to accumulation of large quantities of heterogeneous data, with potential to considerably improve the efficiency of clinical practice and health services [34]. This highlights the importance of novel machine learning techniques for EHR data, which can be integrated with mobile apps to provide personalized guidance for purposes ranging from early diagnosis to support for lifestyle change [12,27]. The latter is specifically relevant to reduce the cost of chronic diseases in the face of the aging population. For instance, the annual economic cost of diabetes in the U.S. is approximately \$250 billion [1].

Inferring relationships between correlated variables is essential in many fields. An important question is to estimate a patient’s response to a given treatment, comparing the patient’s data from before and after the treatment. A traditional solution is to use randomized controlled trials, which, however may be infeasible due to the cost or ethical considerations. One possibility is to use mechanistic models, specifically tailored for the

problem, and use data to learn about their unknown parameters. However, these models require substantial expert knowledge and are not applicable if the underlying mechanism is unknown. On the other hand, data-driven methods, trained on observational EHR data, provide a promising alternative.

Estimating the impact of a treatment is particularly challenging when the response is a continuous curve consisting, for example, of a time-series of measurements of a biological marker. In such cases, the outcome is typically modeled by a Gaussian process [35], but also neural networks have been considered [21]. The treatment may either be a continuous dose function [42], or a discrete event in time [45,41]. The latter approach is often relevant in practice when treatments are recorded as discrete events, even if their true duration is not exactly zero. Treatment data are usually sparse, and hence it is essential to share relevant information in a probabilistic model. A latent trajectory model of [39] uses additive components to explain variation on population and individual levels. Conditional random fields can be incorporated to further capture correlations between different treatment types [40], and multivariate response curves can be modeled by learning latent structure [42] shared across the outcomes.

Despite increased recent attention, there are still crucial issues in treatment-response estimation that have not been addressed when the response is continuous. Most importantly, the treatments are consistently assumed to be exactly measured and known, while in reality the treatment input may be severely perturbed by numerous factors. This problem dramatically escalates for user-reported treatment data, which potentially results in complete discredit of the findings [20]. The erroneous effect is two-fold, i.e., in addition to measurement error in covariates, the actual time of the treatment might be known only approximately. Another issue arises from the competing relationship between a counterfactual trend, i.e., the evolution of the outcome assuming no treatment, and the treatment response. When modeled and trained jointly, it easily happens that a flexible trend completely overrides the treatment response, and therefore these two components are often trained separately in practice.

To address the mentioned shortcomings, we introduce errors-in-variables (EIV) framework for modeling of continuous treatment-response trajectories. The EIV models account for measurement errors not only in the output variable, as common regression, but also in the inputs [13,14]. They are closely related to latent-variable models in machine learning [28,4], and based on modeling the unobserved true values from which noisy observations are obtained. Our contributions can be summarized as follows:

- We formulate an EIV model for personalized treatment-response trajectories, where a treatment comprises a vector of noisy covariates and treatment times are uncertain.
- We introduce an interpretable hierarchical prior on the treatment effects that efficiently shares information between individuals, and allows training the full model jointly, appropriately balancing between the trend and the responses.
- In a challenging topic, representative of the current technological mega-trend on self-monitoring data from wearable devices, we show our method can meaningfully estimate the personalized impact of diet on continuous blood glucose measurements.

The code and data used in the analyses are available at [link added upon acceptance, reviewers can view the material in Supplement] and allow fully reproducing our results.

2 Related work

Treatment response: Besides machine learning, the problem of treatment response estimation has been studied in various fields, including informatics for medicine and social sciences, where the data-driven approach can bring advantages compared to the experimental trials [30]. For example, individual-level treatment response prediction has been studied for schizophrenia [5] and depression [33]. An empirical comparison of classifiers for treatment-response prediction for chemoradiotherapy appears in [9]. Topics studied in social sciences include the effect of a discrete treatment, years of education, on an individuals' income [6], and allowing a response to depend on social interactions and treatments for other individuals [24].

Mechanistic models: In contrast to the data-driven approaches used in machine learning, mechanistic models use substantial knowledge of a specific problem to characterize the system with differential equations, and inference is done for example using filtering algorithms. Similar to our application, [3] and [2] study blood glucose dynamics, affected by nutrition and other factors. Another example is a computational model of the physiological mechanisms for type-2 diabetes, aiming to quantify factors useful for prevention of the disease [37].

EIV models for treatment response: In contrast to the vast body of research concerning data-driven methods for the prediction of a treatment response, very little is known about their performance when measurement error is present. Authors of [46] study a method for inferring causal directions using EIV models, but they do not focus on the response conditioned on specific treatments. Regression on various student covariates, incorporating EIV, has been used to predict standardized test scores [22]. In [26], the effect of measurement error on a binary treatment response is analyzed, underlining the devastating impact of ignoring such errors. A model for measurement errors has been used to quantify uncertainty in order to increase the confidence in detecting genuine treatment changes for liver metastases [31].

None of these works address the problem of estimating the impact of a multivariate vector of covariates on a continuous response with measurement error in covariates and uncertainty in treatment timing, the topic of this paper.

3 Methods

In this section, we first review EIV models on a general level. Then we describe three essential components of our model for personalized treatment-response trajectories: a hierarchical prior on parametric treatment-responses functions, a Gaussian process model for the trend, and measurement error models. Throughout the section, we present the model in generic terms, but also outline the specific model that we use in Section 4.2 to estimate the impact of diet, recorded as nutrient contents of different meals, on continuous blood glucose measurements.

Our model is fully Bayesian, yielding uncertainty estimates for all parameters, essential in scientific applications. Inference is done using Markov chain Monte Carlo (MCMC) with the state-of-the-art No-U-Turn (NUTS) sampler [17] implemented in software PyMC3 [36]. Implementation details are discussed below and in the supplementary, and can be viewed in full in the published code.

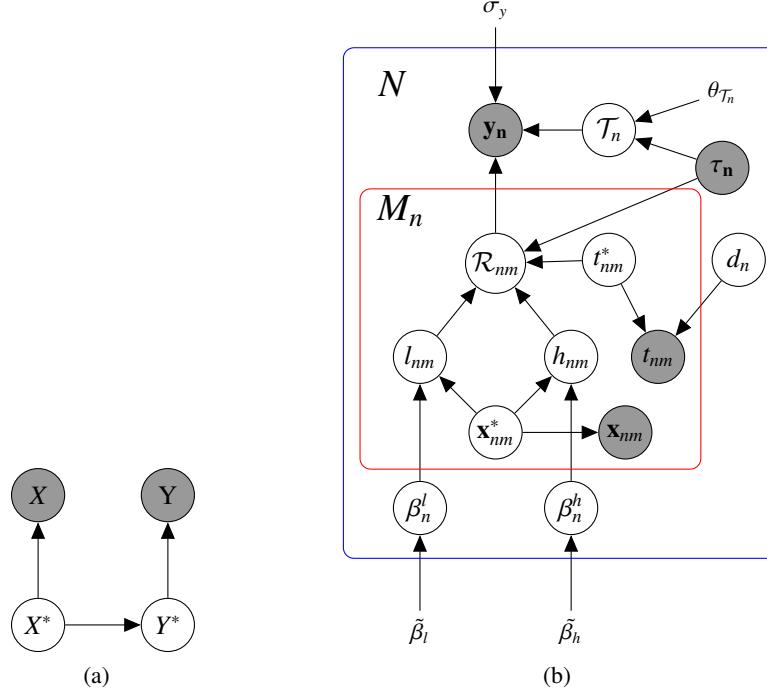


Fig. 1: a) General formulation of the EIV model. For clarity, parameters associated with the distributions are not shown. (b) Model for personalized treatment-response trajectories. Details of the model are discussed in the text.

3.1 Errors-in-variables models

EIV models, a.k.a. measurement error models, are regression or classification models that, in contrast to most existing models, account for errors not only in the output variable but also in inputs [7,38,15,10]. Though commonly neglected, input mismeasurement may be extremely harmful. For example in simple linear regression it leads to biased estimates that can not be corrected for even with an infinite sample, while, on the other hand, unbiased homoscedastic error in the output variable only induces additional variability [7]. A graphical model for a general EIV model is presented in Figure 1a, where X^* and Y^* represent the true values of the inputs and the output, and X and Y are the corresponding noisy observations. The most important type of mismeasurement is *classical error*, which corresponds to independence of an error term from the true value.

Except for the simplest case of linear regression [29], EIV modeling almost always requires auxiliary information or data to correct the mismeasurement bias in estimation. For problems that have an analytical solution, the bias can be corrected by a multiplication or addition of external terms [18], e.g., an estimated reliability ratio [7]. For nonlinear models, auxiliary data, e.g., instrumental variables or repeated measurements, can be exploited to help correct bias, e.g., by estimating the density function of the true variable

using a deconvolution technique [38]. However, without additional data, Bayesian EIV modeling is currently the most powerful and flexible approach, as it allows incorporating additional information in the form of distributional assumptions [15]. In this work, we adopt the Bayesian approach.

Mathematically, the measurement error mechanism is defined as the distribution of the noisy observed input, X , given the true unobserved input, X^* . The joint distribution of the model factorizes accordingly as:

$$P(X^*, Y^*, X, Y, \Theta) = P(X|X^*, \theta_M)P(Y|Y^*, \theta_N)P(Y^*|X^*, \theta_R)P(X^*|\theta_E)P(\Theta), \quad (1)$$

where $P(X|X^*, \theta_M)$ and $P(Y|Y^*, \theta_N)$ are called *error or measurement models*, $P(Y^*|X^*, \theta_R)$ is a *response or outcome model*, $P(X^*|\theta_E)$ is an *exposure model*, and $\Theta = (\theta_M, \theta_N, \theta_R, \theta_E)$ are the corresponding parameters. Bayes theorem can be used to infer the unknown parameters and unobserved true values of the variables.

$$P(X^*, Y^*, \Theta|X, Y) \propto P(\Theta) \prod_i^N P(X_i|X_i^*, \theta_M)P(Y_i|Y_i^*, \theta_N)P(Y_i^*|X_i^*, \theta_R)P(X_i^*, \theta_E). \quad (2)$$

If the exposure model is noninformative and the measurement model is symmetric, i.e., $P(X_i|X_i^*, \theta_M) = P(X_i^*|X_i, \theta_M)$, then the Bayesian modeling of classical error is equivalent to another class of mismeasurement techniques known as *Berkson error modeling* [7].

$$P(X^*, Y^*, \Theta|X, Y) \propto P(\Theta) \prod_i^N P(X_i^*|X_i, \theta_M)P(Y_i|Y_i^*, \theta_N)P(Y_i^*|X_i^*, \theta_R).$$

A well-known difficulty with EIV models is that they are often nonidentifiable [15], i.e. there are more than one set of values for the unknowns leading to the same model. This can be understood intuitively by noticing that the model stays the same if we multiply the linear regression coefficients by a constant factor and at the same time divide the estimated true values of inputs by the same factor. Therefore, to achieve identifiability, some crucial information about measurement model has to be assumed or estimated, e.g., the variance of a classical additive error in simple linear regression [7]. The Bayesian paradigm offers a unique solution to the nonidentifiability of the EIV models, as long as mismeasurement is modest and the prior is sufficiently good [16].

3.2 Model for treatment-response trajectories

Notation: A graph of our model for treatment-response trajectories is presented in Figure 1b. We assume there are N patients, and a trajectory consisting of a time series of length G_n of the outcome (e.g. blood glucose) is observed for each individual:

$$\mathbf{y}_n = (y_{n1}, \dots, y_{nG_n})^T, n = 1, \dots, N.$$

These measurements have been taken at times

$$\tau_n = (\tau_{n1}, \dots, \tau_{nG_n})^T, n = 1, \dots, N.$$

Furthermore, each patient has M_n observed treatments (e.g. meals eaten), indexed by $m \in 1, \dots, M_n$, where each treatment is characterized by P covariates:

$$\mathbf{x}_{nm} = (x_{nm1}, \dots, x_{nmp})^T, \text{ for all } m, n,$$

and the corresponding recorded treatment times are

$$\mathbf{t}_n = (t_{n1}, \dots, t_{nM_n})^T, \text{ for all } n.$$

Here, \mathbf{x}_{nm} and t_{nm} are assumed to be noisy observations of the treatment covariates and timings, and their true unobserved values are denoted by \mathbf{x}_{nm}^* and t_{nm}^* , respectively.

Outcome model: We model the observed outcome trajectory of individual n , \mathbf{y}_n , as

$$\mathbf{y}_n = \mathcal{T}_n + \sum_m \mathcal{R}_{nm} + \mathbf{e},$$

where $\mathcal{T}_n \in \mathbb{R}^{G_n}$ is a counterfactual trend (i.e. it describes the evolution of the outcome had the treatment not been taken), $\mathcal{R}_{nm} \in \mathbb{R}^{G_n}$ is the additive response to the m th treatment, and $\mathbf{e} = (e_1, \dots, e_{G_n})^T$ is the vector of errors with $e_i \sim N(0, \sigma_y^2)$. We note that the sum of the trend and the responses can be viewed as a trajectory for a 'clean' outcome (omitted from Figure 1b), of which a version \mathbf{y}_n corrupted by Gaussian noise is observed. Additive response functions can be seen as a continuous extension of scalar *average treatment effect* (ATE) which is defined as the expected difference of outcomes before and after treatment.

Response function: Response functions specify how treatments affect the outcome over time, and they should be specified to suit the application at hand, balancing flexibility, interpretability, etc. For example, if interpretability is not needed and the amount of data is large, non-parametric functions that learn the shape of the response are attractive. On the other hand, parametric functions are suitable when data are scarce, and they are often interpretable, which is valuable in itself but also helps specifying prior knowledge to improve accuracy. In the application of learning the impact of meals on blood glucose (Section 4.2), we model the treatment response using a bell-shaped parametric function

$$\mathcal{R}_{nm} := f(\mathcal{A}_{nm}, h_{nm}, l_{nm}) := h_{nm} \exp \left\{ \frac{-0.5(\mathcal{A}_{nm} - 3l_{nm})^2}{l_{nm}^2} \right\}, \quad (3)$$

where a lag vector $\mathcal{A}_{nm} = \tau_n - t_{nm}^*$ represents the time since a specific treatment. The shape of this response is shown in Figure 2a and it is determined by two parameters h_{nm} and l_{nm} with straightforward interpretations: h_{nm} is the height of the response, and l_{nm} is the length-scale which is proportional to the 'width' or 'duration' of the response. The main challenge in our application is scarceness and noisiness of data, with only 13 individuals and on average 10 meals per patient. We also tried a more flexible three-parameter response used in [41], which allows a skewed response (see Figure 2a), but this model suffered from convergence problems, for which reason we selected the simpler alternative.

In applications it is often of interest to measure how the response depends on treatment covariates, and therefore we allow these parameters to depend on the covariates:

$$\begin{aligned} h_{nm} &= (\beta_n^h)^T \mathbf{x}_{nm}^*, \text{ and} \\ l_{nm} &= (\beta_n^l)^T \mathbf{x}_{nm}^*, \text{ for all } n, m. \end{aligned} \quad (4)$$

In Equation (4), the coefficient vectors $\beta_n^h, \beta_n^l \in \mathbb{R}^P$ represent the *personalized impact* of each of the P covariates on the height or width of the response for the n th individual. To share information across individuals, we introduce a Bayesian hierarchical prior, see [11], and assume that the personalized height and length-scale coefficients, β_n^h and β_n^l , are drawn from common distributions:

$$\beta_n^h \sim N_P(\tilde{\beta}_h, \Sigma_h) \quad \text{and} \quad \beta_n^l \sim N_P(\tilde{\beta}_l, \Sigma_l).$$

A hyperprior is further placed on the mean parameters of these distributions:

$$\tilde{\beta}_h \sim N_P(\mathbf{0}, \tilde{\Sigma}_h) \quad \text{and} \quad \tilde{\beta}_l \sim N_P(\mathbf{0}, \tilde{\Sigma}_l)$$

The hierarchical prior introduces shrinkage and facilitates estimation of the personalized coefficients with limited data. Further details are given in the Supplementary material.

Counterfactual trend: A counterfactual trend represents the outcome assuming no treatment has been taken. It has to be sufficiently flexible to handle any variation in the outcome that is not accounted for by the treatments. In this paper, we model the trend $\mathcal{T}_n(t)$ for individual n using a Gaussian Process (GP) [35]:

$$\mathcal{T}_n(t) \sim \mathcal{GP}(0, k(t, t' | \theta_{\mathcal{T}_n})),$$

where $\theta_{\mathcal{T}_n}$ are parameters associated with the kernel function $k(x, x' | \theta_{\mathcal{T}_n})$. GPs are non-parametric regression models with well-known closed-form formulas for posterior estimation, which they inherit from the Normal distribution by assuming all training and test data follow a joint Normal distribution. For example, if

$$\mathcal{S}_n = \mathbf{y}_n - \sum_m \mathcal{R}_{nm}$$

is the residual of the outcome after subtracting the impact of the treatment responses, then

$$\begin{aligned} \mathcal{T}_n(t) | \mathcal{S}_n &\sim N(\mu_*, \Sigma_*), \quad \text{where} \\ \mu_* &= k(\tau_n, t)^T K(\tau_n, \tau_n)^{-1} \mathcal{S}_n, \quad \text{and} \\ \Sigma_* &= k(t, t) - k(\tau_n, t)^T K(\tau_n, \tau_n)^{-1} k(\tau_n, t). \end{aligned}$$

We refer the reader to [35] for more details about GPs. As the kernel, we use the sum of Squared Exponential (SE) and constant kernels, where the former equips the GP with desired smoothness, and the latter enables meaningful extrapolation to regions where no input points have been observed. To speed up computation, we use a sparse GP [35] instead of a full GP, which samples a small set of inducing points uniformly from τ_n to achieve a low-rank approximation of $K(\tau_n, \tau_n)$ and its inverse. A detailed prior specification is provided in the Supplementary material.

Measurement models: Measurement models describe error in observations. With self-reported data both covariates and the timing of a treatment may be uncertain. To account for the uncertainty in treatment timing, we assume:

$$t_{nm} \sim N(t_{nm}^* + d_n, (\sigma_n^t)^2), \quad \text{for all } n, m.$$

In words, the observed time t_{nm} is obtained from the true time t_{nm}^* by shifting it with a bias term d_n , and adding Gaussian noise. The bias term d_n represents reporting habits of different individuals. For example, in the blood glucose application in Section 4.2, some individuals may systematically report their meal after eating, while others may do this before eating.

Different models are possible for treatment covariates, depending on the assumptions and data available [15]. Here we assume a simple perturbation on the *amount* of treatment:

$$\begin{aligned}\mathbf{x}_{nm} &= \mathbf{x}_{nm}^* \delta_{nm}, \quad \text{where} \\ \delta_{nm} &\sim \text{LogNormal}(0, \sigma_x^2), \quad \text{for all } n, m.\end{aligned}\tag{5}$$

The coefficient δ_{nm} represents the error in the m th treatment of the n th individual. Intuition in the blood glucose application is that users are able to report correctly what they have eaten, but not how much. While the model (5) captures our understanding of the type of mismeasurement expected in our data, more complicated models could also be justified, but they would require stronger additional assumptions to resolve the nonidentifiability of the EIV models. The model in (5) is identifiable and can be trained with relatively little data, as we demonstrate in Section 4.

Estimating t_{nm}^* is straightforward as it only shifts the response, but does not change its shape. However, estimating \mathbf{x}_{nm}^* is more complicated, and requires assuming that the counterfactual trend is sufficiently regularized. Otherwise the trend could easily compensate for the perturbation. We solve this by encouraging a large length-scale for the squared exponential kernel in the prior. Further details, e.g., prior distributions for \mathbf{x}_{nm}^* , t_{nm}^* , and d_n , are provided in the Supplementary material.

3.3 A note on causality

We briefly review results related to estimation of causal effects from observational data on treatment-response trajectories [32,25,41], to enable a user of our method to judge to what extent the effects found may or may not be interpreted causally. The causal effect of an action A (e.g. a treatment) on Y is defined as $P(Y = y|do(A = a))$, where the $do(\cdot)$ operator represents a manipulation of A to value a . The key assumption is that there are *no unmeasured confounders* (NUC), such as Z_2 in Figure 2b. Without Z_2 , the causal effect of A on Y can be estimated from observational data using the adjustment formula:

$$P(Y = y|do(A = a)) = \sum_{z_1} P(Y = y|A = a, Z_1 = z_1)P(Z_1 = z_1).$$

Time-varying treatments (Figure 2c) further face an issue of treatment-confounder feedback [25], which means that hidden confounders do not have to affect A directly to create a spurious correlation between an action and future observations. A generalized adjustment formula, g-formula, can still be used to calculate $P(\bar{Y}_{\geq t}|do(A_{t-1}, A_t))$, see [8].

A useful result, applicable with our model, is to use the model to estimate $P(\bar{Y}_{\geq t}|\bar{A}_{\leq t}, \bar{Y}_{<t})$ from observational data. Then, assuming NUC, the following holds [25]:

$$P(\bar{Y}_{\geq t}|do(A_t), \bar{A}_{<t}, \bar{Y}_{<t}) = P(\bar{Y}_{\geq t}|\bar{A}_{\leq t}, \bar{Y}_{<t}).\tag{6}$$

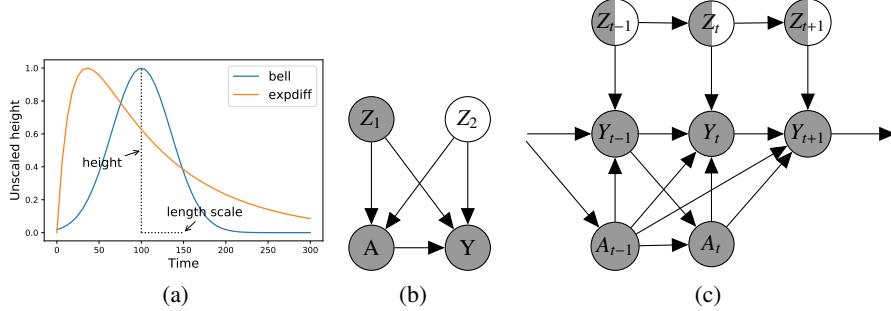


Fig. 2: a) Two Response functions. The blue one is used in this paper, while the orange one is used in [41], (b) Graphical model for a cross-sectional case, showing action (A), response (Y), and observed and hidden confounders (Z_1 and Z_2), and (c) over-time response with a single treatment, where confounders Z can be either observed or hidden.

In words, conditionally on the history of treatments and the outcome (and relevant observed confounders not shown in the formula), the causal impact of the most recent treatment on future outcomes can be estimated from observational data. This *short-term effect* can be used, e.g., to select between alternative treatments available at a certain point in time, when the relevant history of the individual is known.

In Section 4.2 we analyse the impact of diet on blood glucose. Based on domain knowledge, we know that diet is a prominent cause of changes in blood glucose. Furthermore, in our data we often see a rapid increase and decrease in blood glucose after a meal. Therefore, it seems plausible that meals affect blood glucose causally. However, in general, the causal assumptions can not be verified from observational data, and it is possible that some confounder affects both glucose and diet, but the effect of any such confounder is expected to be small compared to the impact of diet. Hence, while interpreting our results causally seems reasonable, we can not make assertions of this. More generally, with emergence of modern wearable self-monitoring devices, it will be possible to measure all relevant factors that could affect blood glucose much more comprehensively, and the NUC assumption is reasonable. Our model is straightforward to extend to such data.

4 Experiments

In this section, we first examine identifiability and accuracy of our model using simulated data, and then use it to analyze a real-world dataset comprising diet and continuous blood glucose measurements. Throughout, we compare four models, in an increasing order of complexity (later models include the previous as special cases):

- \mathcal{M}_{ind} : Separate models for individuals.
- \mathcal{M}_{hier} : Model with the hierarchical prior for the responses to share information across individuals.

- $\mathcal{M}_{hier+time}$: Time uncertainty included.
- $\mathcal{M}_{hier+time+cov}$: Uncertainty in covariates included.

4.1 Identifiability and accuracy of the models on simulated datasets

As a simple experiment, we first study the identifiability of the EIV model when there is measurement error in the covariates. We simulate artificial data using a toy model specified as the sum of a linear trend and the parametric treatment response from Equation (3). The dimension of treatment covariates is here set to 2, and each input is perturbed with an additive term drawn from $N(1, 0.2^2)$. We analyze the data using the EIV model that assumes measurement error, and a model that disregards the noise in the covariates. Results and further details for this simple setup are presented in the Supplementary material, and they show that the EIV model recovers all true inputs and effect sizes with high accuracy, while the model that neglects the noise leads to biased coefficient estimates with wide confidence intervals.

To study the accuracy and identifiability of our method in a more realistic simulated setup, we first fit our model to the real-world data from Subsection 4.2, and use the fitted model to simulate responses and trends for two individuals. We perturb half of the inputs according to Equation (5) and let the model use the perturbed inputs and try to recover the true inputs and parameters. The performance of all models depends on the relative contributions of the trend and responses, and we scale up the response with a factor of 5, which facilitates a meaningful comparison.

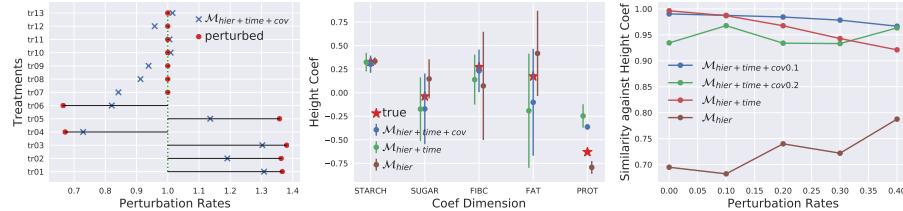


Fig. 3: Simulation results. *Left:* true and estimated perturbations for one individual (the other one shown in Supplement); *Center:* true and estimated coefficients for the height of the response for 5 covariates; *Right:* Cosine similarity of concatenated height coefficient vectors from all individuals against the true value with different levels of perturbation (higher value is better). Two different prior SDs, 0.1 and 0.2, were considered for model $\mathcal{M}_{hier+time+cov}$.

Results for one individual are shown in Figure 3, and for the other in the Supplementary material. We see that the direction of each non-zero perturbation is estimated correctly (left panel), and this is true also for the other individual (Supplementary). On the other hand, if there is no perturbation, the model may even then estimate non-zero perturbations, introducing additional noise. This reflects the trade-off between flexibility and overfitting, and highlights the importance of carefully validating the model to suit the

amount and complexity of data. We also see that the regression coefficients are estimated accurately by the EIV model (center), and that the benefit from using EIV becomes more significant when the size of the perturbation increases (right). However, a too loose EIV prior (large SD) may actually harm the performance by introducing additional noise, when the true perturbation is small.

4.2 Experiments on real-world glucose data

The data contain blood glucose measurements and dietary records. These anonymized data were provided by the Obesity Research Unit at the University of Helsinki, Finland, and they are available for 13 non-diabetic individuals across three days. Diabetic individuals were excluded because their metabolism differs extensively from healthy individuals, and detailed modeling of that is beyond the scope of this work. The real-valued blood glucose measurements were collected by a portable continuous glucose monitoring system approximately every fifteen minutes. The dietary records consist of user-reported contents and times of all meals eaten during the 3-day study period. Each meal has been processed into amounts of five nutrients: starch, sugar, fiber, fat, and protein. The goal of the analysis is to learn how these nutrients influence blood glucose. Both the exact amounts of food eaten and the exact meal times are uncertain, as they are based on values estimated and reported by users, which motivates the use of EIV models for these data. A visualization of the data (and results) for one individual is shown in Figure 4, and for all other individuals in the Supplement. Some markers may be missing due to device errors or when a user has removed the device.

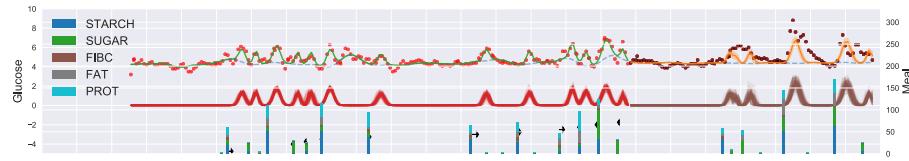


Fig. 4: Visualization of the 3-day time series for one patient. Red and brown dots represent glucose markers in the training and test sets, respectively. Meals are indicated by vertical bars, colored according to amounts of different nutrients in the meal. The green curve is the final fitted trajectory, and it is a combination of the dashed blue line, a counterfactual trend, and the mean of red lines, which are posterior samples of treatment responses. Horizontal arrows associated with the meals show the estimated difference between true and observed meal times.

Metrics: The models are trained using data from the first two days, and the third day is used for testing. The performance of treatment-response estimation is quantified using five metrics $M_i, i \in \{1, \dots, 5\}$. M_1 is the proportion of variance explained by the trend:

$$M_1 = \frac{1}{N} \sum_n \frac{\text{Var}(\mathcal{T}_n)}{\text{Var}(y_n)}.$$

M_2 indicates how much more is explained when also the treatment responses are included:

$$M_2 = \frac{1}{N} \sum_n \frac{\text{Var}(\mathcal{T}_n + \sum_m \mathcal{R}_{nm})}{\text{Var}(\mathbf{y}_n)} - M_1.$$

In detail, a large M_1 means that the outcome is mostly explained by the trend, and a small M_2 represents an inactive treatment response. These metrics are computed in regions of non-zero treatment response. Metrics M_3 and M_4 are simply the mean squared errors in the training and test data. They are calculated for all individuals for whom M_2 indicates that the response has been properly learned. Thus one patient, shown in Figure 5, with $M_2 \approx 0.05$ for the baseline model \mathcal{M}_{hier} is excluded from MSE calculations (other patients have $M_2 > 0.3$).

Because M_4 measures pointwise error, it may be misleadingly low when the response shape is correct if its location is inaccurate. Metric M_5 is insensitive to the inaccuracy of location, and it measures the absolute error in variance between predicted response and outcome:

$$M_5 = \frac{1}{N} \sum_n |\text{Var}\left(\sum_m \mathcal{R}_{nm}\right) - \text{Var}(\mathbf{y}_n)|$$

Because our interest is in estimation of the treatment response, and not the trend, we calculate M_4 and M_5 in windows from one hour before to three hours after each meal.

We use the Mann–Whitney U-test [23] to test if other models are better than \mathcal{M}_{hier} in terms of test error M_4 . The reason for using \mathcal{M}_{hier} as the baseline is the main argument of this article that EIV modeling is beneficial when estimating treatment-response trajectories, and \mathcal{M}_{hier} is otherwise the same as $\mathcal{M}_{hier+time}$ and $\mathcal{M}_{hier+time+cov}$ except that it does not include the EIV components. We also compare the models using an information criterion for predictive accuracy. The state-of-the-art criterion is leave-one-out cross-validation (LOO) [43], which is used here.

	M_1 PVE Trend	M_2 PVE Resp	M_3 MSE Train	M_4 MSE Test	M_5 ΔVar Test	p-value U-test	LOO	pLOO	SE LOO
\mathcal{M}_{ind}	0.361	0.342	0.149	1.695	0.927	1.00	3549.64	246.64	318.8
\mathcal{M}_{hier}	0.359	0.339	0.159	0.752	0.391	-	3587.87	214.62	317.28
$\mathcal{M}_{hier+time}$	0.350	0.424	0.098	0.738	0.377	3.24e-4	2869.91	342.24	265.09
$\mathcal{M}_{hier+time+cov}$	0.344	0.428	0.098	0.743	0.366	4.66e-3	2994.98	465.47	333.7

Table 1: Comparison of models using the real-world glucose data. The metrics M_1 through M_5 are defined in text, where PVE means *Proportion of Variance Explained*. p-value tests if other models are better than \mathcal{M}_{hier} in terms of M_4 . LOO stands for leave-one-out cross-validation, pLOO is the estimated effective number of parameters, and SE-LOO records the standard error in the LOO computations.

Results: Result are shown in Table 1. We see that all models outperform the non-hierarchical baseline \mathcal{M}_{ind} by a large margin. Furthermore, taking treatment time inaccuracy into account in $\mathcal{M}_{hier+time}$ improves significantly over the non-EIV model

\mathcal{M}_{hier} . In fact, estimation of the response fails completely for some individuals without time EIV; the results with and without time uncertainty modeling for one such case are shown in Figure 5. On the other hand, taking uncertainty in covariates into account does not notably improve accuracy, which is likely caused by a combination of increased flexibility and limited amount of data. Overall, models with EIV component outperform the model without EIV in all metrics.

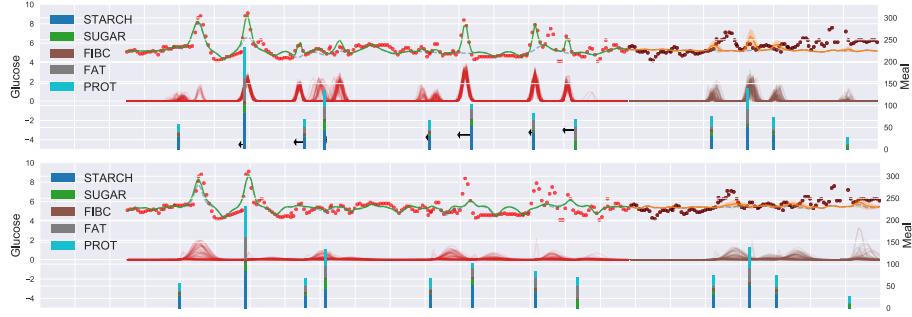


Fig. 5: Demonstration of time uncertainty modeling for one individual. *Upper:* Results using $\mathcal{M}_{hier+time}$, where arrows indicate the estimated difference between the true and observed meal times; *Bottom:* Results using \mathcal{M}_{hier} .

Interpretability of personalized treatment response is also of great interest; for instance, understanding how an individual's glucose level changes if she eats one more unit of sugar. The overall goal of glucose monitoring is to keep the glucose level in a given range, and both the amount of excess and the duration of the hyperglycemic state are clinically important. Hence, a sensible parameter to consider is the impact of different nutrients on the *area* of the response curve. Though this is not a parameter of our model, it is straightforward to derive the personalized increase in response area due to one unit increase of a specific nutrient ΔA_{np} ($n \in 1, \dots, N$, $p \in \{1 \dots P\}$), using coefficients for height and width, which are modeled explicitly (see Supplement).

Overall, starch and sugar have the strongest positive impact on glucose (Figure 6a), consistent with the understanding that carbohydrates increase blood glucose [44]. Protein, on the other hand, has a negative impact, which has been observed before and might represent a complex short-term interaction between nutrients [19]. An advantage of our model is that we get *personalized* coefficients for each individual, as shown for starch in Figure 6b, and for the other nutrients in the Supplement. Finally, posterior uncertainty of personalized starch coefficients is shown in Figure 6c. Importantly, models with EIV have much narrower confidence intervals, meaning that they are estimated more accurately, thanks to increased flexibility that allows fitting the complex data.

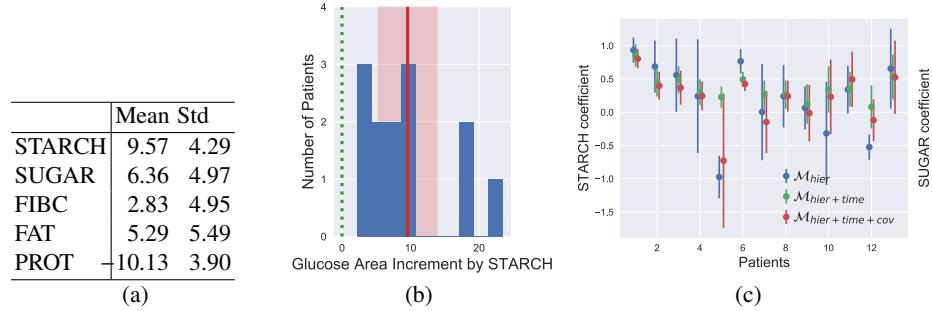


Fig. 6: a). Average impact on response area ΔA_{np} by different nutrients; b) Histogram of personalized starch coefficients and their mean (\pm one SD) (red); c) Posterior uncertainty in the personalized starch coefficients.

5 Conclusion

We presented a hierarchical model with EIV components to estimate personalized treatment-response trajectories when the covariates and timing of a treatment are imprecise. Our model demonstrates superior performance in both simulated and real-world data on various metrics, and allows extracting interpretable and meaningful estimates of the personalized impacts of treatment covariates, valuable in applications. Future directions include extensions and identifiability of EIV modelling, and extending the model to include interactions between covariates and other unmeasured confounders, such as physical activity, for causal completeness.

References

- ADA: Economic costs of diabetes in the U.S. in 2012. *Diabetes Care* **36**(4), 1033–1046 (March 2013). <https://doi.org/10.2337/dc12-2625>, American Diabetes Association
- Albers, D.J., Levine, M., Gluckman, B., Ginsberg, H., Hripcsak, G., Mamikina, L.: Personalized glucose forecasting for type 2 diabetes using data assimilation. *PLoS computational biology* **13**(4), e1005232 (2017)
- Balakrishnan, N.P., Samavedham, L., Rangaiah, G.P.: Personalized mechanistic models for exercise, meal and insulin interventions in children and adolescents with type 1 diabetes. *Journal of Theoretical Biology* **357**, 62 – 73 (2014). <https://doi.org/https://doi.org/10.1016/j.jtbi.2014.04.038>
- Bishop, C.M.: *Pattern Recognition and Machine Learning*. Springer-Verlag, Berlin, Heidelberg (2006)
- Cao, B., Cho, R.Y., Chen, D., Xiu, M., Wang, L., Soares, J.C., Zhang, X.Y.: Treatment response prediction and individualized identification of first-episode drug-naïve schizophrenia using brain functional connectivity. *Molecular Psychiatry* (June 2018). <https://doi.org/10.1038/s41380-018-0106-5>
- Card, D.: The causal effect of education on earnings. In: *Handbook of Labor Economics*, vol. 3, Part A, chap. 30, pp. 1801–1863. Elsevier, 1 edn. (1999)
- Carroll, R.J., Ruppert, D., Crainiceanu, C.M., Stefanski, L.A.: *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC (2006)

8. Daniel, R.M., Cousens, S., De Stavola, B., Kenward, M.G., Sterne, J.: Methods for dealing with time-dependent confounding. *Statistics in medicine* **32**(9), 1584–1618 (2013)
9. Deist, T.M., Dankers, F.J.W.M., Valdes, G., Wijsman, R., Hsu, I.C., Oberije, C., Lustberg, T., van Soest, J., Hoebers, F., Jochems, A., El Naqa, I., Wee, L., Morin, O., Raleigh, D.R., Bots, W., Kaanders, J.H., Belderbos, J., Kwint, M., Solberg, T., Monshouwer, R., Bussink, J., Dekker, A., Lambin, P.: Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. *Medical Physics* **45**(7), 3449–3459 (2018). <https://doi.org/10.1002/mp.12967>
10. Fuller, W.: *Measurement Error Models*. Wiley, New York (1987)
11. Gelman, A., Stern, H.S., Carlin, J.B., Dunson, D.B., Vehtari, A., Rubin, D.B.: *Bayesian data analysis*. Chapman and Hall/CRC (2013)
12. Ghassemi, M., Naumann, T., Schulam, P., Beam, A.L., Ranganath, R.: Opportunities in machine learning for healthcare. *CoRR* **abs/1806.00388** (2018), <http://arxiv.org/abs/1806.00388>
13. Griliches, Z.: Errors in Variables and Other Unobservables. *Econometrica* **42**(6), 971–998 (November 1974), <https://ideas.repec.org/a/ecm/emetrp/v42y1974i6p971-98.html>
14. Griliches, Z., Hausman, J.A.: Errors in variables in panel data. *Journal of Econometrics* **31**(1), 93 – 118 (1986). [https://doi.org/https://doi.org/10.1016/0304-4076\(86\)90058-8](https://doi.org/https://doi.org/10.1016/0304-4076(86)90058-8)
15. Gustafson, P.: *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. CRC Press, New York, 1 edn. (2004)
16. Gustafson, P., Le, N.D., Saskin, R.: Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics* **57**(2), 598–609 (2001)
17. Hoffman, M.D., Gelman, A.: The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**(1), 1593–1623 (2014)
18. Hwang, J.T.: Multiplicative errors-in-variables models with applications to recent data released by the us department of energy. *Journal of the American Statistical Association* **81**(395), 680–688 (1986)
19. Karamanlis, A., Chaikomin, R., Doran, S., Bellon, M., Bartholomeusz, F.D., Wishart, J.M., Jones, K.L., Horowitz, M., Rayner, C.K.: Effects of protein on glycemic and incretin responses and gastric emptying after oral glucose in healthy subjects. *The American Journal of Clinical Nutrition* **86**(5), 1364–1368 (November 2007). <https://doi.org/10.1093/ajcn/86.5.1364>
20. Kreider, B.: Regression coefficient identification decay in the presence of infrequent classification errors. *The Review of Economics and Statistics* **92**(4), 1017–1023 (2010)
21. Lim, B.: Forecasting treatment responses over time using recurrent marginal structural networks. In: *Advances in Neural Information Processing Systems*. pp. 7494–7504 (2018)
22. Lockwood, J.R., McCaffrey, D.F.: Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics* **39**(1), 22–52 (2014). <https://doi.org/10.3102/1076998613509405>
23. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* pp. 50–60 (1947)
24. Manski, C.F.: Identification of treatment response with social interactions. *The Econometrics Journal* **16**(1), S1–S23 (2013). <https://doi.org/10.1111/j.1368-423X.2012.00368.x>
25. Miguel A. Hernán, J.M.R.: Causal inference (2018), preprint on webpage at <https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/>
26. Millimet, D.: The elephant in the corner: A cautionary tale about measurement error in treatment effects models. *IZA Discussion Papers* 5140, Institute for the Study of Labor (IZA) (2010)
27. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics* **19**(6), 1236–1246 (2018). <https://doi.org/10.1093/bib/bbx044>

28. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT Press, 1 edn. (Aug 2013)
29. Pal, M.: Consistent moment estimators of regression coefficients in the presence of errors in variables. *Journal of Econometrics* **14**(3), 349–364 (1980)
30. Passos, I.C., Mwangi, B.: Machine learning-guided intervention trials to predict treatment response at an individual patient level: an important second step following randomized clinical trials. *Molecular Psychiatry* (September 2018). <https://doi.org/10.1038/s41380-018-0250-y>
31. Pathak, R., Ragheb, H., Thacker, N.A., Morris, D.M., Amiri, H., Kuijer, J., deSouza, N.M., Heerschap, A., Jackson, A.: A data-driven statistical model that estimates measurement uncertainty improves interpretation of adc reproducibility: a multi-site study of liver metastases. *Scientific Reports* **7**(1), 14084–14094 (2017). <https://doi.org/10.1038/s41598-017-14625-0>
32. Pearl, J.: Causality. Cambridge university press (2009)
33. Pearson, R., Pisner, D., Meyer, B., Shumake, J., Beevers, C.G.: A machine learning ensemble to predict treatment outcomes following an internet intervention for depression. *Psychological Medicine* p. 1–12 (2018). <https://doi.org/10.1017/S003329171800315X>
34. Powell, J., Buchan, I.: Electronic health records should support clinical research. *Journal of medical Internet research* **7**(1), 93 – 118 (2005). <https://doi.org/10.2196/jmir.7.1.e4>
35. Rasmussen, C.E.: Gaussian processes in machine learning. In: Advanced lectures on machine learning, pp. 63–71. Springer (2004)
36. Salvatier, J., Wiecki, T.V., Fonnesbeck, C.: Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* **2**, e55 (2016)
37. Sarkar, J., Dwivedi, G., Chen, Q., Sheu, I.E., Paich, M., Chelini, C.M., D’Alessandro, P.M., Burns, S.P.: A long-term mechanistic computational model of physiological factors driving the onset of type 2 diabetes in an individual. *PLOS ONE* **13**(2), 1–37 (02 2018). <https://doi.org/10.1371/journal.pone.0192472>
38. Schennach, S.M., Hu, Y., Lewbel, A.: Nonparametric identification of the classical errors-in-variables model without side information. Tech. rep., cemmap working paper (2007)
39. Schulam, P., Saria, S.: A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. In: Advances in Neural Information Processing Systems. pp. 748–756 (2015)
40. Schulam, P., Saria, S.: Integrative analysis using coupled latent variable models for individualizing prognoses. *The Journal of Machine Learning Research* **17**(1), 8244–8278 (2016)
41. Schulam, P., Saria, S.: Reliable decision support using counterfactual models. In: Advances in Neural Information Processing Systems. pp. 1697–1708 (2017)
42. Soleimani, H., Subbaswamy, A., Saria, S.: Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions. arXiv preprint arXiv:1704.02038 (2017)
43. Vehtari, A., Gelman, A., Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27**(5), 1413–1432 (2017)
44. Wolever, T.M., Miller, J.B.: Sugars and blood glucose control. *The American Journal of Clinical Nutrition* **62**(1), 212S–221S (July 1995). <https://doi.org/10.1093/ajcn/62.1.212s>
45. Xu, Y., Xu, Y., Saria, S.: A non-parametric bayesian approach for estimating treatment-response curves from sparse time series. In: Proceedings of the 1st Machine Learning for Healthcare Conference. vol. 56, pp. 282–300 (August 2016)
46. Zhang, Y., Luo, G.: Inferring causal directions in errors-in-variables models. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (July 2014)