IISE Transactions

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uiie21

Efficient estimation and selection for regularized dynamic logistic regression

Sumin Shen, Zhiyang Zhang, Ran Jin & Xinwei Deng

To cite this article: Sumin Shen, Zhiyang Zhang, Ran Jin & Xinwei Deng (2025) Efficient estimation and selection for regularized dynamic logistic regression, IISE Transactions, 57:6, 639-654, DOI: 10.1080/24725854.2024.2359991

To link to this article: https://doi.org/10.1080/24725854.2024.2359991

IISE TRANSACTIONS
2025, VOL. 57, NO. 6, 639–654
https://doi.org/10.1080/24725854.2024.2359991

Efficient estimation and selection for regularized dynamic logistic regression

Sumin Shen[a], Zhiyang Zhang[a], Ran Jin[b], and Xinwei Deng[a]

[a]Department of Statistics, Virginia Tech, Blacksburg, VA, USA; [b]Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA, USA

ABSTRACT

In various data science applications, the relationship between predictor variables and the response is dynamic in the sense that the corresponding model parameters are varying coefficients. Estimation and variable selection for such dynamic models are challenging with a large number of parameters and complex optimization. In this work, we propose a regularized dynamic logistic regression for efficient variable selection and model estimation. The proposed method considers a combination of fused and group regularization to estimate varying effects of important predictors on responses in the presence of irrelevant predictors. Specifically, we select the important variable with dynamic impact on responses through the selection of the entire group of piecewise constant functions for parameters, which can characterize dynamic impacts of predictor variables. Moreover, we develop an efficient algorithm based on the alternating direction method of multipliers for parameter estimation. The performance of the proposed method is evaluated by both simulations and real case studies.

ARTICLE HISTORY
Received 2 November 2022
Accepted 17 April 2024

KEYWORDS
Dynamic model; fused group Lasso; variable selection; varying coefficients

1. Introduction

Advances in technology and computation have provided opportunities to explore large and high-dimensional datasets. The collected high-dimensional dataset presents a variety of details for problems of interest, but frequently contains irrelevant or redundant information. To extract important predictors relevant to responses and estimate their effects for statistical inference, variable selection techniques, such as the best subset selection, the non-negative garrote (Breiman, 1995; Yuan and Lin 2007; Xiong, 2010), and the LASSO (Tibshirani, 1996), have become popular in many applications. Variable selection supervises model complexity and produces sparse interpretable models. Studies of variable selection are often conducted for the regression model due to its well-established properties. However, the relationship between predictor variables and the response can be dynamic in the sense that the corresponding model parameters are varying coefficients. It is challenging to estimate and select key predictor variables with dynamic effects when irrelevant variables are present.

Studies of variable selection for varying coefficient models have not attracted full attention. The Varying Coefficient Model (VCM) is an important tool for studying the dynamic impacts of predictors on responses (Cleveland *et al.,* 2017; Hastie and Tibshirani, 1993) with broad applications (Aguilar and West, 2000; Cai *et al.,* 2000; Beaulieu *et al.,* 2012; Hong *et al.,* 2015). Our work is motivated by the crystal growth manufacturing problem (Zhang *et al.,* 2014; Sun *et al.,* 2016; Jin *et al.,* 2019). In order to produce high-quality crystal ingots, the major procedure in crystal growth manufacturing is to pull the crystal ingot out of the melted polycrystalline silicon upwards and rotate simultaneously and slowly. This process typically lasts more than 20 hours. During the crystal growth, many process variables, such as the pulling speed and the power of the heater, affect the quality of the crystal ingots. These effects of process variables on the quality variable (i.e., the diameter of the ingot) are dynamic, due to the different growth phases and the inevitable equipment degradation. Considering the impact of different process variables at different process phases or equipment status during the crystal growth, it is more sensible to model these dynamic effects in a VCM framework.

The VCM investigates dynamic patterns of variables by allowing the coefficient of variables to be functional coefficients. In the literature, the VCM is linear in terms of predictor variables, of which the coefficients are dependent on other variables. The dependency has been characterized by various coefficient formats, such as the polynomial splines (Hastie and Tibshirani, 1993; Fan and Zhang, 1999) and the smoothing splines (Hoover *et al.,* 1998; Fan and Zhang, 2008). The functional coefficients of the VCM provide great flexibility in modeling, but are frequently implicit and complicated. Moreover, the applications of the VCM were constrained by the assumption that coefficients vary in the forms of smoothing functions. In other words, the functional coefficients of the VCM may not be suited for handling discontinuities or abrupt structural changes in dynamic predictor variables.

CONTACT Xinwei Deng xdeng@vt.edu

Supplemental data for this article can be accessed online at https://doi.org/10.1080/24725854.2024.2359991.

To address the limitations of implicit smoothing functional coefficients, works based on penalized likelihood estimation have been developed to estimate varying coefficient structures. Along this direction, the sparse graphical regression model with parameter fusion was developed to recover temporal structures in time-varying Markov random field networks (Ahmed and Xing, 2009). The multinomial fused LASSO regression model was presented to solve longitudinal classification problems (Adhikari *et al.,* 2019). The varying-coefficient varying-structure model incorporating the fused LASSO and the LASSO penalty was introduced to illustrate dynamic coefficient structures (Kolar *et al.,* 2009). The dynamic quality-process model was proposed to describe the dynamic effects of variables with piecewise linear functions (Jin *et al.,* 2019). The main idea is to apply the penalty term to the magnitude of jumps in coefficients to encourage consecutive segments in coefficients to have similar estimation values (Yao, 1988; Lavielle, 2005; Tibshirani *et al.,* 2005). However, these works did not address the issue of selecting important predictor variables with dynamic effects in the presence of irrelevant variables. The correct identification of key predictor variables with dynamic coefficients can enhance the model interpretability and improve the computation efficiency.

In this work, we propose a regularized Dynamic Logistic Regression (rDLR) model for variable selection and estimation of the varying effects of key predictor variables in the presence of irrelevant variables. The proposed method uses a set of piecewise constant functions as an entire group to characterize the dynamic effects of each predictor variable. Under such a formulation, the selection of each predictor variable with dynamic effects is equivalent to the selection of the entire group of piecewise constant functions. We propose an appropriate regularization to select important predictor variables at the level of individual predictors while allowing dynamic effects for predictor variables. The development is described in detail in Section 2. To efficiently estimate coefficient parameters, we develop an algorithm based on the Alternating Direction Method of Multipliers (ADMM, Boyd *et al.*, 2011) coupled with the Newton–Raphson method. The proposed model has three key advantages. First, the proposed method can yield sparse and interpretable models with predictor variables in high-dimensional data sets. Second, the dynamic effects of variables are represented by a set of piecewise constant coefficients. The piecewise constant functions are adaptable and able to accommodate multiple changes in coefficient structures. Last, the developed ADMM-based method for parameter estimation is applicable to generalized LASSO penalties in the generalized linear model framework and is effective for dealing with large-scale data sets. We used simulation studies and two real case studies (crystal growth manufacturing and a Hong Kong environmental study) to demonstrate that it is beneficial to consider both the VCM structure and the selection of variables with dynamic effects when addressing these problems.

It is worth pointing out that there are several different aspects of the proposed method in comparison with other related methods described in Christoffersen (2021), which includes the static logistic regression model, the generalized additive model, the time varying effects Cox model, and the dynamic hazard method with a logistic link function. Compared with the static logistic regression model, our proposed method studies the varying effects of variables. The generalized additive model uses cubic regression splines with knots spread evenly through the covariate values, whereas our proposed method performs variable selection and estimation of dynamic effects by a set of piecewise constant coefficients to accommodate multiple changes in coefficient structures. The response of interest in the time-varying effects Cox-model is the natural log of the hazard ratio whereas our proposed method considers the natural log of the odds ratio.

The remainder of this article is organized as follows. In Section 2, we detail the proposed regularized dynamic logistic regression model. In Section 3, we describe the estimation method based on ADMM. In Sections 4 and 5, we present the simulations and real case studies. We conclude this work with discussion in Section 6.

2. Regularized dynamic logistic regression

Let y_t denote the binary response at time t, $y_t \in \{0, 1\}, t = 1, ..., n$, and $\boldsymbol{x}_t$ denote the data point at time t, $\boldsymbol{x}_t = (x_{t,1}, ..., x_{t,p})^T$, $p \geq 1$. Without loss of generality, we assume that the p variables are continuous variables with measurements at n time points. We denote $p(\boldsymbol{x}_t) = Pr(y_t = 1|\boldsymbol{x}_t)$. That is, we consider

$$y_t|\boldsymbol{x}_t = \begin{cases} 1, & \text{w.p.} \quad p(\boldsymbol{x}_t), \\ 0, & \text{w.p.} \quad 1 - p(\boldsymbol{x}_t). \end{cases}$$

The conditional probability $p(\boldsymbol{x}_t)$ with the logistic regression model $\log(p(\boldsymbol{x}_t)/(1 - p(\boldsymbol{x}_t))) = \boldsymbol{x}_t^T \boldsymbol{\beta}_t$ with $\boldsymbol{\beta}_t = (\beta_{t,1}, ..., \beta_{t,p})^T$. Moreover, we consider each data point $\boldsymbol{x}_t$ has its own coefficient parameters $\boldsymbol{\beta}_t$ at each time point. In other words, the coefficient parameter $\boldsymbol{\beta}_t$ is allowed to vary over time. Thus, there are np parameters in the proposed model. It is over-parameterized since the number of unknown coefficient parameters exceeds the sample size n. Therefore, we need to consider appropriate penalties in the estimation procedure to enable variable selection and parsimonious model.

It is challenging to estimate the over-parameterized model without additional constraints on parameters. To address the estimation problem, we consider a regularized dynamic logistic regression model, which is expressed as

$$\begin{aligned} \text{logit}(\boldsymbol{x}_t) = \log \frac{p(\boldsymbol{x}_t)}{1 - p(\boldsymbol{x}_t)} = \boldsymbol{x}_t^T \boldsymbol{\beta}_t, t = 1, ..., n \\ s.t. \sum_{j=1}^{p} \sum_{t=2}^{n} |\beta_{t,j} - \beta_{t-1,j}| \leq M_1, \\ \sum_{j=1}^{p} \sqrt{\beta_{1,j}^2 + \cdots + \beta_{n,j}^2} \leq M_2, \end{aligned} \tag{1}$$

where $M_1 \geq 0$ and $M_2 \geq 0$ are the tuning parameters for the l_1-norm fused LASSO and l_2-norm group LASSO

penalties, respectively. Both penalty terms are beneficial to reduce the number of unknown parameters in the proposed model. The l_1-norm fused LASSO penalty on the consecutive coefficient parameters, $\sum_{j=1}^{p}\sum_{t=2}^{n}|\beta_{t,j}-\beta_{t-1,j}|$, encourages that the adjacent coefficient parameters to have similar values than the distant coefficient parameters. In other words, the l_1-norm fused LASSO penalty favors piecewise constant functional coefficients to approximate the dynamic coefficients. This idea of parameter fusion is comparable to those presented by Kolar *et al.* (2009) and Ahmed and Xing (2009). Moreover, the data-driven detection of numbers and magnitudes of change-points in piecewise constant functions is flexible enough to estimate dynamic coefficients even when abrupt structural changes in coefficients exist. In real situations, such a formulation implies an overall dynamic model structure while a static model within a short period of time.

The l_2-norm group LASSO penalty, $\sum_{j=1}^{p}\sqrt{\beta_{1,j}^2+\cdots+\beta_{n,j}^2}$, takes into account the coefficient parameters for the jth predictor variable as a group. Selecting an important variable is essentially equivalent to selecting the entire group. Note that the group LASSO penalty is reduced to the LASSO when $\beta_{1,j}=\cdots=\beta_{n,j}$. The variable selection via the group LASSO penalty yields a sparse and interpretable model, revealing the relationship between the response and the important predictors in the presence of irrelevant variables. It is worthwhile to differentiate the variable selection feature of the proposed method from that of previous works (Kolar *et al.*, 2009; Adhikari *et al.*, 2019). The use of the l_1-norm LASSO penalty in their methods results in a sparse structure in the coefficient parameters but selects no important predictor variables.

The combination of the l_1-norm fused LASSO penalty and the l_2-norm group LASSO penalty yields models that are sparse at the variable level and fused within each important variable. This combination allows simultaneously the selection of important variables and the incorporation of dynamic effects of variables. This idea is similar to the sparse group LASSO (Friedman *et al.*, 2010; Simon *et al.*, 2013), which yields the group-wise sparsity and the within-group sparsity. Note that the idea of combining LASSO, fused LASSO, and group LASSO have been used in linear and logistic regression models (Meier *et al.*, 2008; Zhou *et al.*, 2012; Lee *et al.*, 2014) with the focus on fixed parameters for predictors. In contrast, the proposed method focuses on dynamic parameters by appropriately using a combination of fused LASSO and group LASSO. Moreover, we develop an efficient ADMM algorithm for parameter estimation as described in the next section.

We remark that the proposed method is different from methods in McCormick *et al.* (2012) and Fahrmeir (1992), which uses a combination of state-space model and Markov chain model to allow parameters to vary over time. Their parameter transition equation assumes a dependency between the current value and its previous value. Thus, the transition equation suggests gradual changes in parameter effects. In addition, our proposed method enables the automatic variable selection through the penalization in the parameter estimation. Furthermore, the Bayesian approach in McCormick *et al.* (2012) considered different models having the highest posterior probability at different times. The final obtained model coefficients are dynamic in the sense that they are not only dynamic within each candidate model, but also due to the model selection at different stages. In contrast, the output in our proposed method is a model with selected variables allowing dynamic effects.

3. Efficient model estimation

To estimate the parameter matrix $\boldsymbol{B}=(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_n)^T$ of size $n\times p$, we minimize the logistic regression loss function combined with the l_1-norm fused LASSO penalty and the l_2-norm group LASSO penalty. That is,

$$\underset{\boldsymbol{B}}{\text{minimize}}\; -l(\boldsymbol{B})+\gamma_1\sum_{j=1}^{p}\sum_{t=2}^{n}|\beta_{t,j}-\beta_{t-1,j}| + \gamma_2\sum_{j=1}^{p}\sqrt{\beta_{1,j}^2+\cdots+\beta_{n,j}^2} \tag{2}$$

with

$$l(\boldsymbol{B}) = \log\left\{\prod_{t=1}^{n}\left[p(\boldsymbol{x}_t)^{y_t}(1-p(\boldsymbol{x}_t))^{1-y_t}\right]\right\} = \sum_{t=1}^{n}\{y_t\boldsymbol{x}_t^T\boldsymbol{\beta}_t-\log(1+\exp(\boldsymbol{x}_t^T\boldsymbol{\beta}_t))\}, \tag{3}$$

where $\gamma_1\geq 0$ and $\gamma_2\geq 0$ are tuning parameters. Note that we implicitly assume the data points are independent in the log-likelihood function, which is commonly used in works such as Kolar *et al.* (2009), Gibberd and Nelson (2017), and Adhikari *et al.* (2019).

The objective function in (2) is convex, but the two penalties are not separable in $\boldsymbol{B}$. That is, both the associated l_1-norm fused LASSO penalty and the l_2-norm group LASSO penalty contain the parameter matrix $\boldsymbol{B}$. It is well-studied to optimize a convex objective function associated with either l_1-norm fused LASSO penalty or the l_2-norm group LASSO penalty. However, it is challenging to optimize the objective function directly with the presence of both the l_1-norm fused LASSO penalty and the l_2-norm group LASSO penalty. To address this challenge, we consider an algorithm based on ADMM (Boyd *et al.*, 2011). The ADMM method has been successfully implemented and applied to the generalized LASSO problems (Wahlberg *et al.*, 2012; Zhu, 2017). The main idea is to convert the objective function so that we can deal with the two penalties separately. To update variables in an alternative way, we rewrite the problem (2) in an equivalent form as

$$\begin{aligned}\underset{\boldsymbol{B}}{\text{minimize}}\;\; & l_{\gamma_1,\gamma_2}(\boldsymbol{B})\triangleq -l(\boldsymbol{B})+\gamma_1\sum_{j=1}^{p}||\boldsymbol{Z}_j^{(1)}||_1+\gamma_2\sum_{j=1}^{p}||\boldsymbol{Z}_j^{(2)}||_2,\\ & \text{subject to } F\boldsymbol{B}_j-\boldsymbol{Z}_j^{(1)}=0,\\ & \boldsymbol{B}_j-\boldsymbol{Z}_j^{(2)}=0, j=1,\ldots,p,\end{aligned} \tag{4}$$

where $||\cdot||_1$ is the l_1-norm, $||\cdot||_2$ is the l_2-norm, $\boldsymbol{Z}_j^{(1)}$ and $\boldsymbol{Z}_j^{(2)}$ are the jth columns in the matrix $\boldsymbol{Z}^{(1)}$ of size $(n-1)\times p$ and the matrix $\boldsymbol{Z}^{(2)}$ of size $n\times p$ respectively. $\boldsymbol{B}_j$ is the jth column in the matrix $\boldsymbol{B}$, and $\boldsymbol{F}$ is the first-order difference matrix of size $(n-1)\times n$, written as

$$\mathbf{F}=\begin{bmatrix}1&-1&0&0&0&0&\ldots&0&0&0\\0&1&-1&0&0&0&\ldots&0&0&0\\0&0&1&-1&0&0&\ldots&0&0&0\\&&&\ldots\ldots\ldots\ldots&&&&&&\\0&0&0&0&0&0&\ldots&0&1&-1\end{bmatrix}.$$

The optimization in (4) is different from the original problem in (2) due to the fact that the penalty terms now involve $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$, which are completely decoupled. Therefore, we can solve the optimization by alternating minimization of $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$. The augmented Lagrangian for the problem (4) is as follows:

$$\begin{aligned}l_{al}(\boldsymbol{B})&\triangleq l_{\gamma_1,\gamma_2}(\boldsymbol{B})+\sum_{j=1}^{p}\tilde{\lambda}_j^{(1),T}(F\boldsymbol{B}_j-\boldsymbol{Z}_j^{(1)})+\sum_{j=1}^{p}\tilde{\lambda}_j^{(2),T}(\boldsymbol{B}_j-\boldsymbol{Z}_j^{(2)})\\&\quad+\frac{\rho_1}{2}\sum_{j=1}^{p}||F\boldsymbol{B}_j-\boldsymbol{Z}_j^{(1)}||_2^2+\frac{\rho_2}{2}\sum_{j=1}^{p}||\boldsymbol{B}_j-\boldsymbol{Z}_j^{(2)}||_2^2\\&=l_{\gamma_1,\gamma_2}(\boldsymbol{B})+\frac{\rho_1}{2}\sum_{j=1}^{p}||F\boldsymbol{\beta}_j-\boldsymbol{Z}_j^{(1)}+\boldsymbol{u}_j^{(1)}||_2^2+\frac{\rho_2}{2}\sum_{j=1}^{p}||\boldsymbol{\beta}_j-\boldsymbol{Z}_j^{(2)}\\&\quad+\boldsymbol{u}_j^{(2)}||_2^2-\frac{\rho_1}{2}\sum_{j=1}^{p}||\boldsymbol{u}_j^{(1)}||_2^2-\frac{\rho_2}{2}\sum_{j=1}^{p}||\boldsymbol{u}_j^{(2)}||_2^2,\end{aligned}$$

where $\boldsymbol{\mu}_j^{(1)}=\frac{\tilde{\lambda}_j^{(1)}}{\rho_1}$ and $\boldsymbol{\mu}_j^{(2)}=\frac{\tilde{\lambda}_j^{(2)}}{\rho_2}$. Here, ρ_1 and ρ_2 are the augmented Lagrangian parameters for the l_2-norm group LASSO penalty and the l_1-norm fused LASSO penalty, respectively, and $\tilde{\lambda}_j^{(1)}$ and $\tilde{\lambda}_j^{(2)}$ are the Lagrangian multipliers.

Then we can obtain the iterative updating scheme as

$$\begin{aligned}\boldsymbol{\beta}^{k+1}&=\underset{\boldsymbol{B}}{\operatorname{argmin}}\left(-l(\boldsymbol{B})+\frac{\rho_1^k}{2}\sum_{j=1}^{p}||F\boldsymbol{\beta}_j-\boldsymbol{Z}_j^{(1)}+\boldsymbol{u}_j^{(1)}||_2^2\right.\\&\quad\left.+\frac{\rho_2^k}{2}\sum_{j=1}^{p}||\boldsymbol{\beta}_j-\boldsymbol{Z}_j^{(2)}+\boldsymbol{\mu}_j^{(2)}||_2^2\right),\\\boldsymbol{Z}_j^{(2),k+1}&=f_{ss}(\boldsymbol{B}_j^{k+1}+\boldsymbol{u}_j^{(2),k},\gamma_2/\rho_2^k),j=1,\ldots,p;\\\boldsymbol{Z}_j^{(1),k+1}&=f_{ss}(F\boldsymbol{B}_j^{k+1}+\boldsymbol{u}_j^{(1),k},\gamma_1/\rho_1^k),j=1,\ldots,p;\\\boldsymbol{u}_j^{(2),k+1}&=\boldsymbol{u}_j^{(2),k}+\boldsymbol{B}_j^{k+1}-\boldsymbol{Z}_j^{(2),k+1},j=1,\ldots,p;\\\boldsymbol{u}_j^{(1),k+1}&=\boldsymbol{u}_j^{(1),k}+F\boldsymbol{B}_j^{k+1}-\boldsymbol{Z}_j^{(1),k+1},j=1,\ldots,p,\end{aligned}\tag{5}$$

with the soft-shrinkage function f_{ss} given as

$$f_{ss}(a,b)=\operatorname{sign}(a)\max((|a|-b),0),$$

where ρ_1^k and ρ_2^k are the augmented Lagrangian parameters for the fused LASSO penalty and the group penalty at iteration k, respectively.

Given that updating the dual variables, $\boldsymbol{u}_j^{(1)}$ and $\boldsymbol{u}_j^{(2)}$, and the primal variables, $\boldsymbol{Z}_j^{(1)}$ and $\boldsymbol{Z}_j^{(2)}$, is straightforward, the efficiency of developed algorithm depends on the minimization of $\boldsymbol{B}$ in (5), which is essentially the minimization problem of the classic logistic regression with two additional quadratic terms. It is well known that no analytical solution exists for the classic logistic regression problem of $\boldsymbol{B}$. Thus, we apply the Newton–Raphson method to solve the minimization problem. Specifically, we approximate $l(\boldsymbol{B})$ with its second-order Taylor series as

$$\begin{aligned}l(\boldsymbol{B})\approx\sum_{t=1}^{n}\left[l(\boldsymbol{\beta}_t^{(0)})+\left[\frac{\partial l(\boldsymbol{\beta}_t)}{\partial\boldsymbol{\beta}_t}(\boldsymbol{\beta}_t^{(0)})\right]^T(\beta_t-\boldsymbol{\beta}_t^{(0)})\right.\\\left.+\frac{1}{2}(\beta_t-\boldsymbol{\beta}_t^{(0)})^T\frac{\partial^2 l(\boldsymbol{\beta}_t)}{\partial\boldsymbol{\beta}_t\partial^T\boldsymbol{\beta}_t}(\boldsymbol{\beta}_t^{(0)})(\beta_t-\boldsymbol{\beta}_t^{(0)})\right],\end{aligned}$$

where

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta}_t)}{\partial\boldsymbol{\beta}_t}(\boldsymbol{\beta}_t^{(0)})&=\boldsymbol{x}_t(y_t-p(\boldsymbol{x}_t)),\\p(\boldsymbol{x}_t)&=\frac{\exp(\boldsymbol{x}_t^T\boldsymbol{\beta}_t)}{1+\exp(\boldsymbol{x}_t^T\boldsymbol{\beta}_t)},\\\frac{\partial^2 l(\boldsymbol{\beta}_t)}{\partial\boldsymbol{\beta}_t\partial^T\boldsymbol{\beta}_t}(\boldsymbol{\beta}_t^{(0)})&=-\sum_{t=1}^{n}\boldsymbol{x}_t\boldsymbol{x}_t^Tp(\boldsymbol{x}_t)(1-p(\boldsymbol{x}_t)).\end{aligned}$$

We apply the standard ADMM stopping criterion based on primal and dual residuals, which are defined at iteration $k+1$ (Boyd *et al.*, 2011) as:

$$\begin{aligned}\boldsymbol{r}_j^{(2),k+1}&=\boldsymbol{B}_j^{k+1}-\boldsymbol{Z}_j^{(2),k+1},\\\boldsymbol{s}_j^{(2),k+1}&=\rho_{gr}^k(\boldsymbol{Z}_j^{(2),k+1}-\boldsymbol{Z}_j^{(2),k}),\\\boldsymbol{r}_j^{(1),k+1}&=F\boldsymbol{B}_j^{k+1}-\boldsymbol{Z}_j^{(1),k+1},\\\boldsymbol{s}_j^{(1),k+1}&=\rho_{fuse}^kF^T(\boldsymbol{Z}_j^{(1),k+1}-\boldsymbol{Z}_j^{(1),k}),\end{aligned}$$

where $\boldsymbol{r}_j^{(i)}$ and $\boldsymbol{s}_j^{(i)}$ are the jth columns in the $n\times p$ matrices $\boldsymbol{r}^{(i)}$ and $\boldsymbol{s}^{(i)}$, respectively, $i=1,2$. The suggested termination criterion is that the primal and dual residuals must be small as the ADMM algorithm proceeds, i.e.,

$$||\boldsymbol{r}_{vec}^{(i),k}||_2\leq\epsilon^{(i),pri}\text{ and }||\boldsymbol{s}_{vec}^{(i),k}||_2\leq\epsilon^{(i),dual},i=1,2$$

with

$$\begin{aligned}\epsilon^{(2),pri}&=\sqrt{np}\epsilon^{abs}+\epsilon^{rel}\max\left(||vec(\boldsymbol{B}^k)||_2,||\boldsymbol{Z}_{vec}^{(2),k}||_2\right),\\\epsilon^{(2),dual}&=\sqrt{np}\epsilon^{abs}+\epsilon^{rel}||vec(\rho_2^k\boldsymbol{u}^{(2),k})||_2,\\\epsilon^{(1),pri}&=\sqrt{(n-1)p}\epsilon^{abs}+\epsilon^{rel}\max\left(||vec(F\boldsymbol{B}^k)||_2,||\boldsymbol{Z}_{vec}^{(1),k}||_2\right),\\\epsilon^{(1),dual}&=\sqrt{np}\epsilon^{abs}+\epsilon^{rel}||vec(\rho_1^kF^T\boldsymbol{u}^{(1),k})||_2,\end{aligned}$$

where $\boldsymbol{Z}_{vec}^{(1)}=(\boldsymbol{Z}_1^{(1),T},\ldots,\boldsymbol{Z}_p^{(1),T},)^T$, $\boldsymbol{Z}_{vec}^{(2)}=(\boldsymbol{Z}_1^{(2),T},\ldots,\boldsymbol{Z}_p^{(2),T},)^T$, $\boldsymbol{r}_{vec}^{(1)}=(\boldsymbol{r}_1^{(1),T},\ldots,\boldsymbol{r}_p^{(1),T},)^T$, $\boldsymbol{r}_{vec}^{(2)}=(\boldsymbol{r}_1^{(2),T},\ldots,\boldsymbol{r}_p^{(2),T})^T$, $vec(\cdot)$ is an operator for the vectorization of a matrix, ϵ^{abs} is the absolute tolerance, and ϵ^{abs} is the relative tolerance.

Algorithm 1 summarizes the developed computational algorithm for parameter estimation of the proposed method.

Algorithm 1 An ADMM-based algorithm for parameter estimation

Input: $\boldsymbol{X}$, $\boldsymbol{y}$, γ_1 and γ_2
Initialize $\boldsymbol{B}$, $\boldsymbol{Z}^{(1)}$, $\boldsymbol{Z}^{(2)}$, and $\boldsymbol{u}^{(1)}$, $\boldsymbol{u}^{(2)} = \boldsymbol{0}$.
for iteration k **do**
  $\boldsymbol{B}$-update
    Compute the $\frac{\partial^2 \boldsymbol{L}}{\partial \boldsymbol{B}\partial \boldsymbol{B}^T}$ and $\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{B}}$.
    Find appropriate step size η_s by the backtracking line search strategy.
    $\boldsymbol{B}^{k+1} = \boldsymbol{B}^k + \eta_s \times [-\left(\frac{\partial^2 \boldsymbol{L}}{\partial \boldsymbol{B}\partial \boldsymbol{B}^T}\right)^{-1} \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{B}}]$.
  $\boldsymbol{Z}^{(2)}$- and $\boldsymbol{Z}^{(1)}$-update in (5).
  $\boldsymbol{U}^{(2)}$- and $\boldsymbol{U}^{(1)}$-update in (5).
  if $||\boldsymbol{r}||_2^{(g)} \leq \epsilon^{(g),pri}$ $||\boldsymbol{s}||_2^{(g)} \leq \epsilon^{(g),dual}$, g = 1,2 **then** stop.
  end if
end for
Return $\boldsymbol{Z}^{(2)}$.

We would like to point out that the convergence rate is highly dependent on the selection of the augmented Lagrangian parameters ρ_1 and ρ_2. Zhu (2017) proposed a strategy of varying penalties for updating the parameter ρ. That is,

$$\rho^{k+1} = \begin{cases} \eta\rho^k & \text{if } \|\boldsymbol{r}_{vec}^k\|_2/\epsilon^{pri} \geq \mu\|\boldsymbol{s}_{vec}^k\|_2/\epsilon^{dual}, \\ \eta^{-1}\rho^k & \text{if } \|\boldsymbol{s}_{vec}^k\|_2/\epsilon^{dual} \geq \mu\|\boldsymbol{r}_{vec}^k\|_2/\epsilon^{pri}, \\ \rho^k & \text{otherwise,} \end{cases}$$

where ρ denotes either ρ_1 or ρ_2, η and μ are suggested to be 2 and 10, respectively (Boyd *et al.*, 2011). This strategy aims to improve the algorithm convergence when primal and dual feasibilities are on different scales.

The ADMM algorithm has been demonstrated to be applicable to a wide variety of large-scale statistical estimation problems (Boyd *et al.*, 2011). Ye and Xie (2011) developed a split Bregman method, which is essentially equivalent to ADMM, for large-scale fused LASSO problems. In addition, the ADMM method is flexible to extend the fused LASSO problem to higher-order trend filtering problems (Ramdas and Tibshirani, 2016) or other types of LASSO penalties. Note that in the iterative estimation procedure of Algorithm 1, we use the the Newton–Raphson algorithm for the predictor parameters. To scale-up the proposed method for big data, one can consider other methods such as the Limited-memory BFGS (LBFGS), Newton with Conjugate Gradient (Newton-CG) and Coordinate Descent (CD) algorithms for more efficient computation.

We would like to remark that it will be very valuable to conduct the post-selection inference on the estimated parameters. In the literature, the post-selection inferences have been conducted mainly on the l_1 regularization. For example, Lee *et al.* (2016) derived a truncated normal distribution for the estimated coefficient under the LASSO. The approach proposed by Taylor and Tibshirani (2018) generalizes the post-selection framework in Lee *et al.* (2016) with p-values and confidence intervals. Although the approaches in Taylor and Tibshirani (2018) can be applied for the l_1-regularized logistic regression, it is not straightforward to extend their approach to dynamic logistic regression with both fused LASSO and group LASSO. Alternatively, one may use the non-parametric bootstrap method to estimate the distribution of estimated coefficients for inference, which is known to be computationally expensive.

3.1. Selection of regularization parameters

The regularization parameters, (γ_1, γ_2), are determined over a search grid using the structured K-fold cross-validation technique (Arnold and Tibshirani, 2016). That is, the data set is ordered (for example, by time) and divided into k folds such that every kth point is in the same fold. We train the proposed model on all data points except those in the kth fold, and compute the Mean Squared Cross-Validation (MSCV) error defined as

$$\text{MSCV} = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{n_k}\sum_{i\in F_k}(y_i - \hat{y}_{i,-k})^2\right],$$

where n_k is the number of data points in the kth fold F_k and $\hat{y}_{i,-k}$ is the prediction at the ith point. In particular, $\hat{y}_{i,-k}$ is obtained as the weighted average of the fitted values at the i-1 and $i+1$ positions in this study. In our numerical studies, we use $K = 5$ for cross validation.

3.2. Convergence property of the estimator

Note that the proposed estimation procedure is within the framework of ADMM (Boyd *et al.*, 2011). Under the regularized dynamic logistic models, the objective and the dual variable under the ADMM will converge to the optimal values when satisfying the following two conditions. First, the loss function in the logistic regression and the norm-based penalty terms are closed, proper, and convex. Second, the loss function in the logistic regression with penalty terms has a minima point. Clearly, the loss function in proposed method (i.e., the negative log-likelihood function) satisfies these two conditions and thus the following update

$$\boldsymbol{\beta}^{k+1} = \underset{\boldsymbol{B}}{\text{argmin}}\left(-l(\boldsymbol{B}) + \frac{\rho_1^k}{2}\sum_{j=1}^{p}||\boldsymbol{F}\boldsymbol{\beta}_j - \boldsymbol{Z}_j^{(1)} + \boldsymbol{u}_j^{(1)}||_2^2) + \frac{\rho_2^k}{2}\sum_{j=1}^{p}||\boldsymbol{\beta}_j - \boldsymbol{Z}_j^{(2)} + \boldsymbol{\mu}_j^{(2)}||_2^2\right)$$

converges in the sense that the primal objective function along the sequence of primal variable converges to the optimal value:

$$-l(\boldsymbol{B}^k) + \gamma_1\sum_{j=1}^{p}||F\boldsymbol{B}_j^k||_1 + \gamma_2\sum_{j=1}^{p}||\boldsymbol{B}_j^k||_2 \rightarrow \inf_{\boldsymbol{B}}\left\{-l(\boldsymbol{B}) + \gamma_1\sum_{j=1}^{p}||F\boldsymbol{B}_j||_1 + \gamma_2\sum_{j=1}^{p}||\boldsymbol{B}_j||_2\right\}.$$

Next, we can also establish the following properties to study the estimator from the proposed method.

Theorem 1. *Let $\hat{\boldsymbol{\beta}}_t$ be the estimator from the minimization of the penalized loss function in (2). Denote $H(\boldsymbol{\beta}_t^{(0)})$ to be the Hessian matrix of the negative log likelihood function in the logistic regression at $\boldsymbol{\beta}_t^{(0)}$. Here the logistic regression is in the form of $\log(p(\boldsymbol{x}_t)/(1-p(\boldsymbol{x}_t))) = \boldsymbol{x}_t^T\boldsymbol{\beta}_t$ with $\boldsymbol{\beta}_t = (\beta_{t,1},...,\beta_{t,p})^T$. We assume $\boldsymbol{x}_t$ is fixed and $x_{t,j}^2 \le 1$ for $t = 1,...,n$ and $j = 1,...,p$. Then the obtained estimator $\hat{\boldsymbol{\beta}}_t$ is consistent in probability to the true estimator value $\boldsymbol{\beta}_t^{(0)}$. That is,*

$$\frac{1}{n}\sum_{t=1}^{n}(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^{(0)})^T H(\boldsymbol{\beta}_t^{(0)})(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^{(0)}) \le 6\sigma\sqrt{\frac{2\log(enp/\delta)}{n^2}}\sum_{t=1}^{n}||\boldsymbol{\beta}_t^{(0)}||_1$$

with probability at least 1-δ. Here σ is the upper bound of $\sigma_t = \sqrt{p(\boldsymbol{x}_t)(1-p(\boldsymbol{x}_t))}$.

The estimator for the minimization of the penalized loss function is used so that we can apply the property of optimality. The Hessian matrix is used because we used the second-order Taylor expansion as its approximation. Typically one would like to consider $\boldsymbol{x}_t$ as fixed, and $x_{t,j}^2 \le 1$ so that one can have the upper bound in the standard maximal inequality for the Gaussian distribution. Such conditions are widely used in studying the properties of estimators from the penalized methods (Greenshtein, 2006; Candes and Plan, 2009). The detailed proof can be found in Appendix B.

4. Simulation

In this section, we conduct the simulation studies to evaluate the performance of the proposed regularized dynamic logistic regression model. The response, y_t, follows a Bernoulli distribution with the conditional probability $Pr(y_t = 1|\boldsymbol{x}_t)$. We assume that the underlying model is logit $(\boldsymbol{x}_t) = \boldsymbol{x}_t^T\boldsymbol{\beta}_t$. We fix the number of significant variables with a nonzero coefficient to five. The remaining insignificant variables are noise variables used for high-dimensional settings. In the simulation, we consider two scenarios for generating predictor variables $\boldsymbol{X}_j$. In scenario 1 (S1), the predictor matrix $\boldsymbol{X}$ follows a multivariate normal distribution $N(\boldsymbol{0}, \Sigma)$, where Σ is a covariance matrix of size $p \times p$. We take the absolute value of $\boldsymbol{X}_j$ such that the probability $Pr(y_t = 1|\boldsymbol{x}_t)$ is greater than 0.5 when the sign of coefficient $\boldsymbol{\beta}_t$ is positive. Note that $\boldsymbol{X}_j$ is independent over time. In scenario 2 (S2), the predictor variable $\boldsymbol{X}_j$ is an autocorrelated sequence over time. The instance $x_{t,j}$ is generated from the AR(1) model $x_{t+1,j} = 0.7x_{t,j} + w_{t,j}$, where $t = 1,...,n-1$, $x_{1,j} = 0$. and $\boldsymbol{W} = (w_{t,j})_{n\times p}$ follows a multivariate normal distribution $N(\boldsymbol{0}, \Sigma)$. Note that Σ in both scenarios are equal to $(\rho^{|i-j|})_{p\times p}$ with $\rho^{|i-j|}$ as the element in the ith row and jth column in Σ and the correlation parameter $\rho \ge 0$.

To conduct a comprehensive simulation study, we vary a number of settings, including the sample size, the correlations of predictor variables, and the patterns of coefficients over time. Specifically, we consider five different sample sizes, $n = 100$, 200, 300, 400, and 500; two different scenarios, S1 and S2; three different correlations, $\rho = 0$, 0.35, and 0.7; and four different cases in Figure 1 for the coefficients of significant variables: (a) piecewise constant coefficients over time segments, (b) smooth functional coefficients, (c) constant coefficients, and (d) a mix of smooth functional coefficients and constant coefficients. Specially, in case (a), the magnitudes of $\boldsymbol{\beta}_t$ follow a uniform distribution. The sign of $\boldsymbol{\beta}_t$ alternates between adjacent segments. In order to generate the piecewise constant $\boldsymbol{\beta}_t$, we first partition the time interval into segments of pre-defined lengths. Then, from the uniform distribution $U(0, 1)$ we sample values for $\beta_{t,j}$ in each segment. Clearly, the coefficient $\beta_{t,j}$ is piecewise constant over time. Note that the coefficients have alternating signs in adjacent segments. In case (b), the smooth functions for the five significant variables are $f(t) = \exp(-2t+1)$, $f(t) = 4t(1-t)$, $f(t) = 2\sin^2(2\pi t)$, $f(t) = 2\sin(2\pi t)$, and $f(t) = 2\cos(\pi t) + 1$. In case (c), the constant magnitudes of $\boldsymbol{\beta}_t$ follow a uniform distribution. The sign of $\boldsymbol{\beta}_t$ alternates between adjacent segments. In case (d), the first three of five significant variables are identical to those in case (b), and the remaining two significant variables have constant magnitudes following a uniform distribution and alternating signs. The number of variables p is fixed at 20. For each simulation setting, we perform 30 replications.

The proposed rDLR is compared with five benchmark models, which are: (i) Least Absolute Shrinkage and Selection Operator (LASSO), (ii) Multivariate Adaptive Regression Splines (MARS), (iii) Varying Coefficient Model with smoothing splines basis (VCM1), (iv) Varying Coefficient Model with a polynomial of degree ≤ 2 basis (VCM2), and (v) dynamic logistic regression with fused LASSO penalty (FUSE). These benchmark methods focus on different strategies for formulating coefficients within a varying coefficient model framework.

The LASSO is an l_1-norm regularized method that is widely used in high-dimensional problems where the standard linear regression fails (Tibshirani, 1996). The LASSO tends to produce an interpretable model with a certain sparsity structure. The LASSO solves the following optimization problem as follows:

$$\underset{\boldsymbol{\theta}\in R^p}{\text{minimize}} \; -l(\boldsymbol{X},\boldsymbol{\theta}) + \lambda||\boldsymbol{\theta}||_1,$$

where $l(\boldsymbol{X},\boldsymbol{\theta}) = \sum_{t=1}^{n} y_t\theta^T x_t - \log(1+e^{\theta^T x_t})$ is the log-likelihood function of the logistic regression with response being either zero or one, $\lambda \ge 0$ is the tuning parameter, and $||\boldsymbol{\theta}||_1$ is the l_1-norm penalty on the coefficient vector $\boldsymbol{\theta} = (\theta_1,...,\theta_p)^T$. Note that we consider the main effects only in the logistic regression model.

The MARS is a piecewise linear regression model and allows flexible fitting by the automatic selection of spline basis functions and knots (Friedman, 1991). The MARS solves the regression problem as follows:

$$\log\frac{p(\boldsymbol{x}_t)}{1-p(\boldsymbol{x}_t)} = \sum_{j=1}^{p}\sum_{m=1}^{M_j} B_{m,j}(\boldsymbol{X}_j)a_{m,j},$$

where $B_{m,j}(\boldsymbol{X}_j)$ is the basis function for the predictor $\boldsymbol{X}_j$, $a_{m,j}$ is the mth coefficient of the basis function

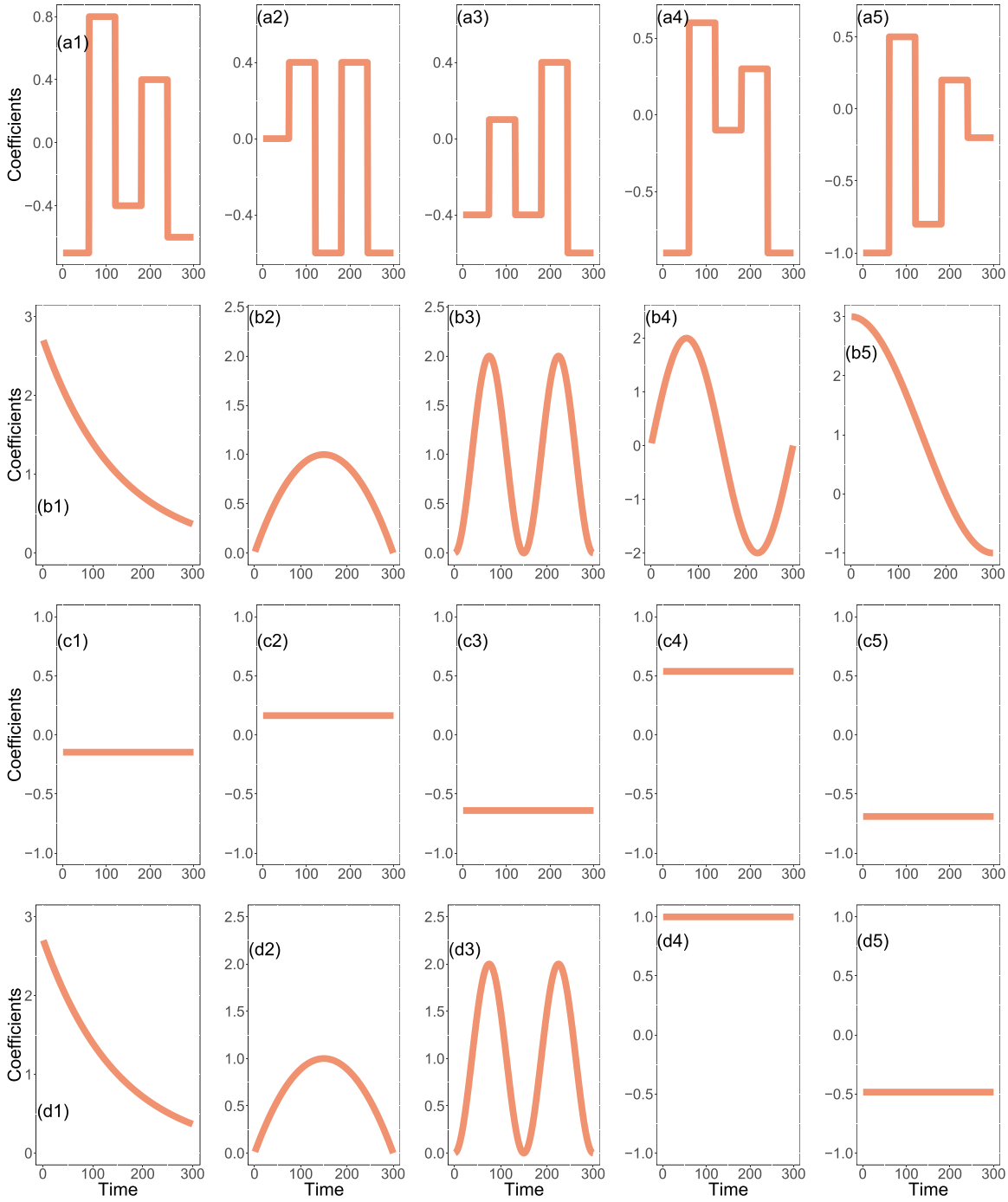

Figure 1. Illustrative plots for the simulated dynamic coefficients of the five significant variables when $n = 300$: (from top to bottom) the piecewise constant coefficients in case (a), the smooth functional coefficients in case (b), the constant coefficients in case (c), and a mix of smooth functional coefficients and constant coefficients in case (d).

$B_{m,j}(\boldsymbol{X}_j)$, and M_j is the number of knots determined for the predictor $\boldsymbol{X}_j$.

The VCM1 solves the regression problem as follows:

$$\log \frac{p(\boldsymbol{x}_t)}{1-p(\boldsymbol{x}_t)} = \sum_{j=1}^{p} f_j(t)\boldsymbol{x}_j, \;\; s.t. \int f''(u)^2 du \le \lambda,$$

where f is represented by the smoothing spline basis. In contrast with the VCM1, the VCM2 uses a polynomial basis with degree less than two and solves the regression problem as follows:

$$\log \frac{p(\boldsymbol{x}_t)}{1-p(\boldsymbol{x}_t)} = \sum_{j=1}^{p} f_j(t)\boldsymbol{x}_j.$$

The FUSE solves the optimization problem as follows:

$$\underset{\boldsymbol{B}}{\text{minimize}} \quad -l(\boldsymbol{B}) + \lambda \sum_{j=1}^{p} ||F\boldsymbol{B}_j||_1, j = 1, ..., p.$$

Compared to the proposed rDLR, the FUSE lacks the l_2-norm group penalty term.

To evaluate the performance of model prediction and parameter estimation, we employ two sets of evaluation metrics: prediction metrics and parameter estimation metrics. The prediction metrics are DViance (DEV) and Misclassification Error Rate (MER). The DEV and MER are computed in the 5-fold cross-validation. Specifically, the data set is divided into five folds, among which four folds are used to train the model, while the remaining portion is used to test the model and compute the metrics DEV and MER. Note that there are two types of Cross-Validation (CV) mechanisms used in this study: the normal random CV and the structured CV. The normal random CV divides the partitions by random sampling. The structured CV is described in detail in Section 3. We apply the normal random CV to the LASSO and MARS, and the structured CV to the rDLR, VCM1, VCM2 and FUSE.

The DEV measures the prediction performance and it is defined in terms of the difference between the negative log-likelihood of the model (nll_1) and the negative log-likelihood of the saturated model (nll_2):

$$\text{DEV} = \frac{2 \times (nll_1 - nll_2)}{m},$$

where m is the number of data points used to compute nll_1. The saturated model has a free parameter for each data point. The MER provides an overview of the method's accuracy. MER is defined as the proportion of incorrectly-classified data points to the total number of data points:

$$\text{MER} = 1 - \frac{\text{TP} + \text{TN}}{n},$$

where TP is the number of true positives and TN is the number of true negatives. The smaller the DEV and MER values, the higher the prediction accuracy.

The parameter estimation metrics are Performance Measurement (PM), correctly identified coefficient rate (CICR), number of Non-Zeros (NZ), and F_1 score. The PM evaluates the estimation accuracy of estimated $\hat{\beta}_{t,j}$ relative to the true $\beta^*_{t,j}$. The PM is defined as

$$\text{PM} = \frac{\sum_{t=1}^{n}\sum_{j=1}^{p}|\hat{\beta}_{t,j} - \beta^*_{t,j}|}{p\sum_{t=1}^{n}\sum_{j=1}^{p}|\beta^*_{t,j}|},$$

The smaller the PM, the higher the estimation accuracy of coefficient parameters. The CICR, evaluating the ability to identify correct variables, is defined as the ratio of the total number of correctly identified zero and non-zero coefficients to the total number of coefficients. The larger the CICR, the better the correct variables are identified. The NZ measures the number of estimated coefficients with non-zero values. For LASSO, the NZ is equal to the model size. For rDLR, VCM1, VCM2, and FUSE, the NZ is an approximation of the model size after normalization with respect to the number of data points. The F_1 score evaluates the ability of parameter identification. The F_1 score is defined as

$$F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}},$$

where FN is the number of false negatives and FP is the number of false positives. Note that all TP, TN, FP, and FN are normalized by the total number of data points in the methods rDLR and FUSE. The greater the F_1 score, the more accurate the identification of variables. The perfect F_1 score is one.

We compare the performance of the proposed rDLR method and benchmark methods in DEV (Table 1) and in MER (Table 2). The proposed rDLR has the smallest DEV and MER values in cases (a), (b), and (d) in both S1 and S2. The proposed rDLR considers dynamic impacts of predictor variables and can characterize coefficients in the form of piecewise constant functions, outperforming the LASSO and the MARS which assume constant coefficients. While the FUSE assumes predictors with dynamic coefficients as well, the lack of variable selection features negatively affects their performances. The variable selection functionality reduces model complexity and drops noisy variables, which might explain why the rDLR outperformed the VCM1 and VCM2. The rDLR has a similar performance in DEV and MER as the LASSO in case (c) where the coefficients are constant. The variable selection capability of the rDLR and the LASSO contributes to their superior performance in comparison with the other benchmarks that do not perform variable selection. The sample size can affect the performance of the rDLR in DEV and MER. The larger the sample sizes, the smaller the DEV and MER values. The simulation results for the other settings are quite similar to those in Table 1 and Table 2, and are available in the supplemental materials.

Figure 2 compares the coefficient-estimation performance of methods when the correlation ρ is 0.35, the scenario is S2, and the dimension p is 20. In all cases, the rDLR and the LASSO have the smallest PM values, suggesting the highest accuracy in coefficient estimation. The VCM1 and VCM2 have the worst PM performance. The rDLR has the highest F_1 values in case (a) while the LASSO has the highest F_1 values in cases (b), (c) and (d). The FUSE, VCM1, and VCM2 have constant F_1 values because these methods selected all variables. The rDLR has smaller CICR values than the LASSO, which is primarily because the rDLR selects more variables than the LASSO, indicated by the nonzero values. Note that the method MARS is not included because it does not compute the estimates of coefficients for the original variables.

5. Case studies

In this section, we evaluate the model performance of the proposed rDLR model by analyzing two real case problems: the crystal growth manufacturing (Jin *et al.*, 2019) and the Hong Kong environmental study (Fan and Chen, 1999; Cai *et al.*, 2000).

5.1. Crystal growth manufacturing

The quality of the silicon ingot produced from crystal growth manufacturing is fundamental to its downstream

Table 1. Performance comparisons of models in terms of deviance (DEV) in the simulation study when the correlation ρ is 0.35 and the number of variables p is 20 for two scenarios, S1 and S2, at four different cases of underlying functional coefficients, (a), (b), (c), (d), from 30 simulation replications (mean and standard errors (in parenthesis)).

		S1				*S2*			
Method	*n*	*(a)*	*(b)*	*(c)*	*(d)*	*(a)*	*(b)*	*(c)*	*(d)*
rDLR	100	1.36 (0.04)	**1.05** (0.11)	**1.27** (0.12)	**1.06** (0.10)	**1.30** (0.07)	**0.97** (0.15)	**1.07** (0.14)	**1.01** (0.13)
	200	**1.28** (0.06)	**0.96** (0.08)	**1.26** (0.09)	**0.99** (0.08)	**1.23** (0.07)	**0.83** (0.09)	**1.04** (0.14)	**0.94** (0.09)
	300	**1.25** (0.06)	**0.90** (0.07)	**1.26** (0.07)	**0.96** (0.07)	**1.20** (0.08)	**0.79** (0.08)	**1.05** (0.15)	**0.91** (0.07)
	400	**1.22** (0.08)	**0.88** (0.04)	**1.21** (0.07)	**0.97** (0.06)	**1.18** (0.07)	**0.78** (0.06)	**1.02** (0.09)	**0.92** (0.06)
	500	**1.21** (0.06)	**0.87** (0.04)	**1.22** (0.07)	**0.95** (0.05)	**1.15** (0.06)	**0.78** (0.05)	**1.01** (0.11)	**0.86** (0.08)
FUSE	100	1.91 (0.17)	1.78 (0.26)	1.94 (0.33)	2.69 (1.81)	1.86 (0.24)	1.97 (0.91)	2.33 (1.13)	2.39 (1.18)
	200	1.52 (0.07)	1.23 (0.15)	1.39 (0.09)	1.17 (0.09)	1.44 (0.08)	1.21 (0.14)	1.16 (0.15)	1.09 (0.09)
	300	1.43 (0.05)	1.11 (0.19)	1.32 (0.08)	1.04 (0.07)	1.38 (0.08)	1.13 (0.25)	1.11 (0.14)	1.00 (0.09)
	400	1.37 (0.08)	1.10 (0.23)	1.25 (0.07)	1.04 (0.07)	1.30 (0.07)	1.07 (0.28)	1.06 (0.09)	0.99 (0.05)
	500	1.34 (0.07)	0.95 (0.15)	1.24 (0.07)	1.00 (0.06)	1.23 (0.07)	0.96 (0.25)	1.04 (0.11)	0.93 (0.08)
LASSO	100	1.40 (0.04)	1.26 (0.09)	1.28 (0.12)	1.10 (0.10)	1.36 (0.05)	1.14 (0.13)	1.10 (0.14)	1.05 (0.13)
	200	1.39 (0.03)	1.20 (0.07)	**1.26** (0.09)	1.01 (0.08)	1.36 (0.04)	1.13 (0.09)	1.06 (0.14)	0.97 (0.09)
	300	1.38 (0.02)	1.18 (0.06)	**1.26** (0.07)	1.00 (0.06)	1.37 (0.02)	1.13 (0.06)	1.06 (0.14)	0.97 (0.09)
	400	1.38 (0.01)	1.17 (0.05)	**1.21** (0.08)	1.01 (0.06)	1.35 (0.03)	1.15 (0.06)	1.03 (0.08)	0.98 (0.07)
	500	1.38 (0.02)	1.18 (0.03)	**1.22** (0.06)	0.99 (0.06)	1.37 (0.02)	1.15 (0.07)	**1.01** (0.11)	0.94 (0.09)
MARS	100	3.15 (1.75)	4.00 (3.02)	3.56 (2.59)	3.80 (2.30)	4.15 (2.88)	5.71 (4.42)	3.98 (2.62)	4.36 (3.52)
	200	1.80 (0.20)	1.69 (0.32)	1.74 (0.25)	1.46 (0.29)	1.81 (0.21)	1.68 (0.30)	1.53 (0.26)	1.45 (0.28)
	300	1.70 (0.13)	1.45 (0.11)	1.49 (0.15)	1.26 (0.14)	1.62 (0.13)	1.40 (0.14)	1.29 (0.16)	1.20 (0.16)
	400	1.56 (0.08)	1.34 (0.09)	1.36 (0.09)	1.17 (0.09)	1.53 (0.13)	1.36 (0.12)	1.21 (0.12)	1.15 (0.10)
	500	1.50 (0.05)	1.32 (0.11)	1.35 (0.10)	1.12 (0.10)	1.50 (0.05)	1.28 (0.09)	1.16 (0.13)	1.09 (0.13)
VCM1	100	2.38 (0.32)	2.14 (0.42)	2.41 (0.38)	2.34 (0.40)	2.06 (0.24)	1.87 (0.31)	2.16 (0.36)	2.06 (0.25)
	200	1.62 (0.11)	1.52 (0.13)	1.73 (0.11)	1.69 (0.09)	1.52 (0.13)	1.27 (0.15)	1.52 (0.12)	1.47 (0.11)
	300	1.49 (0.07)	1.33 (0.09)	1.58 (0.06)	1.55 (0.05)	1.41 (0.08)	1.18 (0.09)	1.51 (0.08)	1.40 (0.07)
	400	1.43 (0.05)	1.30 (0.06)	1.52 (0.04)	1.48 (0.04)	1.35 (0.05)	1.12 (0.08)	1.44 (0.04)	1.37 (0.06)
	500	1.40 (0.04)	1.26 (0.05)	1.49 (0.03)	1.46 (0.03)	1.31 (0.06)	1.10 (0.08)	1.40 (0.05)	1.33 (0.05)
VCM2	100	4.77 (1.11)	4.89 (1.57)	4.98 (1.06)	5.25 (1.25)	4.00 (0.96)	3.36 (1.57)	4.37 (1.26)	3.78 (1.18)
	200	1.91 (0.13)	1.92 (0.18)	2.02 (0.15)	2.00 (0.12)	1.81 (0.16)	1.69 (0.21)	1.80 (0.13)	1.78 (0.13)
	300	1.62 (0.08)	1.50 (0.12)	1.71 (0.08)	1.69 (0.07)	1.52 (0.09)	1.34 (0.10)	1.61 (0.07)	1.53 (0.08)
	400	1.52 (0.05)	1.40 (0.07)	1.59 (0.04)	1.57 (0.04)	1.44 (0.04)	1.23 (0.09)	1.50 (0.04)	1.44 (0.06)
	500	1.48 (0.04)	1.33 (0.05)	1.54 (0.04)	1.52 (0.04)	1.38 (0.06)	1.18 (0.08)	1.45 (0.04)	1.40 (0.04)

Table 2. Performance comparisons of models in terms of misclassification error rate (MER) in the simulation study when the correlation ρ is 0.35 and the number of variables p is 20 for two scenarios, S1 and S2, at four different cases of underlying functional coefficients, (a), (b), (c), (d), from 30 simulation replications (mean and standard errors (in parenthesis)).

		S1				*S2*			
Method	*n*	*(a)*	*(b)*	*(c)*	*(d)*	*(a)*	*(b)*	*(c)*	*(d)*
rDLR	100	**0.38** (0.04)	**0.24** (0.04)	**0.33** (0.06)	**0.22** (0.04)	**0.33** (0.04)	**0.21** (0.05)	**0.25** (0.05)	**0.21** (0.05)
	200	**0.34** (0.05)	**0.22** (0.03)	**0.34** (0.05)	**0.22** (0.03)	**0.32** (0.04)	**0.18** (0.03)	**0.25** (0.05)	**0.21** (0.03)
	300	**0.34** (0.04)	**0.20** (0.02)	**0.34** (0.04)	**0.21** (0.02)	**0.31** (0.03)	**0.17** (0.03)	**0.26** (0.06)	**0.20** (0.02)
	400	**0.33** (0.04)	**0.20** (0.02)	**0.32** (0.04)	**0.22** (0.02)	**0.30** (0.03)	**0.17** (0.02)	**0.25** (0.03)	**0.21** (0.02)
	500	**0.31** (0.04)	**0.19** (0.02)	**0.32** (0.04)	**0.22** (0.02)	**0.29** (0.03)	**0.17** (0.02)	**0.24** (0.04)	**0.19** (0.02)
FUSE	100	0.48 (0.06)	0.37 (0.04)	0.38 (0.06)	0.30 (0.05)	0.42 (0.07)	0.30 (0.06)	0.31 (0.06)	0.27 (0.05)
	200	0.45 (0.05)	0.34 (0.10)	0.37 (0.04)	0.27 (0.04)	0.39 (0.06)	0.35 (0.12)	0.28 (0.06)	0.25 (0.04)
	300	0.44 (0.04)	0.31 (0.13)	0.36 (0.05)	0.24 (0.02)	0.39 (0.06)	0.35 (0.16)	0.28 (0.06)	0.24 (0.03)
	400	0.41 (0.05)	0.33 (0.14)	0.33 (0.04)	0.24 (0.02)	0.36 (0.04)	0.32 (0.15)	0.26 (0.03)	0.23 (0.02)
	500	0.41 (0.05)	0.24 (0.08)	0.33 (0.04)	0.24 (0.02)	0.33 (0.03)	0.27 (0.16)	0.26 (0.04)	0.21 (0.03)
LASSO	100	0.45 (0.05)	0.33 (0.05)	**0.33** (0.06)	0.25 (0.04)	0.38 (0.05)	0.27 (0.05)	0.27 (0.06)	0.24 (0.04)
	200	0.45 (0.05)	0.30 (0.03)	**0.34** (0.05)	0.23 (0.03)	0.41 (0.04)	0.28 (0.04)	**0.25** (0.06)	0.23 (0.03)
	300	0.45 (0.03)	0.31 (0.03)	**0.34** (0.04)	0.23 (0.02)	0.41 (0.04)	0.28 (0.03)	**0.26** (0.06)	0.23 (0.03)
	400	0.45 (0.03)	0.30 (0.03)	**0.32** (0.04)	0.23 (0.02)	0.42 (0.03)	0.29 (0.03)	**0.25** (0.03)	0.23 (0.03)
	500	0.46 (0.03)	0.30 (0.02)	**0.32** (0.03)	0.23 (0.02)	0.43 (0.03)	0.30 (0.03)	**0.24** (0.04)	0.21 (0.03)
MARS	100	0.51 (0.06)	0.42 (0.05)	0.41 (0.06)	0.34 (0.07)	0.47 (0.06)	0.35 (0.08)	0.35 (0.08)	0.30 (0.06)
	200	0.50 (0.05)	0.36 (0.05)	0.39 (0.06)	0.29 (0.05)	0.44 (0.05)	0.32 (0.05)	0.32 (0.05)	0.28 (0.04)
	300	0.49 (0.03)	0.35 (0.04)	0.39 (0.05)	0.27 (0.02)	0.44 (0.04)	0.32 (0.03)	0.30 (0.06)	0.25 (0.04)
	400	0.49 (0.03)	0.33 (0.04)	0.36 (0.04)	0.27 (0.03)	0.44 (0.03)	0.32 (0.03)	0.28 (0.04)	0.26 (0.02)
	500	0.48 (0.04)	0.33 (0.02)	0.35 (0.04)	0.26 (0.02)	0.45 (0.03)	0.32 (0.03)	0.27 (0.05)	0.24 (0.03)
VCM1	100	0.49 (0.06)	0.34 (0.07)	0.49 (0.07)	0.44 (0.06)	0.40 (0.06)	0.27 (0.07)	0.39 (0.08)	0.36 (0.06)
	200	0.43 (0.03)	0.35 (0.04)	0.49 (0.04)	0.46 (0.04)	0.38 (0.05)	0.26 (0.04)	0.38 (0.06)	0.34 (0.04)
	300	0.43 (0.03)	0.33 (0.04)	0.49 (0.04)	0.45 (0.03)	0.37 (0.04)	0.27 (0.03)	0.42 (0.04)	0.36 (0.04)
	400	0.42 (0.03)	0.33 (0.02)	0.49 (0.03)	0.44 (0.03)	0.36 (0.03)	0.26 (0.03)	0.41 (0.04)	0.36 (0.04)
	500	0.42 (0.03)	0.32 (0.02)	0.48 (0.03)	0.45 (0.03)	0.35 (0.03)	0.26 (0.03)	0.41 (0.04)	0.36 (0.03)
VCM2	100	0.49 (0.08)	0.36 (0.10)	0.48 (0.07)	0.46 (0.08)	0.41 (0.08)	0.26 (0.10)	0.39 (0.11)	0.35 (0.10)
	200	0.45 (0.04)	0.38 (0.04)	0.50 (0.04)	0.47 (0.04)	0.42 (0.05)	0.29 (0.04)	0.41 (0.05)	0.38 (0.05)
	300	0.46 (0.03)	0.34 (0.04)	0.50 (0.04)	0.46 (0.03)	0.41 (0.04)	0.29 (0.04)	0.45 (0.04)	0.40 (0.04)
	400	0.45 (0.03)	0.35 (0.03)	0.50 (0.03)	0.46 (0.03)	0.41 (0.03)	0.28 (0.04)	0.44 (0.04)	0.39 (0.03)
	500	0.45 (0.04)	0.33 (0.02)	0.49 (0.03)	0.46 (0.04)	0.40 (0.03)	0.28 (0.03)	0.44 (0.04)	0.39 (0.03)

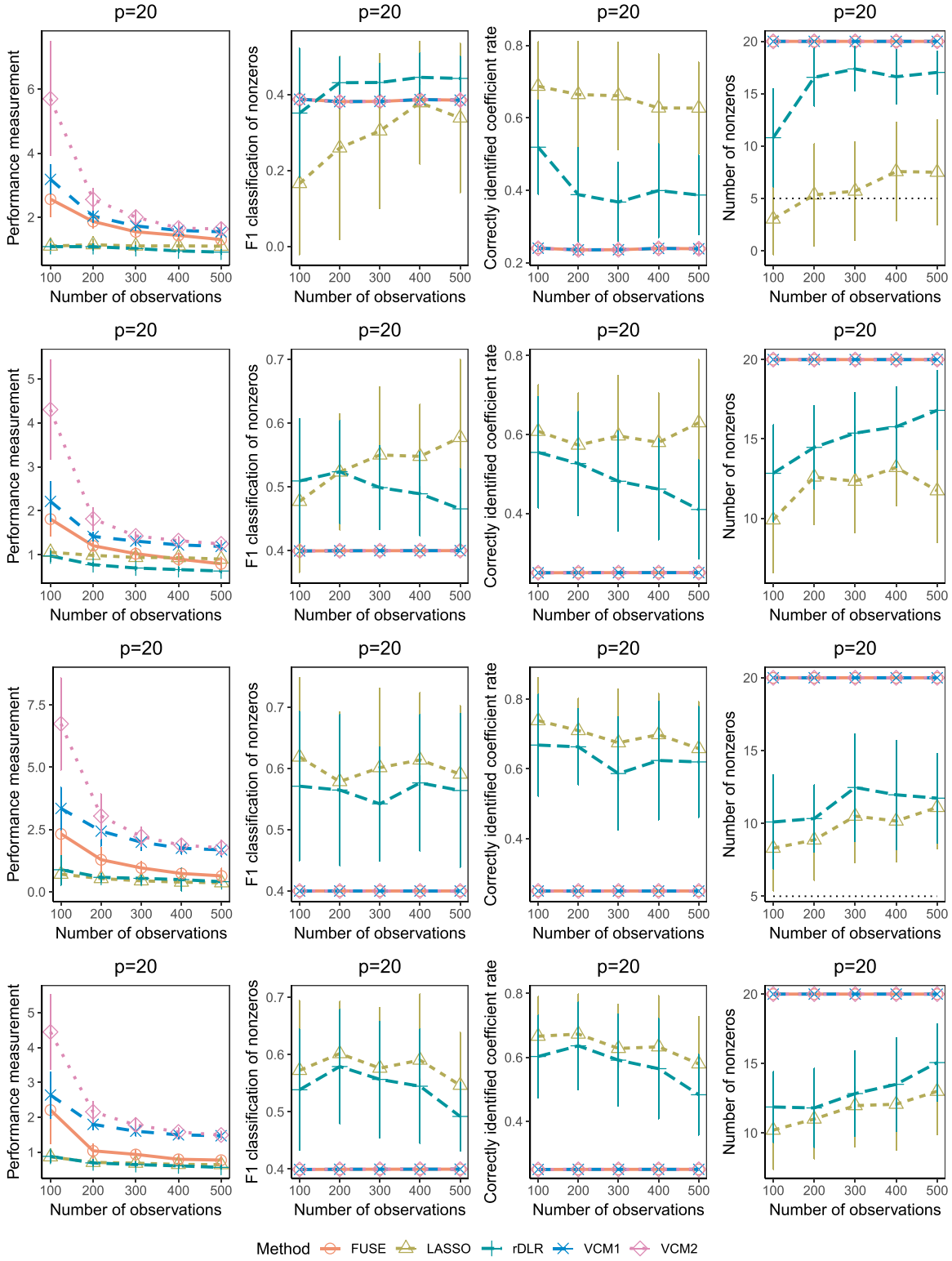

Figure 2. Performance comparison of models in the coefficient estimation for the simulation cases (from top to bottom) (a), (b), (c), and (d) when the correlation ρ is 0.35, the scenario is S2, and the number of variables p is 20.

products such as wafers and solar cells. The goal of this study is to identify key process variables and characterize the dynamic effects of those process variables on the binary quality response during the body growth stage in the crystal growth manufacturing. In this study, the binary response is defined as one when the continuous ingot diameter over time falls within the lower and upper specification limits and as zero otherwise. In total, there are 15 process variables labeled U1 through U15, including the power and the temperature of the heater, the pulling speed, and the rotation speed. The process variables are positive continuous and normalized between zero and one. The number of total data

points exceeds 1600. The process variables and response are measured and aligned at each time instance.

In Table 3, we compare the prediction performance of the rDLR, FUSE, VCM1, VCM2, and LASSO in terms of DEV, MER, and model size. The rDLR has the smallest DEV and MER values. This may be explained by the fact that the effects of important process variables are stage-wise during the manufacturing. The fused LASSO penalty in the rDLR allows the dynamic coefficients at different process stages. The piecewise constant formulation of dynamic coefficient effects in the rDLR improves its prediction performance in comparison with the LASSO, VCM1, and VCM2. The group LASSO penalty in the rDLR provides the variable selection functionality, resulting in a smaller model size and better prediction performance in comparison with the FUSE. For dynamic models, model size is defined as the number of predictor variables that have at least one nonzero estimated coefficient at any data point.

Table 3. Prediction performance of models in the crystal growth manufacturing.

	DEV	*sd*	*MER*	*sd*	*Size*
rDLR	**0.132**	0.011	**0.016**	0.003	14
FUSE	0.675	0.014	0.133	0.005	15
VCM1	0.599	0.011	0.108	0.003	15
VCM2	0.543	0.009	0.096	0.006	15
LASSO	0.651	0.035	0.133	0.011	14

Figure 3 shows the estimated coefficients over time for each process variable in the rDLR. It is evident that the estimated coefficients from the rDLR vary over time. At different time intervals, different sets of process variables contribute together to the crystal growth of the ingot. At the beginning, the variables U5, U7, and U11 have large estimated coefficients, but their effects diminish to zero as the growth progresses. The variables U1, U10, and U14, which

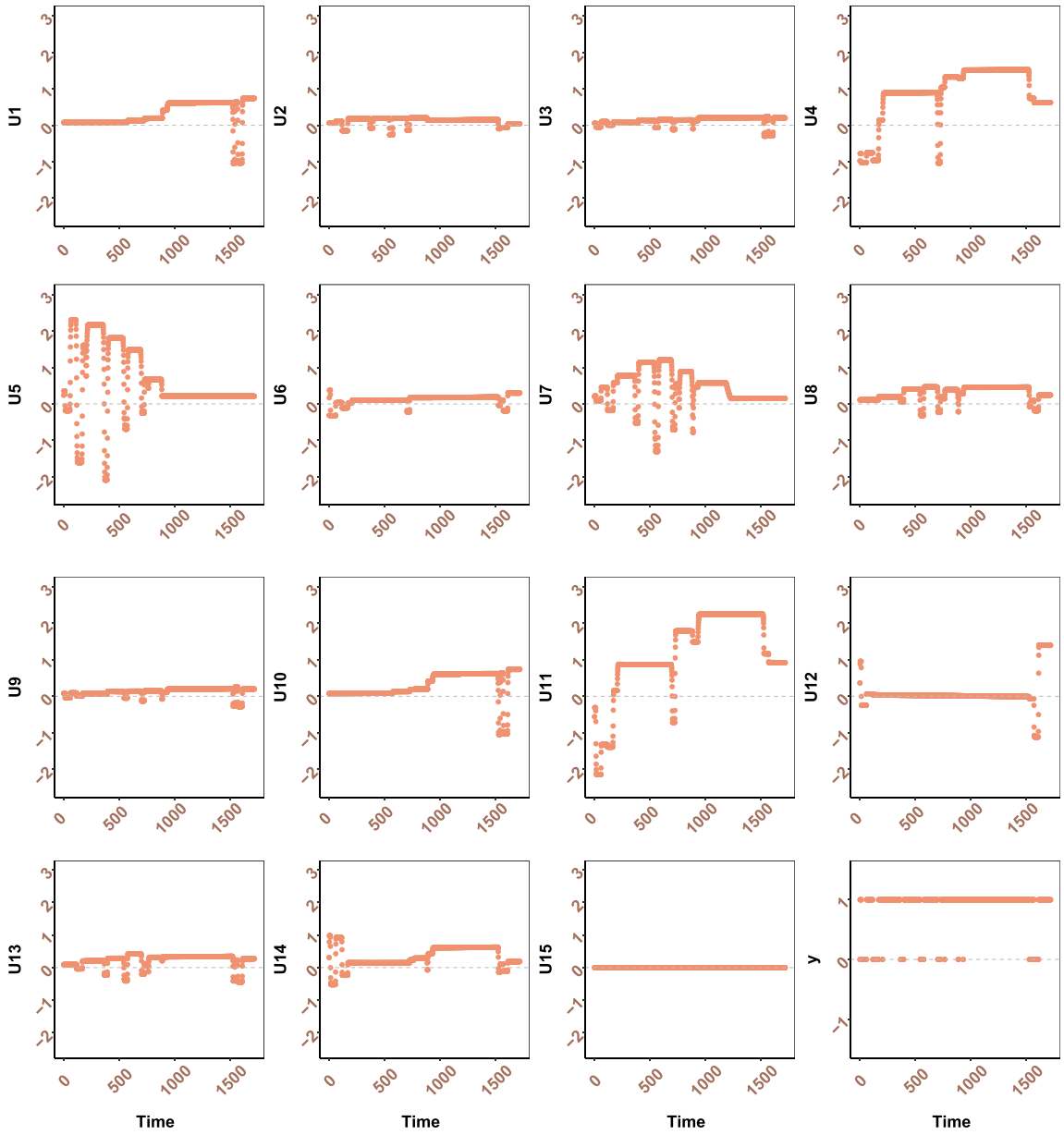

Figure 3. The estimated coefficients from the proposed regularized dynamic logistic regression model (rDLR) with fused LASSO penalty and group LASSO penalty and the response (bottom right) for the crystal growth manufacturing.

have small impacts at the beginning, have large coefficient estimators by the end. On the contrary, the variables U4 and U11 contribute to the crystal growth throughout the duration of the process. The varying impacts of process variables can be explained by different growth phases and the degradation of equipment (Jin *et al.*, 2019). Taking the process variable pulling speed as an example, from the engineering knowledge we learn that the faster the pulling speed, the greater the shrinkage effect on the ingot's diameter. At the early stage of the growth, the ingot is short and can be easily shrunk by increasing the pulling speed. However, at the late stage when the ingot is nearly fully grown, the effect of the pulling speed is limited and it becomes hard to effectively shrink the ingot's diameter. The estimated coefficients of the variable U15 are zero, indicating that the variable U15 is irrelevant to the response. In summary, the proposed rDLR method reveals the dynamic effects of the predictor variables with the estimated varying coefficients in the presence of irrelevant variables.

We also note that there are some estimated coefficients having spikes and back to a stable value. A possible explanation is because of the control of the process. There can be abrupt changes of the process conditions due to the stochastic nature of the crystal growth physical process. Process control is implemented to optimize the objectives to target and adapt to the stochastic process in manufacturing. Therefore, there can be frequent changes of the underlying model coefficients due to frequent control. During the transition stage of the changes, the effect of the model coefficients can be different from the stable condition.

Table 4. Performance comparison of models in the Hong Kong environmental study.

	DEV	*sd*	*MER*	*sd*	*Size*
rDLR	**1.237**	0.042	**0.319**	0.043	14
FUSE	1.345	0.048	0.385	0.027	14
VCM1	1.394	0.029	0.464	0.026	14
VCM2	1.401	0.047	0.462	0.028	14
LASSO	1.340	0.008	0.414	0.022	10

5.2. Hong Kong environmental study

The primary objective of the Hong Kong environmental study is to understand the relationship between the levels of pollutants and the number of daily hospital admissions for circulation and respiration problems. Cai *et al.* (2000) reported that the relationship between the number of hospital admissions and the pollutant levels varies over time. In this work, the response is binary and is defined as one when the number of hospital admission is greater than the median number of daily hospital admission during each whole calendar year and as zero otherwise. The predictors are levels of pollutants including the sulfur dioxide (SO_2) and nitrogen dioxide (NO_2) in Hong Kong between January 1, 1994, and December 31, 1995. Both the SO_2 and NO_2 are positive. In addition, 12 additional noise variables denoted as V1 through V12, are simulated and included in the data set. Six of these variables follow the normal distribution $N(0, 1)$ and the others follow the AR(1) model. The total number of data points is 730.

Table 4 compares the prediction performance of models including the rDLR, FUSE, VCM1, VCM2, and LASSO. The rDLR has the smallest DEV and MER values, suggesting that the piecewise constant formulation of dynamic variables impacts the model performance. Figure 4 shows the estimated coefficients over time for each variable in the rDLR. The rDLR is able to select the significant variables SO_2 and NO_2. Moreover, the rDLR identifies more than five abrupt

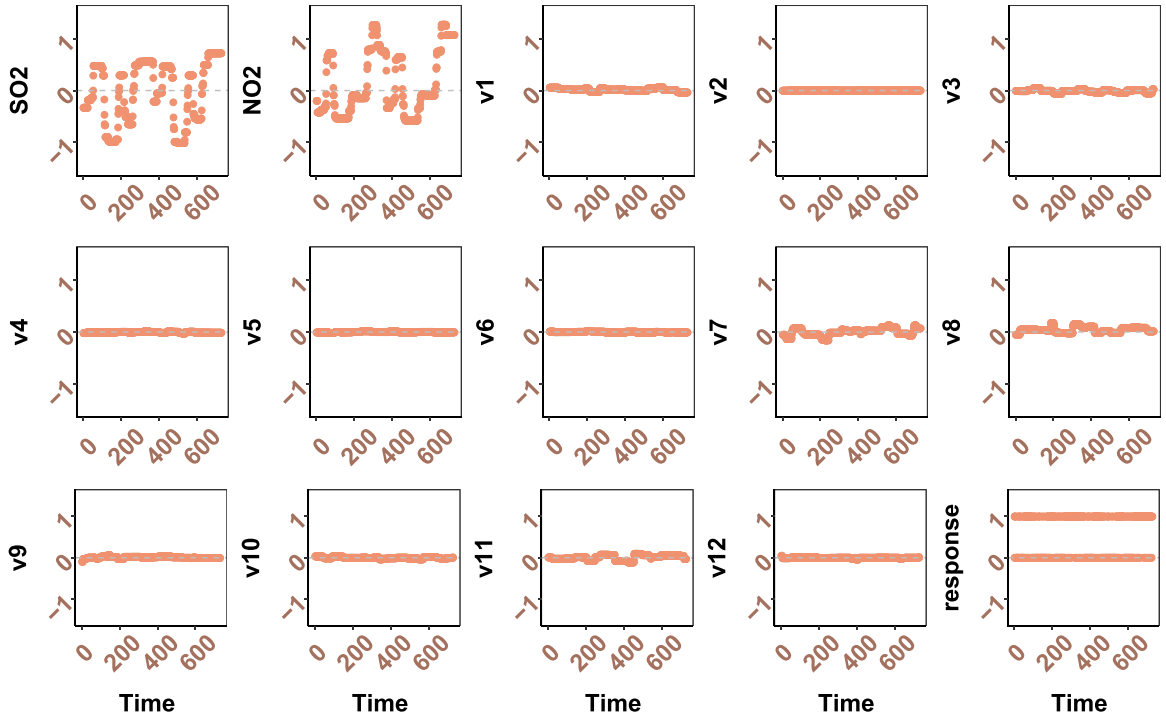

Figure 4. The estimated coefficients from the regularized dynamic logistic regression model with fused LASSO penalty and group LASSO penalty (rDLR) and the response (bottom right) for the Hong Kong environmental study.

changes in their structures of estimated dynamic effects, resulting in coefficient with opposite signs at different time windows. However, variables NO_2 and SO_2 concurrently have the same direction of coefficients. This is in agreement with the knowledge that both SO_2 and NO_2 are hazards to respiratory health. A closer examination of the estimated functional coefficient curves (Cai *et al.,* 2000) shows that the coefficient signs of NO_2 and SO_2 from their model are opposite at some time points. The noise variables are identified with their estimated coefficients.

6. Discussion

Variable selection plays a key role in extracting important predictor variables relevant to responses in high-dimensional statistical inference, but studies of variable selection are often conducted in the context of regression models. Motivated by crystal growth manufacturing, we propose a rDLR model in the framework of a varying coefficient model. The proposed method considers a combination of fused and group regularization to estimate varying effects of key predictors on responses in the presence of irrelevant variables. The fused LASSO encourages functional coefficients to be piecewise constant functions to approximate the dynamic coefficients. The group LASSO considers the coefficient parameters for one particular variable as an entire group. The proposed framework is extensible to accommodate other penalties, such as the generalized fused LASSO, the adaptive LASSO, and the zero-order fused LASSO penalty (Land and Friedman, 1997). The proposed method can also be extended to other types of the response (e.g., counting type) under the generalized linear models. A brief description of a regularized dynamic Poisson regression model can be found in the Appendix A.

It is possible to use some continuous function (such as natural cubic splines and smoothing splines) approximation methods to re-estimate parameters after the variable selection, especially when the domain knowledge suggest that continuous functions are preferred. For the future work, it will be interesting to investigate the spline-based methods combined with proper regularization for variable selection and accommodation of multiple changes in coefficient structures.

We also would like to remark that the computational bottleneck of the proposed ADMM algorithm is in estimating parameters in (5), which is essentially the minimization problem of the classic logistic regression with ridge-type regularization. For the large-scale problem, one can consider using recent techniques for fast logistic regression to enhance the scalability of the proposed algorithm. For example, the batching L-BFGS method (Bollapragada *et al.,* 2018) for fast computation in the logistic regression appears to be one promising direction. Another possibility is to consider the optimal subsampling strategy to handle logistic regression with large sample size (Wang *et al.,* 2018).

Note that the proposed model assumes that the response at any given time point is only affected by important variables at the current time. Responses are likely affected by past values of important variables as well. That is, both current and past values of important variables contribute to the current response values. Future work will focus on determining the order of lags in variables in the dynamic regression model. For variable selection, currently we consider all coefficient parameters of one variable as a group and use the group LASSO penalty. In the applications with unknown and changing groups, the proposed algorithm needs to be extended with more complex penalties (Chu *et al.,* 2021). One may also consider possible nonlinear effects in the model, which will make the variable selection more challenging.

Acknowledgments

The authors acknowledge the editor, associate editor, and reviewers for their insightful comments for us to revise this article.

Funding

Support from the National Science Foundation under contract CMMI-1435996 is gratefully acknowledged.

Notes on contributors

Dr. Sumin Shen obtained his PhD degree in statistics from Virginia Tech in 2019 under the supervision of Xinwei Deng. His research interests include variable selection, design and analysis of online experiments.

Dr. Zhiyang Zhang is an Instructor of Statistics at Virginia Tech. She obtained her PhD degree in chemistry in 2013 and MS degree in statistics in 2016 from Virginia Tech. Her research interests include data science pedagogy, design and analysis of online experiments, covariance matrix estimation, and portfolio optimization.

Dr. Ran Jin is an associate professor and the Director of Laboratory of Data Science and Visualization at the Grado Department of Industrial and Systems Engineering at Virginia Tech. He received his PhD degree in industrial engineering from Georgia Tech, Atlanta, his Master's degrees in industrial engineering, and in statistics, both from the University of Michigan, Ann Arbor, and his bachelor's degree in electronic engineering from Tsinghua University, Beijing. His research focuses on data fusion in smart manufacturing, computation services, and cognitive-based interactive visualization. He is currently serving as an associate editor for *IISE Transactions,* associate editor for *Journal of Manufacturing Science* and Engineering, and associate editor for *INFORMS Journal on Data Science.* He has been working with many leading manufacturing companies in aerospace, semiconductor, personal care, optical fiber industries.

Dr. Xinwei Deng is a Professor of Statistics and Data Science Faculty Fellow at Virginia Tech. He is also a co-director of VT Statistics and Artificial Intelligence Laboratory (VT-SAIL). He obtained his PhD degree from Georgia Institute of Technology in 2009 under the supervision of Jeff Wu and Ming Yuan. He joined in the statistics department at Virginia Tech in 2011.

ORCID

Ran Jin http://orcid.org/0000-0003-3847-4538
Xinwei Deng http://orcid.org/0000-0002-1560-2405

References

Adhikari, S., Lecci, F., Becker, J.T., Junker, B.W., Kuller, L.H., Lopez, O.L. and Tibshirani, R.J. (2019) High dimensional longitudinal classification with the multinomial fused lasso. *Statistics in Medicine*, **38**(12), 2184–2205.

Aguilar, O. and West, M. (2000) Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, **18**(3), 338–357.

Ahmed, A. and Xing, E.P. (2009) Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, **106**(29), 11878–11883.

Arnold, T. and Tibshirani, R. (2016) Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*, **25**(1), 1–27.

Beaulieu, C., Chen, J. and Sarmiento, J.L. (2012) Change-point analysis as a tool to detect abrupt climate variations. *Philosophical Transactions of the Royal Society A*, **370**(1962), 1228–1249.

Bollapragada, R., Nocedal, J., Mudigere, D., Shi, H. and Tang, P. (2018) A progressive batching L-BFGS method for machine learning, in *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, pp. 620–629.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, **3**(1), 1–122.

Breiman, L. (1995) Better subset regression using the nonnegative garrote. *Technometrics*, **37**(4), 373–384.

Cai, Z.W., Fan, J.Q. and Li, R.Z. (2000) Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, **95**(451), 888–902.

Candes, E.J. and Plan, Y. (2009) Near-ideal model selection by L1 minimization. *Annals of Statistics*, **37**(5A), 2145–2177.

Christoffersen, B. (2021) Dynamichazard: Dynamic hazard models using state space models. *Journal of Statistical Software*, **99**(7), 1–38.

Chu, S., Jiang, H., Xue, Z. and Deng, X. (2021) Adaptive convex clustering of generalized linear models with application in purchase likelihood prediction. *Technometrics*, **63**(2), 171–183.

Cleveland, W.S., Grosse, E. and Shyu, W.M. (2017) Local regression models, in *Statistical Models*, Routledge, New York, NY, pp. 309–376.

Fahrmeir, L. (1992) Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association*, **87**(418), 501–509.

Fan, J and Chen, J. (1999) One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**, 927–943.

Fan, J. and Zhang, W. (1999) Statistical estimation in varying coefficient models. *Annals of Statistics*, **27**(5), 1491–1518.

Fan, J. and Zhang, W. (2008) Statistical methods with varying coefficient models. *Statistics and its Interface*, **1**(1), 179–195.

Friedman, J. (1991) Multivariate adaptive regression splines. *Annals of Statistics*, **19**(1), 1–67.

Friedman, J., Hastie, T. and Tibshirani, T. (2010) A note on the group lasso and a sparse group lasso. *ArXiv Preprint*, *ArXiv:1001.0736*.

Gibberd, A.J. and Nelson, J.D. (2017) Regularized estimation of piecewise constant Gaussian graphical models: The group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, **26**(3), 623–634.

Greenshtein, E. (2006) Best subset selection, persistence in high-dimensional statistical learning and optimization under l1 constraint. *Annals of Statistics*, **34**, 2367–2386.

Gu, X., Stanley, D., Byrd, W.E., Dickens, B., Vaca-Trigo, I., Meeker, W.Q., Nguyen, T., Chin, J.W. and Martin, J.W. (2009) Linking accelerating laboratory test with outdoor performance results for a model epoxy coating system, in *Service Life Prediction of Polymeric Materials*, Springer, Boston, MA, pp. 3–28.

Hastie, T.J. and Tibshirani, R.J. (1993) Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, **55**(4), 757–796.

Hong, Y., Duan, Y., Meeker, W.K., Stanley, D.L. and Gu, X. (2015) Statistical methods for degradation data with dynamic covariates information and an application to outdoor weathering data. *Technometrics*, **57**(2), 180–193.

Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.P. (1998) Non-parametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**(4), 809–822.

Jin, R., Deng, X., Chen, X., Zhu, L and Zhang, J. (2019) Dynamic quality-process model in consideration of equipment degradation. *Journal of Quality Technology*, **51**, 217–229.

Kolar, M., Song, L. and Xing, E.P. (2009) Sparsistent learning of varying-coefficient models with structural changes. *Advances in Neural Information Processing Systems*, **22**, 1006–1014.

Land, S.R. and Friedman, J.H. (1997) Variable fusion: A new adaptive signal regression method. Technical Report 656, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

Lavielle, M. (2005) Using penalized contrasts for the change-point problem. *Signal Processing*, **85**(8), 1501–1510.

Lebarbier, E. (2005) Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, **85**(4), 717–736.

Lee, S.H., Yu, D., Bachman, A.H., Lim, J. and Ardekani, B.A. (2014) Application of fused lasso logistic regression to the study of corpus callosum thickness in early Alzheimer's disease. *Journal of Neuroscience Methods*, **221**, 78–84.

Lee, J., Sun, D., Sun, Y. and Taylor, J. (2016) Exact post-selection inference with the lasso. *Annals of Statistics*, **44**(3), 907–927.

McCormick, T.H., Raftery, A.E., Madigan, D. and Burd, R.S. (2012) Dynamic logistic regression and dynamic model averaging for binary classification. *Biometrics*, **68**(1), 23–30.

Meier, L., Van De Geer, S. and Buhlmann, P. (2008) The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(1), 53–71.

Ramdas, A. and Tibshirani, R. (2016) Fast and flexible ADMM algorithms for trend filtering. *Journal of Computational and Graphical Statistics*, **25**(3), 839–858.

Scheel, H.J. (2003) Theoretical and experimental solutions of the striation problem, in *Crystal Growth Technology*, Wiley, New York, NY, pp. 69–91.

Simon, N., Friedman, F., Hastie, T. and Tibshirani, R. (2013) A sparse-group lasso. *Journal of Computational and Graphical Statistics*, **22**(2), 231–245.

Sun, H., Deng, X., Wang, K and Jin, R. (2016) Logistic regression for crystal growth process modeling through hierarchical nonnegative garrote-based variable selection. *IIE Transactions*, **48**, 787–796.

Taylor, J. and Tibshirani, R.J. (2018) Post-selection inference for l_1-penalized likelihood models. *Canadian Journal of Statistics*, **46**(1), 41–61.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J and Knight, K. (2005) Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **67**, 91–108.

Wahlberg, B., Boyd, S., Annergren, M. and Wang, Y. (2012) An ADMM algorithm for a class of total variation regularized estimation problems. *IFAC Proceedings Volumes*, **45**(16), 83–88.

Wang, H., Zhu, R. and Ma, P. (2018) Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, **113**(522), 829–844.

Xiong, S. (2010) Some notes on the nonnegative garrote. *Technometrics*, **52**(3), 349–361.

Yao, Y. (1988) Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, **6**(3), 181–189.

Ye, G. and Xie, X. (2011) Split Bregman method for large scale fused lasso. *Computational Statistics & Data Analysis*, **55**(4), 1552–1569.

Yuan, M. and Lin, Y. (2007) On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(2), 143–161.

Zhang, J., Li, W., Wang, K and Jin, R. (2014) Process adjustment with an asymmetric quality loss function. *Journal of Manufacturing Systems*, **33**, 159–165.

Zhou, J., Liu, J., Narayan, V.A. and Ye, J. (2012) Modeling disease progression via fused sparse group lasso, in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery*, New York, NY, pp. 1095–1103.

Zhu, Y. (2017) An augmented ADMM algorithm with application to the generalized lasso problem. *Journal of Computational and Graphical Statistics*, **26**(1), 195–204.

Appendix

A. Brief description of regularized dynamic Poisson regression

The conditional probability $p(\boldsymbol{x}_t) = Pr(y_t|\boldsymbol{x}_t) = \exp(-\lambda_t)\lambda_t^{y_t}/y_t!$ with the Poisson regression model $\ln\lambda(x_t) = \boldsymbol{x}_t^T\boldsymbol{\beta}_t$ with $\boldsymbol{\beta}_t = (\beta_{t,1},...,\beta_{t,p})^T$.

A regularized dynamic Poisson regression model is expressed as

$$\begin{aligned}&\ln\lambda(x_t) = \boldsymbol{x}_t^T\boldsymbol{\beta}_t, t = 1,...,n\\ \text{s.t. }&\sum_{j=1}^{p}\sum_{t=2}^{n}|\beta_{t,j}-\beta_{t-1,j}| \le M_1,\\ &\sum_{j=1}^{p}\sqrt{\beta_{1,j}^2+\cdots+\beta_{n,j}^2} \le M_2,\end{aligned}$$

where $M_1 \ge 0$ and $M_2 \ge 0$ are the tuning parameters for the l_1-norm fused LASSO and l_2-norm group LASSO penalties, respectively.

To estimate the parameter matrix $\boldsymbol{B} = (\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_n)^T$ of size $n \times p$, we minimize the Poisson regression loss function combined with the l_1-norm fused LASSO penalty and the l_2-norm group LASSO penalty. That is,

$$\underset{\boldsymbol{B}}{\text{minimize}} -l(\boldsymbol{B}) + \gamma_1\sum_{j=1}^{p}\sum_{t=2}^{n}|\beta_{t,j}-\beta_{t-1,j}| + \gamma_2\sum_{j=1}^{p}\sqrt{\beta_{1,j}^2+\cdots+\beta_{n,j}^2}$$

with

$$\begin{aligned}l(\boldsymbol{B}) &= \log\left\{\prod_{t=1}^{n}\left[\exp(-\lambda_t)\lambda_t^{y_t}/y_t!\right]\right\}\\ &= \sum_{t=1}^{n}\{-\exp(x_t^T\boldsymbol{\beta}_t) + y_t\boldsymbol{x}_t^T\boldsymbol{\beta}_t - \log(y_t!)\},\end{aligned}$$

where $\gamma_1 \ge 0$ and $\gamma_2 \ge 0$ are tuning parameters. Similarly, we obtain the iterative updating scheme as

$$\begin{aligned}\boldsymbol{\beta}^{k+1} = \underset{\boldsymbol{B}}{\text{argmin}}\Bigg(&-l(\boldsymbol{B}) + \frac{\rho_1^k}{2}\sum_{j=1}^{p}||F\boldsymbol{\beta}_j - \boldsymbol{Z}_j^{(1)} + \boldsymbol{u}_j^{(1)}||_2^2)\\ &+\frac{\rho_2^k}{2}\sum_{j=1}^{p}||\boldsymbol{\beta}_j - \boldsymbol{Z}_j^{(2)} + \boldsymbol{\mu}_j^{(2)}||_2^2\Bigg),\\ \boldsymbol{Z}_j^{(2),k+1} &= f_{ss}(\boldsymbol{B}_j^{k+1} + \boldsymbol{u}_j^{(2),k}, \gamma_2/\rho_2^k), j = 1,...,p;\\ \boldsymbol{Z}_j^{(1),k+1} &= f_{ss}(F\boldsymbol{B}_j^{k+1} + \boldsymbol{u}_j^{(1),k}, \gamma_1/\rho_1^k), j = 1,...,p;\\ \boldsymbol{u}_j^{(2),k+1} &= \boldsymbol{u}_j^{(2),k} + \boldsymbol{B}_j^{k+1} - \boldsymbol{Z}_j^{(2),k+1}, j = 1,...,p;\\ \boldsymbol{u}_j^{(1),k+1} &= \boldsymbol{u}_j^{(1),k} + F\boldsymbol{B}_j^{k+1} - \boldsymbol{Z}_j^{(1),k+1}, j = 1,...,p,\end{aligned}$$

and we can apply the Newton–Raphson method to solve the minimization problem. Specifically, we approximate $l(\boldsymbol{B})$ with its second-order Taylor series as

$$\begin{aligned}l(\boldsymbol{B}) \approx \sum_{t=1}^{n}\Bigg[&l(\boldsymbol{\beta}_t^{(0)}) + \left[\frac{\partial l(\boldsymbol{\beta}_t)}{\partial\boldsymbol{\beta}_t}(\boldsymbol{\beta}_t^{(0)})\right]^T(\beta_t - \boldsymbol{\beta}_t^{(0)})\\ &+\frac{1}{2}(\beta_t - \boldsymbol{\beta}_t^{(0)})^T\frac{\partial^2 l(\boldsymbol{\beta}_t)}{\partial\boldsymbol{\beta}_t\partial^T\boldsymbol{\beta}_t}(\boldsymbol{\beta}_t^{(0)})(\beta_t - \boldsymbol{\beta}_t^{(0)})\Bigg],\end{aligned}$$

where

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta}_t)}{\partial\boldsymbol{\beta}_t}(\boldsymbol{\beta}_t^{(0)}) &= \boldsymbol{x}_t(y_t - \exp(\boldsymbol{x}_t^T\boldsymbol{\beta}_t)),\\ \frac{\partial^2 l(\boldsymbol{\beta}_t)}{\partial\boldsymbol{\beta}_t\partial^T\boldsymbol{\beta}_t}(\boldsymbol{\beta}_t^{(0)}) &= -\sum_{t=1}^{n}x_t x_t^T\exp(\boldsymbol{x}_t^T\boldsymbol{\beta}_t)).\end{aligned}$$

B. Proof of Theorem 1

Proof. Recall the objective function is

$$\min_{\boldsymbol{B}} l(\boldsymbol{B}) + h(\boldsymbol{B}),$$

where the negative log likelihood function $l(\boldsymbol{B})$, the penalty function $h(\boldsymbol{B})$, the first-order derivative ∇l, and the Hessian H are expressed as

$$\begin{aligned}l(\boldsymbol{B}) &= -\sum_{t=1}^{n}\{y_t\boldsymbol{x}_t^T\boldsymbol{\beta}_t - \log(1+\exp(\boldsymbol{x}_t^T\boldsymbol{\beta}_t))\},\\ h(\boldsymbol{B}) &= \gamma_1\sum_{j=1}^{p}\sum_{t=2}^{n}|\beta_{t,j}-\beta_{t-1,j}| + \gamma_2\sum_{j=1}^{p}\sqrt{\beta_{1,j}^2+\cdots+\beta_{n,j}^2},\\ \nabla l &= -\sum_{t=1}^{n}\boldsymbol{x}_t(y_t - p(\boldsymbol{x}_t)),\\ H &= \sum_{t=1}^{n}\boldsymbol{x}_t\boldsymbol{x}_t^T p(\boldsymbol{x}_t)(1-p(\boldsymbol{x}_t)).\end{aligned}$$

With the Taylor series expansion for the negative log-likelihood function around $\boldsymbol{B}^{(0)}$ (with the tth row as $\boldsymbol{\beta}_t^{(0)}$), and the nature of optimal estimator denoted as $\hat{\boldsymbol{B}}$ (with the tth row as $\hat{\boldsymbol{\beta}}_t$), any other arbitrary estimator denoted as $\hat{\boldsymbol{B}}^{(a)}$ (with the tth row as $\hat{\boldsymbol{\beta}}_t^{(a)}$), we have

$$\begin{aligned}&\sum_{t=1}^{n}\left(\left[l(\boldsymbol{\beta}_t^{(\boldsymbol{0})}) + (\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^{(\boldsymbol{0})})^T\nabla l(\boldsymbol{\beta}_t^{(\boldsymbol{0})}) + \frac{1}{2}(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^{(\boldsymbol{0})})^T H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^{(\boldsymbol{0})})\right] + h(\hat{\boldsymbol{B}})\right)\\ &\le \sum_{t=1}^{n}\left(\left[l(\boldsymbol{\beta}_t^{(\boldsymbol{0})}) + (\hat{\boldsymbol{\beta}}_t^{(a)} - \boldsymbol{\beta}_t^{(\boldsymbol{0})})^T\nabla l(\boldsymbol{\beta}_t^{(\boldsymbol{0})}) + \frac{1}{2}(\hat{\boldsymbol{\beta}}_t^{(a)} - \boldsymbol{\beta}_t^{(\boldsymbol{0})})^T H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})(\hat{\boldsymbol{\beta}}_t^{(a)} - \boldsymbol{\beta}_t^{(\boldsymbol{0})})\right] + h(\hat{\boldsymbol{B}}^{(a)}).\end{aligned}$$

That is,

$$\begin{aligned}&\sum_{t=1}^{n}(\nabla l(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^T\hat{\boldsymbol{\beta}}_t + \frac{1}{2}||H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\hat{\boldsymbol{\beta}}_t - H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\boldsymbol{\beta}_t^{(0)}||_2^2) + h(\hat{\boldsymbol{B}})\\ &\le\sum_{t=1}^{n}(\nabla l(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^T\hat{\boldsymbol{\beta}}_t^{(a)} + \frac{1}{2}||H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\hat{\boldsymbol{\beta}}_t^{(a)} - H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\boldsymbol{\beta}_t^{(0)}||_2^2) + h(\hat{\boldsymbol{B}}^{(a)}).\end{aligned}$$

Note that

$$\begin{aligned}&||H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\hat{\boldsymbol{\beta}}_t - H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\beta_t^{(0)}||_2^2\\ &= ||H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\hat{\boldsymbol{\beta}}_t - H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\hat{\boldsymbol{\beta}}_t^{(a)}||_2^2 + ||H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\hat{\boldsymbol{\beta}}_t^{(a)} - H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\beta_t^{(0)}||_2^2\\ &+2(H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\hat{\boldsymbol{\beta}}_t - H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\hat{\boldsymbol{\beta}}_t^{(a)})^T(H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\hat{\boldsymbol{\beta}}_t^{(a)} - H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\beta_t^{(0)}).\end{aligned}$$

After re-arrangements and algebra, we have

$$\sum_{t=1}^{n}\left(\frac{1}{2}||H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\hat{\boldsymbol{\beta}}_t - H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\hat{\boldsymbol{\beta}}_t^{(a)}||_2^2\right)$$
$$\leq \sum_{t=1}^{n}\left(\left[\nabla l(\boldsymbol{\beta}_t^{(0)}) + H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})(\hat{\boldsymbol{\beta}}_t^{(a)} - \boldsymbol{\beta}_t^{(0)})\right]^T\left(\hat{\boldsymbol{\beta}}_t^{(a)} - \hat{\boldsymbol{\beta}}_t\right)\right) + h(\hat{\boldsymbol{B}}^{(a)}) - h(\hat{\boldsymbol{B}}). \tag{A1}$$

From the dual norm inequality, any two vectors $\boldsymbol{w}$ and $\boldsymbol{\beta}$, we have $|\boldsymbol{w}^T\boldsymbol{\beta}| \leq ||\boldsymbol{w}||_\infty||\boldsymbol{\beta}||_1$. Then,

$$\left|\left[\nabla l(\boldsymbol{\beta}_t^{(\boldsymbol{0})}) + H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})(\hat{\boldsymbol{\beta}}_t^{(a)} - \boldsymbol{\beta}_t^{(0)})\right]^T(\hat{\boldsymbol{\beta}}_t^{(a)} - \hat{\boldsymbol{\beta}}_t)\right|$$
$$\leq \left|\left|\left[\nabla l(\boldsymbol{\beta}_t^{(\boldsymbol{0})}) + H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})(\hat{\boldsymbol{\beta}}_t^{(a)} - \boldsymbol{\beta}_t^{(0)})\right]\right|\right|_\infty||\hat{\boldsymbol{\beta}}_t^{(a)} - \hat{\boldsymbol{\beta}}_t||_1$$
$$= ||-\boldsymbol{x}_t(y_t - p(\boldsymbol{x}_t)) + H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})(\hat{\boldsymbol{\beta}}_t^{(a)} - \boldsymbol{\beta}_t^{(0)})||_\infty||\hat{\boldsymbol{\beta}}_t^{(a)} - \hat{\boldsymbol{\beta}}_t||_1.$$

By taking $\hat{\boldsymbol{\beta}}_t^{(a)} = \boldsymbol{\beta}_t^0$, one has

$$\left|\left[\nabla l(\boldsymbol{\beta}_{\boldsymbol{0}}) + H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})(\hat{\boldsymbol{\beta}}_t^{(a)} - \boldsymbol{\beta}_t^{(0)})\right]^T(\hat{\boldsymbol{\beta}}_t^{(a)} - \hat{\boldsymbol{\beta}}_t)\right|$$
$$\leq ||-\boldsymbol{x}_t(y_t - p(\boldsymbol{x}_t))||_\infty||\boldsymbol{\beta}_t^{(0)} - \hat{\boldsymbol{\beta}}_t||_1.$$

For the studentized residuals, we have the approximation $\frac{y_t - p(\boldsymbol{x}_t)}{\sigma_t} \sim N(0,1)$, and denote σ to be the upper bound of σ_t. With the standard maximal inequality for the Gaussian distribution, we have

$$||-\boldsymbol{x}_t(y_t - p(\boldsymbol{x}_t))||_\infty = \max_j|x_{t,j}(y_t - p(\boldsymbol{x}_t))| \leq \sigma\sqrt{2\log(enp/\delta)}$$

with probability at least 1-δ. Then we have

$$\left|\left[\nabla l(\boldsymbol{\beta}_{\boldsymbol{0}}) + H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})(\hat{\boldsymbol{\beta}}_t^{(a)} - \boldsymbol{\beta}_t^{(0)})\right]^T(\hat{\boldsymbol{\beta}}_t^{(a)} - \hat{\boldsymbol{\beta}}_t)\right|$$
$$\leq \sigma\sqrt{2\log(enp/\delta)}\ ||\boldsymbol{\beta}_t^{(0)} - \hat{\boldsymbol{\beta}}_t||_1$$

Thus in the inequality in (A1) becomes

$$\sum_{t=1}^{n}\left(\frac{1}{2}||H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\hat{\boldsymbol{\beta}}_t - H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\boldsymbol{\beta}_t^{(0)}||_2^2\right)$$
$$\leq \sum_{t=1}^{n}\sigma\sqrt{2\log(enp/\delta)}\ ||\boldsymbol{\beta}_t^{(0)} - \hat{\boldsymbol{\beta}}_t||_1 + \gamma_1\sum_{j=1}^{p}\sum_{t=2}^{n}|\beta_{t,j}^{(0)} - \beta_{t-1,j}^{(0)}|$$
$$+\gamma_2\sum_{j=1}^{p}\sqrt{\beta_{1,j}^{(0),2} + \cdots + \beta_{n,j}^{(0),2}}$$
$$-\left(\gamma_1\sum_{j=1}^{p}\sum_{t=2}^{n}|\hat{\beta}_{t,j} - \hat{\beta}_{t-1,j}| + \gamma_2\sum_{j=1}^{p}\sqrt{\hat{\beta}_{1,j}^2 + \cdots + \hat{\beta}_{n,j}^2}\right)$$
$$\leq \sum_{t=1}^{n}\sigma\sqrt{2\log(enp/\delta)}\ ||\boldsymbol{\beta}_t^{(0)} - \hat{\boldsymbol{\beta}}_t||_1 + \gamma_1'\sum_{t=1}^{n}(||\boldsymbol{\beta}_t^{(0)}||_1 - ||\hat{\boldsymbol{\beta}}_t||_1)$$
$$+\gamma_2\sum_{j=1}^{p}\sqrt{\beta_{1,j}^{(0),2} + \cdots + \beta_{n,j}^{(0),2}}$$
$$-\gamma_2\sum_{j=1}^{p}\sqrt{\hat{\beta}_{1,j}^2 + \cdots + \hat{\beta}_{n,j}^2}$$
$$\leq \sum_{t=1}^{n}\sigma\sqrt{2\log(enp/\delta)}\ ||\boldsymbol{\beta}_t^{(0)} - \hat{\boldsymbol{\beta}}_t||_1 + \gamma_1'\sum_{t=1}^{n}(||\boldsymbol{\beta}_t^{(0)}||_1 - ||\hat{\boldsymbol{\beta}}_t||_1)$$
$$+\gamma_2\sum_{t=1}^{n}||\boldsymbol{\beta}_t^{(0)}||_1.$$

By using the triangle inequality and set the tuning parameters γ_1' and γ_2 as $\sigma\sqrt{2\log(enp/\delta)}$, we then have

$$\frac{1}{n}\sum_{t=1}^{n}(||H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\hat{\boldsymbol{\beta}}_t - H(\boldsymbol{\beta}_t^{(\boldsymbol{0})})^{\frac{1}{2}}\boldsymbol{\beta}_t^{(0)}||_2^2) \leq 6\sigma\sqrt{\frac{2\log(enp/\delta)}{n^2}}\sum_{t=1}^{n}||\boldsymbol{\beta}_t^{(0)}||_1$$

with probability at least 1-δ. □