Journal of Machine Learning Research 17 (2016) 1-24 Submitted 8/15; Revised 12/15; Published 6/16

The Factorized Self-Controlled Case Series Method: An Approach for Estimating the Effects of Many Drugs on Many Outcomes

Ramin Moghaddass RAMIN@MIAMI.EDU
Department of Industrial Engineering
University of Miami
Coral Gables, FL, USA

Cynthia Rudin CYNTHIA@CS.DUKE.EDU
Department of Computer Science
Department of Electrical and Computer Engineering
Duke University
Durham, NC, USA

David Madigan MADIGAN@STAT.COLUMBIA.EDU
Department of Statistics
Columbia University
New York, NY, USA

Editor: Benjamin M. Marlin, C. David Page, and Suchi Saria

Abstract

We provide a hierarchical Bayesian model for estimating the effects of transient drug exposures on a collection of health outcomes, where the effects of all drugs on all outcomes are estimated simultaneously. The method possesses properties that allow it to handle important challenges of dealing with large-scale longitudinal observational databases. In particular, this model is a generalization of the self-controlled case series (SCCS) method, meaning that certain patient specific baseline rates never need to be estimated. Further, this model is formulated with layers of latent factors, which substantially reduces the number of parameters and helps with interpretability by illuminating latent classes of drugs and outcomes. We believe our work is the first to consider multivariate SCCS (in the sense of multiple outcomes) and is the first to couple latent factor analysis with SCCS. We demonstrate the approach by estimating the effects of various time-sensitive insulin treatments for diabetes.

Keywords: Bayesian Analysis, Drug Safety, Self-Controlled Case Series, Matrix Factorization, Effect Size Estimation

1. Introduction

The medical community, the pharmaceutical industry, and health authorities are obligated to confirm that marketed medical products and prescription drugs have acceptable benefit-risk profiles; in fact, these entities have come under increasing scientific, regulatory, and public scrutiny to accurately estimate the effects of drugs. The increasing availability of large-scale longitudinal observational healthcare databases (LODs) opens up exciting new

opportunities to add to the evidence base concerning these issues, though the complexity and scale of some of the available databases presents interesting statistical and computational challenges. In what follows we focus on using longitudinal observational databases to make inference about the effects of many drugs with respect to many outcomes simultaneously.

Many research studies have attempted to characterize the relationship between time-varying drug exposures and adverse events (AEs) related to health outcomes (e.g., in Madigan et al., 2011; Greene et al., 2011; Benchimol et al., 2013; Simpson et al., 2013; Chui et al., 2014) and the use of LODs to study *individual* drug-adverse effect combinations has become routine. The medical literature provides many examples and many different epidemiological and statistical approaches, often tailored to the specific drug and specific adverse effect. There is a major flaw in these approaches of estimating the effect of one drug on one outcome, which is that it is very clear that many drugs are closely related to each other (there are dozens of antibiotics for instance), and many health outcomes are closely related to each other (e.g., strokes, heart attacks, and other vascular diseases). In this work, we borrow strength across both drugs and outcomes in order to obtain better estimates for each individual drug and outcome. Since we are interested in the effects of drugs, and not in the patient-specific baseline rate of the outcome, we use the ideas of the self-controlled case series (SCCS) method of Farrington (1995), which is a conditional Poisson regression approach wherein each patient serves as his or her own control. The SCCS method has been widely applied, especially in vaccine studies (see the tutorial of Whitaker et al., 2006). SCCS controls for all fixed patient-level covariates but remains susceptible to time-varying confounding. The standard SCCS method focuses on one drug and one outcome. Simpson et al. (2013) introduced the high-dimensional multiple self-controlled case series (MSCCS) method that simultaneously provides effect estimates for multiple drugs and a single outcome. In fact, the MSCCS provides a self-controlled approach that can control for many time-varying covariates, drugs being a special case. Bayesian implementations of both SCCS and MSCCS provide significant advantages, especially in high-dimensional settings with thousands or even tens of thousands of drugs and outcomes and even larger numbers of interactions. Suchard et al. (2013a) and Madigan et al. (2014) describe large-scale empirical evaluations of SCCS and MSCCS in comparison with other standard methods for effect size estimation.

Neither SCCS nor MSCCS account for the fact that many drugs/treatments naturally form classes and therefore regression coefficients for drugs from within a single class might reasonably be modeled as arising exchangeably from a common prior distribution. Adverse events and health conditions can also be organized hierarchically, again affording an opportunity to "borrow strength" across related outcomes. For both drugs and outcomes, the hierarchy could extend to multiple levels. In what follows, we formalize these ideas within the framework of latent factor Bayesian hierarchical models.

Factor models, which have been traditionally used in behavioral sciences and bioinformatics, provide a flexible framework for modeling multivariate data via unobserved latent factors (e.g., Ghosh and Dunson, 2009; Carvalho et al., 2008). In this paper, we do not impose specific latent structure *a priori*. However, our approach can also be used for cases where classes of drugs and conditions are known *a priori*. We will show that the latent factor approach not only brings more interpretability to our model, but also can significantly contribute to reducing the computational complexity. To our knowledge, only a

few authors have previously considered matrix factorization-based data analysis techniques for drug safety and surveillance (for example, Zitnik and Zupan 2014, for drug-induced liver injury prediction and Cobanoglu et al. 2013, for predicting drug-target interactions in neurobiological disorders, which are both very different from our study).

We introduce three models for predicting the effects of multiple drugs on multiple outcomes that use hierarchical Bayesian analysis. The first model (Model 0) does not use latent factors, and borrows strength across all drugs and outcomes. The second model (Model 1) uses one set of latent drugs and one set of latent outcomes, through a single matrix factorization. The third model (Model 2) uses two sets of latent factors, by factoring the matrix of coefficients into three matrices; one for converting drugs to latent drugs, another for converting outcomes to latent outcomes, and the third for modeling the effects of latent drugs on latent outcomes. By allowing for latent factors, the second and third models provide an increased level of interpretability, use fewer variables, and are thus more computationally efficient to estimate.

The rest of this paper is organized as follows: Section 2 provides an overview of the self-controlled case series (SCCS) method. In Sections 3, 4.1, 4.2, and 4.3 we describe the model and the Bayesian inference procedure. We then use a series of simulations in Section 5 to show that we can recover the true generating parameters from data. Finally, we demonstrate the approach in Section 6 for estimating the effects of various insulin treatments for diabetes. Our proposed methodology has broader applicability beyond estimating the effects of drugs considered in this paper.

2. Background: Overview of the Self-Controlled Case Series (SCCS)

The self-controlled case series method (Farrington, 1995) models the event rate during drug exposure in comparison to the baseline event rate while unexposed (see Whitaker et al., 2006; Madigan et al., 2010; Suchard et al., 2013a). In the self-controlled case series method, each individual also acts as their own control. Each treatment observation, which is a period of time that someone is drug-exposed, is considered with respect to other periods of time in which the same person is not exposed. This way of matching gracefully avoids patient-level selection bias; it controls for all fixed confounders, such as the individual's underlying frailty, the severity of their underlying disease, genetics, socioeconomic status, and so on. Further, because of the way the SCCS model is designed around this choice, the non-time dependent factors for each person cancel within the formula for the likelihood, and do not appear in the likelihood at all. This allows us to focus our modeling efforts on the time-dependent terms that involve the effects of the drugs.

To obtain SCCS's benefits, we also suffer its disadvantages and assumptions. First, SCCS is susceptible to bias due to potential unmeasured time-varying confounders. (However, SCCS does account for non-time-varying confounding.) This means we should include all features that affect the outcome and vary over time. Second, SCCS assumes that treatment effects are homogeneous across subjects. This avoids having to model patient-specific effects. However, it is possible that patients experience different effects from the various treatments. It is possible to create extensions of our approach that include patient specific random effects if desired. Third, the basic version of SCCS assumes that future outcomes are independent of past ones, but this can be changed, as discussed later. Conditional on

the model parameters, outcomes are assumed to be independent of each other, although because we are using latent factors, there can be marginal dependences among the outcomes.

In the SCCS, events are modeled as arising from a non-homogeneous Poisson process. The event rate varies over time, based on exposure to drugs. Each patient $i = 1, \cdots, N$ carries an unknown individual baseline event rate of e^{ϕ_i}. The exposure to drug $j = 1, ..., J$ measured each day results in a multiplicative effect of e^{β_j} to this baseline rate e^{ϕ_i}. The historical data for patient i on day d ($d = 1, \cdots, \tau_i$) includes a vector of drug exposure as $\boldsymbol{x}_{id} = [x_{id1}, x_{id2}, ..., x_{idJ}]^\top$, where $x_{idj} = 1$ if patient i is exposed to drug j on day d and 0 otherwise. The SCCS defines $\lambda_{id} = \exp(\phi_i + \boldsymbol{x}_{id}^\top \boldsymbol{\beta})$ as the Poisson event rate for patient i on interval d, where $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_J]^\top$ are regression coefficients. We denote y_{id} as the number of events that patient i experiences on day d. Conditioning on the total number of events for patient i, denoted by n_i, nuisance quantities ϕ_i cancel out of the SCCS likelihood, leaving log-likelihood as follows:

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_i^N \left[\sum_d^{\tau_i} y_{id}\, \boldsymbol{x}_{id}^\top \boldsymbol{\beta} - n_i \log \left(\sum_d^{\tau_i} e^{\boldsymbol{x}_{id}^\top \boldsymbol{\beta}} \right) \right]. \tag{1}$$

Since larger LODs can contain millions of patients, avoiding estimation of the patient-specific baseline rates represents a significant computational and statistical advantage.

The most basic version of the SCCS deals with one drug and estimates a single unknown, β_1, the effect estimate for the target drug of direct interest. However, most patients in longitudinal healthcare databases often take multiple drugs and treatments throughout the course of their observation and also experience multiple health outcomes. This motivates us to use a multiple-drug, multiple-outcome analysis.

3. Multi-drug, Multi-Outcome Self-Controlled Case Series - Notation and Inference

The methods proposed here generalize the self-controlled case series to handle multiple drugs/treatments and multiple outcomes/conditions. We describe the extended SCCS/MSCCS where there are J drugs and O health outcomes. The notation used throughout the paper is as follows:

N: number of patients (i indexes individuals from 1 to N).
x_{idj}: binary indicator reflecting whether patient i is exposed to drug j on interval d.
$\boldsymbol{x}_{id} = [x_{id1}, x_{id2}, ..., x_{idJ}]^\top$: the vector of exposed drugs for patient i on interval d.
J: number of drugs (treatments).
O: number of health outcomes (adverse events).
D_i^o: the set of observation intervals where patient i has outcome o.
τ_i^o: the number of observation intervals where patient i has outcome o (the size of D_i^o).
y_{id}^o: binary indicator reflecting whether patient i has outcome o on interval d.
$\boldsymbol{y}_i^o = [y_{i1}^o, y_{i2}^o, ..., y_{i\tau_i^o}^o]^\top$: the vector of observed outcomes o for patient i.
ϕ_i^o: baseline incidence of outcome o for patient i.

$\boldsymbol{\Phi} = \begin{pmatrix} \phi_1^1 & ... & \phi_1^O \\ : & : & : \\ \phi_N^1 & ... & \phi_N^O \end{pmatrix}$: baseline incidence matrix.

β_j^o: regression coefficients associated with outcome o and drug j.
$\boldsymbol{\beta}^o = [\beta_1^o, \beta_2^o, ..., \beta_J^o]^\top$: regression coefficients associated with outcome o.

$\mathbf{B} = \begin{pmatrix} \beta_1^1 & ... & \beta_1^O \\ : & : & : \\ \beta_J^1 & ... & \beta_J^O \end{pmatrix}$: drug-outcome coefficient matrix.

$\lambda_{id}^0 = \exp(\phi_i^o + \boldsymbol{x}_{id}^\top \boldsymbol{\beta}^o)$: the Poisson event rate of outcome o, for patient i, on interval d.

Similar to the SCCS, outcomes occur according to a nonhomogeneous Poisson process, where drug exposure can modulate the rate over time. Patient i has an individual baseline rate of $\exp(\phi_i^o)$ for outcome o that remains constant over time. Drug j has a multiplicative effect of $\exp(\beta_j^o)$ on the individual baseline rate $\exp(\phi_i^o)$ during its exposure period. The Poisson event rate for outcome o and patient i on interval d according to the SCCS is

$$\lambda_{id}^o = \exp(\phi_i^o + \boldsymbol{x}_{id}^\top \boldsymbol{\beta}^o).$$

The key benefit of the SCCS is that the ϕ_i^o terms do not need to be modeled, since we are interested in the ratio of Poisson intensities with and without the drug. For instance, considering only one drug j, comparing the intensity ratio for day d_1 to a different day d_2 with no exposure to the drug, we have

$$\frac{\lambda_{id_1}^o}{\lambda_{id_2}^o} = \frac{\exp(\phi_i^o + 1\beta_j^o)}{\exp(\phi_i^o + 0\beta_j^o)} = \exp(\beta_j^o).$$

As the Poisson rate is assumed to be constant within each interval, the number of outcomes o observed for patient i on interval d is distributed as a Poisson random variable (r.v.) denoted by Y_{id}^o as

$$\Pr(Y_{id}^o = y_{id}^o | \boldsymbol{x}_{id}) = \frac{e^{-\lambda_{id}^o} \lambda_{id}^{o\, y_{id}^o}}{y_{id}^o!}.$$

Based on the above, the contribution to the likelihood for patient i and outcome o for the observed sequence of events $\boldsymbol{y}_i^o = [y_{i1}^o, y_{i2}^o, ..., y_{i\tau_i^o}^o]^\top$, conditioned on the observed exposures $\boldsymbol{x}_i = [\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{i\tau_i^o}]$ is

$$\begin{aligned} \mathcal{L}_i^o = \Pr(\boldsymbol{y}_i^o | \boldsymbol{x}_i) &= \prod_{d \in D_i^o} \Pr(y_{id}^o | \boldsymbol{x}_{id}) = \exp\left(- \sum_{d \in D_i^o} e^{\phi_i^o + \boldsymbol{x}_{id}^\top \boldsymbol{\beta}^o} \right) \prod_{d \in D_i^o} \frac{\left(e^{\phi_i^o + \boldsymbol{x}_{id}^\top \boldsymbol{\beta}^o} \right)^{y_{id}^o}}{y_{id}^o!} \qquad (2) \\ &= \exp\left(-e^{\phi_i^o} \sum_{d \in D_i^o} e^{\boldsymbol{x}_{id}^\top \boldsymbol{\beta}^o} \right) \prod_{d \in D_i^o} e^{\phi_i^o y_{id}^o} \prod_{d \in D_i^o} \frac{\left(e^{\boldsymbol{x}_{id}^\top \boldsymbol{\beta}^o} \right)^{y_{id}^o}}{y_{id}^o!} \\ &= \exp\left(\phi_i^o n_i^o - e^{\phi_i^o} \sum_{d \in D_i^o} e^{\boldsymbol{x}_{id}^\top \boldsymbol{\beta}^o} \right) \prod_{d \in D_i^o} \frac{\left(e^{\boldsymbol{x}_{id}^\top \boldsymbol{\beta}^o} \right)^{y_{id}^o}}{y_{id}^o!}, \end{aligned}$$

where $n_i^o = \sum_d y_{id}^o$.

Two key assumptions underly the above likelihood. First, the model assumes that future outcomes are independent of past outcomes. For certain outcomes (e.g., myocardial infarction) this may not be reasonable. Simpson (2013), Schuemie et al. (2014), and Farrington et al. (2011) consider SCCS generalizations that allow for such dependence; in future work it is possible to consider similar generalizations of the method proposed here. The SCCS model also assumes that conditional on the parameters, outcomes are independent of each other. The latent structure, however, allows for arbitrary marginal dependence among outcomes.

One could form the full likelihood to estimate the unknown parameters $(\mathbf{\Phi}, \mathbf{B})$. In order to avoid estimating the nuisance parameter set $\mathbf{\Phi}$, we can condition on its sufficient statistic, which removes the dependence on $\mathbf{\Phi}$. The cumulative intensity is a sum (rather than an integral) since we assume a constant intensity over each interval. Conditioning on n_i^o yields the following likelihood for person i:

$$\mathcal{L}_i^o = \Pr(\boldsymbol{y}_i^o | \boldsymbol{x}_i, n_i^o) = \frac{\prod_{d \in D_i^o} \Pr(y_{id}^o | \boldsymbol{x}_{id})}{\Pr(n_i^o | \boldsymbol{x}_i)} = \frac{\prod_{d \in D_i^o} \Pr(y_{id}^o | \boldsymbol{x}_{id})}{\left[\frac{\left(\exp\left(-\sum_{d \in D_i^o} \lambda_{id}^o \right) \right) \left(\sum_{d \in D_i^o} \lambda_{id}^o \right)^{n_i^o}}{n_i^o!} \right]} \tag{3}$$

$$\propto \exp \prod_{d \in D_i^o} \left(\frac{e^{\boldsymbol{x}_{id}^\top \boldsymbol{\beta}^o}}{\sum_{d'} e^{\boldsymbol{x}_{id'}^\top \boldsymbol{\beta}^o}} \right)^{y_{id}^o}.$$

Notice that because n_i^o is sufficient, the individual likelihood in the above expression no longer contains $\mathbf{\Phi}$. Assuming that patients are independent and outcomes are conditionally independent, the full conditional likelihood for event o is simply the product of the individual likelihoods (i.e. $\mathcal{L}^o = \prod_{i=1}^{N} \mathcal{L}_i^o$). Intuitively it follows that if i has no outcomes of type o, it cannot provide any information about the relative rate of outcome o.

Using the notation and the formula for the likelihood established in this section, we next present three hierarchical models called Factorized Self-Controlled Case Series methods, for multiple drug, multiple outcome analysis and discuss how to estimate the drug-outcome coefficient matrix $\mathbf{B}$. Two of the models have latent factors that allow $\mathbf{B}$ to be expressed in a simpler and more interpretable way. In our experiments, the empirical performance of these methods is approximately the same.

4. Factorized Self-Controlled Case Series (FSCCS)

Building on the notation in the previous section, this section describes the proposed self-controlled case series methods within the three following subsections.

4.1 Model 0 - Hierarchical Model With No Latent Factors

Instead of estimating each coefficient independently, we borrow strength over both drugs and outcomes, which adds substantial regularization. This is particularly relevant when

considering a set of related outcomes and drugs, e.g., heart-disease related outcomes and the set of drugs one might prescribe for heart-related conditions. We take a hierarchical Bayesian approach. By analogy with ridge regression, we use normal priors for the regression parameters (sparsifying priors such as the double exponential could be used instead). We shrink the coefficients for drug j for all outcomes o to μ_j by placing an independent normal prior on each β_j^o as $\beta_j^o \sim \mathcal{N}(\mu_j, \sigma_j^2), \forall (j,o)$, where $\mu_j \sim \mathcal{N}(0, \gamma^2), \forall j$. This prior helps with numerical instability, overfitting, and makes the model more interpretable. We assume uniform priors for hyperparameters σ_j and γ as $\sigma_j \sim \mathbf{U}(0,a), \forall j$ and $\gamma \sim \mathbf{U}(0,a)$, where hyperparameter a is a user-defined constant, which can also be determined through cross-validation. A natural extension of this model (not explored here) would be to have drugs belong to certain classes of drugs, so that priors can be defined based on each class of drugs; similarly with outcomes. The posterior density is as follows:

$$\begin{aligned}\Pr(\mathbf{B}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \gamma | \boldsymbol{y}, a) \;&\propto \Pr(\boldsymbol{y}|\mathbf{B}) \times \Pr(\mathbf{B}|\boldsymbol{\mu}, \boldsymbol{\sigma}) \times \Pr(\boldsymbol{\mu}|\gamma) \times \Pr(\gamma|a) \times \Pr(\boldsymbol{\sigma}|a) \qquad (4)\\ &\propto \prod_o \prod_i \prod_{d \in D_i} \left(\frac{\exp\left(\boldsymbol{x}_{id}^\top \boldsymbol{\beta}^o\right)}{\sum_{d'} \exp\left(\boldsymbol{x}_{id'}^\top \boldsymbol{\beta}^o\right)} \right)^{y_{id}^{(o)}} \\ &\times \prod_j \prod_o \mathcal{N}(\beta_j^o | \mu_j, \sigma_j^2) \;\times \prod_j \mathcal{N}(\mu_j | 0, \gamma^2) \times \prod_j \Pr(\sigma_j | a) \times \Pr(\gamma | a).\end{aligned}$$

The negative log-posterior (which can be used for finding the MAP solution if desired) is:

$$\mathcal{L}_1 = -\log\left(\Pr(\mathbf{B}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \gamma | \boldsymbol{y}, a)\right).$$

The graphical representation of this model is shown in Figure 1.

4.2 Model 1 - One Level of Latent Factors

Two considerations motivate this model. First, modeling the full posterior distribution of Model 0 can be computationally expensive, particularly for large N, J, and O, where J and O determine the number of variables to be estimated within the $\mathbf{B}$ matrix. Second, Model 0 overlooks the fact that drugs and outcomes might come from a smaller number of latent classes; for instance, there are commonly several drugs that are extremely similar to each other for treating a set of highly related illnesses. We consider F latent factors for drugs and outcomes. We model the $J \times O$ matrix $\mathbf{B}$ as $\mathbf{B} = \mathbf{L}^{(D)} \times \mathbf{L}^{(O)}$, where

$$\mathbf{L}^{(D)} = \begin{pmatrix} L_{1,1}^{(D)} & \dots & L_{1,F}^{(D)} \\ \vdots & \vdots & \vdots \\ L_{J,1}^{(D)} & \dots & L_{J,F}^{(D)} \end{pmatrix}, \mathbf{L}^{(O)} = \begin{pmatrix} L_{1,1}^{(O)} & \dots & L_{1,O}^{(O)} \\ \vdots & \vdots & \vdots \\ L_{F,1}^{(O)} & \dots & L_{F,O}^{(O)} \end{pmatrix}.$$

This way, we do not assume we know in advance which drugs have similar effects on which outcomes, instead we estimate this from data. The number of latent factors F can be determined by cross-validation. The total number of latent factors is $J \times F + F \times O$, which can be substantially less than $J \times O$. The coefficient β_j^o associated with outcome o and drug j can be calculated as $\beta_j^o = \sum_{f=1}^{F} L_{j,f}^{(D)} \times L_{f,o}^{(O)}$.

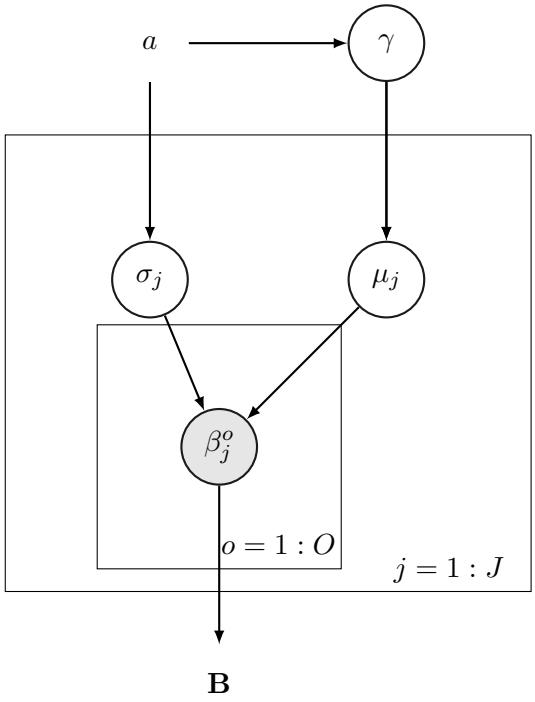

Figure 1: Graphical representation of Model 0

For drug latent factors, we place independent normal priors on the entries of $L^{(D)}$ as

$$L_{jf}^{(D)} \sim \mathcal{N}(\mu_f^{(D)}, \sigma_f^{(D)2}), \forall (j,f), \text{ where } \mu_f^{(D)} \sim \mathcal{N}(0, {\gamma^{(D)}}^2), \forall f.$$

Similarly, we define normal priors on the entries of $L^{(O)}$ as

$$L_{fo}^{(O)} \sim \mathcal{N}(\mu_f^{(O)}, \sigma_f^{(O)2}), \forall (f,o), \text{ where } \mu_f^{(O)} \sim \mathcal{N}(0, {\gamma^{(O)}}^2), \forall f.$$

We assume uniform priors for hyperparameters σ_j and γ as

$$\sigma_f^{(D)} \sim \mathbf{U}(0,a), \forall f, \sigma_f^{(O)} \sim \mathbf{U}(0,a), \forall f, \gamma^{(D)} \sim \mathbf{U}(0,b), \gamma^{(O)} \sim \mathbf{U}(0,b),$$

where (a,b) are known parameters. The posterior over the parameters is now defined as

$$
\begin{aligned}
&\Pr(\mathbf{L}^{(D)}, \mathbf{L}^{(O)}, \boldsymbol{\mu}^{(D)}, \boldsymbol{\mu}^{(O)}, \boldsymbol{\sigma}^{(D)}, \boldsymbol{\sigma}^{(O)}, \gamma^{(D)}, \gamma^{(O)} | \boldsymbol{y}) \propto \Pr(\boldsymbol{y} | \mathbf{L}^{(D)}, \mathbf{L}^{(O)}) \\
&\times \Pr(\mathbf{L}^{(D)} | \boldsymbol{\mu}^{(D)}, \boldsymbol{\sigma}^{(D)}) \times \Pr(\mathbf{L}^{(O)} | \boldsymbol{\mu}^{(O)}, \boldsymbol{\sigma}^{(O)}) \times \Pr(\boldsymbol{\mu}^{(D)} | \gamma^{(D)}) \times \Pr(\boldsymbol{\mu}^{(O)} | \gamma^{(O)}) \\
&\times \Pr(\boldsymbol{\sigma}^{(D)} | a) \times \Pr(\boldsymbol{\sigma}^{(O)} | a) \times \Pr(\gamma^{(D)} | b) \times \Pr(\gamma^{(O)} | b) \\
&\propto \prod_{o} \prod_{i} \prod_{d \in D_i^o} \left(\frac{\exp\left(\boldsymbol{x}_{id}^{\top} \boldsymbol{\beta}^o\right)}{\sum_{d'} \exp\left(\boldsymbol{x}_{id'}^{\top} \boldsymbol{\beta}^o\right)} \right)^{y_{id}^{(o)}} \\
&\times \prod_{j=1}^{J} \prod_{f=1}^{F} \mathcal{N}(L_{jf}^{(D)} | \mu_f^{(D)}, \sigma_f^{(D)2}) \times \prod_{f=1}^{F} \prod_{o=1}^{O} \mathcal{N}(L_{fo}^{(O)} | \mu_f^{(O)}, \sigma_f^{(O)2}) \\
&\times \prod_{f=1}^{F} \mathcal{N}(\mu_f^{(D)} | 0, {\gamma^{(D)}}^2) \times \prod_{f=1}^{F} \mathcal{N}(\mu_f^{(O)} | 0, {\gamma^{(O)}}^2) \\
&\times \Pr(\sigma_f^{(D)} | b) \times \Pr(\sigma_f^{(O)}) | b) \times \Pr(\gamma^{(D)} | a) \times \Pr(\gamma^{(D)} | a).
\end{aligned} \tag{5}
$$

The graphical representation of this hierarchical Bayesian model is given in Figure 2.

4.3 Model 2 - Two Levels of Latent Factors

Here we represent $\mathbf{B}$ as

$$\mathbf{B} = \mathbf{L}^{(D)} \times \mathbf{L}^{(F)} \times \mathbf{L}^{(O)},$$

where

$$\mathbf{B} = \begin{pmatrix} L_{1,1}^{(D)} & \dots & L_{1,F_1}^{(D)} \\ \vdots & \vdots & \vdots \\ L_{J,1}^{(D)} & \dots & L_{J,F_1}^{(D)} \end{pmatrix} \times \begin{pmatrix} L_{1,1}^{(F)} & \dots & L_{1,F_2}^{(F)} \\ \vdots & \vdots & \vdots \\ L_{F_1,1}^{(F)} & \dots & L_{F_1,F_2}^{(F)} \end{pmatrix} \times \begin{pmatrix} L_{1,1}^{(O)} & \dots & L_{1,O}^{(O)} \\ \vdots & \vdots & \vdots \\ L_{F_2,1}^{(O)} & \dots & L_{F_2,O}^{(O)} \end{pmatrix}.$$

The number of latent factors is thus $J \times F_1 + F_1 \times F_2 + F_2 \times O$, which can be less than the number of variables of Model 1 in many cases. Its major benefit is interpretability, since now the number of latent drug factors and the number of latent outcome factors can be estimated differently. $\mathbf{L}^{(D)}$ represents the relationship between drugs and latent drug-related factors, $\mathbf{L}^{(F)}$ represents the relationship between latent drug-related factors and latent health-outcome-related factors, and $\mathbf{L}^{(O)}$ represents the relationship between latent health-outcome-related factors and health-outcome-related factors. $\mathbf{L}^{(F)}$ is really the core set of variables since they relate the latent treatments to the latent health outcomes.

The priors are $L_{jf_1}^{(D)} \sim \mathcal{N}(\mu_{f_1}^{(D)}, \sigma_{f_1}^{(D)2})$, $L_{f_2o}^{(O)} \sim \mathcal{N}(\mu_{f_2}^{(O)}, \sigma_{f_2}^{(O)2})$, $L_{f_1f_2}^{(F)} \sim \mathcal{N}(\mu^{(F)}, {\sigma^{(F)}}^2)$, $\mu_{f_1}^{(D)} \sim \mathcal{N}(0, {\gamma^{(D)}}^2)$, $\mu^{(F)} \sim \mathcal{N}(0, {\gamma^{(F)}}^2)$, $\mu_{f_2}^{(O)} \sim \mathcal{N}(0, {\gamma^{(O)}}^2)$, $\sigma_{f_1}^{(D)} \sim \mathbf{U}(0, a)$, $\sigma^{(F)} \sim \mathbf{U}(0, a)$, $\sigma_{f_2}^{(O)} \sim \mathbf{U}(0, a)$, $\gamma^{(D)} \sim \mathbf{U}(0, b)$, $\gamma^{(F)} \sim \mathbf{U}(0, b)$, and $\gamma^{(O)} \sim \mathbf{U}(0, b)$ for all (f_1, f_2, j, o).

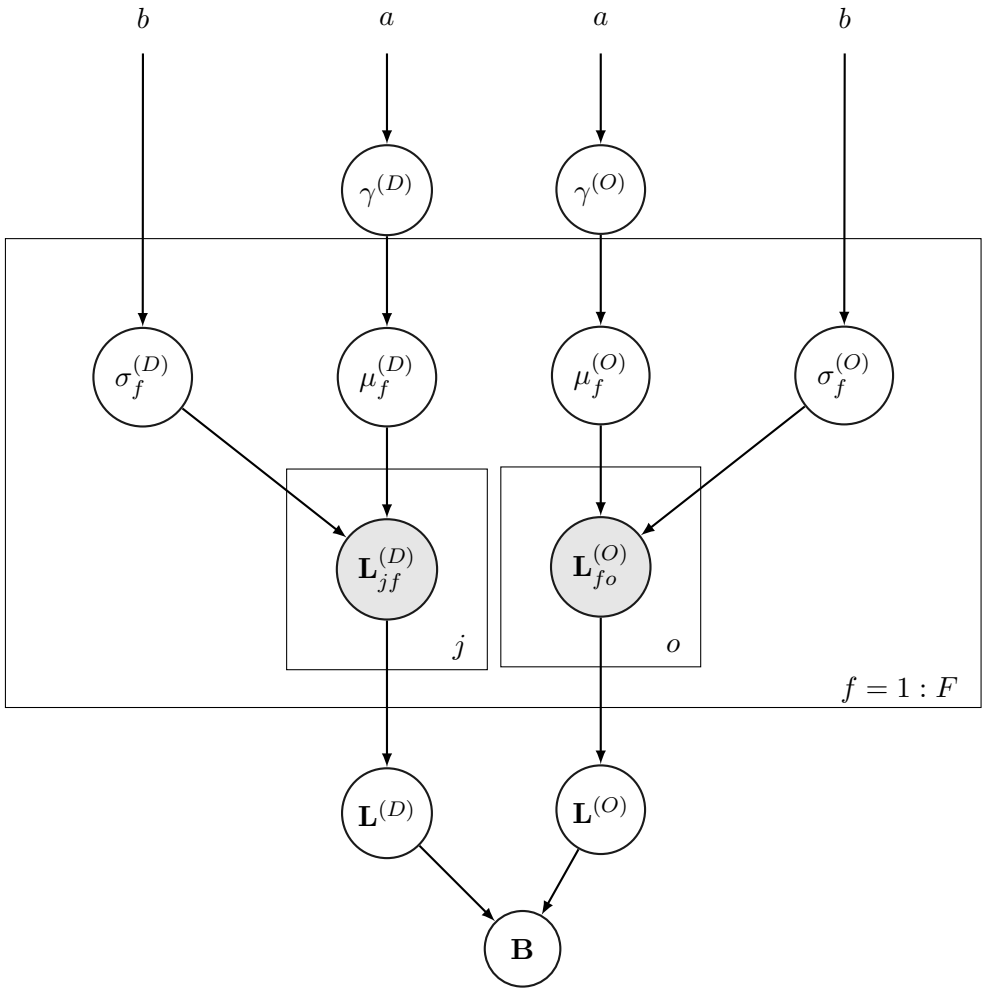

Figure 2: The Graphical Framework for Hierarchical Bayesian Model with one level of latent factors

The posterior density is

$$\begin{aligned}
&\Pr(\mathbf{L}^{(D)}, \mathbf{L}^{(F)}, \mathbf{L}^{(O)}, \boldsymbol{\mu}^{(D)}, \boldsymbol{\mu}^{(F)}, \boldsymbol{\mu}^{(O)}, \boldsymbol{\sigma}^{(D)}, \boldsymbol{\sigma}^{(F)}, \boldsymbol{\sigma}^{(O)}, \gamma^{(D)}, \gamma^{(F)}, \gamma^{(O)} | \boldsymbol{y}) \\
&\propto \Pr(\boldsymbol{y}|\mathbf{L}^{(D)}, \mathbf{L}^{(F)}, \mathbf{L}^{(O)}) \times \Pr(\mathbf{L}^{(D)}|\boldsymbol{\mu}^{(D)}, \boldsymbol{\sigma}^{(D)}) \times \Pr(\mathbf{L}^{(F)}|\mu^{(F)}, \boldsymbol{\sigma}^{(D)}) \times \Pr(\mathbf{L}^{(O)}|\boldsymbol{\mu}^{(O)}, \boldsymbol{\sigma}^{(O)}) \\
&\times \Pr(\boldsymbol{\mu}^{(D)}|\gamma^{(D)}) \times \Pr(\boldsymbol{\mu}^{(F)}|\gamma^{(F)}) \times \Pr(\boldsymbol{\mu}^{(O)}|\gamma^{(O)}) \\
&\times \Pr(\boldsymbol{\sigma}^{(D)}|a) \times \Pr(\boldsymbol{\sigma}^{(F)}|a) \times \Pr(\boldsymbol{\sigma}^{(O)}|a) \times \Pr(\gamma^{(D)}|b) \times \Pr(\gamma^{(F)}|b) \times \Pr(\gamma^{(O)}|b).
\end{aligned} \tag{6}$$

Table 1 compares the number of parameters in each of the three models. Models 1 and 2 have much fewer parameters when F, F_1, and F_2 are lower than J and O. We use Markov Chain Monte Carlo (MCMC) to approximate the entries of $\mathbf{B}$, specifically random walk Metropolis (RWM) Hasting. The algorithm employs a Gaussian proposal distribution

Model Name	# of Parameters	# of Hyperparameters	Total
Model 0	$J * O$	$2 * J + 1$	$J * O + 2 * J + 1$
Model 1	$J * F + F * O$	$4 * F + 2$	$J * F + F * O + 4 * F + 2$
Model 2	$J * F1 + F1 * F2 + F2 * O$	$2 * F_1 + 2 * F_2 + 5$	$J * F_1 + F_1 * F2 + F2 * O + 2 * F_1 + 2 * F_2 + 5$

Table 1: The number of parameters and hyperparameters in each model.

$J_t(x, x')$ which proposes a new parameter set x' given the current parameter set x. We denote $\mathbf{\Theta}$ as the set of all parameters in the model excluding $\mathbf{B}$.

Step 1. Generate an initial state $\{\mathbf{B}^0, \mathbf{\Theta}^0\}$ with positive probability $\Pr(\mathbf{B}^0, \mathbf{\Theta}^0|\boldsymbol{y})$ and set $t = 1$.

Repeat the following until stationary distribution and the desired number of samples are reached considering optional burn-in and/or thinning.

Step 2. Sample $\{\mathbf{B}^*, \mathbf{\Theta}^*\}$ from the symmetric proposal distribution $J_t(\{\mathbf{B}^{t-1}, \mathbf{\Theta}^{t-1}\}, \{\mathbf{B}^*, \mathbf{\Theta}^*\})$.

Step 3. Calculate the acceptance probability

$$\alpha = \min\left(1, \frac{\Pr(\mathbf{B}^*, \mathbf{\Theta}^*|\boldsymbol{y})}{\Pr(\mathbf{B}^{t-1}, \mathbf{\Theta}^{t-1}|\boldsymbol{y})}\right).$$

Step 4. Draw a random number u from Unif$(0, 1)$. If $u \leq \alpha$, accept the proposal state $\{\mathbf{B}^*, \mathbf{\Theta}^*\}$ and set $\mathbf{B}^t = \mathbf{B}^*, \mathbf{\Theta}^t = \mathbf{\Theta}^*$, else set $\mathbf{B}^t = \mathbf{B}^{t-1}, \mathbf{\Theta}^t = \mathbf{\Theta}^{t-1}$. Set $t : t + 1$.

Our implementation uses a component-wise sampling approach. For truly large-scale applications, blocked sampling approaches may be necessary.

5. Simulation Study

As a sanity check, we will show that for data generated from our model, the true data-generating parameters $\mathbf{B}$ can be recovered. We simulated sample trajectories of drug exposure and health outcomes for 600 patients over 60 days. We set the number of drugs to $J = 4$, and the number of health conditions to $O = 4$. Each patient randomly took between 1 and J drugs over the past 60 days. The average exposure period was assumed to be 20% of the study interval for each patient (that is, on average, each patient was exposed to one or more drugs for at least 12 days). The exposure intervals are randomly selected, so these intervals could be multiple non-consecutive days, multiple consecutive days, or a combination of both. Drugs can have positive or negative contributions to the likelihood and intensity rate of each outcome. For each model (Model 0, Model 1, Model 2), we generated the elements of $\mathbf{B}$ according to the model's hierarchy. Figure 3 and Figures 12-13 (which are given in the Appendix) show the posterior density for each parameter of $\mathbf{B}$, for Models 0, 1 and 2, as estimated by MCMC sampling. These figures show that the posterior samples were concentrated around the true values and the posterior mean of each variable was generally close to its true value. We summarize Figures 3, 12, and 13 in Figure 4, which provides a scatter plot of the posterior means and true values for each of the three models. It can be observed that each of the posterior means are very close to their true values.

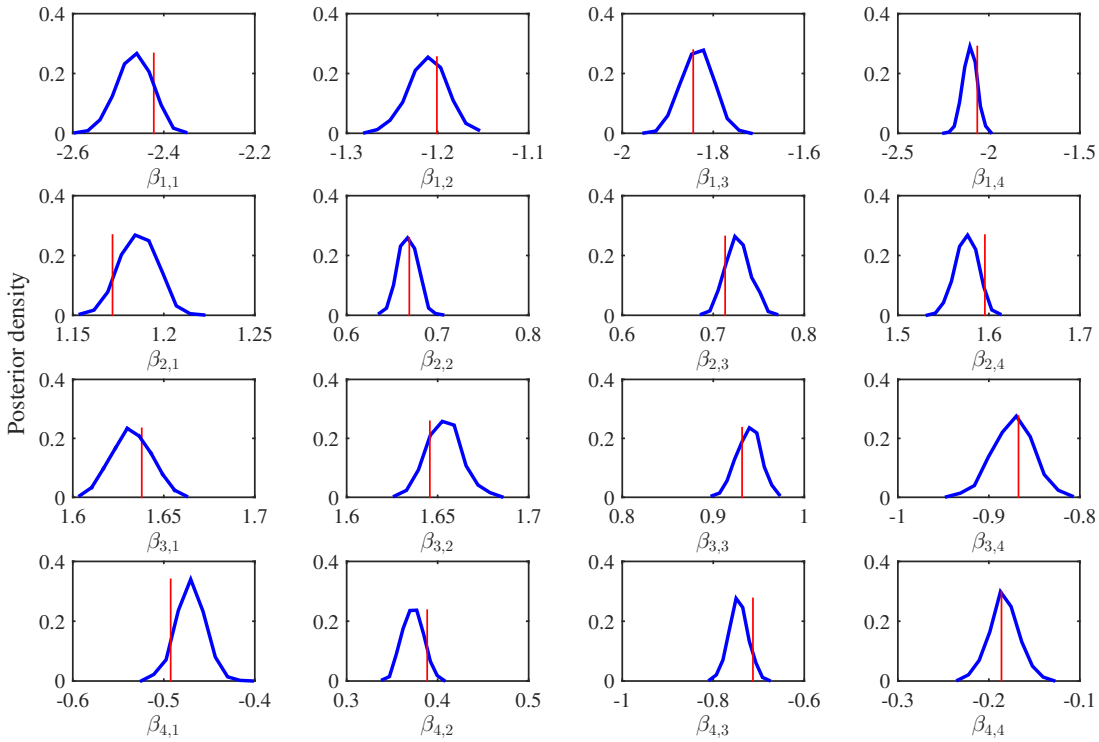

Figure 3: Normalized histograms of posterior samples for each element of **B** in Model 0. The vertical line indicates the true value.

6. Application to Blood Glucose Analysis for Diabetes

We consider an application to Insulin-Dependent Diabetes Mellitus, where our goal is to predict blood glucose level outcomes under different circumstances of a patient's daily life, including their recent eating history, exercise, and insulin injections. Our data are longitudinal measurements taken multiple times per day from 70 patients (this is the AIM-94 data set provided by Michael Kahn, MD, PhD, Washington University, St. Louis, MO, Bache and Lichman, 2013). We aim mainly to illustrate (i) how the models we introduce can be used with complex longitudinal data to predict outcomes, (ii) the prediction power, and (iii) interpretability of the proposed models. It is well known that current therapies for regulating glucose level in diabetics are challenging and often frustrating, as they require patients to continuously regulate diet, exercise, and various medications – any deviations can be dangerous (Benchimol et al., 2013). Blood glucose measurements, symptoms and insulin treatments were recorded with timestamps for each patient, over the course of several weeks to months. The two main classes of health outcomes considered here are hyperglycemia (high blood glucose) and hypoglycemia (low blood glucose). All other health outcomes we define later are related to these two classes. Figure 5 provides a schematic of the type of data we are considering for one patient over a course of day.

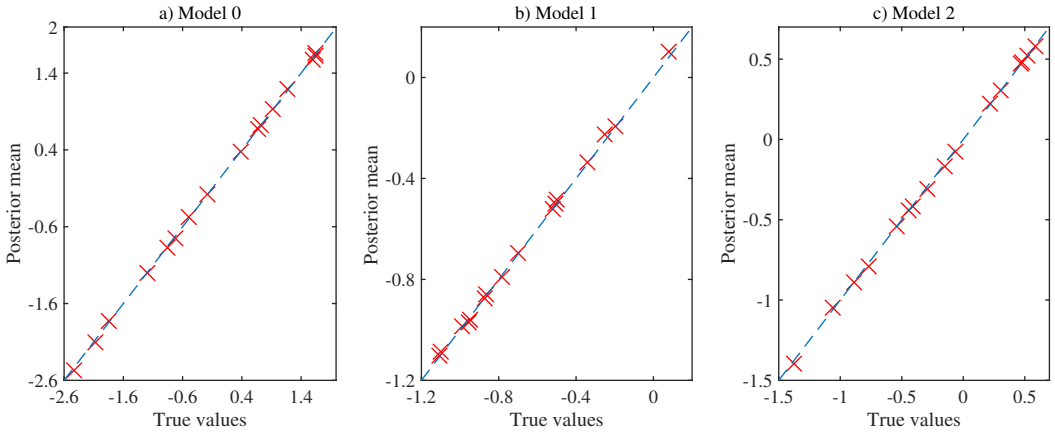

Figure 4: Scatter plot of posterior means vs. true parameter values of elements of $\mathbf{B}$ for Models 0, 1 and 2.

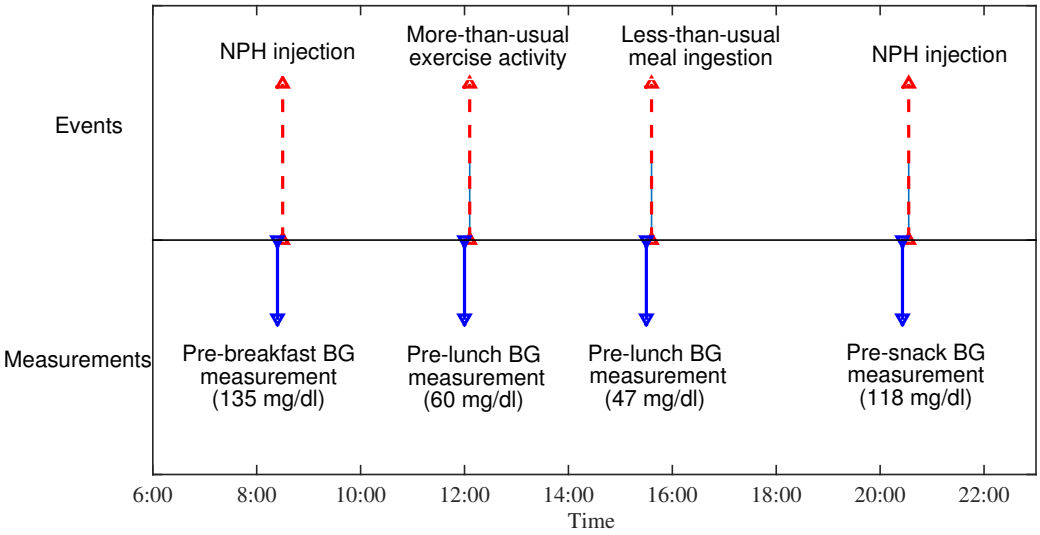

Figure 5: Sample longitudinal traces for a patient with multiple drug/treatments exposures and blood glucose measurements over a day. Downwards arrows indicate glucose measurements. Upwards arrows indicate treatments.

We will describe the setup in more detail:
Drugs/Treatments: Diet, exercise, and injected insulin were treated as three different classes of treatments. It is obvious that interactions among these treatments are important. Insulin doses are given one or more times a day, typically before meals and sometimes also at bedtime. Three types of insulin were considered: (1) regular, (2) Neutral Protamine Hagedorn (NPH), and (3) Ultralente. Each insulin type has its own characteristic time of

onset (O), time of peak action (P), and effective duration (D). The exposure time intervals for peak, and duration used in this paper, which were provided with the data set, are shown in Table 2 in the "Exposure Time" column in the bottom several rows of the table labeled "Insulin."

Treatment			Exposure Time	**Health Outcome**						
				Low Glucose Level				High Glucose Level		
				1. Too low	2. Low	3. $D1$	4. Hypo Symptom	5. Too High	6. High	7. $D10$
Exercise		1. Normal	0-4 h	–	–	–	–	–	–	–
		2. Too High	0-4 h	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$
		3. Low	0-4 h	–	–	–	–	–	–	–
Diet		4. Normal	0-4 h	–	–	–	–	–	–	–
	Level	5. Too High	0-4 h	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$
		6. Low	0-4 h	–	–	–	–	–	–	–
	Time	7. After Meal	0-4 h	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$
		8. Before Meal	0-4 h	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$
Insulin	Regular	9. Peak	1-3 h	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$
		10. Duration	0-5 h	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$
	NPH	11. Peak	4-6 h	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$
		12. Duration	0-12 h	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$
	Ultralente	13. Peak	14-24 h	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$
		14. Duration	0-27 h	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$

Table 2: The list of drugs/treatments and health outcomes and their known correlations. $\mathcal{P}$ means strong positive correlation, and $\mathcal{N}$ means strong negative correlation, $D1$ is lower decile and $D10$ means highest decile.

Based on the actual time of injection, we determined the intervals at which the patient is at peak and/or within the duration of an insulin injection. Based on this, six types of treatments were considered, (1) regular insulin on peak, (2) regular insulin on duration, (3) NPH insulin on peak, (4) NPH insulin on duration, (5) Ultralente insulin on peak, and (6) Ultralente insulin on duration. At each interval of time, the patient can be either insulin free or subject to one of the above six exposures.

The second class of treatment is exercise, which may have complex effects on the glucose level. For example, glucose levels can fall during exercise but also quite a few hours afterwards. Three types of exercise are reported, (1) normal exercise, (2) lower than normal exercise, and (3) higher than normal exercise. Each type of exercise was considered separately as a single treatment.

The third class of treatment is for diet, which also can have complex effects on the glucose level. For example, a larger meal may lead to a longer and possibly higher elevation of blood glucose. Missing a meal may put the patient at risk for low glucose levels in the hours that follow. Three types of diet are reported: (1) normal diet, (2) higher than normal diet, and (3) lower than normal diet. Each of these types of diet were taken as a single treatment. Since measurements were collected before a meal, after a meal, and at other times, we considered two other features in the model, (1) before meal measurement of blood glucose and (2) after meal measurement, and we treated them as binary features. These extra features allow us to distinguish whether the measurement was made before or after the meal (there is a big difference between glucose measurements taken before a meal and after a meal).

Based on all of the treatments described, the total number of variables associated with treatments in the model is 14 ($J = 14$). The variables are all listed on Table 2 in the "Treatment" column on the left.

Health Outcomes. The outcomes are divided into categories, based on glucose level. Given that normal pre-meal blood glucose ranges from approximately 80-120 mg, and post-meal blood glucose ranges from 80-140 mg/dl (Bache and Lichman, 2013), we considered seven health outcomes for glucose level: extremely low (below 40 mg/dl), low (between 40-80 mg/dl), high (over 140 mg/dl), extremely high (over 180 mg/dl), lower decile (lower 10% of glucose level for each patient), upper decile (upper 10% of glucose level for each patient), and hypoglycemic (low glucose) symptoms. Thus, the total number of outcomes considered for our analysis is $O = 7$. We can perform an evaluation only on intervals where we have glucose measurements, thus we only use those intervals. Note that more than one outcome can occur in each interval.

True Relationships Between Drugs and Outcomes. We wanted to determine whether our model reproduces known relationships between treatments and glucose levels from the data alone. The information about true relationships within Table 2 mainly come from material accompanying the data set and *www.diabetes.org*. We denoted known positive effects in Table 2 by $\mathcal{P}$, strong negative effects by $\mathcal{N}$, and relationships that were unknown were denoted by dashes "–". For example, we expect NPH injection on peak to decrease the likelihood of having "Too High" glucose level, so the correlation between NPH on peak and "Too High" glucose level is known to be negative ($\mathcal{N}$).

Mixing. We performed cross-validation, dividing our data into five folds, training our models on four folds and testing on the fifth. We removed the first 5000 iterations (as burn-in) of Metropolis-Hastings sampling, and obtained 6000 additional samples to estimate the posterior. Figure 6 shows samples from the posterior of one of the variables, $\beta^{\text{Too high GL}}_{\text{NPH on peak}}$, for five separate model instantiations (Model 0, Model 1 with F=2, Model 1 with F=3, Model 2 with $F_1 = 2, F_2 = 2$, and Model 2 with $F_1 = 3, F_2 = 3$. Recall that in Models 1 and 2, we sample elements of the matrices of latent factors and then calculate $\mathbf{B}$. From this figure, we observe reasonable mixing for all models, and we observe that models with latent factors (Models 1 and 2) have better mixing and convergence, possibly due to the smaller number of variables.

Computation. The number of parameters differs substantially between models, which affects CPU time of the MCMC sampler. In Figure 7, the number of parameters for each model and the associated CPU time for running MCMC are shown. This figure shows a clear correlation between the number of variables and CPU time, that is CPU time increases with the number of parameters. In particular, Model 0 takes a long time to run, because it has substantially more variables than the other models. Interestingly, using latent factors has a purpose beyond interpretability and regularization, in that it helps with tractability.

Interpretation of coefficients in $\mathbf{B}$ ***and comparison with ground truth.*** We compare the estimated coefficients in $\mathbf{B}$ to the ground truth signs of coefficients given in Table 2. It is not necessarily the case that the signs of the estimated coefficients need to agree with the ground truth signs in order for the model to perform well, but it is a reasonable aspect of the model to consider. To perform this comparison, we ranked the estimated coefficients in $\mathbf{B}$ and used these rankings and the true signs of coefficients to generate an ROC curve. That is, the ROC curve was generated by placing thresholds at each point in

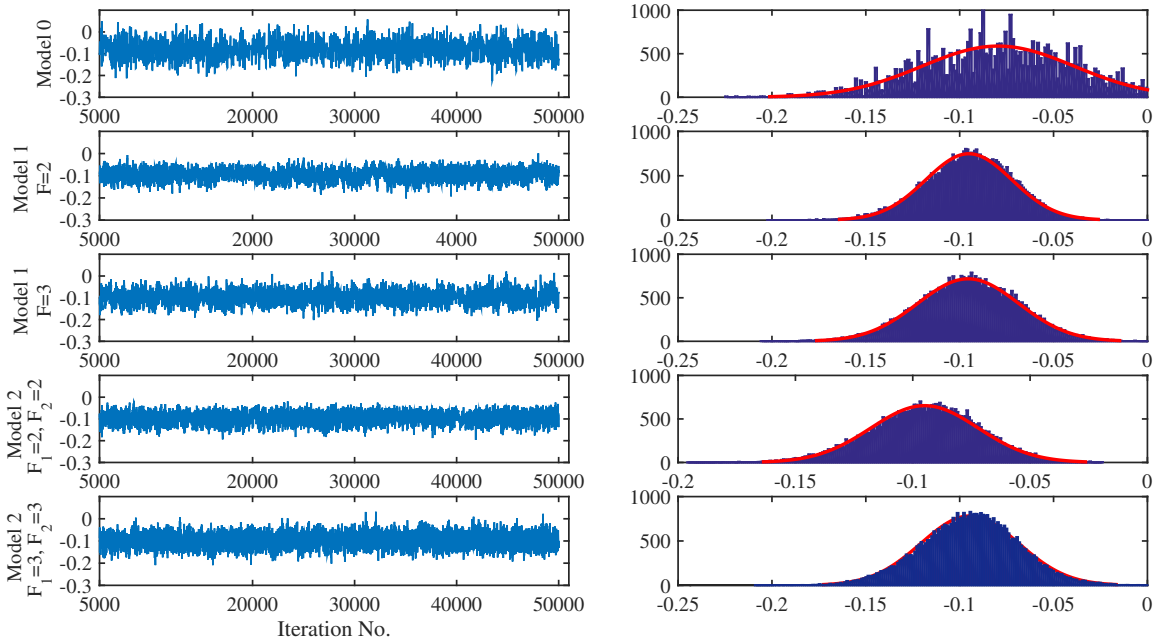

Figure 6: Samples from the posterior of $\beta^{\text{Too high GL}}_{\text{NPH on peak}}$ for five model instantiations (left), and their histograms (right). The first 5000 samples were discarded as burn-in. We observe better mixing and convergence in the models with latent factors (Model 1 and Model 2).

the ranking, and calculating the true positive rate and false positive rate with respect to the true coefficients. We computed these ROC curves for all five model instantiations to obtain Figure 8. These ROC curves indicate that all models performed well, in the sense of estimating reasonable signs for the coefficients. The curves also indicate that models with more latent factors performed better than Model 0. The models with two latent treatment and outcome factors performed slightly better than the models with three latent treatment and outcome factors, though there was no significant difference in performance between Model 1 and Model 2 for the same number of latent treatment and outcome factors.

Drug Surveillance. We evaluated prediction performance of our model as follows: for each patient we calculated the Poisson rate of each condition at each hour considering all drug exposures. For each condition, we then checked whether or not the patient had the condition at that time. The Poisson rate acts as the score of each patient with regards to each condition. In Figure 9, we present the actual glucose level for a patient at 20 measurement points (upper figure) as well as the estimated intensity rate of the Too-High glucose level for the same patient (lower figure). Each point in this figure represents the estimated $\boldsymbol{x}_{id}^{\top}\hat{\boldsymbol{\beta}}^{o}$, where $\hat{\boldsymbol{\beta}}^{o}$ are the estimated regression coefficients associated with too-high glucose level and $\boldsymbol{x}_{id}^{\top}$ is the known vector of drug exposures at interval d. The figure shows that the estimated intensity rate is reasonably close to the actual level of glucose, particularly when the glucose level is actually too high. In Figure 10, we repeated the

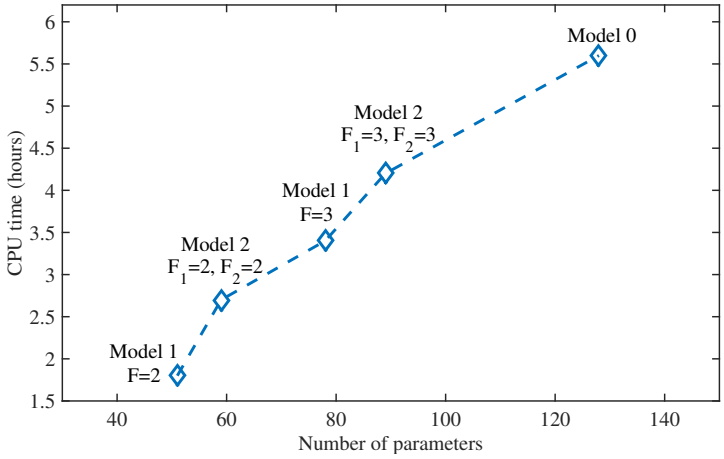

Figure 7: Comparison between the CPU time and number of variables in the five model instantiations. Model 0 has larger number of variables and significantly higher CPU time.

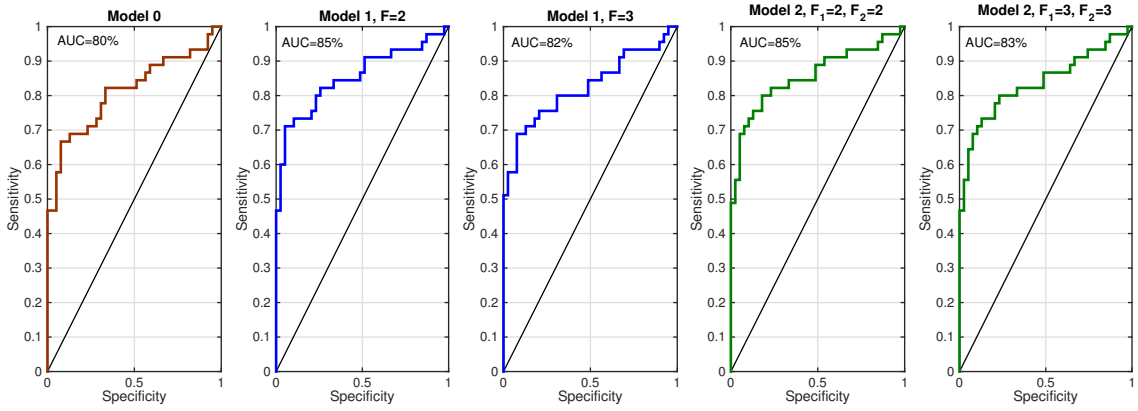

Figure 8: Receiver Operating Characteristic (curve) for evaluating the signs of coefficients against true signs for five model instantiations. See the text for details of how these curves were generated.

analysis for Too-Low glucose level. It can be observed from this figure that the times where this coefficient is particularly large are the same times where the glucose level drops substantially. This kind of dramatic agreement was not observed for all patients nor all conditions, so below we describe a more general evaluation procedure.

In Figure 11, we show the box plots for the estimated Poisson rates of the two conditions of "Too Low" glucose level and "Too High" glucose level on all seventy patients in the test sets. For comparison, we also plotted the box plots of the estimated Poisson rates for normal conditions, where patients did not have too-high or too-low glucose levels. For

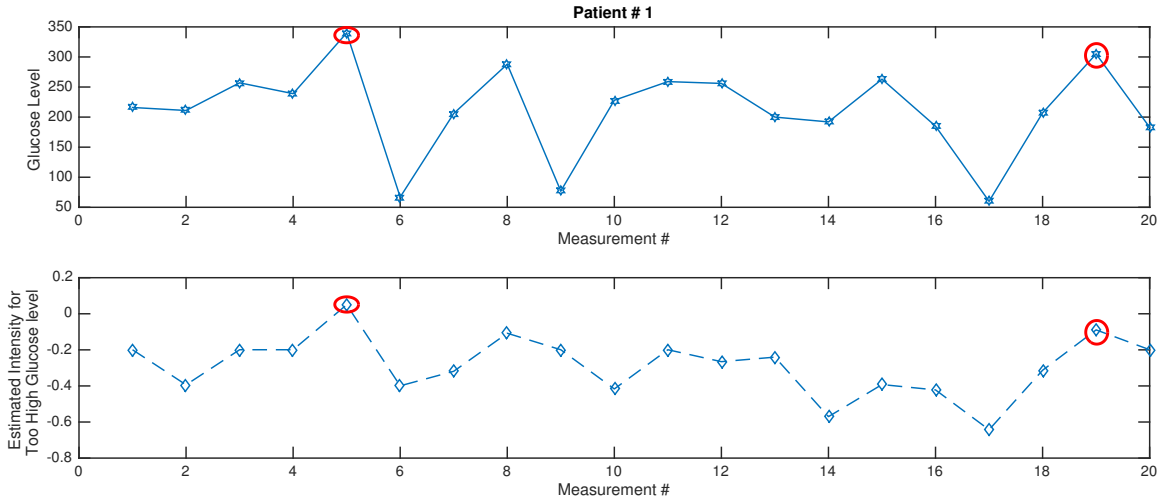

Figure 9: Monitoring Too-High glucose levels over 20 hours. The upper figure shows the true glucose levels obtained by measurement, and the lower figure shows the estimated $\boldsymbol{x}_{id}^{\top}\hat{\boldsymbol{\beta}}^{o}$ for the Too-High glucose outcome over 20 consecutive measurements.

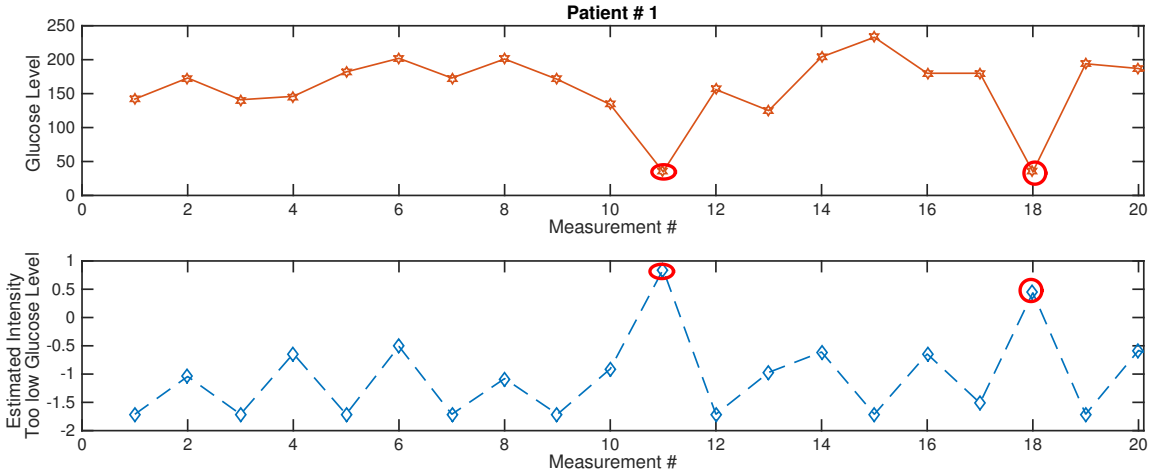

Figure 10: Monitoring too-low glucose levels over 20 hours. The upper figure shows the true glucose levels obtained by measurement, and the lower figure shows the estimated $\boldsymbol{x}_{id}^{\top}\hat{\boldsymbol{\beta}}^{o}$ for the too-low glucose outcome over 20 hours.

clarity and fairness, we normalized the estimated Poisson rates for each patient. It can be observed from this figure that the estimated Poisson rate of these conditions are elevated when patients actually suffer from Too-Low (or Too-High, respectively) glucose. Thus, our model could be a useful approach for monitoring the likelihood of a condition, given the

timing of the drugs recently taken by the patient. The results were consistent across all five model instantiations.

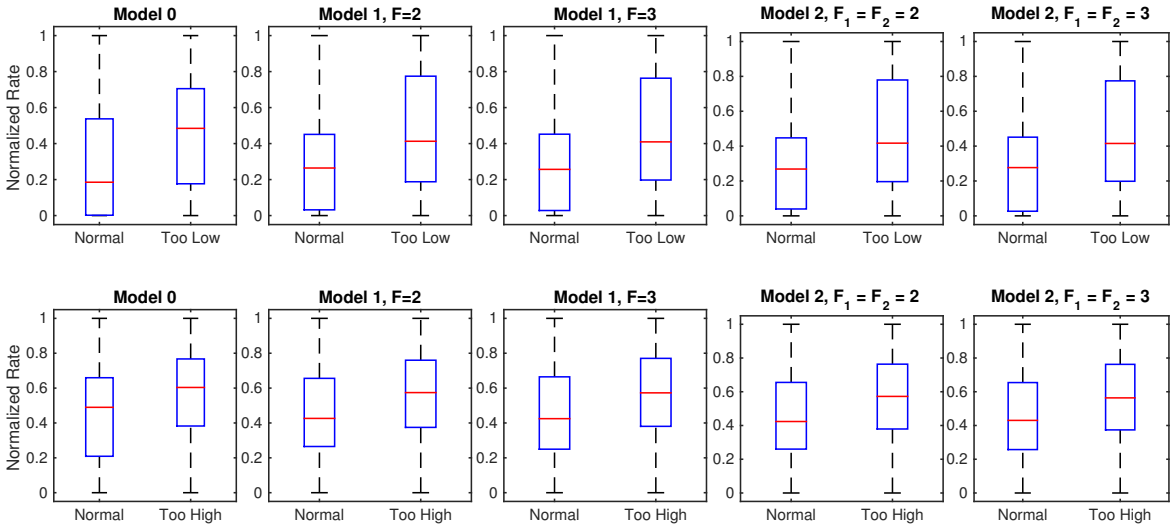

Figure 11: Comparison between the box-plots for Too-High and Too-Low glucose levels and the normal condition for all five model instantiations, where the Poisson rates were normalized for each person.

In Table 3, we report an AUC (area under the ROC curve) value that measures the probability that a method ranks a positive condition timepoint higher than a timepoint with no condition, for the same patient. In particular, we are testing whether the Poisson rate of a patient with a health condition is higher than the estimated rate of the same patient when no condition is present. For each model instantiation, we trained the model on four training folds and then calculated the AUC on the fifth fold, and we repeated this for each condition. We reported the average and standard deviation of AUC over all patients in the test sets. Again D1 is the lower decile (lower 10% of glucose level for each patient), and D10 is the upper decile (upper 10% of glucose level for each patient).

Intuition of* $\mathbf{B}$ *as compared with other methods. We compared the performance of the proposed models with that of the univariate self-controlled case series (SCCS), and multi-variate self-controlled case series (MSCCS) (Simpson et al., 2013) with and without an ℓ_2 regularization term on the coefficients. For each model instantiation and each method, the estimated entries of $\mathbf{B}$ were compared to their known effects (positive or negative) from Table 2. For each method, we provide the area under the curve (AUC) for this comparison in Table 4. We also reported the mean, median, and standard deviation of all estimated coefficients in the group of Table 2's positive group and Table 2's negative group. Better models should have higher AUC, and the estimated coefficients of $\mathbf{B}$ should agree in sign with those in Table 2. From Table 4, we observe that as expected, the MSCCS performed better than the SCCS, the regularized MSCCS worked slightly better than the normal

		Too low	Low	Too high	High	D1	D10	Hypo Symptom
Model 0	Mean	70.30%	62.92%	62.98%	61.26%	58.20%	56.39%	59.17%
	sd	9.56%	2.79%	3.38%	2.70%	3.93%	3.51%	4.95%
Model 1, $F = 2$	Mean	70.86%	62.70%	63.06%	61.58%	57.97%	56.05%	56.71%
	sd	8.28%	2.47%	3.14%	2.11%	3.36%	3.04%	3.19%
Model 1, $F = 3$	Mean	70.66%	62.68%	62.92%	61.51%	57.97%	56.35%	56.33%
	sd	7.89%	2.52%	3.14%	2.12%	3.37%	2.90%	2.87%
Model 2, $F_1 = F_2 = 2$	Mean	70.19%	62.71%	62.97%	61.44%	58.00%	55.91%	56.88%
	sd	9.51%	2.43%	3.18%	2.13%	3.34%	3.06%	3.02%
Model 2, $F_1 = 3, F_2 = 3$	Mean	70.07%	62.73%	62.99%	61.47%	58.05%	56.05%	56.68%
	sd	9.52%	2.44%	3.16%	2.07%	3.31%	3.05%	2.79%

Table 3: Average and standard deviation (sd) of AUC over 5 folds. The entities that were ranked for each AUC calculation are measurements for a patient including time points when the patient had a condition and time points when the patient did not have a condition. The AUC indicates whether our method ranks a randomly chosen time point where a patient had a condition higher than a randomly chosen time point where a patient did not have a condition.

MSCCS, and all FSCCS Bayesian model instantiations (Model 0-2) performed better than all of the traditional models, yielding higher AUC's and better agreement in the mean signs of coefficients; further the standard deviations for the coefficient values were substantially lower. These performance benefits come in addition to the other benefits discussed earlier, including computational tractability and interpretability of the latent factors.

Measure		Method							
		Model 0	Mode1 $F = 2$	Model 1 $F = 3$	Model 2 $F_1 = 2, F_2 = 2$	Model 2 $F_1 = 3, F_2 = 3$	SCCS	MSCCS	Regularized MSCCS
	AUC	0.813	0.852	0.824	0.856	0.836	0.693	0.773	0.774
$\mathbf{B}^-$	Mean	-0.190	-0.281	-0.276	-0.278	-0.282	-0.492	-0.663	-0.349
	Median	-0.077	-0.136	-0.143	-0.130	-0.141	-0.047	-0.122	-0.122
	sd	0.353	0.417	0.410	0.422	0.424	2.362	2.462	0.640
$\mathbf{B}^+$	Mean	0.071	0.099	0.096	0.105	0.096	-0.552	-0.561	-0.023
	Median	0.030	0.103	0.111	0.106	0.106	0.065	0.126	0.125
	sd	0.368	0.369	0.385	0.368	0.378	3.158	3.149	0.771

Table 4: Comparison with existing models.

7. Concluding Remarks

The novel elements of this work are as follows. (1) We estimate the effects of many drugs on many health outcomes simultaneously. Borrowing strength across similar drugs and outcomes allows us to create better estimates across both drugs and outcomes. (2) We use latent factors to encode latent classes of drugs and outcomes, to help with interpretability, and to provide a computational benefit. Another type of computational benefit is provided naturally by using the SCCS's framework, since we do not need to estimate the baseline rates of outcomes for each patient. This approach is scalable to large longitudinal observational databases, is applicable to problems beyond healthcare, and provides a level of interpretability to physicians and patients that was not previously possible. Fully Bayesian inference via MCMC may not be feasible for truly large-scale problems. Recent developments in cyclic coordinate descent algorithms (see, for example, in Suchard et al., 2013b) would apply in our context and represent one possible approach for very scale MAP estimation.

Acknowledgments

We gratefully acknowledge funding from the Natural Science and Engineering Research Council of Canada (NSERC) and the MIT Big Data Initiative.

Appendix A. Figures 12 and 13.

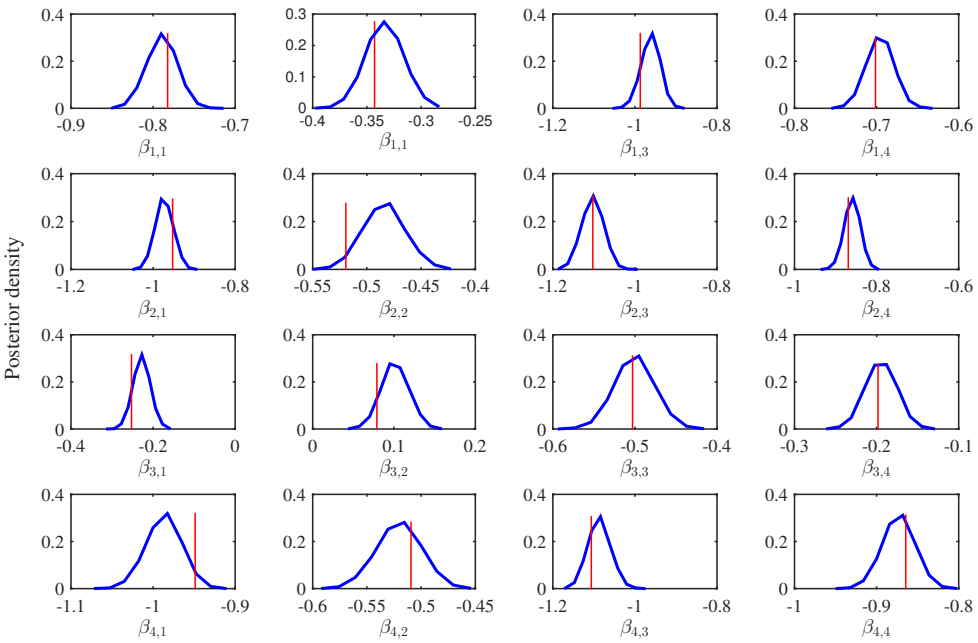

Figure 12: Normalized histograms of posterior samples for each element of **B** in Model 1. The vertical line indicates the true value.

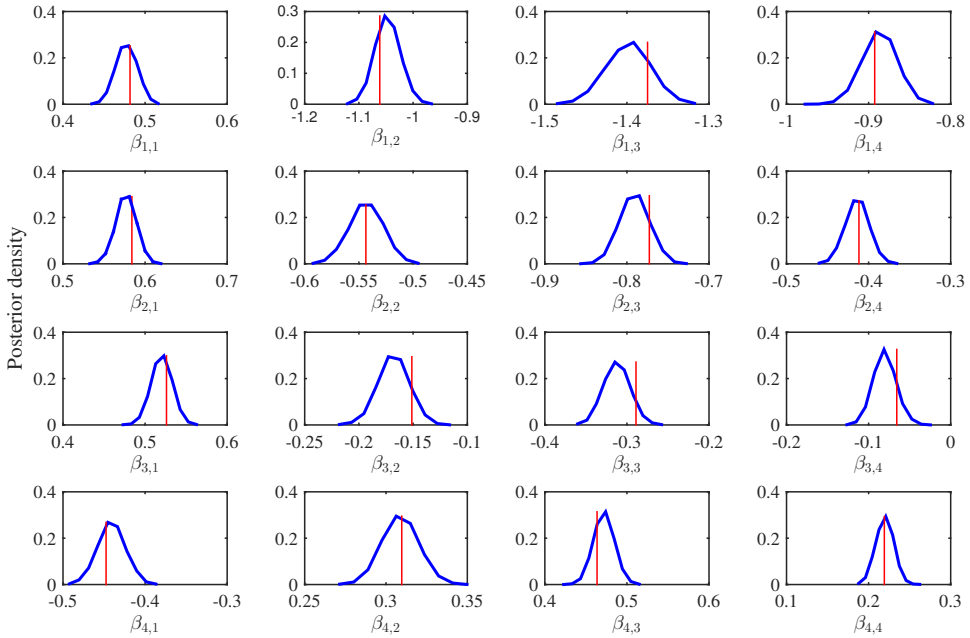

Figure 13: Normalized histograms of posterior samples for each element of **B** in Model 2. The vertical line indicates the true value.

References

K. Bache and M. Lichman. UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, 2013.

E.I. Benchimol, S. Hawken, J.C. Kwong, and K. Wilson. Safety and utilization of influenza immunization in children with inflammatory bowel disease. *Pediatrics*, 131(6):1811–1820, 2013.

C.M. Carvalho, J. Chang, J.E. Lucas, J.R. Nevins, Q. Wang, and M. West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484), 2008.

C.S.L. Chui, K.K.C. Man, C.L. Cheng, E.W. Chan, W.C.Y. Lau, V.C.C. Cheng, D.S.H. Wong, Y.H. Yang Kao, and I.C.K. Wong. An investigation of the potential association between retinal detachment and oral fluoroquinolones: A self-controlled case series study. *Journal of Antimicrobial Chemotherapy*, 69(9):2563–2567, 2014.

M.C. Cobanoglu, C. Liu, F. Hu, Z.N. Oltvai, and I. Bahar. Predicting drug-target interactions using probabilistic matrix factorization. *Journal of Chemical Information and Modeling*, 53(12):3399–3409, 2013.

P. Farrington. Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics*, 51:228–235, 1995.

P. C. Farrington, K. Anaya-Izquierdo, H.J. Whitaker, M.N. Hocine, I. Douglas, and L. Smeeth. Self-controlled case series analysis with event-dependent observation periods. *Journal of the American Statistical Association*, 106(494), 2011.

J. Ghosh and D.B. Dunson. Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2): 306–320, 2009.

S.K. Greene, M. Kulldorff, R. Yin, W.K. Yih, T.A. Lieu, E.S. Weintraub, and G.M. Lee. Near real-time vaccine safety surveillance with partially accrued data. *Pharmacoepidemiology and Drug Safety*, 20(6):583–590, 2011.

D. Madigan, P. Ryan, S.E. Simpson, and I. Zorych. *Bayesian methods in pharmacovigilance*, volume 9. Oxford University Press, 2010.

D. Madigan, S. Simpson, W. Hua, A. Paredes, B. Fireman, and M. Maclure. The self-controlled case series: Recent developments, 2011.

D. Madigan, P.E. Stang, J.A. Berlin, M. Schuemie, J.M. Overhage, M.A. Suchard, B. Dumouchel, A.G. Hartzema, and P.B. Ryan. A systematic statistical approach to evaluating evidence from observational studies. *Annual Review of Statistics and Its Application*, 1: 11–39, 2014.

MJ. Schuemie, G. Trifirò, PM. Coloma, PB. Ryan, and D. Madigan. Detecting adverse drug reactions following long-term exposure in longitudinal observational data: The exposure-adjusted self-controlled case series. *Statistical Methods in Medical Research*, 2014.

SE. Simpson. A positive event dependence model for self-controlled case series with applications in postmarketing surveillance. *Biometrics*, 69(1):128–136, 2013.

S.E. Simpson, D. Madigan, I. Zorych, M.J. Schuemie, P.B. Ryan, and M.A. Suchard. Multiple self-controlled case series for large-scale longitudinal observational databases. *Biometrics*, 69(4):893–902, 2013.

MA. Suchard, I. Zorych, SE. Simpson, MJ. Schuemie, PB. Ryan, and D. Madigan. Empirical performance of the self-controlled case series design: lessons for developing a risk identification and analysis system. *Drug Safety*, 36(1):83–93, 2013a.

Marc A. Suchard, Shawn E. Simpson, Ivan Zorych, Patrick Ryan, and David Madigan. Massive parallelization of serial inference algorithms for a complex generalized linear model. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 23(1):10, 2013b.

H. J. Whitaker, C. P. Farrington, B. Spiessens, and P. Musonda. Tutorial in biostatistics: the self-controlled case series method. *Statistics in Medicine*, 25(10):1768–1797, 2006.

M. Zitnik and B. Zupan. Matrix factorization-based data fusion for drug-induced liver injury prediction. *Systems Biomedicine*, 2(1):16–22, 2014.