Statistica Sinica **32** (2022), 2023-2046
doi:https://doi.org/10.5705/ss.202020.0251

POISSON REGRESSION WITH ERROR CORRUPTED HIGH DIMENSIONAL FEATURES

Fei Jiang and Yanyuan Ma

The University of California, San Francisco and Pennsylvania State University

Abstract: Features extracted from aggregated data are often contaminated with errors. Errors in these features are usually difficult to handle, especially when the feature dimension is high. We construct an estimator of the feature effects in the context of a Poisson regression with a high dimensional feature and additive measurement errors. The procedure penalizes a target function that is specially designed to handle measurement errors. We perform optimization within a bounded region. Benefiting from the convexity of the constructed target function in this region, we establish the theoretical properties of the new estimator in terms of algorithmic convergence and statistical consistency. The numerical performance is demonstrated using simulation studies. We apply the method to analyze the possible effect of weather on the number of COVID-19 cases.

Key words and phrases: Composite gradient descent, COVID-19, non-convex optimization, Poisson regression, measurement error.

1. Introduction

Measurement errors frequently occur to features extracted from aggregated data sets, such as average temperatures from multiple sensors, owing to the loss of raw data information after the data aggregation. The measurement error issue for count outcome prediction has gained great attention in infectious disease studies where numerous data are collected to predict disease spread. For example, with the recent outbreak of the COVID-19 pandemic, there is some hope that the pandemic will ease when the weather becomes warmer. However, conclusions on the association between climate and COVID-19 infection are varied and controversial. For example, Tosepu et al. (2020) showed that temperature has a positive association with COVID-19 cases, whereas Jüni et al. (2020) showed that there is no significant association between climate and COVID-19 cases. Nevertheless, none of these studies considered the potential error contamination of the climate data. For example, weather components such as temperature and precipitation vary within a county, whereas the COVID-19 cases are usually summarized at

Corresponding author: Fei Jiang, School of Medicine, UCSF, CA 94158, USA. E-mail: fei.jiang @ucsf.edu.

the county-level. Therefore, a natural approach is to aggregate the covariates at different locations into a county level summary weather covariate, and then to study the relation between the number of COVID-19 cases and the weather in the previous several days to account for the virus incubation period. Thus, to study the relation between the number of COVID-19 cases and the weather, we face errors in the covariate due to the aggregation.

Although the issue of measurement error has been acknowledged, when the covariate dimension is high, it is handled only in linear models. In infection disease studies, counts are the most frequently collected outcomes, and the Poisson model is widely used to model these data. Hence, there is an urgent need to develop statistically valid methods to handle measurement error models in high-dimensional covariate Poisson models. The potential obstacles are as follows. (I) The Poisson regression function is nonlinear, and hence it is not straightforward to construct a legitimate loss function that approximates the error-free objective function or its second derivative. In a linear model $Y = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon$, where ϵ is a regression error independent of $\mathbf{X}$, the least squares estimator is the solution of minimizing the loss function $n^{-1}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})^2$. When $\mathbf{W} = \mathbf{X} + \mathbf{U}$ is observed instead of $\mathbf{X}$, under the assumption $\mathbf{U} \perp\!\!\!\perp \mathbf{X}$ and $\mathbf{U} \perp\!\!\!\perp \epsilon$, an approximation to the loss function is $n^{-1}\sum_{i=1}^{n}(Y_i - \mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})^2 - \boldsymbol{\beta}(n^{-1}\sum_{i=1}^{n}\mathbf{U}_i\mathbf{U}_i^{\mathrm{T}})\boldsymbol{\beta} \approx n^{-1}\sum_{i=1}^{n}(Y_i - \mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})^2 - \boldsymbol{\beta}^{\mathrm{T}}\mathrm{var}(\mathbf{U})\boldsymbol{\beta}$. Similarly, an approximation to the second derivative of the loss function is $\mathrm{var}(\mathbf{W}) - \mathrm{var}(\mathbf{U})$. Thus, we can use the new loss function $n^{-1}\sum_{i=1}^{n}(Y_i - \mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta})^2 - \boldsymbol{\beta}^{\mathrm{T}}\mathrm{var}(\mathbf{U})\boldsymbol{\beta}$ to obtain an estimator. However, in a Poisson model, the relation between the response Y and the covariate $\mathbf{X}$ is $\mathrm{pr}(Y = y \mid \mathbf{X}) = \exp(-e^{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}})e^{y\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}/y!$, for $y = 0, 1, \ldots$, which is nonlinear. A loglikelihood based loss function is $\sum_{i=1}^{n}(e^{\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}} - y_i\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})$. It is not obvious how to correct this loss function when $\mathbf{X}_i$ is replaced by $\mathbf{W}_i$, owing to the term $e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}$. (II) As derived in detail in Section 3, the second derivative of the loss function contains random quantities with heavy tailed distributions. Therefore the standard Poisson Lasso regression, which requires a bounded regression function (Shi et al. (2019)), cannot guarantee recovering the true parameters. In fact, the second derivative turns out to be $n^{-1}\sum_{i=1}^{n} e^{\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}}\mathrm{var}(\mathbf{U})\boldsymbol{\beta}}[\{\mathbf{W}_i\mathrm{var}(\mathbf{U})\boldsymbol{\beta}\}^{\otimes 2} - \mathrm{var}(\mathbf{U})\}$, which has a heavy tail because $\mathbf{W}$ is not bounded. (III) In a Poisson regression, the conditional mean of the outcome is $e^{\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}$. Thus, the conditional mean increases much faster than the linear predictor $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ and will easily explode, even if $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ is moderately large. This issue is more prominent in Poisson measurement error models, owing to the wider range of $\mathbf{W}_i$ than $\mathbf{X}_i$ and the term involving $\mathrm{var}(\mathbf{U})$. In fact, even when all $\mathbf{X}_i$ are zero, the term $\boldsymbol{\beta}^{\mathrm{T}}\mathrm{var}(\mathbf{U})\boldsymbol{\beta}$ can explode, which leads to a nearly singular second derivative. Hence, controlling the magnitude of the linear predictor and $\boldsymbol{\beta}$ is crucial for the algorithm to converge.

Therefore, to study count outcomes with error-corrupted covariates with an unknown distribution, we develop a novel optimization procedure to address the three complications raised from the high-dimensional Poisson measurement error models, and evaluate its theoretical guarantee. These techniques provide foundations on which to study the statistical and numerical properties of estimation procedures that involve nonlinear regression functions under high-dimensional measurement error settings. For problem (I), we construct an objective function, the expectation of which reduces to that in the canonical Poisson regression. Furthermore, this objective function is shown to nonconvex; hence we propose a restricted l_1 penalty procedure and use the composite gradient descent algoritm to handle the high-dimensional case under the sparsity assumption. For problem (II), we discover that, conditional on the error prone covariates, the first and second derivatives of the objective function can be sub-Gaussian and sub-exponential, respectively. Hence, we define the conditional sub-Gaussian and sub-exponential distributions, and derive their tail properties. Furthermore, we separate the response variable from the regression function, and the error contaminated covariates from the unobservable error-free covariates using these conditional arguments. Then we show that the conditional exponential decline implies the marginal exponential decline of the probability measures under weak conditions. For problem (III), we restrict the parameter searching space to an l_2 ball so that the regression function and the eigenvalues of the Hessian matrix do not explode in this set.

Benefiting from the conditional exponentially decayed tail and the constraints imposed by the feasible set, the Hessian matrix of the objective function is locally positive definite. This, with a specific choice of the penalty term, satisfies the restricted eigenvalue conditions introduced in Section 4, and eventually leads to the consistency of the estimators.

We rigorously establish the theoretical properties of the new estimator. This includes showing the statistical properties of the ideal optimizer in Section 4, and showing the algorithmic convergence of the numerical optimizer to the ideal optimizer in Section 5. We demonstrate the numerical performance of the new estimator using extensive simulations. In Section 6, we apply the algorithm to analyze a COVID19 data set. We conclude the paper in Section 7. Necessary conditions for the theoretical guarantees are presented in an Appendix. Owing to space limitations, we provide only intuitive explanations about the lemmas and theorems. Detailed proofs are included in the supplementary material.

2. Related Works

Measurement error models are notoriously difficult to handle. The only measurement error model that has received in-depth studies in both low and high-dimensional feature cases is the linear model. Methods in low-dimensional covariate case are well documented in Fuller (1987). Recently, many works have appeared for the high-dimensional case (Loh and Wainwright (2012); Belloni and Rosenbaum (2016); Datta and Zou (2017)) as well. However, to the best of our knowledge, no existing work handles high-dimensional features with measurement errors in nonlinear models. For example, when the response is count data, although consistent estimators exist separately for the low-dimensional case (Carroll et al. (2006)) and for the error-free case Negahban et al. (2009), no research considers both simultaneously.

In the measurement error-free case, many recommendation system algorithms have been proposed to explore the feature effects on outcomes, where linear (Srebro, Rennie and Jaakkola (2004); Van den Oord, Dieleman and Schrauwen (2013); Volkovs, Yu and Poutanen (2017)) and logistic (Dziugaite and Roy (2015); Wang et al. (2016); He et al. (2017b)) relations are often used to model the mean structure of the continuous and binary responses, respectively. To handle count outcomes, Poisson factorization models have been studied (Gopalan, Hofman and Blei (2013); Gopalan, Charlin and Blei (2014)). Here to avoid the difficulties related to nonlinear regression, the mean structure of the count response is assumed to be linear in the feature, while the coefficients are forced to be positive. This is quite awkward because not all features have positive effects, and the nature of the positive count data is basically ignored.

Low-dimensional measurement error models are studied extensively in the statistical community. Measurement error treatments of the Poisson model are either approximate, such as the regression calibration methods (Carroll et al. (2006)) and simulation extrapolation methods (Cook and Stefanski (1994)), or based on estimating equations, such as the efficient score and conditional score estimators (Stefanski and Carroll (1987)). Most treatments to high dimensional covariate models are based on penalizing a convex target function, the optimization of which in the low dimensional case yields a consistent estimator. This raises difficulties in transporting the existing estimators to the corresponding high-dimensional case for Poisson measurement error models, because we can no longer guarantee the convexity of the objective functions.

3. Model and Estimator

Let Y_i be a count random variable following a Poisson distribution, and let $E(Y_i \mid \mathbf{X}_i) = \exp(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)$, where $\boldsymbol{\beta}_0$ is the true covariate effect. Here, $\mathbf{X}_i$ is a p-dimensional covariate subject to measurement error, the distribution of which is unspecified. Let $\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i$, where $\mathbf{U}_i$ is a normally distributed measurement error with covariance matrix $\boldsymbol{\Omega}$. Following the common practice in the measurement error literature, we assume that $\boldsymbol{\Omega}$ is known or can be estimated externally. Without loss of generality, assume $E(\mathbf{X}_i) = \mathbf{0}$. We consider the case where $p >> n$, and assume that the p dimensional parameter $\boldsymbol{\beta}_0$ has at most k nonzero entries. Let $k << n$. The observations are written as $(\mathbf{W}_i, Y_i)$ for $i = 1, \ldots, n$.

When $\mathbf{X}_i$ is observable, $\boldsymbol{\beta}_0$ can be estimated easily using a maximum likelihood estimation (MLE), that is by minimizing $-n^{-1}\sum_{i=1}^{n}\{Y_i\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} - \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}$. Note that for a normally distributed error $\mathbf{U}_i$, it holds that

$$E\left\{\exp\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i - \frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\beta}}{2}\right)\Big| \mathbf{X}_i\right\} = \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i). \tag{3.1}$$

Taking advantage of these relations, we get

$$E\left\{Y_i\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta} - \exp\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i - \frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\beta}}{2}\right)\Big| \mathbf{X}_i, Y_i\right\} = Y_i\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} - \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i).$$

Motivated by this relation, to handle the high-dimensional covariate issue, we propose estimating $\boldsymbol{\beta}_0$ by solving the following constrained minimization problem:

$$\widehat{\boldsymbol{\beta}} = \underset{\|\boldsymbol{\beta}\|_1 \leqslant B_0\sqrt{K}, \|\boldsymbol{\beta}\|_2 \leqslant B_0}{\operatorname{argmin}} \left\{\mathcal{L}_1(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1\right\}, \tag{3.2}$$

at suitable constants B_0, K_0, where

$$\mathcal{L}_1(\boldsymbol{\beta}) = -n^{-1}\sum_{i=1}^{n}\left\{Y_i\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\beta} - \exp\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i - \frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\beta}}{2}\right)\right\}.$$

Note that $\|\boldsymbol{\beta}\|_1 \leqslant B_0\sqrt{K}$ is imposed to guarantee that the objective function satisfies the restricted eigenvalue condition defined in Section 4; $\|\boldsymbol{\beta}\|_2 \leqslant B_0$ is stressed to avoid the explosion of the regression function. If we could perform the optimization with the restriction $\|\boldsymbol{\beta}\|_0 \leqslant K$, then the constraint $\|\boldsymbol{\beta}\|_1 \leqslant B_0\sqrt{K}$ would be redundant, given the constraint $\|\boldsymbol{\beta}\|_2 \leqslant B_0$. However, using $\|\boldsymbol{\beta}\|_0 \leqslant K$ as an active constraint is infeasible computationally. The method (3.2) is closely linked to Loh and Wainwright (2012), with the difference that we have the additional constraint on the l_2 norm of $\boldsymbol{\beta}$.

To solve for $\widehat{\boldsymbol{\beta}}$ in (3.2), we first choose a large value B_0 that is guaranteed to satisfy $\|\boldsymbol{\beta}_0\|_2 \leqslant B_0$. We also set K to be a sufficiently large value. We then obtain $\widehat{\boldsymbol{\beta}}$ using the composite gradient descent method. Specifically, we update $\boldsymbol{\beta}$ recursively through

$$\boldsymbol{\beta}^{t+1} = \underset{\|\boldsymbol{\beta}\|_1 \leqslant B_0\sqrt{K}, \|\boldsymbol{\beta}\|_2 \leqslant B_0}{\operatorname{argmin}} \left\{ \frac{\partial \mathcal{L}_1(\boldsymbol{\beta}^t)}{\partial \boldsymbol{\beta}^{\mathrm{T}}}(\boldsymbol{\beta} - \boldsymbol{\beta}^t) + \frac{\eta}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^t\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \quad (3.3)$$

where $\eta > 0$ is a stepsize parameter. To solve (3.3), it is easy to see that, ignoring the constraints on the norms of $\boldsymbol{\beta}$, (3.3) is a typical quadratic function plus a Lasso penalty. Thus, taking into account the restrictions, we can use existing algorithms to first obtain

$$\widetilde{\boldsymbol{\beta}}^{t+1} \equiv \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{\partial \mathcal{L}_1(\boldsymbol{\beta}^t)}{\partial \boldsymbol{\beta}^{\mathrm{T}}}(\boldsymbol{\beta} - \boldsymbol{\beta}^t) + \frac{\eta}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^t\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}. \quad (3.4)$$

We then project $\widetilde{\boldsymbol{\beta}}^{t+1}$ onto the l_1 ball with radius $B_0\sqrt{K}$ to obtain $\breve{\boldsymbol{\beta}}^{t+1}$ using the simplex projection method discussed in Duchi et al. (2008). Finally, if $\|\breve{\boldsymbol{\beta}}^{t+1}\|_2 > B_0$, we shrink it to get $\boldsymbol{\beta}^{t+1} = \breve{\boldsymbol{\beta}}^{t+1}B_0/\|\breve{\boldsymbol{\beta}}^{t+1}\|_2$. Otherwise, we let $\boldsymbol{\beta}^{t+1} = \breve{\boldsymbol{\beta}}^{t+1}$.

In the above algorithm, the bound B_0 restricts the total search range of the optimization procedure. One can perform a naive analysis treating $\mathbf{W}$ as $\mathbf{X}$, and use c times the l_2 norm of the naive Poisson Lasso regression estimator as B_0. In practice, choosing $c \geqslant 2$ is a secure practice that we recommend. Of course, if empirical knowledge is available on B_0, one can use it as well. Similarly, K serves as a sparsity restriction. Note that the upper bound of the true sparseness is in the order of $\sqrt{n/\log(p)}$, as assumed in Theorem 1. Therefore, we set $K = \{n/\log(p)\}^{1/2+\epsilon}$, where $\epsilon > 0$, which is guaranteed to be greater than the true sparseness asymptotically. In (3.4), the stepsize η is usually chosen in an ad hoc way (Girshick (2015); He et al. (2017a)). Here, we choose η so that the l_2 norm of the consecutive two outputs is less than 0.01 within the first 100 iterations. Finally, the tuning parameter λ in (3.4) can be chosen using cross-validation (Friedman, Hastie and Tibshirani (2010)) using $\sum_{i=1}^{n}\{Y_i - \exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{W}_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\Omega}\widehat{\boldsymbol{\beta}}/2)\}^2$ as the loss function. We summarize the algorithm as follows, where the initial value $\widehat{\boldsymbol{\beta}}^0$ can also be the regression calibration estimator.

We point out two key properties of the proposed estimator (3.2). First, the derivative of $\mathcal{L}_1(\boldsymbol{\beta})$ has mean zero at $\boldsymbol{\beta}_0$. Second, the mean of $\mathcal{L}_1(\boldsymbol{\beta})$ has the form of a convex function and, hence, the optimization procedure has statistical and algorithmic convergence guaranteed asymptotically. These two properties

Algorithm 1 Algorithm

```
Inputs: Given $(\mathbf{W}_i, Y_i), i = 1, \ldots, n, \mathbf{\Omega}, \eta, \lambda, M$, tol.
Obtain $\widehat{\boldsymbol{\beta}}^0$ from the naive Poisson Lasso regression.
Set $B_0 = c\|\widehat{\boldsymbol{\beta}}^0\|_2$ and $K = \{n/\log(p)\}^{1/2+\epsilon}$.
for $t$ in 0 to $M$ do
  1. Obtain $\widetilde{\boldsymbol{\beta}}^{t+1}$ from solving (3.4).
  2. Project $\widetilde{\boldsymbol{\beta}}^{t+1}$ to the $l_1$ ball with radius $B_0\sqrt{k}$ to get $\breve{\boldsymbol{\beta}}^{t+1}$.
  3. $\boldsymbol{\beta}^{t+1} = \breve{\boldsymbol{\beta}}^{t+1} \min(B_0, \|\breve{\boldsymbol{\beta}}^{t+1}\|_2)/\|\breve{\boldsymbol{\beta}}^{t+1}\|_2$.
  if $\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|_2 \leqslant$ tol, stop.
end for
```

jointly lead to the consistency property of our estimator under suitable sparsity and other regularity conditions, as we establish below.

4. Statistical Properties

4.1. Definition and regularity conditions

To prepare for analyzing the theoretical properties of the proposed estimator $\widehat{\boldsymbol{\beta}}$ in (3.2), we first introduce some notation. We name the set of all $\boldsymbol{\beta}$ that satisfy $\|\boldsymbol{\beta}\|_1 \leqslant b_0\sqrt{k}, \|\boldsymbol{\beta}\|_2 \leqslant b_0$ the feasible set, where b_0 is a constant. For a matrix $\mathbf{M}$, let $\|\mathbf{M}\|_{\max}$ be the matrix maximum norm, $\|\mathbf{M}\|_\infty$ be the l_∞ norm, and $\|\mathbf{M}\|_p$ be the l_p norm. For a general vector $\mathbf{a}$, let $\|\mathbf{a}\|_\infty$ be the vector sup-norm, and $\|\mathbf{a}\|_p$ be the vector l_p-norm. Let $\alpha_{\min}(\mathbf{M})$ and $\alpha_{\max}(\mathbf{M})$ be the minimal and maximal eigenvalues of the matrix $\mathbf{M}$, respectively. To simplify the notation, we define

$$\alpha_{\min}(\boldsymbol{\beta}) \equiv \alpha_{\min}[E\{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\mathbf{X}\mathbf{X}^{\mathrm{T}}\}],$$

and

$$\alpha_{\max}(\boldsymbol{\beta}) \equiv \alpha_{\max}[E\{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\mathbf{X}\mathbf{X}^{\mathrm{T}}\}].$$

We use the following restricted eigenvalue (RE) conditions to show the consistency of the estimators, whose various versions are introduced in Bickel, Ritov and Tsybakov (2009); Van De Geer and Bühlmann (2009), and Loh and Wainwright (2012).

Definition 1. (Lower-RE condition). A matrix $\mathbf{\Gamma}$ satisfies a lower RE condition with curvature $a_1 > 0$ and tolerance $\tau(n, p) > 0$ if

$$\boldsymbol{\beta}^{\mathrm{T}}\mathbf{\Gamma}\boldsymbol{\beta} \geqslant a_1\|\boldsymbol{\beta}\|_2^2 - \tau(n,p)\|\boldsymbol{\beta}\|_1^2, \quad \forall \boldsymbol{\beta} \in \mathbb{R}^p.$$

Definition 2. (Upper-RE condition). A matrix $\mathbf{\Gamma}$ satisfies an upper RE condition with smoothness $a_2 > 0$ and tolerance $\tau(n, p) > 0$ if

$$\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\beta} \leqslant a_2\|\boldsymbol{\beta}\|_2^2 + \tau(n,p)\|\boldsymbol{\beta}\|_1^2, \quad \forall \boldsymbol{\beta} \in \mathbb{R}^p.$$

Consider the situation when $\boldsymbol{\Gamma}$ is the second derivative of the loss function $\mathcal{L}_1(\boldsymbol{\beta})$. When the features are correctly measured, $\mathbf{W}_i = \mathbf{X}_i$, $\boldsymbol{\Omega} = \mathbf{0}$, and $\boldsymbol{\Gamma} = n^{-1}\sum_{i=1}^n \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}$. Thus, $\boldsymbol{\Gamma}$ is positive definite. However, when measurement errors occur, the second deriviate of the loss function becomes $\boldsymbol{\Gamma} = n^{-1}\sum_{i=1}^n \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\beta}/2)\{(\mathbf{W}_i - \boldsymbol{\Omega}\boldsymbol{\beta})^{\otimes 2} - \boldsymbol{\Omega}\}$, which is no longer guaranteed to be positive definite. In this case, the lower-RE condition ensures that $\boldsymbol{\Gamma}$, even though it may not be positive definite globally, can still be shown to achieve positive definite properties in the feasible set. The upper-RE condition, together with lower-RE condition, guarantees the computational convergency of the composite gradient descent algorithm.

The theoretical properties of our estimator $\widehat{\boldsymbol{\beta}}$ are based on the mild Conditions (C1)–(C6) in the Supplementary Material. To save space, we provide a brief discussion on the conditions here. Condition (C1) guarantees the boundedness and the invertibility of the Hessian matrix $E\{\exp(\boldsymbol{\beta}^{\mathrm{T}})\mathbf{X}\mathbf{X}^{\mathrm{T}}\}$ for $\boldsymbol{\beta}$ in the feasible set, that is the second derivative of the log likelihood without a measurement error. Condition (C2) bounds the total variability of both the response Y and the measurement error U marginally. Condition (C3) essentially controls the order of $|X_{ij}|$. A similar requirement is also assumed in Loh and Wainwright (2012). Condition (C4) constrains the dimensionality in relation to the sample size. Conditions (C5) and (C6) are not very stringent, and we provide examples that satisfy the conditions in Section S.2. The lower bounds in (i) of Condition (C5) and in (S.1) and (S.2) of Condition (C6) are assumed to avoid zero denominators. In the Supplementary Material, we show that all other requirements in both conditions are satisfied when, for example, $\|\boldsymbol{\Omega}\|_2 = O(1)$, $\|\mathrm{cov}(\mathbf{X})\|_2 = O(1)$, and the moments of the conditional sub-Gaussian and sub-exponential norms are uniformly bounded.

4.2. Statistical consistency

In this section, we establish the main theorem on the statistical consistency of $\widehat{\boldsymbol{\beta}}$. To control the error between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, we use the optimality of the loss function that $\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \mathcal{L}(\boldsymbol{\beta}_0) \leqslant 0$, for any $\boldsymbol{\beta} \neq \widehat{\boldsymbol{\beta}}$. Writing $\widehat{\mathbf{v}} \equiv \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$, by the Taylor expansion, this leads to

$$-\widehat{\mathbf{v}}^{\mathrm{T}}\frac{\partial \mathcal{L}_1(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \geqslant \frac{1}{2}\widehat{\mathbf{v}}^{\mathrm{T}}\frac{\partial^2 \mathcal{L}_1(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}}\widehat{\mathbf{v}} + \lambda\|\boldsymbol{\beta}_0 + \widehat{\mathbf{v}}\|_1 - \lambda\|\boldsymbol{\beta}_0\|_1, \tag{4.1}$$

where $\boldsymbol{\beta}^*$ is the point on the line connecting $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. The inequality in (4.1) serves as the foundation for all the following derivations. To show the convergence of $\hat{\mathbf{v}}$ to zero, we must establish the upper bound for the left-hand side and the lower bound for the right-hand side of (4.1). However, the marginal distribution of $\mathbf{W}_i$ is unknown, because the distribution of $\mathbf{X}_i$ is unspecified. Furthermore, the distribution of $\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i)$ could be heavy tail even when $\mathbf{W}_i$ has a multivariate Gaussian distribution. Hence, we cannot apply the commonly used joint analysis of $Y_i, \mathbf{W}_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i$ to establish the consistency. To overcome these difficulties, we discuss the tail property of $\|\partial\mathcal{L}_1(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\|_\infty$, conditioning on $\mathbf{W}_i$, for $i = 1, \ldots, n$, and the tail property of $\|\partial^2\mathcal{L}_1(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}\|_2$, conditional on $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i$, for $i = 1, \ldots, n$ and $\|\boldsymbol{\beta}\|_2 \leqslant 2b_0$. The conditional tail properties are defined based Lemma 1, 3 and on 2, 4 for the conditional sub-Gaussian and sub-exponential variables, respectively.

Lemma 1. *Let $\mathcal{F}$ be a sub-field of the sigma-field generated by X. Let $K_j(\mathcal{F}) > 0, j = 1, \ldots, 4$ be functions of random variables in $\mathcal{F}$. Then, the following properties 1, 2, and 3 are equivalent, and when $E(X|\mathcal{F}) = 0$, they are further equivalent to property 4. In addition, $K_j(\mathcal{F})$ can be chosen to satisfy $0 < c < K_j(\mathcal{F})/K_k(\mathcal{F}) < C < \infty$, for all $k, j \in \{1, 2, 3, 4\}$, where c, C are absolute constants.*

1. *Tail: There exists $K_1(\mathcal{F})$ such that $\Pr(|X| > t|\mathcal{F}) \leqslant \exp\{1 - t^2/K_1^2(\mathcal{F})\}$;*

2. *Moments: There exists $K_2(\mathcal{F})$ such that $E(|X|^k|\mathcal{F})^{1/k} \leqslant K_2(\mathcal{F})\sqrt{k}$, for all $k \geqslant 1$;*

3. *Super-exponential moment: There exists $K_3(\mathcal{F})$ such that $E[\exp\{X^2/K_3^2(\mathcal{F})\}|\mathcal{F}] \leqslant e$;*

4. *Let $E(X|\mathcal{F}) = 0$. There exists $K_4(\mathcal{F})$ such that $E\{\exp(tX)|\mathcal{F}\} \leqslant \exp\{t^2 K_4^2(\mathcal{F})\}$.*

Definition 3. A random variable X that satisfies one of the equivalent properties in Lemma 1 is named a conditional sub-Gaussian random variable with respect to the sub-sigma field $\mathcal{F}$. The conditional sub-gaussian norm of X with respect to $\mathcal{F}$, denoted by $\|X\|_{\psi_2(\mathcal{F})}$, is defined as the smallest $K_2(\mathcal{F})$ in property 2. That is,

$$\|X\|_{\psi_2(\mathcal{F})} = \sup_{k \geqslant 1} k^{-1/2} E(|X|^k|\mathcal{F})^{1/k}.$$

Lemma 2. *Let $\mathcal{F}$ be a sub-field of the sigma-field generated by X. Let $K_j(\mathcal{F}) > 0, j = 1, 2, 3$ be functions of the random variables in $\mathcal{F}$. Then, the following properties 1, 2, and 3 are equivalent. In addition, $K_j(\mathcal{F})$ can be chosen to satisfy $0 < c < K_j(\mathcal{F})/K_k(\mathcal{F}) < C < \infty$, for all $k, j \in \{1, 2, 3\}$, where c, C are absolute constants.*

1. *Tail: There exists $K_1(\mathcal{F})$ such that $\Pr(|X| > t|\mathcal{F}) \leqslant \exp\{1 - t/K_1(\mathcal{F})\}$;*
2. *Moments: There exists $K_2(\mathcal{F})$ such that $E(|X|^k|\mathcal{F})^{1/k} \leqslant K_2(\mathcal{F})k$, for all $k \geqslant 1$;*
3. *Super-exponential moment: There exists $K_3(\mathcal{F})$ such that $E[\exp\{|X|/K_3(\mathcal{F})\}|\mathcal{F}] \leqslant e$;*

Definition 4. A random variable X that satisfies one of the equivalent properties in Lemma 2 is named a conditional sub-exponential random variable with respect to sub-sigma field $\mathcal{F}$. The conditional sub-exponential norm of X with respect to $\mathcal{F}$, denoted by $\|X\|_{\psi_1(\mathcal{F})}$, is defined as the smallest $K_2(\mathcal{F})$ in property 2. That is,

$$\|X\|_{\psi_1(\mathcal{F})} = \sup_{k \geqslant 1} k^{-1} E(|X|^k|\mathcal{F})^{1/k}.$$

Specifically, based on the above definitions and properties, we first establish the upper bound of the left side of (4.1). It is easy to see that the left side of (4.1) can be split into three terms:

$$n^{-1}\sum_{i=1}^{n}\{Y_i - \exp(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)\}\hat{\mathbf{v}}^{\mathrm{T}}\mathbf{X}_i, n^{-1}\sum_{i=1}^{n} Y_i\hat{\mathbf{v}}^{\mathrm{T}}(\mathbf{W}_i - \mathbf{X}_j),$$
$$n^{-1}\sum_{i=1}^{n}\exp\left(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{W}_i - \frac{\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\beta}_0}{2}\right)\hat{\mathbf{v}}^{\mathrm{T}}(\mathbf{W}_i - \boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\Omega}) - \exp(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)\hat{\mathbf{v}}^{\mathrm{T}}\mathbf{X}_i.$$

The three terms are conditional sub-exponential, conditional sub-Gaussian and conditional sub-Gaussian respectively. Note that to show the third term is conditional sub-Gaussian, we use the property that $\mathbf{W}_i$, given $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i$ and $\mathbf{X}_i$, is still a Gaussian variable with mean linearly in $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i$, which is crucial to deriving the tail properties.

Following the properties of the conditional sub-Gaussian and sub-exponential distributions, the sup-norms of these terms are bounded in the order of $O_p\{\sqrt{n/\log(p)}\}$, as shown in Lemma S.5 in the Supplementary Material. The bounds of the three terms combined lead to the sup-norm bound in Lemma 3, which allows us to bound the left side of (4.1) by the product of an $O_p\{\sqrt{n/\log(p)}\}$ term and $\|\mathbf{v}\|_1$.

Lemma 3. *Under Conditions* (C1)–(C4),

$$\Pr\left[\left\|\frac{\partial\mathcal{L}_1(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right\|_\infty > 3\max\left\{4e\sqrt{M_1}, 8eM_2C,\right.\right.$$
$$\left.\left.\frac{2c_{10}M_3Q_1(1+r)}{m_3}, \sqrt{72e^2M_0}, 1\right\}\sqrt{\frac{log(p)}{n}}\right] \leqslant 6p^{-1}.$$

Next, we show that the quadratic form on the right side of (4.1) satisfies the lower-RE condition by using the fact that it is sub-exponential conditioning on $\boldsymbol{\beta}^{*^{\mathrm{T}}}\mathbf{W}_i, \mathbf{X}_i$, for $i = 1, \ldots, n$. Although $\exp(\boldsymbol{\beta}^{*^{\mathrm{T}}}\mathbf{W}_i - \boldsymbol{\beta}^{*^{\mathrm{T}}}\boldsymbol{\Omega}\boldsymbol{\beta}^*/2)$, for $i = 1, \ldots, n$, can be treated as constant weights, it is necessary to consider their growth under the high-dimensional setting. We first show the tail property of the Hessian matrix under the finite dimensional settings. Then, using the covering argument, we show in Lemma S.7 that for growing dimensions, $\partial^2\mathcal{L}_1(\boldsymbol{\beta}^*)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}$ still converges to $E\{\exp(\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\}$ for $\boldsymbol{\beta}^*$ in the feasible set in the l_2 distance. This relation, together with the positive-definite property of $E\{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\}$, helps to establish the lower-RE condition in Lemma 4.

Lemma 4. *Assume that Conditions* (C1), (C4), *and* (C6) *hold. Then, for sufficiently large n and p, with probability going to one and $\|\boldsymbol{\beta}\|_2 \leqslant 2b_0$, $\partial^2\mathcal{L}_1(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}$ satisfies the lower and upper-RE conditions with*

$$a_1 = \alpha_{\min}(\boldsymbol{\beta})\left\{1 - \frac{1}{2c}\right\},\ a_2 = \alpha_{\max}(\boldsymbol{\beta})\left\{1 + \frac{1}{2c}\right\}$$

and

$$\tau(n,p) = \sup_{\{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_2\leqslant 2b_0\}}\left\{\frac{\alpha_{\min}(\boldsymbol{\beta})}{2c}\right\}\left[\frac{1}{32C\max(M_4, M_5)}\sqrt{\frac{n}{log(p)}}\right.$$
$$\left.\min\left\{\left(\sup_{\{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_2\leqslant 2b_0\}}\frac{\alpha_{\min}(\boldsymbol{\beta})}{486ce}\right)^2, 1\right\}\right]^{-1}.$$

Here, c is a constant.

The detailed statement of the above results and their proofs are presented in the Supplementary Material Sections S.6 and S.7. The above derivations allow us to bound from above the left-hand side of (4.1) by expressions containing $\|\mathbf{v}\|_1$, and to bound from below the right-hand side of (4.1) by expressions containing $\|\mathbf{v}\|_1$ and $\|\mathbf{v}\|_2$. Combining these results and using the sparsity properties of $\|\boldsymbol{\beta}_0\|$ and the feasible set property, we further obtain the convergence of $\mathbf{v}$ to zero in both the l_1 and l_2 norms in Theorem 1.

Theorem 1. *Assume that Conditions* (C1)–(C6) *hold. Define*

$$\phi \equiv 3\max\left\{4e\sqrt{M_1}, 8eM_2C, \frac{2c_{10}M_3Q_1(1+r)}{m_3},\right.$$
$$\left.\sqrt{72e^2M_0}, \sup_{\{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_2\leqslant 2b_0\}}\frac{\alpha_{\min}(\boldsymbol{\beta})b_0}{2c\sqrt{c_1}}, 1\right\},$$

where $c = 128$,

$$c_{10} = \max\left[\sqrt{\frac{18e^2m_3^2}{M_3Q_1^2(1+r)r}}, 1\right],$$

and

$$c_1 = \min\left\{\left(\sup_{\{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_2\leqslant 2b_0\}}\frac{\alpha_{\min}(\boldsymbol{\beta})}{486ce}\right)^2, 1\right\}/32C\max(M_4, M_5),$$

C is the constant defined in Condition (C4)*, M_0 is the constant defined in Condition* (C2)*, M_1 and M_2 are the constants defined in Condition* (C3)*, m_3, M_3, and Q_1 are the constants defined in Condition* (C5)*, M_4 and M_5 are the constants defined in Condition* (C6)*, and r is an arbitrary positive constant. Further let $\lambda \geqslant 8/3\phi\{log(p)/n\}^{1/4}$ and let*

$$\sqrt{s}\tau(n,p) = \min\left[\sup_{\{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_2\leqslant 2b_0\}}\frac{\alpha_{\min}(\boldsymbol{\beta})}{(2c\sqrt{s})}, \frac{\phi}{b_0}\left\{\frac{log(p)}{n}\right\}^{1/4}\right],$$

where $s = c_1\sqrt{n/log(p)}$ Then, fora vector $\boldsymbol{\beta}_0$ with sparsity at most k, $k \leqslant s$, $\|\boldsymbol{\beta}_0\|_2 \leqslant b_0$, the global minimizer $\widehat{\boldsymbol{\beta}}$ of (3.2) *satisfies the bounds*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \leqslant \frac{2^{13}\sqrt{k}\lambda}{191\alpha_{\min}(\boldsymbol{\beta}^*)} \quad \textit{and} \quad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leqslant \frac{2^{16}k\lambda}{191\alpha_{\min}(\boldsymbol{\beta}^*)}$$

where $\boldsymbol{\beta}^*$ is a point on the line connecting $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$.

A detailed proof of Theorem 1 is provided in the Appendix. We can set $\lambda = 8/3\phi\{\log(p)/n\}^{1/4}$ in Theorem 1. Because ϕ is of order $O_p(1)$, we can see that when $k = o_p[\{n/\log(p)\}^{1/2}]$, $\widehat{\boldsymbol{\beta}}$ is consistent to $\boldsymbol{\beta}_0$ in the l_2 norm, and when $k = o_p[\{n/\log(p)\}^{1/4}]$, $\widehat{\boldsymbol{\beta}}$ is consistent in the l_1 norm. Here, we have a larger penalty than the usual requirement that $\lambda \geqslant 2\|\partial\mathcal{L}_1(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\|_\infty$, where the latter is of order $O_p[\{n/\log(p)\}^{1/2}]$ (Agarwal, Negahban and Wainwright (2012); Loh and Wainwright (2012)), and the convergence rate is slower than that of the linear model. This is because the maximum term, such as $\max_i |A(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i)|K_{gvi}(\boldsymbol{\beta})$ in

(S.2), grows at a $\log(n)$ rate, while the corresponding quantity in the linear model is bounded, owning to the property of the sub-Gaussian feature.

5. Algorithmic Convergence

The composite gradient descent algorithm was first proposed to solve convex optimization problems (Agarwal, Negahban and Wainwright (2012)). Carefully examining Theorem 2 in Agarwal, Negahban and Wainwright (2012), we found that the algorithm converges when the loss function is convex in the feasible set, which holds naturally following the lower-RE condition and the definition of the feasible set in our problem. Hence, in Theorem 2, we relax the global convexity required in Agarwal, Negahban and Wainwright (2012). Further, in Theorem 3, we establish the numerical convergence with specific choices of the contract factors. Note again that the choice of λ is more stringent than the one in the linear model, because $\tau(n,p)$ approaches zero more slowly so that a larger penalty is required to establish convexity in the feasible set.

Now, define $\mathcal{M}$ as the subspace of all vectors with support contained within the support of $\boldsymbol{\beta}_0$. Recall that the support of a vector is defined as the set that contains the indices of the nonzero elements of the vector. Let $\psi(\mathcal{M}) \equiv \sup_{\boldsymbol{\beta}:\boldsymbol{\beta}\in\mathcal{M},\boldsymbol{\beta}\neq\mathbf{0}} \|\boldsymbol{\beta}\|_1/\|\boldsymbol{\beta}\|_2 = \sqrt{k}$, $\boldsymbol{\beta}^*$ be the point between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$,

$$\bar{\gamma}_l \equiv 2\alpha_{\min}(\boldsymbol{\beta}^*)\left\{1-\frac{1}{2c}\right\} - 64\tau(n,p)\psi^2(\mathcal{M}),$$

and the contraction coefficients be

$$\kappa(\mathcal{M}) \equiv \xi(\mathcal{M})\left\{1-\frac{\bar{\gamma}_l}{8\alpha_{\max}(\boldsymbol{\beta}^*)\{1+(1/2c)\}}+\frac{64\psi^2(\mathcal{M})\tau(n,p)}{\bar{\gamma}_l}\right\},$$

where $\tau(n,p)$ is defined in Definition 1, $\xi(\mathcal{M}) \equiv \{1-64\psi^2(\mathcal{M})\tau(n,p)/\bar{\gamma}_l\}^{-1}$. Define

$$\beta(\mathcal{M}) \equiv 2\tau(n,p)\left(\frac{\bar{\gamma}_l}{8\alpha_{\max}(\boldsymbol{\beta}^*)\{1+(1/2c)\}}+\frac{128\tau(n,p)\psi^2(\mathcal{M})}{\bar{\gamma}_l}\right)+10\tau(n,p),$$

and $\epsilon^2(\mathcal{M}) \equiv 8\xi(\mathcal{M})\beta(\mathcal{M})\{6\psi(\mathcal{M})\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2\}^2$.

Theorem 2. *Assume that Conditions* (C1)–(C6) *hold. Consider the optimization problem* (3.2) *with the regularization parameter* $\lambda \geqslant 8/3\|\partial\mathcal{L}_1(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}_0\|_\infty$*, and suppose that for any* $\boldsymbol{\beta}$ *in the feasible set,* $\partial^2\mathcal{L}_1(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}$ *satisfies the lower-RE and upper-RE conditions with parameters* $\{a_1,\tau(n,p)\}$ *and* $\{a_2,\tau(n,p)\}$*, respectively, where*

$$a_1 = \alpha_{\min}(\boldsymbol{\beta}^*)\left\{1 - \frac{1}{2c}\right\}, a_2 = \alpha_{\max}(\boldsymbol{\beta}^*)\left\{1 + \frac{1}{2c}\right\},$$

$\tau(n,p)$ is defined in Definition 1 and $\boldsymbol{\beta}^*$ is a point on the line connecting $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. Assume $\kappa(\mathcal{M}) \in [0,1)$ and

$$\lambda \geqslant \frac{32 b_0 \sqrt{k}}{1-\kappa(\mathcal{M})}\xi(\mathcal{M})\beta(\mathcal{M}).$$

Then, for any $\delta^2 \geqslant \epsilon^2(\mathcal{M})/(1-\kappa(\mathcal{M}))$, when

$$t \geqslant \frac{2log[\{\mathcal{L}_1(\boldsymbol{\beta}^0) + \lambda\|\boldsymbol{\beta}^0\|_1 - \mathcal{L}_1(\widehat{\boldsymbol{\beta}}) - \lambda\|\widehat{\boldsymbol{\beta}}\|_1\}/\delta^2]}{log\{1/\kappa(\mathcal{M})\}} + \left(1 + \frac{log(2)}{log\{1/\kappa(\mathcal{M})\}}\right) log_2 log_2\left(\frac{b_0\sqrt{k}\lambda}{\delta^2}\right),$$

we have

$$\|\boldsymbol{\beta}^t - \widehat{\boldsymbol{\beta}}\|_2^2 \leqslant \frac{2\delta^2}{\bar{\gamma}_l} + \frac{16\delta^2\tau(n,p)}{\bar{\gamma}_l\lambda^2} + \frac{4\tau(n,p)\{6\psi(\mathcal{M})\}^2}{\bar{\gamma}_l}. \tag{5.1}$$

Theorem 3. *Assume that Conditions (C1)–(C6) hold and $k = o[\{n/log(p)\}^{1/2}]$. Let*

$$a_1 = \alpha_{\min}(\boldsymbol{\beta}^*)\left\{1 - \frac{1}{2c}\right\}, a_2 = \alpha_{\max}(\boldsymbol{\beta}^*)\left\{1 + \frac{1}{2c}\right\}$$

in Definitions 1 and 2, respectively. Let $\tau(n,p)$ be as defined in Theorem 1. Let

$$\lambda \geqslant \frac{8}{3}\phi\left\{\frac{log(p)}{n}\right\}^{1/4}, \quad \lambda = O\left[\left\{\frac{log(p)}{n}\right\}^{1/4}\right].$$

Then, for $\delta^2 = \epsilon^2(\mathcal{M})/(1-\kappa(\mathcal{M}))$ and

$$t \geqslant \frac{2log[\{\mathcal{L}_1(\boldsymbol{\beta}^0) + \lambda\|\boldsymbol{\beta}^0\|_1 - \mathcal{L}_1(\widehat{\boldsymbol{\beta}}) - \lambda\|\widehat{\boldsymbol{\beta}}\|_1\}/\delta^2]}{log\{1/\kappa(\mathcal{M})\}} + \left(1 + \frac{log(2)}{log\{1/\kappa(\mathcal{M})\}}\right) log_2 log_2\left(\frac{b_0\sqrt{k}\lambda}{\delta^2}\right),$$

we have

$$\|\boldsymbol{\beta}^t - \widehat{\boldsymbol{\beta}}\|_2^2 = o(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2) + O(k\lambda^2). \tag{5.2}$$

Here, we impose the condition that $\lambda = O[\{\log(p)/n\}^{1/4}]$ so that $\lambda^2 = O\{\tau(n,p)\}$, and hence the last term in (5.1) has order $O(k\lambda^2)$ as shown in (5.2).

The results in (5.2) shows that the error between $\boldsymbol{\beta}^t$ and $\widehat{\boldsymbol{\beta}}$ consists of two terms: one term has smaller order than the statistical error, and the other term has order $\sqrt{k}\lambda$. Further, by Theorem 1, the l_2 norm of the statistical error is upper bounded by a quantity of order $O(\sqrt{k}\lambda)$. Hence, the algorithmic convergence and the statistical consistency are achieved simultaneously when $\sqrt{k}\lambda \to 0$.

Remark 1. Unsurprisingly, the number of iterations needed, t, depends on the initial value $\boldsymbol{\beta}^0$. When $\boldsymbol{\beta}^0$ satisfies

$$\log_2\left(\frac{b_0\sqrt{k}\lambda}{\delta^2}\right) = O_p\left[\frac{\mathcal{L}_1(\boldsymbol{\beta}^0) + \lambda\|\boldsymbol{\beta}^0\|_1 - \mathcal{L}_1(\widehat{\boldsymbol{\beta}}) - \lambda\|\widehat{\boldsymbol{\beta}}\|_1}{\delta^2}\right],$$

then we can stop the iteration and declare convergence when

$$t \geqslant \frac{d_1 \log[\{\mathcal{L}_1(\boldsymbol{\beta}^0) + \lambda\|\boldsymbol{\beta}^0\|_1 - \mathcal{L}_1(\widehat{\boldsymbol{\beta}}) - \lambda\|\widehat{\boldsymbol{\beta}}\|_1\}/\delta^2]}{\log\{1/\kappa(\mathcal{M})\}},$$

for some positive constant d_1. On the other hand, if $\boldsymbol{\beta}^0$ satisfies

$$\frac{\mathcal{L}_1(\boldsymbol{\beta}^0) + \lambda\|\boldsymbol{\beta}^0\|_1 - \mathcal{L}_1(\widehat{\boldsymbol{\beta}}) - \lambda\|\widehat{\boldsymbol{\beta}}\|_1}{\delta^2} = o_p\left\{\log_2\left(\frac{b_0\sqrt{k}\lambda}{\delta^2}\right)\right\},$$

the convergence is achieved when

$$t \geqslant d_2 \log_2 \log_2\left(\frac{b_0\sqrt{k}\lambda}{\delta^2}\right),$$

for some positive constant d_2.

6. Numerical Performance

6.1. Simulations

We evaluate the proposed estimator for the measurement error Poisson regression model (MPR). We simulated $\mathbf{X}$ from the uniform distribution in the interval $(1, 2)$. Further, the measurement error $\mathbf{U}$ was simulated from the multivariate Gaussian with variance $\boldsymbol{\Omega}$, where the (i, j) element of $\boldsymbol{\Omega}$ is $0.04 \times 0.5^{|i-j|}$. We selected $\boldsymbol{\Omega}$ so that the variance of $\mathbf{U}$ is $1/2$ of the variance of $\mathbf{X}$. The number of nonzero $\boldsymbol{\beta}_0$ are $4, 6, \ldots, 18$. The nonzero elements in $\boldsymbol{\beta}_0$ are k values equally spaced in the interval $[1, 2]$. We then simulated Y from the Poisson distribution with mean $\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X} - \alpha)$, where α is selected to keep the variance of Y at 500 to avoid too many zero values. In all estimations, we choose

Table 1. Performance of the estimators over 100 simulation runs. The result is calculated as the mean of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ over 100 simulation times. Naive, Regcal stand for the naive and regression calibration estimator, respectively.

		$k = 6$			$k = 10$		
n	p	Naive	Regcal	MPR	Naive	Regcal	MPR
100	64	2.165	1.901	1.268	3.427	3.224	2.598
	128	2.261	1.994	1.403	3.489	3.309	2.801
	256	2.451	2.184	1.612	3.628	3.415	3.071
200	64	1.837	1.404	0.775	2.939	2.549	1.771
	128	1.983	1.611	0.905	2.933	2.584	1.803
	256	2.092	1.727	1.013	3.034	2.700	1.978

Table 2. Performance of the estimators over 100 simulation runs for sample size $n = 200$. The result is calculated as the mean of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ over 100 simulation times.

k	$p = 64$	$p = 128$	$p = 256$
18	3.443	3.574	3.736
16	2.914	3.008	3.459
14	2.516	2.932	2.899
12	1.955	2.351	2.544
10	1.771	1.803	1.978
8	1.115	1.387	1.474
6	0.775	0.905	1.013

$B_0 = b_0\{n/\log(p)\}^{1/4+0.01}$ and $K = 0.25k\{n/\log(p)\}^{1/2+0.02}$, where b_0, and k are the l_2 and l_0 norms, respectively, of the initial estimator. Furthermore, we set $\lambda = 20\{n/\log(p)\}^{1/4}$ and $\eta = 2e4, 4e4, 8e4$ for $p = 64, 128, 256$, respectively. We present results with different choices of n, p, c in Table 1, Table 2, and Figure 1. Clearly, when $\sqrt{n/\log(p)}/k$ increases, the mean l_2 error decreases. Furthermore, increasing $n/\log(p)$ also reduces the estimation error.

We compare the MPR with the naive and regression calibration methods, described as follows. In the naive method, we treat $\mathbf{W}$ as the measurement error free covariate and obtain the estimators through the Poisson regression. For the regression calibration method, we replace $\mathbf{W}$ by $E(\mathbf{X}|\mathbf{W})$ in the Poisson regression to obtain the estimators, where the true distribution of $\mathbf{X}$ is used to obtain $E(\mathbf{X}|\mathbf{W})$. We impose the Lasso penalty in both procedures, where the corresponding regularization parameters are chosen using 10-fold cross-validation. It can be seen that the MPR and the regression calibration methods outperform the naive estimator, and the MPR always performs best among all three estimators in terms of generating the smallest mean l_2 errors.

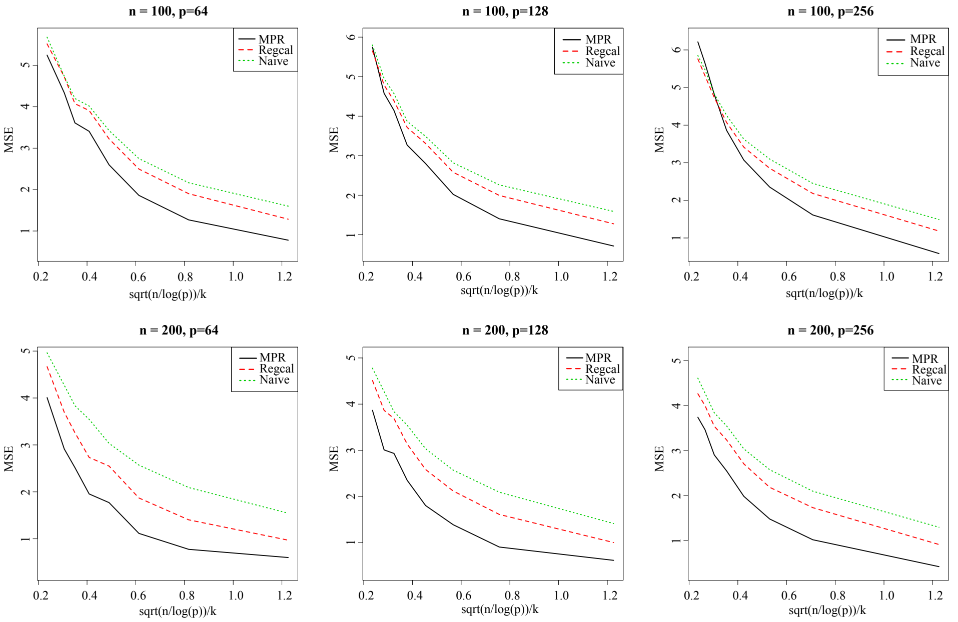

Figure 1. Comparison of the MPR, naive (Naive), and regression calibration (Regcal) estimators.

Next, we study the statistical convergency in Theorem 1 by examining the relation between the empirical average of $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2$, the sample sizes, and the order of $k^{-1}\lambda^{-2}$, that is, $\sqrt{n/\log(p)}/k$ in Figure 2. In this simulation, we choose $k = \sqrt{p}$ and vary n, p values. As we show in Theorem 1, $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2$ converges in the order of $O(\sqrt{k}\lambda)$. Hence, it is expected that $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2$ will decrease along with an increase in both $\sqrt{n}$ and $\sqrt{n/\log(p)}/k$. When we plot $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2^2$ versus $\sqrt{n}$, the estimation error curves corresponding to different p values are well separated, with curves corresponding to larger p values below those corresponding to smaller p values, reflecting the obvious fact that a smaller parameter dimension leads to a better estimation. On the other hand, when we plot $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2^2$ versus $\sqrt{n/\log(p)}/k$, the curves corresponding to different p values align very close to each other, reflecting the results in Theorem 1 that the convergence rate is $\sqrt{n/\log(p)}/k$ and the dependence of the convergence on p is fully captured by $\sqrt{n/\log(p)}/k$.

We further investigate how the MPR performs when $\mathbf{U}$ is not multivariate normal. To reflect the heavy-tailed errors and the nonsymmetric errors, we generate $\mathbf{U}$ from a Student's t distribution with 2 degrees of freedom and from a gamma distribution with both the shape and the scale parameters equal to two, respectively. We then standardize $\mathbf{U}$ to have mean zero and variance $\boldsymbol{\Omega}$, identical

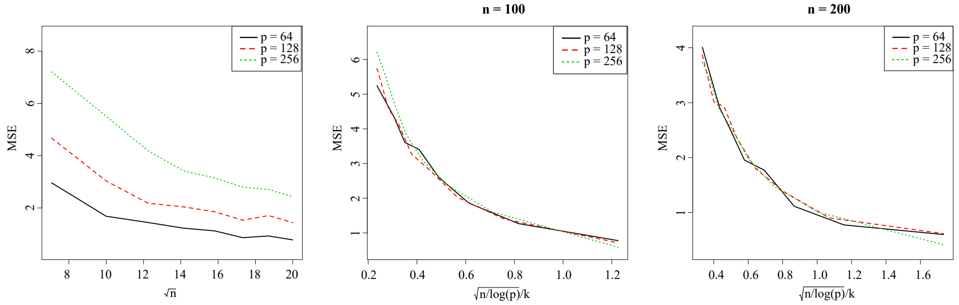

Figure 2. The convergence of the MSE, the average of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2$ over 100 simulations.

to that in the normal error case. We compare $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ from the naive, regression calibration, and the MPR methods. The results in Table 3 suggest that although the MPR is proposed under the assumption that $\mathbf{U}$ has multivariate normal distribution, it still outperforms the naive and the regression calibration method when the normal assumption is violated.

6.2. Analysis of COVID-19 data

The COVID-19 pandemic has significantly impacted our lives, with some people hoping that the situation will improve when summer arrives. Thus, we use the proposed method to study the association between COVID-19 occurrences and climate in $n = 119$ U.S. counties, roughly consisting of the top three to five counties in each state for COVID-19 cases, with sufficient climate records in March. The outcome Y_i is the number of cumulative cases per thousand people on April 1st from the ith county. The corresponding covariate $\mathbf{X}_i$ has dimension 94, with its first component 1 to account for the intercept effect, and the remaining components are 93 climate variables. Specifically, in the period 2020/03/01 to 2020/03/31, we use $X_{i2}, \dots, X_{i32}$ to denote daily precipitation, $X_{i33}, \dots, X_{i63}$ to denote daily average temperature (the mean of the minimal and maximal daily temperatures), and $X_{64}, \dots, X_{94}$ to denote the daily temperature change (the difference between the maximal and minimal temperatures). The true $\mathbf{X}_i$ is usually not available. Based on information from the National Climatic Data Center (`http://www.ncdc.noaa.gov`), we collect data from multiple sensors in each county at different locations. Denote the reading of the sensor at the jth location in county i as $\mathbf{Z}_{ij}$. We assume $\mathbf{Z}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}$, where $\mathbf{U}_{ij}$ is a measurement error. Let $\mathbf{W}_i = \sum_{j=1}^{n_j} \mathbf{Z}_{ij}/n_i = \mathbf{X}_i + \mathbf{U}_i$ be the observed climate data, where $\mathbf{U}_i = \sum_{j=1}^{n_i} \mathbf{U}_{ij}/n_i$ and n_i is the number of sensors in county i. We use the duplicated measurements $\mathbf{Z}_{ij}$ to estimate the covariance of $\mathbf{U}_i$ as

Table 3. Performance of the estimators over 100 simulation runs when the distributions of $\mathbf{U}$ are, respectively Student's t and gamma distributions. The result is calculated as the mean of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ over 100 simulation times. Naive and Regcal stand for the naive and regression calibration estimators, respectively.

		$k = 6$			$k = 10$		
n	p	Naive	Regcal	MPR	Naive	Regcal	MPR
		Student t distribution					
100	64	3.102	2.313	1.901	4.398	3.483	3.021
	128	3.294	2.577	2.041	4.609	3.949	3.492
	256	3.412	2.732	2.290	4.696	4.155	4.198
200	64	3.169	2.281	1.690	4.050	3.024	2.421
	128	3.048	2.082	1.402	4.173	3.032	2.380
	256	3.249	2.289	1.698	4.458	3.314	2.705
		Gamma distribution					
100	64	2.162	1.811	1.194	3.108	2.883	2.111
	128	2.390	1.996	1.424	3.374	3.083	2.499
	256	2.541	2.194	1.611	3.693	3.372	2.866
200	64	2.072	1.568	0.872	2.746	2.252	1.516
	128	1.989	1.525	0.871	3.098	2.659	1.876
	256	2.137	1.667	0.945	2.982	2.560	1.811

$\widehat{\boldsymbol{\Omega}} = \sum_{i=1}^{n} \sum_{j=1}^{n_i} n_i^{-2} \{\mathbf{Z}_{ij} - \sum_{j=1}^{n_j} \mathbf{Z}_{ij}/n_i\}^{\otimes 2}/n$.

We first compare the prediction performance of the MPR method and the naive method (Poisson Lasso regression) using three-fold cross-validation with 70% training data and 30% testing data. For the naive method, we use $\exp(\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{W}_i)$ as a predictor for Y_i in the testing sample, where $\widetilde{\boldsymbol{\beta}}$ is the Poisson Lasso regression estimator based on the training sample. For the MPR method, we use $\exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{W}_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{\beta}}/2)$ as a predictor for Y_i in the testing sample, where $\widehat{\boldsymbol{\beta}}$ is the MPR estimator based on the training sample. The tuning parameters are selected based on the strategies described in Section 3.

We plot the box plots of the mean absolute errors of the two methods from 100 cross-validations in Figure 3. Figure 3 shows that the MPR method outperforms the naive method, with a significantly smaller prediction error. We further show the parameter estimators from the two methods and the corresponding 90% confidence intervals from 1,000 bootstraps in Figures 4 and 5. Figure 5 shows that the naive method picks up the average temperature on March 30th and the temperature changes on March 15th and 30th as important predictors for Y_i. The estimated effects and the 90% confidence intervals are -0.102 (-0.003, -0.216), -0.063 (-0.006, 0.125), and -0.056 (-0.001, -0.114), respectively. However,

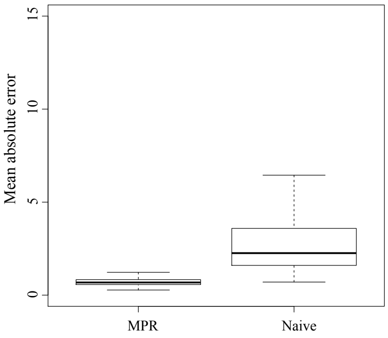

Figure 3. Mean prediction errors from 100 cross validations.

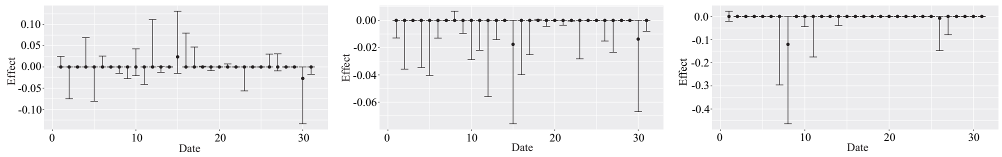

Figure 4. Estimated effect from the MPR method for the average temperature (top left), temperature change (top right), and precipitation (bottom). The error bars represent 90% confidence intervals. No significance is detected.

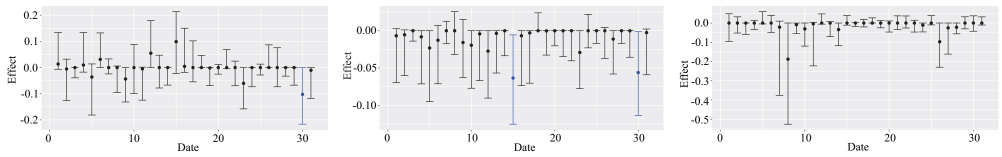

Figure 5. Estimated effect from the naive method for the average temperature (top left), temperature change (top right), and precipitation (bottom). The error bars represent 90% confidence intervals. Blue represents negative significance.

after adjusting the measurement errors, the MPR method suggests that neither the precipitation nor the temperatures have any significant effect on the number of cases on April 1st. Because the significance captured by the naive method may attribute to the measurement error, there is no strong evidence to support that climate changes will mitigate the spread of COVID-19.

To show the robustness of the MPR method, we implement the MPR and the naive methods on a set of error-perturbed artificial data. To generate the artificial data, we simulate the noise from mean zero multivariate normal distributions with covariance matrix $\gamma\widehat{\boldsymbol{\Omega}}$, where $\gamma = 0.7, 1.1, 1.5$. Then, we add the noise to the orignal design matrix to form the new design matrices with increased error

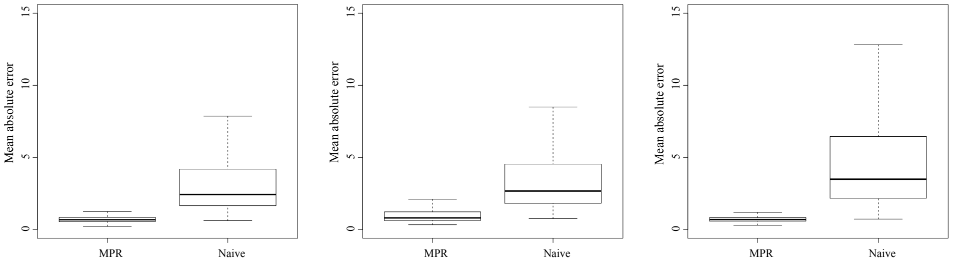

Figure 6. Mean prediction errors from 100 cross-validations for $\gamma = 0.7, 1.1, 1.5$ from left to the right.

corruption. Figure 6 contains the prediction errors from the three-fold cross-validations, which suggests that the MPR is consistently better than the naive method, producing smaller prediction errors. In addition, the MPR is robust to the error corruptions with consistently low prediction errors, while the prediction errors for the naive method increase with the noise levels. Furthermore, we show the estimated covariate effects and their confidence intervals under each setting in Figure 7 and Figure 8, respectively. Figure 7 shows that the results from the MPR are coherent across all settings, suggesting there is no significant association between the outcome and the covariates. On the other hand, Figure 8 shows that the naive method leads to different conclusions when γ changes. These results further justify that the MPR is protected from the covariate error contamination. On the other hand, the naive estimator generates a large deviation from the truth by ignoring the error corruptions.

7. Conclusion

Count outcomes are one of the most frequently used endpoint results in infectious disease studies and are attracting increasing attention in epidemiology. The corresponding regression models often naturally have high-dimensional features, and measurement errors in these features are often unavoidable. However, the Poisson regression model, which is the standard model used to handle count data responses, is not studied in this context, owing to its difficulties.

To fill this gap, we study the Poisson regression under the high-dimensional covariate setting with errors in the features. We construct an explicit objective function and devise a computational algorithm under the sparseness assumption of the parameters. By adding the measurement error structure, the proposed model corrects the erroneous results obtained from the naive error-free treatment.

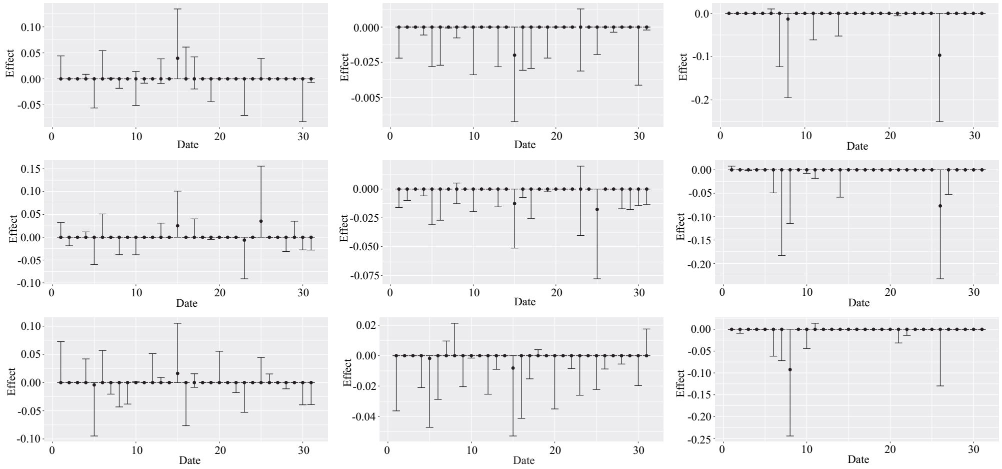

Figure 7. Estimated effect from the MPR method for the average temperature (left), temperature change (middle), and precipitation (right), and for $\gamma = 0.7$ (top), $\gamma = 1.1$ (middle), and $\gamma = 1.5$ (bottom). The error bars represent 90% confidence intervals.

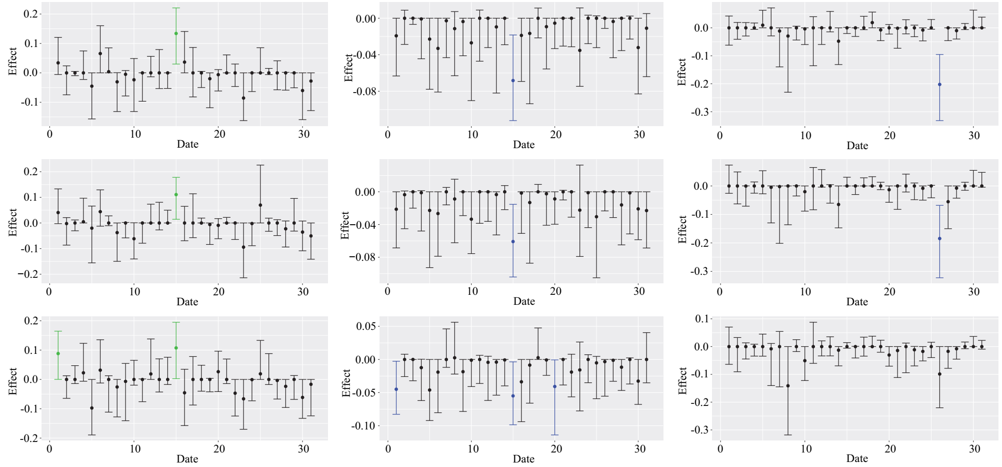

Figure 8. Estimated effect from the naive method for the average temperature (left), temperature change (middle), and precipitation (right), and for $\gamma = 0.7$ (top), $\gamma = 1.1$ (middle), and $\gamma = 1.5$ (bottom). The error bars represent 90% confidence intervals. Green and blue represent the positive and negative significances, respectively.

We hope further research will extend our work to other regression models and more complex measurement error structures.

Supplementary Material

The online Supplementary Material includes the comprehensive proofs of all theoretical results.

References

Agarwal, A., Negahban, S. and Wainwright, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics* **40**, 2452–2482.

Belloni, A. and Rosenbaum, M. (2016). An $\{l_1,\ l_2,\ l_\infty\}$-regularization approach to high-dimensional errors-in-variables models. *Electronic Journal of Statistics* **10**, 1729–1750.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37**, 1705–1732.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M. and Stefanski, L. A. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective.* Chapman and Hall/CRC, New Tork.

Cook, J. and Stefanski, L. A. (1994). A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association* **89**, 1314–1328.

Datta, A. and Zou, H. (2017). Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics* **45**, 2400–2426.

Duchi, J., Shalev-Shwartz, S., Singer, Y. and Chandra, T. (2008). Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, 272–279. ACM, New York.

Dziugaite, G. K. and Roy, D. M. (2015). Neural network matrix factorization. *arXiv preprint arXiv:1511.06443.*

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

Fuller, W. A. (1987). *Measurement Error Models.* Willey, New York.

Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448. IEEE, Washington, D.C.

Gopalan, P., Hofman, J. M. and Blei, D. M. (2013). Scalable recommendation with Poisson factorization. *arXiv preprint arXiv:1311.1704.*

Gopalan, P. K., Charlin, L. and Blei, D. (2014). Content-based recommendations with Poisson factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems* (Edited by Z. Ghahramani, M. Welling, C. Cortes and N. D. Lawrence) **2**, 3176–3184. MIT Press, Cambridge.

He, K., Gkioxari, G., Dollár, P. and Girshick, R. (2017a). Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X. and Chua, T.-S. (2017b). Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, 173–182. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva.

Jüni, P., Rothenbühler, M., Bobos, P., Thorpe, K. E., da Costa, B. R., Fisman, D. N. et al. (2020). Impact of climate and public health interventions on the COVID-19 pandemic: A prospective cohort study. *CMAJ* **192**, E566–E573.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics* **40**, 1637–1664.

Negahban, S., Yu, B., Wainwright, M. J. and Ravikumar, P. K. (2009). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems* (Edited by Y. Bengio, D. Schuurmans, J. D. Lafferty, C K. I. Williams and A. Culotta), 1348–1356. Cuuran Associates Inc., Red Hook.

Shi, C., Song, R., Chen, Z. and Li, R. (2019). Linear hypothesis testing for high dimensional generalized linear models. *The Annals of Statistics* **47**, 2671–2703.

Srebro, N., Rennie, J. and Jaakkola, T. S. (2004). Maximum-margin matrix factorization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems* (Edited by L. K. Saul, Y. Weiss and L. Bottou), 1329–1336. MIT Press, Cambridge.

Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika* **74**, 703–716.

Tosepu, R., Gunawan, J., Effendy, D. S., Lestari, H., Bahar, H., Asfian, P. et al. (2020). Correlation between weather and COVID-19 pandemic in Jakarta, Indonesia. *Science of The Total Environment* **725**, 138436.

Van De Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* **3**, 1360–1392.

Van den Oord, A., Dieleman, S. and Schrauwen, B. (2013). Deep content-based music recommendation. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* (Edited by C. J. C Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger), 2643–2651. Curran Associates Inc., Red Hook.

Volkovs, M., Yu, G. and Poutanen, T. (2017). DropoutNet: Addressing cold start in recommender systems. In *Advances in Neural Information Processing Systems*, 4957–4966.

Wang, Y., Feng, D., Li, D., Chen, X., Zhao, Y. and Niu, X. (2016). A mobile recommendation system based on logistic regression and gradient boosting decision trees. In *2016 IJCNN*, 1896–1902. IEEE, Washington, D.C.

Fei Jiang
School of Medicine, UCSF, San Francisco, CA 94158, USA.
E-mail: fei.jiang@ucsf.edu
Yanyuan Ma
Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA.
E-mail: yanyuanma@gmail.com

(Received June 2020; accepted March 2021)