Temporal Poisson Square Root Graphical Models

Sinong Geng* [1]   **Zhaobin Kuang*** [1]   **Peggy Peissig** [2]   **David Page** [1]

Abstract

We propose temporal Poisson square root graphical models (TPSQRs), a generalization of Poisson square root graphical models (PSQRs) specifically designed for modeling longitudinal event data. By estimating the temporal relationships for all possible pairs of event types, TPSQRs can offer a holistic perspective about whether the occurrences of any given event type could excite or inhibit any other type. A TPSQR is learned by estimating a collection of interrelated PSQRs that share the same template parameterization. These PSQRs are estimated jointly in a pseudo-likelihood fashion, where Poisson pseudo-likelihood is used to approximate the original more computationally-intensive pseudo-likelihood problem stemming from PSQRs. Theoretically, we demonstrate that under mild assumptions, the Poisson pseudo-likelihood approximation is *sparsistent* for recovering the underlying PSQR. Empirically, we learn TPSQRs from Marshfield Clinic electronic health records (EHRs) with millions of drug prescription and condition diagnosis events, for adverse drug reaction (ADR) detection. Experimental results demonstrate that the learned TPSQRs can recover ADR signals from the EHR effectively and efficiently.

1. Introduction

Longitudinal event data (LED) and the analytics challenges therein are ubiquitous now. In business analytics, purchasing events of different items from millions of customers are collected, and retailers are interested in how a distinct market action or the sales of one particular type of item could boost or hinder the sales of another type (Han et al., 2011). In search analytics, web search keywords from bil-

lions of web users are usually mapped into various topics (e.g. travel, education, weather), and search engine providers are interested in the interplay among these search topics for a better understanding of user preferences (Gunawardana et al., 2011). In health analytics, electronic health records (EHRs) contain clinical encounter events from millions of patients collected over decades, including drug prescriptions, biomarkers, and condition diagnoses, among others. Unraveling the relationships between different drugs and different conditions is vital to answering some of the most pressing medical and scientific questions such as drug-drug interaction detection (Tatonetti et al., 2012), comorbidity identification, adverse drug reaction (ADR) discovery (Simpson et al., 2013; Bao et al., 2017; Kuang et al., 2017b), computational drug repositioning (Kuang et al., 2016a;b), and precision medicine (Liu et al., 2013; 2014a).

All these analytics challenges beg the statistical modeling question: *can we offer a comprehensive perspective about the relationships between the occurrences of all possible pairs of event types in longitudinal event data?* In this paper, we propose a solution via temporal Poisson square root graphical models (TPSQRs), a generalization of Poisson square root graphical models (PSQRs, Inouye et al. 2016) made in order to represent multivariate distributions among count variables evolving temporally in LED.

The reason why conventional undirected graphical models (UGMs) are not readily applicable to LED is the lack of mechanisms to address the *temporality* and *irregularity* in the data. Conventional UGMs (Liu and Page, 2013; Liu et al., 2014b; Yang et al., 2015a; Liu et al., 2016; Kuang et al., 2017a; Geng et al., 2018) focus on estimating the co-occurrence relationships among various variables rather than their temporal relationships, that is, how the occurrence of one type of event may affect the future occurrence of another type. Furthermore, existing temporal variants of UGMs (Kolar et al., 2010; Yang et al., 2015b) usually assume that data are regularly sampled, and observations for all variables are available at each time point. Neither assumption is true, due to the irregularity of LED.

In contrast to these existing UGM models, a TPSQR models temporal relationships, by *data aggregation*; a TPSQR extracts a sequence of time-stamped summary count statistics of distinct event types that preserves the relative temporal

Sinong Geng and Zhaobin Kuang contribute equally. Their names are listed in alphabetical order. [1]The University of Wisconsin, Madison [2]Marshfield Clinic Research Institute. Correspondence to: Sinong Geng <sgeng2@wisc.edu>.

order in the raw data for each subject. A PSQR is then used to model the joint distribution among these summary count statistics for each subject. Different PSQRs for different subjects are assumed to share the same *template parameterization* and hence can be learned jointly by estimating the template in a pseudo-likelihood fashion. To address the challenge in temporal irregularity, we compute the exact time difference between each pair of time-stamped summary statistics, and decide whether a difference falls into a particular predefined time interval, hence transforming the irregular time differences into regular timespans. We then incorporate the effects of various timespans into the template parameterization as well as PSQR constructions from the template.

By addressing temporality and irregularity of LED in this fashion, TPSQR is also different from many point process models (Gunawardana et al., 2011; Weiss et al., 2012; Weiss and Page, 2013; Du et al., 2016), which usually strive to pinpoint the exact occurrence times of events, and offer generative mechanisms to event trajectories. TPSQR, on the other hand, adopts a coarse resolution approach to temporal modeling via the aforementioned data aggregation and time interval construction. As a result, TPSQR focuses on estimating stable relationships among occurrences of different event types, and does not model the precise event occurrence timing. This behavior is especially meaningful in application settings such as ADR discovery, where the importance of identifying the occurrence of an adverse condition caused by the prescription of a drug usually outweighs knowing about the exact time point of the occurrence of the ADR, due to the high variance of the onset time of ADRs (Schuemie et al., 2016).

Since TPSQR is a generalization of PSQR, many desirable properties of PSQR are inherited by TPSQR. For example, TPSQR, like PSQR, is capable of modeling both positive and negative dependencies between covariates. Such flexibility cannot usually be taken for granted when modeling a multivariate distribution over count data due to the potential dispersion of the partition function of a graphical model (Yang et al., 2015a). TPSQR can be learned by solving the pseudo-likelihood problem for PSQR. For efficiency and scalability, we use Poisson pseudo-likelihood to approximately solve the original pseudo-likelihood problem induced by a PSQR, and we show that the Poisson pseudo-likelihood approximation can recover the structure of the underlying PSQR under mild assumptions. Finally, we demonstrate the utility of TPSQRs using Marshfield Clinic EHRs with millions of drug prescription and condition diagnosis events for the task of adverse drug reaction (ADR) detection. Our contributions are three-fold:

- TPSQR is a generalization of PSQR made in order to represent the multivariate distributions among count variables evolving temporally in LED. TPSQR can accommodate both positive and negative dependencies among covariates, and can be learned efficiently via the pseudo-likelihood problem for PSQR.

- In terms of advancing the state-of-the-art of PSQR estimation, we propose Poisson pseudo-likelihood approximation in lieu of the original more computationally-intensive conditional distribution induced by the joint distribution of a PSQR. We show that under mild assumptions, the Poisson pseudo-likelihood approximation procedure is sparsistent (Ravikumar et al., 2007) with respect to the underlying PSQR. Our theoretical results not only justify the use of the more efficient Poisson pseudo-likelihood over the original conditional distribution for better estimation efficiency of PSQR but also establish a formal correspondence between the more intuitive but less stringent local Poisson graphical models (Allen and Liu, 2013) and the more rigorous but less convenient PSQRs.

- We apply TPSQR to Marshfield Clinic EHRs to determine the relationships between the occurrences of various drugs and the occurrences of various conditions, and offer more accurate estimations for adverse drug reaction (ADR) discovery, a challenging task in health analytics due to the (thankfully) rare and weak ADR signals encoded in the data, whose success is crucial to improving healthcare both financially and clinically (Sultana et al., 2013).

2. Background

We show how to deal with the challenges in temporality and irregularity mentioned in Section 1 via the use of data aggregation and an influence function for LED. We then define the template parameterization that is central to the modeling of TPSQRs.

2.1. Longitudinal Event Data

Longitudinal event data are time-stamped events of finitely many types collected across various subjects over time. Figure 1 visualizes the LED for two subjects. As shown in Figure 1, the occurrences of different event types are represented as arrows in different colors. No two events for one subject occur at the exact same time. We are interested in modeling the relationships among the occurrences of different types of events via TPSQR.

2.2. Data Aggregation

To enable PSQRs to cope with the temporality in LED, TPSQRs start from extracting relative-temporal-order-preserved summary count statistics from the raw LED via data aggregation, to cope with the high volume and frequent consecutive replications of events of the same type that are
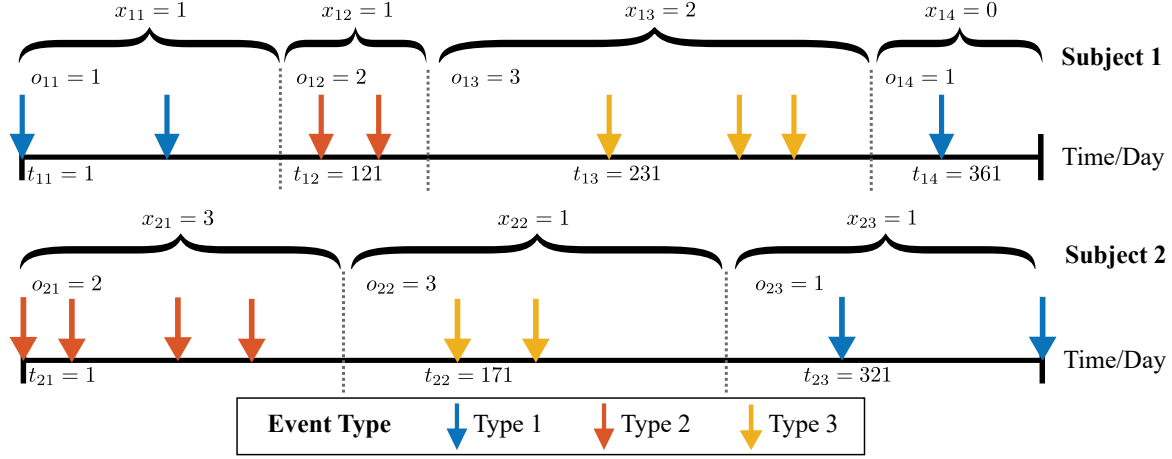
Figure 1: Visualization of longitudinal event data from two subjects. Curly brackets denote the timespans during which events of only one type occur. x_{ij}'s represent the number of subsequent occurrences after the first occurrence. o_{ij}'s are the types of events in various timespans.

commonly observed in LED. Take Subject 1 in Figure 1 as an illustrative example; we divide the raw data of Subject 1 into four timespans by the dashed lines. Each of the four timespans contains only events of the same type. We use three statistics to summarize each timespan: the time stamp of the first occurrence of the event in each timespan: $t_{11} = 1$, $t_{12} = 121$, $t_{13} = 231$, and $t_{14} = 361$; the event type in each timespan: $o_{11} = 1$, $o_{12} = 2$, $o_{13} = 3$, and $o_{14} = 1$; and the counts of subsequent occurrences in each timespan: $x_{11} = 1$, $x_{12} = 1$, $x_{13} = 2$, and $x_{14} = 0$. Note that the reason $x_{14} = 0$ is that there is only one occurrence of event type 1 in total during timespan 4 of subject 1. Therefore, the number of subsequent occurrence after the first and only occurrence is 0.

Let there be N independent subjects and p types of events in a given LED $\mathbb{X}$. We denote by n_i the number of timespans during which only one type of event occurs to subject i, where $i \in \{1, 2, \cdots, N\}$. The jth timespan of the ith subject can be represented by the vector $\mathbf{s}_{ij} := \begin{bmatrix} t_{ij} & o_{ij} & x_{ij} \end{bmatrix}^\top$, where $j \in \{1, 2, \cdots, n_i\}$, and ":=" represents "defined as." $t_{ij} \in [0, +\infty)$ is the time stamp at which the first event occurs during the timespan $\mathbf{s}_{ij}$. Furthermore, $t_{11} < t_{12} < \cdots < t_{1n_i}$. $o_{ij} \in \{1, 2, \cdots, p\}$ represents the event type in $\mathbf{s}_{ij}$. Furthermore, $o_{ij} \neq o_{i(j+1)}$, $\forall i \in \{1, 2, \cdots, N\}$ and $\forall j < n_i$. $x_{ij} \in \mathbb{N}$ is the number of subsequent occurrences of events of the same type in $\mathbf{s}_{ij}$.

2.3. Influence Function

Let $\mathbf{s}_{ij}$ and $\mathbf{s}_{ij'}$ be given, where $j < j' \leq n_i$. To handle the irregularity of the data, we map the time difference $t_{ij'} - t_{ij}$ to a one-hot vector that represents the activation of a time interval using an *influence function* $\phi(\cdot)$, a common mechanism widely used in point process models and signal

processing. In detail, let $L + 1$ user-specified time-threshold values be given, where $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_L$. $\phi(\tau)$ is a $L \times 1$ one hot vector whose l^{th} component is defined as:

$$[\phi(\tau)]_l := \begin{cases} 1, & \tau_{l-1} \leq \tau < \tau_l \\ 0, & \text{otherwise} \end{cases}, \quad (1)$$

where $l \in \{1, 2, \cdots, L\}$. In our case, we let $\tau := t_{ij'} - t_{ij}$ to construct $\phi(\tau)$ according to (1). Widely used influence functions in signal processing include the dyadic wavelet function and the Haar wavelet function (Mallat, 2008); both are piecewise constant and hence share similar representation to (1).

2.4. Template Parameterization

Template parameterization provides the capability of TP-SQRs to represent the effects of all possible (ordered) pairs of event types on all time scales. Specifically, let an ordered pair $(k, k') \in \{1, 2, \cdots, p\}^2$ be given. Let $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_K$ also be given. For the ease of presentation, we assume that $k \neq k'$, which can be easily generalized to $k = k'$. Considering a particular patient, we are interested in knowing the effect of an occurrence of a type k event towards a subsequent occurrence of a type k' event, when the time between the two occurrences falls in the lth time window specified via (1). Enumerating all L time windows, we have:

$$\mathbf{w}_{kk'} := \begin{bmatrix} w_{kk'1} & w_{kk'2} & \cdots & w_{kk'L} \end{bmatrix}^\top. \quad (2)$$

Note that since (k, k') is ordered, $\mathbf{w}_{k'k}$ is different from $\mathbf{w}_{kk'}$. We further define $\mathbf{W}$ as a $(p-1)p \times L$ matrix that stacks up all $\mathbf{w}_{kk'}^\top$'s. In this way, $\mathbf{W}$ includes all possible pairwise temporally bidirectional relationships among the

p variables on different time scales, offering holistic representation power. To represent the intrinsic prevalence effect of the occurrences of events of various types, we further define $\boldsymbol{\omega} := \begin{bmatrix} \omega_1 & \omega_2 & \cdots & \omega_p \end{bmatrix}^\top$. We call $\boldsymbol{\omega}$ and $\mathbf{W}$ the template parameterization, from which we will generate the parameters of various PSQRs as shown in Section 3.

3. Modeling

Let $\mathbf{s}_{ij}$'s be given where $j \in \{1, 2, \cdots, n_i\}$; we demonstrate the use of the influence function and template parameterization to construct a PSQR for subject i.

Let $\mathbf{t}_i := \begin{bmatrix} t_{i1} \, t_{i2} \, \cdots \, t_{in_i} \end{bmatrix}^\top$, $\mathbf{o}_i := \begin{bmatrix} o_{i1} \, o_{i2} \, \cdots \, o_{in_i} \end{bmatrix}^\top$, and $\mathbf{x}_i := \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{in_i} \end{bmatrix}^\top$. Given $\mathbf{t}_i$ and $\mathbf{o}_i$, a TPSQR aims at modeling the joint distribution of counts $\mathbf{x}_i$ using a PSQR. Specifically, under the template parameterization $\boldsymbol{\omega}$ and $\mathbf{W}$, we first define a symmetric parameterization $\boldsymbol{\Theta}^{(i)}$ using $\mathbf{t}_i$ and $\mathbf{o}_i$. The component of $\boldsymbol{\Theta}^{(i)}$ at the jth row and the j'^{th} column is:

$$\theta^{(i)}_{jj'} := [\boldsymbol{\Theta}^{(i)}]_{jj'} := \begin{cases} \omega_{o_{ij}}, & j = j' \\ \mathbf{w}^\top_{o_{ij}o_{ij'}} \phi(|t_{ij'} - t_{ij}|), & j < j' \\ [\boldsymbol{\Theta}^{(i)}]_{j'j}, & j > j' \end{cases} \quad (3)$$

We then can use $\boldsymbol{\Theta}^{(i)}$ to parameterize a PSQR that gives a joint distribution over $\mathbf{x}_i$ as:

$$P\Big(\mathbf{x}_i; \boldsymbol{\Theta}^{(i)}\Big) := \exp\Bigg[\sum_{j=1}^{n_i} \theta^{(i)}_{jj} \sqrt{x_{ij}} + \sum_{j=1}^{n_i-1} \sum_{j'>j}^{n_i} \theta^{(i)}_{jj'} \sqrt{x_{ij}x_{ij'}} \\ - \sum_{j=1}^{n_i} \log(x_{ij}!) - A_{n_i}\Big(\boldsymbol{\Theta}^{(i)}\Big) \Bigg]. \quad (4)$$

In (4), $A_{n_i}(\boldsymbol{\Theta}^{(i)})$ is a normalization constant called the log-partition function that ensures the legitimacy of the probability distribution in question:

$$A_{n_i}\Big(\boldsymbol{\Theta}^{(i)}\Big) := \log \sum_{\mathbf{x} \in \mathbb{N}^{n_i}} \exp \Bigg[\sum_{j=1}^{n_i} \theta^{(i)}_{jj} \sqrt{x_j} \\ + \sum_{j=1}^{n_i-1} \sum_{j'>j}^{n_i} \theta^{(i)}_{jj'} \sqrt{x_j x_{j'}} - \sum_{j=1}^{n_i} \log(x_j!) \Bigg]. \quad (5)$$

Note that in (5) we emphasize the dependency of the partition function upon the dimension of $\mathbf{x}$ using the subscript n_i, and $\mathbf{x} := \begin{bmatrix} x_1 & x_1 & \cdots & x_{n_i} \end{bmatrix}$.

To model the joint distribution of $\mathbf{x}_i$, TPSQR directly uses $\boldsymbol{\Theta}^{(i)}$, which is extracted from $\boldsymbol{\omega}$ and $\mathbf{W}$ via (3) depending on the individual and temporal irregularity of the data characterized by $\mathbf{t}_i$ and $\mathbf{o}_i$. Therefore, $\boldsymbol{\omega}$ and $\mathbf{W}$ serve as a

template for constructing $\boldsymbol{\Theta}^{(i)}$'s, and hence provide a "template parameterization." Since there are N subjects in total in the dataset, and each $\boldsymbol{\Theta}^{(i)}$ offers a personalized PSQR for one subject, TPSQR is capable of learning a collection of interrelated PSQRs due to the use of the template parameterization. Recall the well-rounded representation power of a template shown in Section 2.4; learning the template parameterization via TPSQR can hence offer a comprehensive perspective about the relationships for all possible temporally ordered pairs of event types.

Furthermore, since TPSQR is a generalization of PSQR, it inherits many desirable properties enjoyed by PSQR. A most prominent property is its capability of accommodating both positive and negative dependencies between variables. Such flexibility in general cannot be taken for granted when modeling multivariate count data. For example, a Poisson graphical model (Yang et al., 2015a) can only represent negative dependencies due to the diffusion of its log-partition function when positive dependencies are involved. Yet for example one drug (e.g., the blood thinner Warfarin) can have a positive influence on some conditions (e.g., bleeding) and a negative influence on others (e.g., stroke). We refer interested readers to Allen and Liu 2013; Yang et al. 2013; Inouye et al. 2015; Yang et al. 2015a; Inouye et al. 2016 for more details of PSQRs and other related Poisson graphical models.

4. Estimation

In this section, we present the pseudo-likelihood estimation problem for TPSQR. We then point out that solving this problem can be inefficient, which leads to the proposed Poisson pseudo-likelihood approximation to the original pseudo-likelihood problem.

4.1. Pseudo-Likelihood for TPSQR

We now present our estimation approach for TPSQR based on pseudo-likelihood. We start from considering the pseudo-likelihood for a given i^{th} subject. By (4), the log probability of x_{ij} conditioned on $\mathbf{x}_{i,-j}$, which is an $(n_i - 1) \times 1$ vector constructed by removing the jth component from $\mathbf{x}_i$, is given as:

$$\log P\Big(x_{ij} \mid \mathbf{x}_{i,-j}; \boldsymbol{\theta}^{(i)}_j \Big) = -\log(x_{ij}!) + \\ \Big(\theta^{(i)}_{jj} + \boldsymbol{\theta}^{(i)\top}_{j,-j} \sqrt{\mathbf{x}_{i,-j}} \Big) \sqrt{x_{ij}} - \tilde{A}_{n_i}\Big(\boldsymbol{\theta}^{(i)}_j \Big), \quad (6)$$

where $\boldsymbol{\theta}^{(i)}_j$ is the jth column of $\boldsymbol{\Theta}^{(i)}$ and hence

$$\boldsymbol{\theta}^{(i)}_j := \begin{bmatrix} \theta^{(i)}_{1j} & \cdots & \theta^{(i)}_{j-1,j} & \theta^{(i)}_{jj} & \theta^{(i)}_{j+1,j} & \cdots & \theta^{(i)}_{n_i,j} \end{bmatrix}^\top$$

$$:= \begin{bmatrix} \theta_{1j}^{(i)} & \cdots & \theta_{j-1,j}^{(i)} & \theta_{jj}^{(i)} & \theta_{j,j+1}^{(i)} & \cdots & \theta_{j,n_i}^{(i)} \end{bmatrix}^\top. \quad (7)$$

In (7), by the symmetry of $\boldsymbol{\Theta}^{(i)}$, we rearrange the index after $\theta_{jj}^{(i)}$ to ensure that the row index is no larger than the column index so that the parameterization is consistent with that in (4). We will adhere to this convention in the subsequent presentation. Furthermore, $\boldsymbol{\theta}_{j,-j}^{(i)}$ is an $(n_i - 1) \times 1$ vector constructed from $\boldsymbol{\theta}_j^{(i)}$ by excluding its jth component, and $\sqrt{\mathbf{x}_{i,-j}}$ is constructed by taking the square root of each component of $\mathbf{x}_{i,-j}$. Finally,

$$\tilde{A}_{n_i}\left(\boldsymbol{\theta}_j^{(i)}\right) :=$$
$$\log \sum_{\mathbf{x} \in \mathbb{R}^{n_i}} \exp\left[\left(\theta_{jj}^{(i)} + \boldsymbol{\theta}_{j,-j}^{(i)\top}\sqrt{\mathbf{x}_{i,-j}}\right)\sqrt{x_{ij}} - \log(x_{ij}!)\right], \quad (8)$$

which is a quantity that involves summing up infinitely many terms, and in general cannot be further simplified, leading to potential intractability in computing (8).

With the conditional distribution in (6) and letting $M := \sum_{i=1}^N n_i$, the pseudo-likelihood problem for TPSQR is given as:

$$\max_{\boldsymbol{\omega},\mathbf{W}} \frac{1}{M} \sum_{i=1}^N \sum_{j=1}^{n_i} \log \mathrm{P}\left(x_{ij} \mid \mathbf{x}_{i,-j}; \boldsymbol{\theta}_j^{(i)}\right). \quad (9)$$

(9) is the maximization over all the conditional distributions of all the count variables for all N personalized PSQRs generated by the template. *Therefore, it can be viewed as a pseudo-likelihood estimation problem directly for $\boldsymbol{\omega}$ and $\mathbf{W}$.* However, solving the pseudo-likelihood problem in (9) involves the predicament of computing the potentially intractable (8), which motivates us to use Poisson pseudo-likelihood as an approximation to (9).

4.2. Poisson Pseudo-Likelihood

Using the parameter vector $\boldsymbol{\theta}_j^{(i)}$, we define the conditional distribution of x_{ij} given by $\mathbf{x}_{i,-j}$ via the Poisson distribution as:

$$\hat{\mathrm{P}}\left(x_{ij} \mid \mathbf{x}_{i,-j}; \boldsymbol{\theta}_j^{(i)}\right) \propto$$
$$\exp\left[\left(\theta_{jj}^{(i)} + \boldsymbol{\theta}_{-j}^{(i)\top}\mathbf{x}_{i,-j}\right)x_{ij} - \exp\left(\theta_{jj}^{(i)} + \boldsymbol{\theta}_{-j}^{(i)\top}\mathbf{x}_{i,-j}\right)\right]. \quad (10)$$

Notice the similarity between (6) and (10). We can define the sparse Poisson pseudo-likelihood problem similar to the original pseudo-likelihood problem by replacing $\log \mathrm{P}\left(x_{ij} \mid \mathbf{x}_{i,-j}; \boldsymbol{\theta}_j^{(i)}\right)$ with $\log \hat{\mathrm{P}}\left(x_{ij} \mid \mathbf{x}_{i,-j}; \boldsymbol{\theta}_j^{(i)}\right)$:

$$\max_{\boldsymbol{\omega},\mathbf{W}} \frac{1}{M} \sum_{i=1}^N \sum_{j=1}^{n_i} \log \hat{\mathrm{P}}\left(x_{ij} \mid \mathbf{x}_{i,-j}; \boldsymbol{\theta}_j^{(i)}\right) - \lambda \|\mathbf{W}\|_{1,1}, \quad (11)$$

where $\lambda \geq 0$ is the regularization parameter, and the penalty

$$\|\mathbf{W}\|_{1,1} := \sum_{i=1}^{(p-1)p} \sum_{j=1}^L |[\mathbf{W}]_{ij}|$$

is used to encourage sparsity over the template parameterization $\mathbf{W}$ that determines the interactions between the occurrences of two distinct event types. As mentioned at the end of Section 4.1, TPSQR learning is equivalent to learning a PSQR over the template parameterization. Therefore, the sparsity penalty induced here is helpful to recover the structure of the underlying graphical model.

The major advantage of approximating the original pseudo-likelihood problem with Poisson pseudo-likelihood is the gain in computational efficiency. Based on the construction in Geng et al. 2017, (11) can be formulated as an l_1-regularized Poisson regression problem, which can be solved much more efficiently via many sophisticated algorithms and their implementations (Friedman et al., 2010; Tibshirani et al., 2012) compared to solving the original problem that involves the potentially challenging computation for (8). Furthermore, in the subsequent section, we will show that even though the Poisson pseudo-likelihood is an approximation procedure to the pseudo-likelihood of PSQR, under mild assumptions Poisson pseudo-likelihood is still capable of recovering the structure of the underlying PSQR.

4.3. Sparsistency Guarantee

For the ease of presentation, in this section we will reuse much of the notation that appears previously. The redefinitions introduced in this section only apply to the contents in this section and the related proofs in the Appendix. Recall at the end of Section 4.1, the pseudo-likelihood problem of TPSQR can be viewed as learning a PSQR parameterized by the template. Therefore, without loss of generality, we will consider a PSQR over p count variables $\mathbf{X} = \mathbf{x} \in \mathbb{N}^p$ parameterized by a $p \times p$ symmetric matrix $\boldsymbol{\Theta}^*$, where $\mathbf{X} := \begin{bmatrix} X_1 & X_2 & \cdots & X_p \end{bmatrix}^\top$ is the multivariate random variable, and $\mathbf{x}$ is an assignment to $\mathbf{X}$. We use $\|\cdot\|_\infty$ to represent the infinity norm of a vector or a matrix. Let $\mathbb{X} := \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ be a dataset with n independent and identically distributed (i.i.d.) samples generated from the PSQR. Then the joint probability distribution over $\mathbf{x}$ is:

$$\mathrm{P}(\mathbf{x}; \boldsymbol{\Theta}^*) := \exp\left[\sum_{j=1}^p \theta_{jj}^* \sqrt{x_{ij}} + \sum_{j=1}^{p-1} \sum_{j'>j}^p \theta_{jj'}^* \sqrt{x_{ij}x_{ij'}} \right.$$
$$\left. - \sum_{j=1}^p \log(x_{ij}!) - A(\boldsymbol{\Theta}^*)\right],$$

where $A(\boldsymbol{\Theta}^*)$ is the log-partition function, and the corresponding Poisson pseudo-likelihood problem is:

$$\hat{\boldsymbol{\Theta}} := \arg\min_{\boldsymbol{\Theta}} F(\boldsymbol{\Theta}) + \lambda \|\boldsymbol{\Theta}\|_{1,\mathrm{off}}, \quad (12)$$

where

$$F(\mathbf{\Theta}) := \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \Big[-(\theta_{jj} + \boldsymbol{\theta}_{j,-j}^{\top} \mathbf{x}_{i,-j}) x_{ij} \tag{13}$$
$$+ \exp(\theta_{jj} + \boldsymbol{\theta}_{j,-j}^{\top} \mathbf{x}_{i,-j}) \Big],$$

and $\|\mathbf{\Theta}\|_{1,\text{off}}$ represents imposing l_1 penalty over all but the diagonal components of $\mathbf{\Theta}$.

Sparsistency (Ravikumar et al., 2007) addresses whether $\hat{\mathbf{\Theta}}$ can recover the structure of the underlying $\mathbf{\Theta}^*$ with high probability using n i.i.d. samples. In what follows, we will show that $\hat{\mathbf{\Theta}}$ is indeed sparsistent under mild assumptions.

We use $\mathbb{E}[\cdot]$ to denote the expectation of a random variable under $P(\mathbf{x}; \mathbf{\Theta}^*)$. The first assumption is about the boundedness of $\mathbb{E}[\mathbf{X}]$, and the boundedness of the partial second order derivatives of a quantity related to the log-partition $A(\mathbf{\Theta}^*)$. This assumption is standard in the analysis of pseudo-likelihood methods (Yang et al., 2015a).

Assumption 1. $\|\mathbb{E}[\mathbf{X}]\|_{\infty} \le C_1$ *for some* $C_1 > 0$. *Let*

$$B(\mathbf{\Theta}, \mathbf{b}) := \log \sum_{\mathbf{x} \in \mathbb{N}^p} \exp \Bigg[\sum_{j=1}^{p} \theta_{jj} \sqrt{x_j} + \mathbf{b}^{\top} \mathbf{x}$$
$$+ \sum_{j=1}^{p-1} \sum_{j'>j}^{p} \theta_{jj'} \sqrt{x_j x_{j'}} - \sum_{j=1}^{p} \log(x_j!) \Bigg].$$

For some $C_2 > 0$, *and* $\forall k \in [0,1]$,

$$\forall j \in \{1, 2, \cdots, p\}, \quad \frac{\partial^2 B(\mathbf{\Theta}, \mathbf{0} + k\mathbf{e}_j)}{\partial^2 b_j} \le C_2,$$

where $\mathbf{e}_j$ *is the one-hot vector with the* j^{th} *component as* 1.

The following assumption characterizes the boundedness of the conditional distributions given by the PSQR under $\mathbf{\Theta}^*$ and by the Poisson approximation using the same $\mathbf{\Theta}^*$.

Assumption 2. *Let* $\lambda_{ij}^* := \exp\left(\theta_{jj}^* + \boldsymbol{\theta}_{j,-j}^{*\top} \mathbf{x}_{i,-j}\right)$ *be the mean parameter of a Poisson distribution. Then* $\forall i \in \{1, 2 \cdots, n\}$ *and* $\forall j \in \{1, 2, \cdots, p\}$, *for some* $C_3 > 0$ *and* $C_4 > 0$, *we have that* $\mathbb{E}[X_j \,|\, \mathbf{x}_{i,-j}] \le C_3$ *and* $\left| \lambda_{ij}^* - \mathbb{E}[X_j \,|\, \mathbf{x}_{i,-j}] \right| \le C_4$.

The third assumption is the mutual incoherence condition vital to the sparsistency of sparse statistical learning with l_1-regularization. Also, with a slight abuse of notation, in the remaining of Section 4.3 as well as in the corresponding proofs, we should view $\mathbf{\Theta}$ as a vector generated by stacking up $\theta_{jj'}$'s, where $j \le j'$, whenever it is clear from context.

Assumption 3. *Let* $\mathbf{\Theta}^*$ *be given. Define the index sets*

$$A := \left\{ (j, j') \mid \theta_{jj'}^* \ne 0, j \ne j', j, j' \in \{1, 2, \cdots, p\} \right\},$$
$$D := \left\{ (j, j) \mid j \in \{1, 2, \cdots, p\} \right\}, \quad S := A \cup D,$$

$$I := \left\{ (j, j') \mid \theta_{jj'}^* = 0, j \ne j', j, j' \in \{1, 2, \cdots, p\} \right\}.$$

Let $\mathbf{H} := \nabla^2 F(\mathbf{\Theta}^*)$. *Then for some* $0 < \alpha < 1$ *and* $C_5 > 0$, *we have* $\left\| \mathbf{H}_{IS} \mathbf{H}_{SS}^{-1} \right\|_{\infty} \le 1 - \alpha$ *and* $\left\| \mathbf{H}_{SS}^{-1} \right\|_{\infty} \le C_5$, *where we use the index sets as subscripts to represent the corresponding components of a vector or a matrix.*

The final assumption characterizes the second-order Taylor expansion of $F(\mathbf{\Theta}^*)$ at a certain direction $\mathbf{\Delta}$.

Assumption 4. *Let* $R(\mathbf{\Delta})$ *be the second-order Taylor expansion remainder of* $\nabla F(\mathbf{\Theta})$ *around* $\mathbf{\Theta} = \mathbf{\Theta}^*$ *at direction* $\mathbf{\Delta} := \mathbf{\Theta} - \mathbf{\Theta}^*$ *(i.e.* $\nabla F(\mathbf{\Theta}) = \nabla F(\mathbf{\Theta}^*) + \nabla^2 F(\mathbf{\Theta}^*)(\mathbf{\Theta} - \mathbf{\Theta}^*) + R(\mathbf{\Delta})$*), where* $\|\mathbf{\Delta}\|_{\infty} \le r := 4C_5 \lambda \le \frac{1}{C_5 C_6}$ *with* $\mathbf{\Delta}_I = \mathbf{0}$, *and for some* $C_6 > 0$. *Then* $\|R(\mathbf{\Delta})\|_{\infty} \le C_6 \|\mathbf{\Delta}\|_{\infty}^2$.

With these mild assumptions, the sparsistency result is stated in Theorem 1.

Theorem 1. *Suppose that Assumption 1 - 4 are all satisfied. Then, with probability of at least* $1 - \left((\exp(C_1 + C_2/2) + 8) p^{-2} + p^{-1/C_2} \right)$, *$\hat{\mathbf{\Theta}}$ shares the same structure with $\mathbf{\Theta}^*$, if for some constant* $C_7 > 0$,

$$\lambda \ge \frac{8}{\alpha} \left[C_3 (3 \log p + \log n) + (3 \log p + \log n)^2 \right] \sqrt{\frac{\log p}{n}}$$
$$+ 8C_4 \left(C_1 + \sqrt{\frac{2 \log p}{n}} \right) \alpha, \quad \lambda \le C_7 \sqrt{\frac{\log^5 p}{n}},$$

$$r \le \|\mathbf{\Theta}_S^*\|_{\infty}, \text{ and } n \ge \left(64 C_7 C_5^2 C_6 / \alpha \right)^2 \log^5 p.$$

We defer the proof of Theorem 1 to the Appendix. Note that $\log^5 p$ in Theorem 1 represents a higher sample complexity compared to similar results in the analysis of Ising models (Ravikumar et al., 2010). Such a higher sample complexity intuitively makes sense since the multivariate count variables that we deal with are unbounded and usually heavy-tailed, and we are also considering the Poisson pseudo-likelihood approximation to the original pseudo-likelihood problem induced by PSQRs. The fact that Poisson pseudo-likelihood is a sparsistent procedure for learning PSQRs not only provides an efficient approach to learn PSQRs with strong theoretical guarantees, but also establishes a formal correspondence between local Poisson graphical models (LPGMs, Allen and Liu 2013) and PSQRs. This is because Poisson pseudo-likelihood is also a sparsistent procedure for LPGMs. Compared to PSQRs, LPGMs are more intuitive yet less stringent theoretically due to the lack of a joint distribution defined by the model. Fortunately, with the guarantees in Theorem 1, we are able to provide some reassurance for the use of LPGMs in terms of structure recovery.

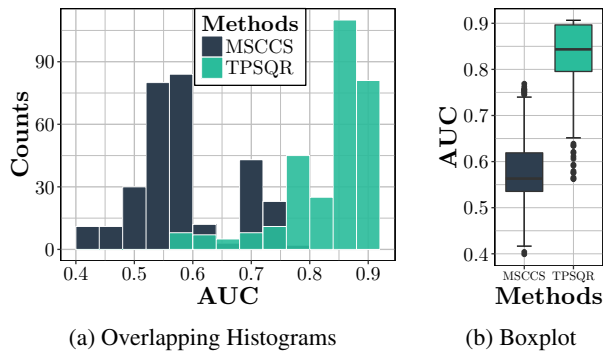(a) Overlapping Histograms (b) Boxplot

Figure 2: Overall performance of TPSQR and MSCCS measured by AUC among 300 different experimental configurations for each of the two methods.

5. Adverse Drug Reaction Discovery

To demonstrate the capability of TPSQRs to capture temporal relationships between different pairs of event types in LED, we use ADR discovery from EHR as an example. ADR discovery is the task of finding unexpected and negative incidents caused by drug prescriptions. In EHR, time-stamped drug prescriptions as well as condition diagnoses are collected from millions of patients. These prescriptions of different drugs and diagnoses of different conditions can hence be viewed as various event types in LED. Therefore, using TPSQR, we can model whether the occurrences of a particular drug k could elevate the possibility of the future occurrences of a condition k' on different time scales by estimating $\mathbf{w}_{kk'}$ defined in (2). If an elevation is observed, we can consider the drug k as a potential candidate to cause condition k' as an adverse drug reaction.

Postmarketing ADR surveillance from EHR is a multi-decade research and practice effort that is of utmost importance to the pharmacovigilance community (Bate et al., 2018), with substantial financial and clinical implication for health care delivery (Sultana et al., 2013). Various ADR discovery methods have been proposed over the years (Harpaz et al., 2012), and a benchmark task is created by the Observational Medical Outcome Partnership (OMOP, Simpson 2011) to evaluate the ADR signal detection performance of these methods. The OMOP task is to identify the ADRs in 50 drug-condition pairs, coming from a selective combination of ten different drugs and nine different conditions. Among the 50 pairs, 9 of them are confirmed ADRs, while the remaining 41 of them are negative controls.

A most successful ADR discovery method using EHR is the multiple self-controlled case series (MSCCS, Simpson et al. 2013), which has been deployed in real-world ADR discovery related projects (Hripcsak et al., 2015). A reason for the success of MSCCS is its introduction of fixed effects to address the heterogeneity among different sub-jects (e.g. patients in poorer health might tend to be more likely to have a heart attack compared to a healthy person, which might confound the effects of various drugs when identifying drugs that could cause heart attacks as an ADR).

Therefore, when using TPSQR, we will also introduce fixed effects to equip TPSQRs with the capability of addressing subject heterogeneity. Specifically, we consider learning a variant of (11):

$$\max_{\boldsymbol{\alpha},\boldsymbol{\omega},\mathbf{W}} \frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\alpha_{io_{ij}}+\log\hat{\mathrm{P}}\left(x_{ij}\mid\mathbf{x}_{i,-j};\boldsymbol{\theta}_j^{(i)}\right)-\lambda\|\mathbf{W}\|_{1,1},$$

where $\boldsymbol{\alpha}$ is the fixed effect parameter vector constructed by $\alpha_{io_{ij}}$'s that depicts the belief that different patients could have different baseline risks of experiencing different types of events.

6. Experiments

In what follows, we will compare the performances of TP-SQR, MSCCS, and Hawkes process (Bao et al., 2017) in the OMOP task. The experiments are conducted using Marshfield Clinic EHRs with millions of drug prescription and condition diagnosis events from 200,000 patients.

6.1. Experimental Configuration

Minimum Duration: clinical encounter sequences from different patients might span across different time lengths. Some have decades of observations in their records while other might have records only last a few days. We therefore consider minimum duration of the clinic encounter sequence as a threshold to determine whether we admit a patient to the study or not. In our experiments, we consider two minimum duration thresholds: *0.5 year* and *1 year*.

Maximum Time Difference: for TPSQR, in (1), τ_L determines the maximum time difference between the occurrences of two events within which the former event might have nonzero influence on the latter event. We call τ_L the maximum time difference to characterize how distant in the past we would like to take previous occurrences into consideration when modeling future occurrences. In our experiments, we consider three maximum time differences: *0.5 year*, *1 year*, and *1.5 years*. $L = 3$ and the corresponding influence functions are chosen according to Bao et al. 2017. In MSCCS, a configuration named *risk window* serves a similar purpose to the maximum difference in TPSQR. We choose three risk windows according to Kuang et al. 2017b so as to ensure that the both TPSQR and MSCCS have similar capability in considering the event history on various time scales.

Regularization Parameter: we use l_1-regularization for TPSQR since it encourages sparsity, and the sparsity pat-
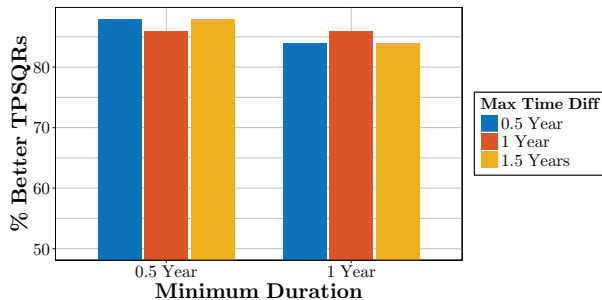
Figure 3: Percentage of better TPSQR models under various minimum duration and maximum time difference designs compared to the best MSCCS model
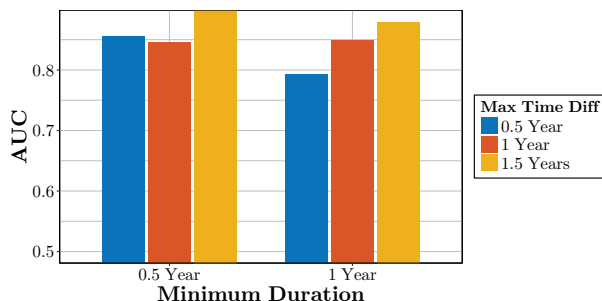

Figure 4: AUC of TPSQR models selected by AIC for given minimum duration and maximum time difference designs

terns learned correspond to the structures of the graphical models. We use L_2-regularization for MSCCS since it yields outstanding empirical performance in previous studies (Simpson et al., 2013; Kuang et al., 2017b). 50 regularization parameters are chosen for both TPSQR and MSCCS.

To sum up, there are $2 \times 3 \times 50 = 300$ experimental configurations respectively for TPSQR and MSCCS.

6.2. Overall Performance

For each of the 300 experimental configurations for TPSQR and MSCCS, we perform the OMOP task using our EHR data. Both TPSQR and MSCCS can be implemented by the R package glmnet (Friedman et al., 2010; Tibshirani et al., 2012). We then use Area Under the Curve (AUC) for the receiver operating characteristic curve to evaluate how well TPSQR and MSCCS can distinguish actual ADRs from negative controls under this particular experimental configuration. The result is 300 AUCs corresponding to the total number of experimental configurations for each of the two methods. For TPSQR, since the effect of drug k on condition k' is estimated over different time scales via $\mathbf{w}_{kk'}$, the score corresponding to this drug-condition pair used to calculate the AUC is computed by the average over

all the components of $\mathbf{w}_{kk'}$. For MSCCS, AUC is computed according to Kuang et al. 2017b. Figure 2 presents the histogram of these two sets of 300 AUCs. The contrast in the performances between TPSQR and MSCCS is obvious. The distribution of TPSQR shifts substantially towards higher AUC values compared to the distribution of MSCCS. Therefore, the overall performance of TPSQR is superior to that of MSCCS in the OMOP task under various experimental configurations in question. As a matter of fact, the top performing TPSQR model reaches an AUC of 0.91, as opposed to 0.77 for MSCCS. Furthermore, the majority of TPSQRs have higher AUCs even compared to the MSCCS model that has the best AUC. We also contrast the performance of TPSQR with the Hawkes process method in Bao et al. 2017, whose best AUC is 0.84 under the same experiment configurations.

6.3. Sensitivity Analysis and Model Selection

To see how sensitive the performance of TPSQR is for different choices of experimental configurations, we compute the percentage of TPSQRs with a given minimum duration and a given maximum time difference design that are better than the best MSCCS model (with an AUC of 0.77). The results are summarized in Figure 3. As can be seen, the percentage of better TPSQRs is consistently above 80% under various scenarios, suggesting the robustness of TPSQRs to various experimental configurations. Given a fixed minimum duration and a fixed maximum time difference, we conduct model selection for TPQSRs by the Akaike information criterion (AIC) over the regularization parameters. The AUC of the selected models are summarized in Figure 4. Note that under various fixed minimum duration and maximum time difference designs, AIC is capable of selecting models with high AUCs. In fact, all the models selected by AIC have higher AUCs than the best performer of MSCCS. This phenomenon demonstrates that the performance of TPSQR is consistent and robust with respect to the various choices of experimental configurations.

7. Conclusion

We propose TPSQRs, a generalization of PSQRs for the temporal relationships between different event types in LED. We propose the use of Poisson pseudo-likelihood approximation to solve the pseudo-likelihood problem arising from PSQRs. The approximation procedure is extremely efficient to solve, and is sparsistent in recovering the structure of the underlying PSQR. The utility of TPSQR is demonstrated using Marshfield Clinic EHRs for adverse drug reaction discovery.

References

Genevera I Allen and Zhandong Liu. A local poisson graphical model for inferring networks from sequencing data. *IEEE Transactions on Nanobioscience*, 12(3):189–198, 2013.

Yujia Bao, Zhaobin Kuang, Peggy Peissig, David Page, and Rebecca Willett. Hawkes process modeling of adverse drug reactions with longitudinal observational data. In *Machine Learning for Healthcare Conference*, pages 177–190, 2017.

Andrew Bate, Robert F Reynolds, and Patrick Caubel. The hope, hype and reality of big data for pharmacovigilance, 2018.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1555–1564. ACM, 2016.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

Sinong Geng, Zhaobin Kuang, and David Page. An efficient pseudo-likelihood method for sparse binary pairwise Markov network estimation. *arXiv preprint arXiv:1702.08320*, 2017.

Sinong Geng, Zhaobin Kuang, Jie Liu, Stephen Wright, and David Page. Stochastic learning for sparse discrete Markov random fields with controlled gradient approximation error. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (2018)*, 2018.

Asela Gunawardana, Christopher Meek, and Puyang Xu. A model for temporal dependencies in event streams. In *Advances in Neural Information Processing Systems*, pages 1962–1970, 2011.

Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

Rave Harpaz, William DuMouchel, Nigam H Shah, David Madigan, Patrick Ryan, and Carol Friedman. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology & Therapeutics*, 91 (6):1010–1021, 2012.

George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in Health Technology and Informatics*, 216:574, 2015.

David Inouye, Pradeep Ravikumar, and Inderjit Dhillon. Square root graphical models: Multivariate generalizations of univariate exponential families that permit positive dependencies. In *International Conference on Machine Learning*, pages 2445–2453, 2016.

David I Inouye, Pradeep K Ravikumar, and Inderjit S Dhillon. Fixed-length poisson MRF: Adding dependencies to the multinomial. In *Advances in Neural Information Processing Systems*, pages 3213–3221, 2015.

Mladen Kolar, Le Song, Amr Ahmed, and Eric P Xing. Estimating time-varying networks. *The Annals of Applied Statistics*, pages 94–123, 2010.

Zhaobin Kuang, James Thomson, Michael Caldwell, Peggy Peissig, Ron Stewart, and David Page. Baseline regularization for computational drug repositioning with longitudinal observational data. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2521–2528. AAAI Press, 2016a.

Zhaobin Kuang, James Thomson, Michael Caldwell, Peggy Peissig, Ron Stewart, and David Page. Computational drug repositioning using continuous self-controlled case series. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 491–500. ACM, 2016b.

Zhaobin Kuang, Sinong Geng, and David Page. A screening rule for l1-regularized ising model estimation. In *Advances in Neural Information Processing Systems*, pages 720–731, 2017a.

Zhaobin Kuang, Peggy Peissig, Vitor Santos Costa, Richard Maclin, and David Page. Pharmacovigilance via baseline regularization with large-scale longitudinal observational data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1537–1546. ACM, 2017b.

Jie Liu and David Page. Bayesian estimation of latently-grouped parameters in undirected graphical models. In *Advances in Neural Information Processing Systems*, pages 1232–1240, 2013.

Jie Liu, David Page, Houssam Nassif, Jude Shavlik, Peggy Peissig, Catherine McCarty, Adedayo A Onitilo, and Elizabeth Burnside. Genetic variants improve breast cancer risk prediction on mammograms. In *AMIA Annual Symposium Proceedings*, volume 2013, page 876. American Medical Informatics Association, 2013.

Jie Liu, David Page, Peggy Peissig, Catherine McCarty, Adedayo A Onitilo, Amy Trentham-Dietz, and Elizabeth Burnside. New genetic variants improve personalized breast cancer diagnosis. *AMIA Summits on Translational Science Proceedings*, 2014:83, 2014a.

Jie Liu, Chunming Zhang, Elizabeth Burnside, and David Page. Learning heterogeneous hidden Markov random fields. In *Artificial Intelligence and Statistics*, pages 576–584, 2014b.

Jie Liu, Chunming Zhang, David Page, et al. Multiple testing under dependence via graphical models. *The Annals of Applied Statistics*, 10(3):1699–1724, 2016.

Stephane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic Press, 2008.

James M Ortega and Werner C Rheinboldt. *Iterative solution of nonlinear equations in several variables*. SIAM, 2000.

Pradeep Ravikumar, Han Liu, John Lafferty, and Larry Wasserman. Spam: Sparse additive models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 1201–1208. Curran Associates Inc., 2007.

Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional ising model selection using 1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Martijn J Schuemie, Gianluca Trifirò, Preciosa M Coloma, Patrick B Ryan, and David 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*, 25(6):2577–2592, 2016.

Shawn E Simpson. *Self-controlled methods for postmarketing drug safety surveillance in large-scale longitudinal data*. Columbia University, 2011.

Shawn E Simpson, David Madigan, Ivan Zorych, Martijn J Schuemie, Patrick B Ryan, and Marc A Suchard. Multiple self-controlled case series for large-scale longitudinal observational databases. *Biometrics*, 69(4):893–902, 2013.

Janet Sultana, Paola Cutroneo, and Gianluca Trifirò. Clinical and economic burden of adverse drug reactions. *Journal of Pharmacology & Pharmacotherapeutics*, 4(Suppl1): S73, 2013.

Nicholas P Tatonetti, P Ye Patrick, Roxana Daneshjou, and Russ B Altman. Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, 4(125): 125ra31–125ra31, 2012.

Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266, 2012.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

Jeremy Weiss, Sriraam Natarajan, and David Page. Multiplicative forests for continuous-time processes. In *Advances in Neural Information Processing Systems*, pages 458–466, 2012.

Jeremy C Weiss and David Page. Forest-based point process for event prediction from electronic health records. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 547–562. Springer, 2013.

Eunho Yang and Pradeep Ravikumar. On the use of variational inference for learning discrete graphical model. In *International Conference on Machine Learning*, pages 1009–1016, 2011.

Eunho Yang, Pradeep K Ravikumar, Genevera I Allen, and Zhandong Liu. On poisson graphical models. In *Advances in Neural Information Processing Systems*, pages 1718–1726, 2013.

Eunho Yang, Pradeep Ravikumar, Genevera I Allen, and Zhandong Liu. Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16(1):3813–3847, 2015a.

Sen Yang, Zhaosong Lu, Xiaotong Shen, Peter Wonka, and Jieping Ye. Fused multiple graphical lasso. *SIAM Journal on Optimization*, 25(2):916–943, 2015b.