Optimal model averaging for divergent-dimensional Poisson regressions

February 16, 2022

Jiahui Zou[1], Wendun Wang[2], Xinyu Zhang[3,4] and Guohua Zou[5]

[1]*School of Statistics, Capital University of Economics and Business*

[2]*Econometric Institute, Erasmus University Rotterdam; Tinbergen Institute*

[3]*Academy of Mathematics and Systems Science, Chinese Academy of Sciences*

[4]*Beijing Academy of Artificial Intelligence*

[5]*School of Mathematical Sciences, Capital Normal University*

Proposed running head: Optimal averaging for Poisson regressions

Corresponding author:
Wendun Wang
Econometric Institute
Erasmus University Rotterdam
3062 PA Rotterdam
The Netherlands
e-mail: wang@ese.eur.nl

Abstract: This paper proposes a new model averaging method to address model uncertainty in Poisson regressions, allowing the dimension of covariates to increase with the sample size. We derive an unbiased estimator of the Kullback-Leibler (KL) divergence to choose averaging weights. We show that when all candidate models are misspecified, the proposed estimate is asymptotically optimal by achieving the least KL divergence among all possible averaging estimators. In another situation where correct models exist in the model space, our method can produce consistent coefficient estimates. We apply the proposed techniques to study the determinants and predict corporate innovation outcomes measured the number of patents.

JEL Classification: C13, C25, C52, C53, G30

Keywords: Asymptotic optimality; Consistency; Divergent dimension; Model averaging; Poisson regression;

1 Introduction

Poisson regressions have become a standard tool to model count dependent variables, and are widely used in many economic and financial studies. For example, researchers and decision makers exhibit a keen interest in understanding what affects innovation outcomes and how to predict the outcomes, because innovation is often regarded as a driving force of countries' or firms' long-run growth and performance (Aghion and Howitt, 1992; Kogan et al., 2017; Hochberg et al., 2018). This interest has yielded a vast body of literature investigating the determinants of innovation from various perspectives, e.g., incentives (Coles et al., 2006), managerial personality (Sunder et al., 2017), and firm characteristics (Fang et al., 2014). Since the most popular measure of innovation outcome is the number of patents or citations, both of which are count variables, Poisson regressions are widely employed to explain and predict the innovation outcomes.

As in many regression analyses, it is often uncertain which covariates should be included in Poisson regression models. In the innovation example, there are typically a large number of covariates that are potentially related with innovation outcomes, but such relations are not always clear and often depend on the research questions and datasets used. A large set of potential covariates introduces great model uncertainty, and thus it is a vital question of how to address model uncertainty in Poisson regressions.

This paper proposes a new method to address model uncertainty in Poisson regressions. We employ a model averaging (MA) technique that combines the estimates from multiple models with certain weights. To choose appropriate weights, we employ a Kullback-Leibler (KL)-based criterion that is an *unbiased* estimator of the KL divergence. When all candidate models are misspecified, we show that minimizing this criterion leads to asymptotically optimal weights that achieve the minimum KL-type divergence as the infeasible best possible model averaging estimator. Such asymptotic optimality provides theoretical ground for our model averaging prediction. Moreover, when the set of candidate models happens to include correct models, we show that our model averaging estimates of slope coefficients are consistent. Importantly, in both cases, we allow the number of covariates to increase as the sample size increases, and thus, the dimension and the number of candidate models also diverge. Hence, our approach is particularly useful in the applications where more covariates can be included in the model as new observations become available and are added to the sample.

Model averaging offers an attractive method to address model uncertainty. The conventional solution to model uncertainty is to select the "best" model based on some data-driven criteria, e.g., information criteria, and the estimators obtained from the selected model are called pretest estimators. Such estimators clearly separate selection and estimation in two steps and thus suffer from unbounded risk (Magnus, 2002). Moreover, the standard inference of these estimators ignores the uncertainty emerging from the selection step (Danilov and Magnus, 2004), although some recent advances may allow us to capture additional sampling variability introduced by the selection step (Berk et al., 2013; Charkhi and Claeskens, 2018). Model averaging (MA) addresses model uncertainty from a different perspective. Rather than relying on a single best model, MA accounts for all candidate models and averages their estimates with weights that can reflect model performance. Thus, it is an integrated procedure where both model and estimation uncertainty are taken into account. Unlike the pretest estimates, model averaging estimates are continuous and unconditional and have substantially less risk (Hansen, 2014).

There are two streams of model averaging approaches: one from the Bayesian perspective and the other on the frequentist basis. Bayesian model averaging is flexible and works for a wide range of models (see Hoeting et al., 1999, for a comprehensive review), but the choice of appropriate priors is often not clear and experiential. Frequentist model averaging (FMA) methods can be further divided into two categories, differing in the purpose of combination. The first category combines for adaption (see Yang, 2001; Yuan and Yang, 2005; Wei and Yang, 2012, among others). More specifically, the purpose of combination for these methods is to approach the performance of the best single model, and thus, the theoretical justification primarily focuses on the risk property of the averaging estimate. The second category of FMA methods combines estimates for the purpose of outperforming any single model, and therefore, they are sometimes referred to as combination for improvement or optimal model averaging. Steel (2020) provided an overview of these two types of model averaging methods and their recent economic applications. For these methods, the optimality of weight choices is often of particular interest. Our approach falls into the second category and is intended to outperform any single model via combination. It is relevant in practice since most (single) empirical models are misspecified.

We contribute to the model averaging literature in terms of two main aspects. First, we propose an *unbiased* criterion to choose averaging weights for Poisson regressions, which allows for a divergent number of covariates. Existing optimal model averaging methods primarily focus on linear models; examples include Mallows model averaging (Hansen, 2007), jackknife model averaging (Hansen and Racine, 2012), heteroskedasticity robust C_p (Liu and Okui, 2013), quantile regression averaging (Lu and Su, 2015), prediction model averaging (Xie, 2015), predictive regression averaging (Liu and Kuo, 2016), and functional data model averaging (Zhang et al., 2018). These techniques cannot be directly applied to Poisson regressions because the model is estimated using the maximum likelihood based on a Poisson distribution function, and the asymptotic optimality of above averaging methods (for linear models) no longer holds. Moreover, these existing studies assume that the dimension of each candidate model is fixed, which could be a restrictive assumption when more variables become available as the sample size increases. We extend optimal model averaging to Poisson regression, a form of generalized linear models (GLMs), and allow the number of candidate models and the dimension of each candidate model to diverge as the sample size increases. Since the KL divergence is a common measure of model performance for generalized linear models (Zhang et al., 2016; Ando and Li, 2017), we employ perturbation techniques and develop an unbiased criterion based on the KL divergence to determine the weights and show their asymptotic optimality.

Recently, Zhang et al. (2016) and Ando and Li (2017) studied optimal model averaging for GLMs. Zhang et al. (2016) proposed a weight choice criterion based on a penalized (negative) likelihood function. This criterion is, however, a biased approximation of KL divergence since it is equivalent to the KL divergence plus a penalty term. Ando and Li (2017) proposed to determine the weights by minimizing leave-one-out cross-validation. Although asymptotic optimality of the resulting weights has been established for both methods, neither of the the criteria are unbiased for the KL divergence, and thus, the choice of weights does not directly minimize the KL divergence. We differ from these two studies by focusing on a specific GLM, i.e., Poisson regression. This setup allows us to develop a weight choice criterion that is a precisely unbiased estimator of the KL divergence. Related studies also include De Luca et al. (2018) and Charkhi et al. (2016). De Luca et al. (2018) proposed a weighted average least squares (WALS) procedure for GLMs, which is a combination of frequentist and Bayesian approaches and thus does not consider the asymptotic properties of the weight choice. Its finite-sample sampling properties have been further studied in De Luca et al. (2020). Charkhi et al. (2016) considered frequentist model averaging for general likelihood models but employed a local misspecification framework that is not assumed in our case. These two averaging estimators mainly concern the bias-variance trade-off in parameter estimation (Hjort and Claeskens, 2003). In contrast, our averaging method targets the asymptotic optimality of prediction when all candidate models are misspecified as well as the consistency of parameters when at least one candidate models is correctly specified. Another important difference from these extant studies on GLM averaging is that we allow the number and the dimension of candidate models to diverge. Specifically, De Luca et al. (2018); Charkhi et al. (2016); Zhang et al. (2016) assumed that the number of covariates and/or the number of candidate models is finite. While Ando and Li (2017) allowed potentially high-dimensional covariates, they first sort and group the covariates based on their bivariate relevance with the outcome variable, and then only average candidate models that contain finite and the most relevant groups of covariates (discarding the less relevant ones). In contrast, our averaging scheme explicitly allows both the dimension of each candidate model and the number of candidate models to diverge.

Our second theoretical contribution is to study the asymptotic property of averaging estimates of slope coefficients in Poisson regressions when the candidate models include correct models. Zhang et al. (2020) showed the consistency of averaging coefficient estimates when at least one correct model exists in the candidate model set, but they only concern linear models. We extend this argument to Poisson regressions. This extension is technically challenging since our estimation is achieved by maximizing the likelihood function and the coefficient estimates do not have analytical solutions. We overcome these challenges by employing a very different technique from Zhang et al. (2020) that relies on the consistency of weights to prove the coefficient consistency. This result complements the asymptotic optimality when all candidate models are misspecified, and demonstrates the validity of our method in the situation with correct models existing in the model space. To the best of our knowledge, no consistency results have been established for GLM model averaging, and the current paper provides the first study regarding Poisson model averaging.

A simulation study containing various designs of experiments confirms our theoretical results and demonstrates the advantages of the proposed method over several popular model selection and averaging methods. We apply the method to study the innovation outcome measured by the number of patents using the U.S. corporate data. Given a large number of potential determinants of innovation, there is great model uncertainty in the Poisson regression of the number of patents on innovation determinants. We show that our KL-based model averaging performs well in predicting corporate innovation outcomes. We also examine the possible determinants of innovation outcome, particularly focusing on the role of CEOs holding a pilot licence (Sunder et al., 2017). Our full model estimation using all covariates confirms the significance of pilot CEO coefficient reported by Sunder et al. (2017). However, when we take into account model uncertainty, the model averaging estimate of pilot CEO effect is less salient. Further examination reveals that pilot CEO is strongly correlated with some of other managerial characteristics, such as her wealth and technical education background, which seem redundant in explaining innovation outcomes. This result suggests a large degree of model uncertainty in the innovation regression, such that the strong and significant association between pilot CEOs and innovation outcomes reported by Sunder et al. (2017) should be interpreted with great caution.

The remainder of the paper is organized as follows. Section 2 sets up the model and presents the model averaging method. Section 3 establishes asymptotic results when all candidate models are misspecified and when correct models are included in the model space. Section 4 evaluates the finite sample performance of the proposed method and compares it with model selection and other averaging methods. Section 5 applies the proposed method to revisit the relation between corporate innovation outcome and managerial risk-taking preference. Section 6 concludes. Finally, technical proofs are provided in the Supplementary Material.

2 Model setup and estimation

This section first presents the setup of our model and then proposes a new model averaging method to estimate this model.

2.1 Averaging estimates in Poisson regression

Suppose we observe a dataset $\{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$, where y_i is the realization of a count variable Y_i measuring the number of occurrences of an event in a given interval for the i^{th} individual, $\mathbf{x}_i$ is a p-dimensional vector, and n denotes the number of observations. We allow $\mathbf{x}_i$ to be divergent-dimensional in the sense that p may increase as $n \to \infty$, but $p < n$. Our interest is to explain y_i with the potential determinants in $\mathbf{x}_i$ and to predict y_i. For these purposes, a Poisson distribution is commonly used, which associates the probability of Y_i events with $\mathbf{x}_i$ as follows:

$$\Pr(Y_i = y_i|\mu_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}, \tag{1}$$

where μ_i is the Poisson incidence rate that depends on $\mathbf{x}_i$ as $\mu_i = \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ and $\boldsymbol{\beta}$ is the associated $p \times 1$ coefficient vector.

In practice, not all covariates are "useful" in predicting y_i, and researchers are typically uncertain regarding which should be included in the model *ex ante*. Hence, a number of candidate models with different specifications of $\mathbf{x}_i$ are considered. Let S be the number of candidate models. Typically, with p covariates in $\mathbf{x}_i$, we have $S = 2^p - 1$ models. However, when p is great, we may consider a model screening step prior to model averaging, resulting in $S < 2^p - 1$. Let Π_s be a $p \times p_s$ selection matrix that consists of 0's and 1's and selects p_s covariates ($p_s \leq p$) for the s^{th} model. Denote the p_s-dimensional covariate vector in the s^{th} model by $\mathbf{x}_{(s),i}^{\mathrm{T}} = \mathbf{x}_i^{\mathrm{T}}\Pi_s$. We model the Poisson incidence rate of this candidate model as

$$\mu_{(s),i} = \exp(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^*), \tag{2}$$

where $\boldsymbol{\beta}_{(s)}^*$ is the quasi-true parameter (White, 1982) that minimizes the KL divergence between the density (1) and the density of the s^{th} model. We estimate $\boldsymbol{\beta}_{(s)}^*$ via the maximum likelihood (ML) method and denote the resulting estimate as $\widehat{\boldsymbol{\beta}}_{(s)}$. Note that since p is allowed to increase with the sample size n, the dimension of each submodel, p_s for $s = 1, 2, \ldots, S$, also diverges when $n \to \infty$, leading to an increasing-dimensional parameter estimation problem for all candidate models as well as a diverging number of candidate models. Thus, our setup is in sharp contrast to conventional averaging techniques for GLMs that restrict the dimension and/or the number of candidate models to be finite (Zhang et al., 2016; Charkhi et al., 2016; Ando and Li, 2017; De Luca et al., 2018).

To account for model uncertainty and avoid the problems caused by pretesting (Magnus, 2002), we propose to average the estimates obtained from each candidate model. First, we need to unify the dimension of the coefficient estimates of each candidate model by $\widehat{\boldsymbol{\beta}}_s = \Pi_s \widehat{\boldsymbol{\beta}}_{(s)}$. Then, we can compute the model averaging estimates of coefficient $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}}(\mathbf{w}) = \sum_{s=1}^{S} w_s \widehat{\boldsymbol{\beta}}_s, \tag{3}$$

where $\mathbf{w} = (w_1, w_2, ..., w_S)^{\mathrm{T}}$ is the weight vector belonging to the set $\mathcal{W}_n = \{\mathbf{w} \in [0,1]^S : \sum_{s=1}^{S} w_s = 1\}$. The model averaging problem that we considered here differs from the literature in that we consider *generalized* linear models with a *divergent* dimension as the sample size increases.

2.2 Weight choice criterion

The model averaging estimates in (3) depends on the choice of weights, and this section provides a feasible and data-driven weight choice. Since each candidate model is estimated via maximum likelihood with a Poisson distribution function, it is natural to consider a KL-type criterion, which measures the divergence between the fitted and true density functions.

Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^{\mathrm{T}}$. The averaging estimator of μ_i can be obtained by

$$\widehat{\mu}_i(\mathbf{w}, \mathbf{y}) = \exp\left\{\mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}(\mathbf{w})\right\}. \tag{4}$$

We stack $\widehat{\mu}_i(\mathbf{w}, \mathbf{y})$ in a vector and denote $\widehat{\boldsymbol{\mu}}(\mathbf{w}, \mathbf{y}) = (\widehat{\mu}_1(\mathbf{w}, \mathbf{y}), \widehat{\mu}_2(\mathbf{w}, \mathbf{y}), ..., \widehat{\mu}_n(\mathbf{w}, \mathbf{y}))^{\mathrm{T}}$. The KL loss function for two independent sets of realizations, $\mathbf{y}$ and $\mathbf{y}^*$, is defined as

$$\begin{aligned}
\mathrm{KL}(\mathbf{w}) &= E_{f(\mathbf{y}^*)} \log\left[f(\mathbf{y}^*)/g\left\{\mathbf{y}^* | \widehat{\boldsymbol{\mu}}(\mathbf{w}, \mathbf{y})\right\}\right] \\
&= E_{f(\mathbf{y}^*)} \log f(\mathbf{y}^*) - E_{f(\mathbf{y}^*)} \log g\left\{\mathbf{y}^* | \widehat{\boldsymbol{\mu}}(\mathbf{w}, \mathbf{y})\right\} \\
&= \sum_{i=1}^{n} \left[\mu_i \log(\mu_i) - \mu_i - \mu_i \log\{\widehat{\mu}_i(\mathbf{w}, \mathbf{y})\} + \widehat{\mu}_i(\mathbf{w}, \mathbf{y})\right],
\end{aligned} \tag{5}$$

where $f(\mathbf{y}^*)$ is the density of $\mathbf{y}^*$, $g\left\{\mathbf{y}^* | \widehat{\boldsymbol{\mu}}(\mathbf{w}, \mathbf{y})\right\}$ is the conditional density of $\mathbf{y}^*$ given the fitted model using observations $\mathbf{y}$, and the expectation $E_{f(\mathbf{y}^*)}$ is taken over $\mathbf{y}^*$. We suggest a weight choice criterion by estimating the KL-type risk function $E_{f(\mathbf{y})}\left[\mathrm{KL}(\mathbf{w})\right]$ using perturbation techniques.

Define $\mathbf{y}^{(y_i-1)} = (y_1, \dots, y_{i-1}, y_i - 1, y_{i+1}, \dots, y_n)^{\mathrm{T}}$, which replaces the i^{th} element of $\mathbf{y}$ by $y_i - 1$. The model averaging estimate of the Poisson incidence rate using the data $\mathbf{y}^{(y_i-1)}$ is then given by

$$\widehat{\boldsymbol{\mu}}(\mathbf{w}, \mathbf{y}^{(y_i-1)}) = \left(\widehat{\mu}_1(\mathbf{w}, \mathbf{y}^{(y_i-1)}), \widehat{\mu}_2(\mathbf{w}, \mathbf{y}^{(y_i-1)}), ..., \widehat{\mu}_n(\mathbf{w}, \mathbf{y}^{(y_i-1)})\right)^{\mathrm{T}}.$$

With the approximated incidence rate readily there, we can estimate the KL-type risk function $E_{f(\mathbf{y})}\left[\mathrm{KL}(\mathbf{w})\right]$ by

$$\mathcal{C}(\mathbf{w}) = \log f(\mathbf{y}) + \sum_{i=1}^{n} \left[\widehat{\mu}_i(\mathbf{w}, \mathbf{y}) + \log(y_i!) - y_i \log\{\widehat{\mu}_i(\mathbf{w}, \mathbf{y}^{(y_i-1)})\}\right]. \tag{6}$$

Note that when $y_i = 0$, $y_i \log\{\widehat{\mu}_i(\mathbf{w}, \mathbf{y}^{(y_i-1)})\} = 0$. By the Stein-Chen lemma (Chen, 1975; Chen et al., 2010), we have the following lemma.

Lemma 1 *If* $\mathbf{y}$ *is generated from a Poisson regression as in* (1)*, we have that*

$$E_{f(\mathbf{y})}\left[\mathcal{C}(\mathbf{w})\right] = E_{f(\mathbf{y})}\left[\mathrm{KL}(\mathbf{w})\right]. \tag{7}$$

Proof. See Appendix A.1.

This lemma states that our weight-choice criteria $\mathcal{C}(\mathbf{w})$ is an unbiased estimator of the KL-type risk $E_{f(\mathbf{y})}\left[\mathrm{KL}(\mathbf{w})\right]$ as long as the underlying distribution of dependent variable is Poisson, and thus minimizing $\mathcal{C}(\mathbf{w})$ is asymptotically equivalent to minimizing the KL-type risk. The unbiasedness property provides the first justification of the validity of our criterion (see Hansen, 2007, for a similar argument for the Mallows criterion in linear regressions).

Removing terms in $\mathcal{C}(\mathbf{w})$ that do not depend on $\mathbf{w}$, we obtain the following feasible criterion:

$$\mathcal{C}^*(\mathbf{w}) = -\log g\{\mathbf{y}|\widehat{\boldsymbol{\mu}}(\mathbf{w}, \mathbf{y})\} + \sum_{i=1}^{n} \left[y_i \log\left\{\widehat{\mu}_i(\mathbf{w}, \mathbf{y})\right\} - y_i \log\left\{\widehat{\mu}_i(\mathbf{w}, \mathbf{y}^{(y_i-1)})\right\}\right]. \tag{8}$$

Thus, we can choose weights by minimizing $\mathcal{C}^*(\mathbf{w})$, i.e.,

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathcal{W}_n} \mathcal{C}^*(\mathbf{w}). \tag{9}$$

Note that our criterion (8) is not a special case of the weight choice criterion for GLM given by (3) of Zhang et al. (2016) since the second term of $\mathcal{C}^*(\mathbf{w})$ depends on the weights nonlinearly, while the penalty term in criterion (3) of Zhang et al. (2016) is a linear function of the weights. With the estimated weights available, we can obtain the averaging estimates of the coefficient and Poisson incidence rate by $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$ and $\widehat{\mu}_i(\widehat{\mathbf{w}}, \mathbf{y}) = \exp\{\mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})\}$, respectively.

3 Asymptotic properties

Next, we examine the asymptotic properties of the proposed averaging estimates. We consider two situations: (1) all candidate models are misspecified and (2) the set of candidate models includes the correct (but not necessarily the true) models. We show that if all candidate models are misspecified, our weight choice is asymptotically optimal in the sense that it yields a KL loss that is asymptotically identical to that resulted from the infeasible best possible MA estimator. In the second situation with correct models included in the set of candidate models, we establish the consistency of MA estimators of slope coefficients.

3.1 Asymptotic optimality

Some regularity conditions are needed to show the asymptotic optimality of the weights.

Condition 1 *For any $s \in \{1, \dots, S\}$, the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_{(s)}$ exists, and the following equation about $\boldsymbol{\beta} \subset \mathcal{R}^{p_s}$ has a solution:*

$$\sum_{i=1}^{n} \left\{ \exp(\mathbf{x}_{(s),i}^{\mathrm{T}} \boldsymbol{\beta}) - \mu_i \right\} \mathbf{x}_{(s),i} = 0. \tag{10}$$

This condition guarantees the existence of maximum likelihood estimate $\widehat{\boldsymbol{\beta}}_{(s)}$. It is a high level condition but can be satisfied under some weak assumptions according to Shao (2003). The solution of (10) is the quasi-true parameter $\boldsymbol{\beta}^*_{(s)}$ since it minimizes the KL divergence between the density (1) and the density of the s^{th} model. The uniqueness of the solution will be discussed by Lemma 3 in Appendix A.1.

Condition 2 *There exist positive constants C_1 and C_2 such that*

$$0 < C_1 < \frac{\|\boldsymbol{\mu}\|^2}{n} < C_2 < \infty. \tag{11}$$

Condition 2 restricts the variability of the Poisson incidence rate, and inequality (11) here is the same as Condition (8) of Ando and Li (2014) and Condition (A4) of Ando and Li (2017). Moreover, if y_i is generated from a Poisson distribution, considering that $n^{-1} \sum_{i=1}^{n} \mathrm{Var}(y_i) = n^{-1} \sum_{i=1}^{n} \mu_i \leq \sqrt{\|\boldsymbol{\mu}\|^2/n}$, (11) also imposes a restriction on the upper bound of the variation of y_i.

Condition 3 *There exist positive constants C_3, C_4, C_5 and ρ such that*

$$\max_{1\le s\le S}\max_{1\le i\le n}\frac{\|\mathbf{x}_{s,i}\|}{\sqrt{p_s}} \le C_3 < \infty, \tag{12}$$

$$\max_{1\le s\le S}\sup_{\boldsymbol{\beta}_{(s)}\in O(\boldsymbol{\beta}^*_{(s)},\rho)}\frac{1}{n}\sum_{i=1}^{n}\exp(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}) \le C_4 < \infty, \tag{13}$$

$$\max_{1\le s\le S}\frac{\|\boldsymbol{\beta}^*_{(s)}\|}{\sqrt{p_s}} \le C_5 < \infty, \tag{14}$$

where $O(\boldsymbol{\beta}^*_{(s)},\rho)$ is a neighborhood of $\boldsymbol{\beta}^*_{(s)}$, i.e., $\{\boldsymbol{\beta}_{(s)}\in\mathcal{R}^{p_s} : \|\boldsymbol{\beta}_{(s)}-\boldsymbol{\beta}^*_{(s)}\|\le\rho\}$.

Condition 4 *Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ be the minimum and maximum eigenvalues, respectively, and define*

$$I_{(s)}(\boldsymbol{\beta}_{(s)}) = \frac{1}{n}\sum_{i=1}^{n}\exp\left(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}\right)\mathbf{x}_{(s),i}\mathbf{x}_{(s),i}^{\mathrm{T}}.$$

There exists a $\rho>0$ and two positive constants $C_{\min}$ and $C_{\max}$ such that

$$\min_{1\le s\le S}\inf_{\boldsymbol{\beta}_{(s)}\in O(\boldsymbol{\beta}^*_{(s)},\rho)}\lambda_{\min}\left\{I_{(s)}(\boldsymbol{\beta}_{(s)})\right\} \ge C_{\min} > 0 \tag{15}$$

and

$$\max_{1\le s\le S}\sup_{\boldsymbol{\beta}_{(s)}\in O(\boldsymbol{\beta}^*_{(s)},\rho)}\lambda_{\max}\left\{I_{(s)}(\boldsymbol{\beta}_{(s)})\right\} \;\le\; C_{\max} < \infty, \tag{16}$$

where $O(\boldsymbol{\beta}^*_{(s)},\rho)$ is a neighborhood of $\boldsymbol{\beta}^*_{(s)}$, i.e., $\{\boldsymbol{\beta}_{(s)}\in\mathcal{R}^{p_s} : \|\boldsymbol{\beta}_{(s)}-\boldsymbol{\beta}^*_{(s)}\|\le\rho\}$.

Conditions 3 and 4 restrict the variability of covariates and coefficients. The assumption stated in (12) resembles Assumption (A2) of Liang and Du (2012). When $\rho=0$, (13) implies that

$$\max_{1\le s\le S} n^{-1}\sum_{i=1}^{n}\exp(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}^*_{(s)}) = \max_{1\le s\le S} n^{-1}\sum_{i=1}^{n}\mu_{(s),i} \le C_4,$$

which reasonably controls $\boldsymbol{\mu}_{(s)}$. Combining (13) and (16) can lead to Condition (C.4) of Zhang et al. (2016). Condition 4 is commonly imposed to show the convergence of $\widehat{\boldsymbol{\beta}}_{(s)}$ and resembles Assumption (A2) of Liang and Du (2012), Conditions 1-2 in Lv and Liu (2014, Theorem 6) and Condition (C.4) of Zhang et al. (2016). When $\rho=0$, the inequality (16) implies that

$$\max_{1\le s\le S}\lambda_{\max}\left(n^{-1}\sum_{i=1}^{n}\mu_{(s),i}\mathbf{x}_{(s),i}\mathbf{x}_{(s),i}^{\mathrm{T}}\right) \le \left(\max_{s,i}\mu_{(s),i}\right)\lambda_{\max}\left(n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right),$$

where the left-hand side of the inequality is bounded when $\sup_{s,i} \mu_{(s),i}$ and $\lambda_{\max}\left(n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)$ are both bounded. The latter one is an assumption widely used in the literature. See, for example, Assumption (A2) of Liang and Du (2012).

Define the KL divergence based on quasi-true coefficients as

$$\mathrm{KL}^*(\mathbf{w}) \;=\; \sum_{i=1}^{n}\{-\mu_i + \mu_i \log(\mu_i)\} + \sum_{i=1}^{n}\left[\mu_i^*(\mathbf{w}) - \mu_i \log\{\mu_i^*(\mathbf{w})\}\right], \tag{17}$$

where $\mu_i^*(\mathbf{w}) = \exp(\sum_{s=1}^{S} w_s \mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^*)$ and $\boldsymbol{\beta}_{(s)}^*$ is the quasi-true parameter defined in (2).

Condition 5 *As $n \to \infty$, $Sp/n \to 0$ and $Sp^2n/\xi_n^2 \to 0$, where $\xi_n = \inf_{\mathbf{w}\in\mathcal{W}_n}\mathrm{KL}^*(\mathbf{w})$.*

Condition 5 concerns how close the candidate models can be to the true model. It requires that $\inf_{\mathbf{w}\in\mathcal{W}_n}\mathrm{KL}^*(\mathbf{w})$ grow at a rate no slower than $S^{1/2}pn^{1/2}$ while allowing p to increase with n at certain rate. This requirement implies that the candidate models cannot be too close to the true model, and it obviously rules out the scenario where the true model is included in the set of candidate models (Flynn et al., 2013). This assumption is similar to Condition (8) of Ando and Li (2014) and Condition (A3) of Ando and Li (2017).

With these conditions, we can establish the asymptotic optimality of the weights in terms of minimizing the KL loss in the following theorem.

Theorem 1 *Under Conditions 1–5,*

$$\frac{\mathrm{KL}(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}_n}\mathrm{KL}(\mathbf{w})} \to 1 \tag{18}$$

in probability as $n \to \infty$.

Proof. See Appendix A.2.

This theorem shows that the model averaging estimate of Poisson incidence rate using weights derived from (9) is asymptotically optimal in the sense that its KL divergence is asymptotically identical to that obtained from the infeasible best possible model averaging estimator. The above holds even when the functional form of μ_i is unknown, as in most applications. This fact implies that any misspecification of μ_i is allowed, including unrestricted form of omitted variables as well as a deviation between the true underlying distribution and the Poisson model. Note that our misspecification framework differs from the widely adopted local misspecification framework (Hjort and Claeskens, 2003), which restricts the order of omitted variable bias to decay as n increases.

3.2 Consistency of averaging coefficient estimates

In some applications, the set of candidate models may include the correct model but not necessarily the true model. In this case, we can establish the consistency of our averaging estimates. To this end, we first distinguish between the true and correct models. Let $\boldsymbol{\beta}_{\text{true}} = (\beta_{\text{true},1}, \beta_{\text{true},2}, ..., \beta_{\text{true},p})^{\mathrm{T}}$ be the true parameters of model (1) and denote the set of the indices of nonzero true coefficients as $\mathcal{T} = \{j : \beta_{\text{true},j} \neq 0\}$. The cardinality of $\mathcal{T}$, p_{true}, is allowed to diverge. The true model contains the covariates whose indices are in $\mathcal{T}$ and does not contain any other covariates, and thus, it is unique. Nevertheless, there may be multiple correct models, and they all nest the true model. More formally, let $\mathcal{M}_s$ be the set of indices of elements in $\mathbf{x}_{(s),i}$ for $s = 1, \ldots, S$. If $\mathcal{M}_s \supseteq \mathcal{T}$, the s^{th} candidate model is called the correct model. In contrast, if $\mathcal{M}_s \not\supseteq \mathcal{T}$, the s^{th} model is called a misspecified model.

Condition 6 *There exist two positive constants c_0 and C_6 such that*

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right) \geq c_0 > 0, \tag{19}$$

$$\max\left\{\max_{1\leq i\leq n}\max_{1\leq s\leq S}|\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_s^*|, \max_{1\leq i\leq n}|\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\text{true}}|\right\} \leq C_6 < \infty. \tag{20}$$

Condition 7 *As $n \to \infty$, $pS^{1/2}n^{-1/2} \to 0$.*

Conditions 6-7 are used to guarantee the consistency of $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$. Inequality (19) is similar to Assumption (A2) of Liang and Du (2012). Considering $\mu_{(s),i} = \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_s^*)$ and $\mu_i = \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\text{true}})$, inequality (20) essentially guarantees that the expectations of the s^{th} candidate model and the true model are bounded. Condition 7 imposes a restriction on the order among p, S and n. Compared with $Sp/n \to 0$ in Condition 5, Condition 7 needs the sample size n to be of a larger order.

Theorem 2 *Under Conditions 1–4 and 6-7, if correct models are included in the set of candidate models, then*

$$\sqrt{\frac{n}{p}}\left\|\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})-\boldsymbol{\beta}_{\text{true}}\right\|=O_P(1). \tag{21}$$

Proof. See Appendix A.3.

This theorem shows that if one or multiple correct models are included in the model space, our model averaging method can produce a consistent estimator for $\boldsymbol{\beta}$, namely $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$. This consistency result further implies that the prediction based on our KL-MA estimator is expected to perform not much worse than that of the correct model, if not better. Theorems 1 and 2 jointly suggest that our method can provide good prediction regardless of whether the model space contains correct models. The consistency of model averaging estimates has been studied by Zhang (2015), but only for linear regressions with finite dimension. Here, we consider a generalized linear model and allow for divergent dimension.

3.3 Model averaging with (ultra-)high dimensional models

So far, we have allowed a divergent dimension of covariates, but we still require that the dimension do not exceed the sample size. In some applications, the number of covariates could be sizeable or even exceed the sample size, such that estimating and averaging over all possible candidate models is (computationally) infeasible. In these cases, we can first order the covariates based on their marginal correlations with the dependent variable, and construct the candidate models by including one extra covariate at each time based on the ordering. The idea of model screening based on bivariate correlation is in a similar spirit of the "sure independence screening" proposed by Fan and Lv (2008) and Fan and Song (2010). Alternatively, we can divide the covariates in several groups based on the magnitude of their marginal correlations with the dependent variable, and then build one candidate model for each group but discard the group with correlations close to 0, which is similar to Ando and Li (2014, 2017). Both approaches are essentially pre-screening model averaging procedures, which first rule out poor candidate models and only average a subset of models with a good model fit.

To justify the pre-screening model averaging in our framework, we first show the asymptotic optimality of the resulting estimates. Let $\mathcal{D}$ be a (random) subset of $\{1, \ldots, S\}$ and $\mathcal{W}_n^s = \{\mathbf{w} \in [0,1]^S : \sum_{s \in \mathcal{D}} w_s = 1$ and $\sum_{s \notin \mathcal{D}} w_s = 0\}$ be a subset of $\mathcal{W}_n$. Note that $\mathcal{W}_n^s$ is also random due to the randomness of $\mathcal{D}$. The pre-screening model-averaging estimator based on the subset $\mathcal{D}$ is obtained by using the weight vector $\widehat{\mathbf{w}}^s = \arg\min_{\mathbf{w} \in \mathcal{W}_n^s} \mathcal{C}^*(\mathbf{w})$. We make an additional assumption:

Condition 8 *There exist a non-negative sequence of* ν_n *and a weight sequence of* $\mathbf{w}_n \in \mathcal{W}_n$ *such that* $\xi_n^{-1}\nu_n \to 0$*,* $\inf_{\mathbf{w} \in \mathcal{W}_n} \mathcal{C}^*(\mathbf{w}) = \mathcal{C}^*(\mathbf{w}_n) - \nu_n$*, and* $\Pr(\mathbf{w}_n \in \mathcal{W}_n^s) \to 1$ *as* $n \to \infty$.

This condition requires that there exists a weight sequence $\{w_n\}$ that achieves the minimum of the averaging criterion $\inf_{\mathbf{w} \in \mathcal{W}_n} \mathcal{C}^*(\mathbf{w})$ relative to the minimum KL loss $\xi_n = \inf_{\mathbf{w} \in \mathcal{W}_n} \mathrm{KL}^*(\mathbf{w})$, i.e., $\{\inf_{\mathbf{w} \in \mathcal{W}_n} \mathcal{C}^*(\mathbf{w}) - \mathcal{C}^*(\mathbf{w}_n)\}/\xi_n \to 0$, and this sequence is also contained in the set of post-screening weights, $\mathcal{W}_n^s$, with probability approaching one. Intuitively, this condition ensures that there still exists a good weight sequence after screening, such that the asymptotic optimality is valid over the entire model space composed of all candidate models. Under Conditions 1–8, we can use the same arguments as Theorem 3 of Zhang et al. (2016) to show that the pre-screening model averaging estimator based on the candidate model set $\mathcal{W}_n^s$ still achieves the asymptotic optimality, namely

$$\frac{\mathrm{KL}(\widehat{\mathbf{w}}^s)}{\inf_{\mathbf{w} \in \mathcal{W}_n^s} \mathrm{KL}(\mathbf{w})} \to 1.$$

Moreover, the consistency result in Theorem 2 also allows for a pre-screening procedure as long as the pre-screened model space contains a correct model. Considering the fact that the sure independence screening guarantees the true model to lie in the set of screened model space in generalized linear regressions shown by Fan and Song (2010), the pre-screening model averaging estimator is expected to be consistent when correct models exist in the set of candidate models, although extra uncertainty may arise and inference would be more complicated due to pre-screening.

4 Monte Carlo simulation

In this section, we evaluate the finite sample performance of the proposed model averaging estimators via simulation. We consider two simulation designs, one with a finite dimension of covariates and the other with a divergent dimension. In each design, we consider two subcases that differ in terms of whether correct models are included as candidate models.

4.1 Simulation designs

Design 1

Our first design considers finite-dimensional covariates. We generate y_i from Poisson(μ_i) as

$$\mu_i = \exp(\beta_0 + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + z_i\theta), \tag{22}$$

where $\mathbf{x}_i$ is a 6×1 vector that follows a multivariate normal distribution with zero means and a variance-covariance matrix Σ_x whose diagonal elements are 1 and off-diagonal elements are all 0.8, and we set $\beta_0 = 0.1$ and $\boldsymbol{\beta} = (0, 0.15, -0.6, 0, 0.7, -0.07)$. With different specifications of which elements of $\mathbf{x}_i$ are included in the model, we have $S = 2^6 - 1$ candidate models containing at least one element of $\mathbf{x}_i$. z_i is an omitted variable that is in the data generating process (DGP) but not included in any candidate model. It also follows a normal distribution with a zero mean and unit variance and is correlated with each element of $\mathbf{x}_i$ with correlations of 0.8. The value of θ determines whether and how severe the candidate models are misspecified, and we consider two settings of θ.

> Design 1.1: We set $\theta = 0.3$, and thus, omitting z_i causes all of the candidate models to be misspecified. In this case, we compare the competing predictions in terms of the relative KL loss (RKL) with respect to that obtained from the best single candidate model, i.e.,
>
> $$\text{RKL} = \text{KL} - \text{KL}_{\text{bs}}, \tag{23}$$
>
> where the KL loss of a method is computed by
>
> $$\text{KL} = n_{\text{eval}}^{-1} \sum_{i=1}^{n_{\text{eval}}} \left[\mu_i \log(\mu_i) - \mu_i - \mu_i \mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}} + \exp(\mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}) \right], \tag{24}$$
>
> $\widehat{\boldsymbol{\beta}}$ is the estimator of the $\boldsymbol{\beta}$ obtained by applying the method to the in-sample dataset (estimation sample), $n_{\text{eval}} = 1000$ is the size of the evaluation sample, and KL_{bs} is the minimum KL loss produced by the best single candidate model. To avoid randomness, we replicate the

estimation and evaluation $D = 1000$ times and report the average and standard error of RKL across replications. To examine the optimality of the KL-based model averaging (KL-MA) method, we also examine how the ratio $\mathrm{KL}(\widehat{\mathbf{w}})/\inf_{\mathbf{w}}\mathrm{KL}(\mathbf{w})$ behaves as the sample size increases, where $\widehat{\mathbf{w}}$ is the weights obtained from (9) and $\inf_{\mathbf{w}}\mathrm{KL}(\mathbf{w})$ is the minimum loss over all possible weight choices.

Design 1.2: We set $\theta = 0$, such that the correct models are included in the set of candidate models. In this case, the true model is to include x_{2i}, x_{3i}, x_{5i}, and x_{6i}, while the correct model can be any that includes at least these four covariates, e.g., the largest model with all six covariates. With correct models included in the model space, we focus on the consistency of model averaging estimates of $\boldsymbol{\beta}_{\text{true}}$, the result shown by Theorem 2. We evaluate the accuracy of coefficient estimates based on the average mean square error (MSE) of $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$ across replications, i.e.,

$$\text{MSE} = D^{-1}\sum_{d=1}^{D}\left\|\widehat{\boldsymbol{\beta}}^{(d)}(\widehat{\mathbf{w}}^{(d)}) - \boldsymbol{\beta}_{\text{true}}\right\|^2, \tag{25}$$

where $\widehat{\boldsymbol{\beta}}^{(d)}(\widehat{\mathbf{w}}^{(d)})$ is the KL-MA coefficient estimates in the d^{th} replication.

Design 2

Next, we consider the case where the number of covariates in each candidate model diverges with the sample size. We let the dimension of $\boldsymbol{\beta}$ depend on n as

$$\boldsymbol{\beta} = (0, 0.15, -0.6, 0, 0.7, -0.07, 0, 0.15, -0.6, 0, 0.7, -0.07, \cdots)^{\mathrm{T}}_{\lfloor n^{0.3}\rfloor}.$$

where the subscript $\lfloor n^{0.3}\rfloor$ is the speed at which the dimension of $\boldsymbol{\beta}$ increases and $\lfloor\cdot\rfloor$ takes the integer part of the number (see Condition 5 for the restriction on the divergent speed of p). The covariates $\mathbf{x}_i$ are generated the same as in Design 1 but with the divergent dimension corresponding to $\boldsymbol{\beta}$. The remaining setting is identical to Design 1, and we also consider two subcases that differ in terms of whether the model space contains correct models:

Design 2.1: $\theta = 0.3$, such that all candidate models are misspecified.

Design 2.2: $\theta = 0$, such that correct models are included in the set of candidate models.

In all designs, we consider the sample size for estimation as $n = 50$, 100, 200, 400, and 800.

4.2 Implementation

We compare our KL-MA method with four prevalent model selection methods: AIC, BIC (Buckland et al., 1997), least absolute shrinkage and selection operator (Lasso) and post-Lasso, and with three other averaging methods: smooth-AIC (SAIC) and smooth-BIC (SBIC) (see, e.g., Hjort and Claeskens, 2003; Claeskens et al., 2006; Zhang et al., 2016) and optimal model averaging (OPT) by Zhang et al. (2016). We also compare with bagging, a popular ensemble learning method, which is related with model averaging since it also combines different base learners.[1]

The KL-MA predicted value is $\widehat{\mu}_i(\widehat{\mathbf{w}}, \mathbf{y}) = \exp\left\{\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})\right\}$, where the estimated weights are obtained by solving the optimization (9). The two information criteria (IC) can be computed as

$$\text{AIC} = -2L_s + 2p_s \quad \text{and} \quad \text{BIC} = -2L_s + p_s \log(n),$$

where L_s is the log likelihood of the s^{th} model. The associated IC-based averaging uses the weights

$$w_s = \exp(-\text{IC}_s/2)/\sum_{s=1}^{S}\exp(-\text{IC}_s/2).$$

where IC represents either AIC or BIC.

The Lasso estimator of $\boldsymbol{\beta}$ for Poisson regressions is obtained by minimizing the following objective function:

$$l_\lambda(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\left\{y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \log(y_i!)\right\} + \lambda\|\boldsymbol{\beta}\|_1,$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{p}|\beta_i|$, and λ is the tunning parameter chosen by 5-fold cross-validation. Specifically, we divide the total observations into five folds. For each candidate value λ_t in the searching range (e.g., 100 grids between 0.001 and 1), we minimize $l_{\lambda_t}(\boldsymbol{\beta})$ using four folds to obtain $\widehat{\boldsymbol{\beta}}$ and compute the value of the negative log-likelihood function with $\widehat{\boldsymbol{\beta}}$ using the remaining fold. We repeat this step for five times with different choices of training folds and sum up the five negative

[1] Also compared includes the ridge estimator, which is also a shrinkage-type estimator like Lasso. The relative KL-divergence of the ridge estimator is similar to that of Lasso, and thus omitted here.

log-likelihoods to obtain $L(\lambda_t)$. Then, the optimal tunning parameter is the one that minimizes $L(\lambda_t)$. Following the idea of Belloni and Chernozhukov (2013), we can implement post-Lasso, which uses the selected covariates by Lasso to estimate the coefficients with GLM.

The OPT method of Zhang et al. (2016) chooses the weights by minimizing the following criterion:

$$\mathcal{P}(\mathbf{w}) = -2\log g\{\mathbf{y}|\widehat{\boldsymbol{\mu}}(\mathbf{w},\mathbf{y})\} + \lambda_n \mathbf{w}^{\mathrm{T}}\mathbf{k},$$

where $\mathbf{k} = (k_1, k_2, ..., k_S)^{\mathrm{T}}$, k_s is the number of covariates used in the s^{th} candidate model and λ_n is a tuning parameter. For this method to work, Zhang et al. (2016) suggested two choices of tuning parameters, $\lambda_n = 2$ (denoted as OPT$_1$) and $\lambda_n = \log(n)$ (denoted as OPT$_2$), which lead to AIC and BIC selection, respectively, when the weights only take values of 0 and 1.

Finally, to implement bagging for Poisson regressions and faciliate comparison, we replace the standard base learners in bagging by the candidate models the same as our KL-MA, and we also attach the weights to each coefficient estimator like KL-MA (rather than to the outcome variable as in the standard bagging) in order to be able to compute the KL loss.

We compare these methods by the relative KL divergence, defined as RKL $= \mathrm{KL} - \mathrm{KL}_{\mathrm{bs}}$, where KL is the KL divergence produced by each method, and $\mathrm{KL}_{\mathrm{bs}}$ is the minimum KL divergence produced by the best single candidate model.

4.3 Simulation results

We first examine the performance of competing methods when all candidate models are misspecified in the finite-dimensional case. Table 1 presents the mean and standard deviation of the relative KL divergence of all methods. The full model performance is reported in the last column as a benchmark. The minimum average relative KL divergence (RKL) in each row is highlighted in bold. In general, we find that the performance of all methods improves as n increases, and KL-MA produces the smallest RKL in almost all cases. In particular, when $n = 50$, KL-MA produces the minimum RKL on average with a fairly small standard deviation, closely followed by OPT$_1$, and then by SAIC and OPT$_2$. Lasso is less preferred than smoothed-IC averaging, and the IC-based model selection, post-Lasso and bagging perform even worse, with the average RKL almost twice as much as that of KL-MA as well as a large standard deviation. As the sample size increases, KL-MA generally remains the best method in terms of the KL divergence, but the difference between KL-MA and OPT$_1$ and OPT$_2$ becomes smaller. The superiority of these three methods over others is more prominent when n is large. The performance of Lasso and post-Lasso also improves as n increases. These results demonstrate the superiority of model averaging over model selection when all candidate models are misspecified. They also illustrate the advantages of using an unbiased criterion of KL divergence to choose the weights, compared to those using a biased criterion.

INSERT TABLE 1 HERE

Figure 1(a) plots the ratio of KL divergence produced by KL-MA over that given by the infeasible optimal averaging, i.e., $\mathrm{KL}(\widehat{\mathbf{w}})/\inf_{\mathbf{w}}\mathrm{KL}(\mathbf{w})$, in the finite-dimensional case. As the estimation sample size increases, the ratio converges to 1 monotonically, confirming the asymptotic optimality of Theorem 1.

INSERT FIGURE 1 HERE

We now examine the case in which there exist correct models in the set of candidate models with finite-dimensional covariates. In this case, Condition 5 is violated, and thus, asymptotic optimality does not hold. Hence, we focus on verifying the consistency of averaging coefficient estimates. The upper panel of Table 2 presents the MSE of coefficient estimates of KL-MA as n increases in the finite-dimensional case. We find that the MSE decreases when $\sqrt{p/n}$ decreases, confirming the consistency result of Theorem 2.

INSERT TABLE 2 HERE

Next, we consider the design when the number of covariates and candidate models diverges with the sample size. Table 3 compares the KL loss of competing methods when all candidate models are misspecified (Design 2.1). In this case, the KL loss of all methods does not necessarily decreases monotonically as the estimation sample size n increases because the dimension of the covariates and candidate models is also increasing, leading to potentially larger loss. When n is small, OPT$_2$

appears to perform best, with the smallest average RKL, and KL-MA follows closely, with only a marginally greater loss. However, the performance of OPT$_2$ is not stable as n increases. In contrast, KL-MA and OPT$_1$ perform rather stably for different sample sizes. These two methods perform similarly and deliver the lowest KL loss when n is at least 200. The popular model averaging methods using smoothed IC do not allow for divergent dimension and thus do not work well in this case. Figure 1(b) shows the behavior of $\mathrm{KL}(\widehat{\mathbf{w}})/\inf_{\mathbf{w}}\mathrm{KL}(\mathbf{w})$ as n increases when all candidate models are misspecified, and we also find a generally converging pattern, although the curve is less smooth than the finite-dimensional case.

INSERT TABLE 3 HERE

The bottom panel of Table 2 reports how the MSE of KL-MA coefficient estimates behaves as n increases in the divergent-dimensional design. Again, we see a convergence pattern, as in the fixed-dimensional case, providence evidence of the validity of Theorem 2 in divergent-dimensional situations.

5 Explaining and predicting corporate innovation outcomes

In this section, we apply the KL-based model averaging to study corporate innovation outcomes. Recently, Sunder et al. (2017) have studied the determinantion of corporate innovation outcomes measured by the number of patents, with a particular focus on the role of CEOs' hobby of flying airplanes. They argued that CEOs with a pilot license are characterized by higher risk-taking propensity, with a desire to pursue novel experiences, and therefore tend to implement more risky managing policies, including high R&D expenditure. As a result, Sunder et al. (2017) found that CEOs' hobby of flying airplanes is associated with significantly better corporate innovation outcomes, measured by more patents and citations.

We revisit the potential effect of pilot CEOs on innovation and also try to predict innovation outcomes given possible determinants suggested by Sunder et al. (2017). We follow Sunder et al. (2017) to measure the innovation by the number of patent applications during the year and construct a pilot CEO dummy that equals 1 if a CEO holds a pilot license and zero otherwise. Other covariates include firm characteristics, namely the logarithm of total assets (log(assets)), the logarithm of the ratio of net property, plant, and equipment to the number of employees (log(PPE/EMP)), stock returns, Tobin's q, and institutional holdings. Also included are CEOs' characteristics to capture their extrinsic incentives, human capital, and other behavioral traits, namely CEOs' tenure status (log(1+tenure)), the sensitivity of CEO wealth to stock volatility (log(1+vega)), the sensitivity of CEO wealth to performance (log(1+delta)), age (log(CEO age)), overconfidence, and dummy variables for CEOs' education background: whether a CEO has top university degrees (top university), technical education, a PhD degree in technical education (PhD in tech. edu), finance education, military experience, and no schooling information. Thus, we have 17 covariates in total.[2] We collect the data from ExecuComp, BoardEx, Compustat, the NBER patent database, and the U.S. Federal Aviation Administration airmen certification records and employ the sample from 1994 to 2003, excluding missing observations. We also exclude financial firms and regulated utilities as in Sunder et al. (2017), and obtain a data set largely similar to that of Sunder et al. (2017), consisting of 5371 observations. To stay close to Sunder et al. (2017) for comparison purposes, we estimate a cross-sectional regression even though panel data are available. Moreover, a large number of missing observations can cause a lack of observations if we wish to obtain a balanced panel. Model averaging for panel Poisson regression is an interesting topic for future research.

With a large number of covariates, model uncertainty is an important issue, because the empirical results may vary across different model specifications. For example, the estimated coefficients of pilot CEO and log(assets) can vary by roughly 10% and 50%, respectively, across specifications, and the estimators of military experience and overconfidence can change from strongly significant to insignificant, when some CEO characteristic variables are included (see Table A.3 in the Appendix for estimation results obtained under a selected set of different specifications). The prediction of innovation outcome also varies across specifications to a large extent. Therefore, it is crucial to account for such model uncertainty when examining the determinants and predicting the number of patent applications. With a count measure as the outcome variable and to address model uncertainty, we employ the KL-based model averaging method for Poisson regressions.

[2] See Tables A.2 in the Appendix for the descriptive statistics of the variables.

Since there are a vast number of candidate models if we consider all possible combinations of covariates, we first screen the candidate models prior to model averaging as discussed in Section 3.3. We consider two model screening procedures to prepare the set of candidate models as summarized in Table A.4 in the Appendix. First, since the pilot CEO dummy is the variable of interest, we include pilot CEO and the intercept in all candidate models. The remaining covariates are arranged in descending order according to the absolute value of their bivariate correlations with the outcome variable. Then, the candidate models include one extra covariate at each time based on the ordering. We refer to this approach as Screening 1. As a second screening method (Screening 2), we do not fix pilot CEO in each candidate model but sort *all* covariates and construct candidate models by including one extra covariate at each time based on the descending order of bivariate correlation, as above. These two model screening procedures based on bivariate correlation are in line with the ideas of sure independence screening proposed by Fan and Lv (2008) and Fan and Song (2010). Section 3.3 provides theoretical justifications of the pre-screening model averaging in both cases with and without correct models in the set of candidate models.

We first study how potential determinants affect corporate innovation outcome based on the KL-MA coefficient estimates. Statistical inference of model averaging estimators is a challenging task and has not been much studied in the literature, because the estimated averaging weights are random and candidate models may be misspecified. A working method is bootstrapping advocated by many studies, e.g., Buckland et al. (1997) and Hansen and Racine (2018), although its theoretical justification warrants future research. We adopt this technique to obtain the distribution of KL-MA coefficient estimators.

INSERT TABLE 4 HERE

Table 4 provides the coefficient estimates of KL-MA using three model screening procedures and their bootstrap p-values based on 1000 resamplings. We also report the estimates of the full model for comparison. The full model coefficient estimates are very similar to those of Sunder et al. (2017), and we find that a CEO having a pilot license is strongly and significantly associated with higher innovation outcomes. Other significant determinants include log(PPE/EMP), Tobin's Q, institutional holdings, CEO's age, CEO's top university experience and finance degree. The KL-MA coefficient estimates are in line with but also differ from the full model estimates to certain extent. Importantly, we find that when we account for model uncertainty, the significance of the effect of pilot CEO is weakened. In particular, if we prepare all candidate models as nested using model screening procedures 1, the estimated coefficient of pilot CEO is less sizeable and its p-value increases to almost 0.05, much greater than that produced by the full model. If we consider more flexible candidate models allowing for a higher degree of model uncertainty, as in screening procedure 2, the variability of estimated coefficient of pilot CEO is even larger, leading to further weaker statistical significance. These results suggest that the strong and significant association between pilot CEO and corporate innovation outcomes may not be very robust. In fact, if the full model does not coincide with the DGP and some covariates that are correlated with pilot CEO but irrelevant for innovation are included in the models, they may contaminate the effect of pilot CEO (either inflate or deflate its estimate), leading to spurious inference. To confirm this explanation, we further examine the association between pilot CEO and insignificant determinants from the full model. The insignificant determinants based on the full model include log(assets), stock return, log(1+delta), log(1+vega), technical education, PhD in technical education, no school information, military, and overconfidence. We regress pilot CEO on those insignificant determinants and find that CEOs who hold pilot license also often hold a PhD in technical education and have higher vega values in their compensation packages (higher CEO's wealth). This outcome is expected because pursuing a pilot license is technically demanding and also costly and thus better suits CEOs with a good technical background who are also financially unconstrained. The high correlation between pilot CEO and these personal characteristics is also confirmed by Sunder et al. (2017) based on a descriptive analysis (see their Table 3). Due to such a correlation, if these personal characteristics are redundant in the innovation regression, including these covariates does not improve the consistency but may largely contaminate the estimate of pilot CEO and introduce much estimation uncertainty, leading to a different conclusion. These findings demonstrate the necessity of controlling for model uncertainty in such a regression with many uncertain covariates.

Next, we evaluate the prediction performance of competing methods using this real dataset. To this end, we randomly divide the sample into two subsamples, one for estimating parameters and weights and the other for evaluation, following the idea of Hansen and Racine (2012) and Lehrer and Xie (2017). Let n_0 and $n_1 = n - n_0$ denote the sizes of the estimation and evaluation samples, respectively. We range n_0 among $\lfloor 0.7n \rfloor$, $\lfloor 0.8n \rfloor$ and $\lfloor 0.9n \rfloor$. We evaluate the prediction performance using two measures. First, we consider the relative KL divergence with respect to the best single model. Unlike in the simulation, the true density function of $\mathbf{y}^*$, $f(\mathbf{y}^*)$ in $E_{f(\mathbf{y}^*)} \log f(\mathbf{y}^*)$ is unknown in real-data applications. Since this term is common across all methods, we omit it in KL and compute the sample version of the relative KL divergence as

$$\text{SRKL} = -n_1^{-1} \sum_{i=1}^{n_1} \left[y_i \mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}) - \exp\{\mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})\} - \log(y_i!) \right].$$

Second, we also measure the prediction performance by the relative mean squared prediction error (RMSPE), i.e.,

$$\text{RMSPE} = \text{MSPE} - \text{MSPE}_{\text{bs}},$$

where $\text{MSPE} = 1/n_1 \sum_{i=n_0+1}^{n} (y_i - \widehat{\mu}_i)^2$ and MSPE_{bs} is the MSPE of the best single model, which is defined as the single candidate model that produces the minimum MSPE. To avoid randomness and abnormally large value due to the exponential function, we repeat the random sample division, prediction, and evaluation for 200 times, and report the trimmed mean and trimmed standard deviation of the SRKL and RMSPE across replications, excluding the lower and upper 10% of the values.

INSERT TABLES 5 and 6 HERE

Tables 5 and 6 present the trimmed mean and trimmed standard deviation of SRKL and RMSPE based on the two screening methods to construct candidate models. The two screening methods lead to similar prediction for KL-MA and IC-based methods, while the prediction by Lasso, post-Lasso and the full model does not depend on pre-screening. In almost all cases, the full model performs worst, suggesting a large degree of efficiency loss when all covariates are included. As discussed above, several firm and CEO characteristics are insignificantly related with the innovation outcome in the full model, and some of these insignificant variables are strongly correlated with pilot CEO. Hence, including these insignificant covariates hardly adds extra information for prediction, but can substantially inflate the prediction variance.

Among the remaining methods, KL-MA and Lasso seem to produce the most accurate prediction. When we use $\lfloor 0.7n \rfloor$ for estimation, Lasso produces the lowest RMSPE, closely followed by KL-MA and post-Lasso. Further examination reveals that Lasso typically favors small models, with the average number of nonzero coefficients across replications around 8. The most frequently selected covariates by Lasso include pilot CEO, log(PPE/EMP), Tobin's q, and top university. These covariates also often appear in the models that receive large weights by KL-MA. In contrast, IC-based selection and averaging all perform poorly in these cases as they tend to choose or assign heavy weights on a large model. These results suggest that when the sample size is relatively small, estimation variance takes a large part of prediction error. Thus Lasso prediction that relies on a small model leads to significant variance reduction and produce low SRKL and RMSPE. Although weight estimation introduces extra uncertainty for KL-MA, it still performs reasonably well with SRKL slightly higher than Lasso, even when the sample size is not very large.

As the estimation sample size increases, the advantages of KL-MA become more obvious, and it outperforms other methods, including Lasso and OPT. The trimmed mean SRKL of KL-MA is more than 15% and 30% less than that of the second-best method when $n_0 = \lfloor 0.8n \rfloor$ and $n_0 = \lfloor 0.9n \rfloor$, respectively, and its RMSPE is even more than 44% lower than that of the second-best method when $n_0 = \lfloor 0.9n \rfloor$. The good performance of KL-MA can be partly explained by increasing accuracy in weight estimation as the sample size increases, and it also confirms the asymptotic optimality of KL-MA as shown in the theory and simulation studies.

The prediction horse race again shows the importance of accounting model uncertainty in empirical innovation regressions and the advantages of our proposed model averaging approach.

6 Conclusion

Poisson regressions are widely used when the dependent variables are count data, such as in corporate finance, where significant interest lies in understanding the determinants of innovation outcomes typically measured by the number of patent or citations. This paper proposes a model averaging estimation method based on the Kullback-Leibler divergence that allows researchers to address model uncertainty in Poisson regressions. Our weighting criterion is an unbiased estimator of the KL divergence. We show that the proposed model averaging estimate is asymptotically optimal in the sense that it yields a KL divergence that is asymptotically identical to that resulting from the infeasible best possible averaging estimator when all candidate models are misspecified. In a different situation where there exist correct models in the candidate models, our model averaging estimates can also produce consistent estimates of the slope coefficients. An important advantage of our techniques is that we allow the number of covariates and the number of candidate models to diverge as the sample size increases. Using the proposed approach to revisit the association between pilot CEO and innovation outcome measured by the number of patent applications, we find that the pilot CEO dummy does associate with higher innovation outcomes to some extent, but this association is less strong when model uncertainty is considered and thus needs to be interpreted with caution. Our KL-MA performs well in predicting the corporate innovation outcome in the presence of great model uncertainty.

Several relevant questions deserve future research. First, this paper focuses on the case of $p < n$, although the dimension of each candidate model is allowed to diverge as n increases. It remains an open and challenging topic to study optimal averaging for Poisson regressions when $p > n$. Second, the current weighting scheme requires that all weights lie between 0 and 1, and that they sum up to one. In some situations with highly similar candidate models, negative weights are likely to appear. Moreover, if some candidate models are not competitive, the sum-to-one constraint may need to be relaxed (Ando and Li, 2014). A comprehensive theoretical analysis on relaxing weight restrictions calls for future research. Finally, while bootstrap provides a feasible and promising way of quantifying the variability of model averaging estimators, inference of optimal model averaging estimators remains an open but important research question.

Disclosure statement

The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest, or non-financial interest in the subject matter or materials discussed in this manuscript.

Funding

Jiahui Zou's work was supported by the Young Teacher Foundation from Capital University of Economics and Business (XJ2021004601); Zhang's work was supported by the National Key R&D Program of China (2020AAA0105200), the National Natural Science Foundation of China (Grant Nos. 71925007, 72091212, 71988101 and 11688101), and the Beijing Academy of Artificial Intelligence; Guohua Zou's work was partially supported by the National Natural Science Foundation of China (Grant nos. 11971323 and 12031016). All remaining errors are ours.

References

P. Aghion and P. Howitt. A model of growth through creative destruction. *Econometrica*, 60:323–351, 1992.

T. Ando and K.-C. Li. A model averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, 109:254–265, 2014.

T. Ando and K. C. Li. A weight-relaxed model averaging approach for high-dimensional generalized linear models. *Annals of Statistics*, 45:2654–2679, 2017.

A. Belloni and V. Chernozhukov. Least squares after model selection in high dimensional sparse models. *Bernoulli*, 19:521–547, 2013.

R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *Annals of Statistics*, 41:802–837, 2013.

S. T. Buckland, K. P. Burnham, and N. H. Augustin. Model selection: An integral part of inference. *Biometrics*, 53:603–618, 1997.

A. Charkhi and G. Claeskens. Asymptotic post-selection inference for Akaike's information criterion. *Biometrika*, 105:645–664, 2018.

A. Charkhi, G. Claeskens, and B. E. Hansen. Minimum mean squared error model averaging in likelihood models. *Statistica Sinica*, 26:809–840, 2016.

L. H. Y. Chen. Poisson approximation for dependent trials. *Annals of Probability*, 3:534–545, 1975.

L. H. Y. Chen, L. Goldstein, and Q.-M. Shao. *Normal Approximation by Stein's Method*. Springer: Berlin, 2010.

G. Claeskens, C. Croux, and J. van Kerckhoven. Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics*, 62:972–979, 2006.

J. L. Coles, N. D. Daniel, and L. Naveen. Managerial incentives and risk-taking. *Journal of Financial Economics*, 79:431–468, 2006.

D. Danilov and J. R. Magnus. On the harm that ignoring pretesting can cause. *Journal of Econometrics*, 122:27–46, 2004.

G. De Luca, J. R. Magnus, and F. Peracchi. Weighted-average least squares estimation of generalized linear models. *Journal of Econometrics*, 204:1–17, 2018.

G. De Luca, J. R. Magnus, and F. Peracchi. Sampling properties of the Bayesian posterior mean with an application to WALS estimation. Tinbergen Institute Discussion Papers 20-015/III, Tinbergen Institute, 2020.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70:849–911, 2008.

J. Fan and R. Song. Sure independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics*, 38:3567–3604, 2010.

V. W. Fang, X. Tian, and T. Sheri. Does stock liquidity enhance or impede firm innovation? *The Journal of Finance*, 69:2085–2125, 2014.

C. J. Flynn, C. M. Hurvich, and J. S. Simonoff. Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models. *Journal of the American Statistical Association*, 108:1031–1043, 2013.

B. E. Hansen. Least squares model averaging. *Econometrica*, 75:1175–1189, 2007.

B. E. Hansen. Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics*, 5: 495–530, 2014.

B. E. Hansen and J. Racine. Jacknife model averaging. *Journal of Econometrics*, 167:38–46, 2012.

B. E. Hansen and J. Racine. Bootstrap model averaging unit root inference. *Working paper*, 2018.

N. L. Hjort and G. Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association*, 98:879–899, 2003.

Y. V. Hochberg, C. J. Serrano, and R. H. Ziedonis. Patent collateral, investor commitment, and the market for venture lending. *Journal of Financial Economics*, 130:74–94, 2018.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14:382–417, 1999.

L. Kogan, D. Papanikolaou, A. Seru, and N. Stoffman. Technological innovation, resource allocation, and growth. *The Quarterly Journal of Economics*, 132:665–712, 2017.

S. Lehrer and T. Xie. Box office buzz: Does social media data steal the show from model uncertainty when forecasting for hollywood? *The Review of Economics and Statistics*, 99:749–755, 2017.

H. Liang and P. Du. Maximum likelihood estimation in logistic regression models with a diverging number of covariates. *Electronic Journal of Statistics*, 6:1838–1846, 2012.

C.-A. Liu and B.-S. Kuo. Model averaging in predictive regressions. *The Econometrics Journal*, 19: 203–231, 2016.

Q. Liu and R. Okui. Heteroscedasticity-robust C$_p$ model averaging. *The Econometrics Journal*, 16: 463–472, 2013.

X. Lu and L. Su. Jackknife model averaging for quantile regressions. *Journal of Econometrics*, 188: 40–58, 2015.

J. Lv and J. S. Liu. Model selection principles in misspecified models. *Journal of the Royal Statical Society: Series B*, 76:141–167, 2014.

J. R. Magnus. Estimation of the mean of a univariate normal distribution with known variance. *Econometrics Journal*, 5:225–236, 2002.

J. Shao. *Mathematical Statistics*. Springer, USA, 2 edition, 2003.

M. F. J. Steel. Model averaging and its use in economics. *Journal of Economic Literature*, 58: 644–719, 2020.

J. Sunder, S. V. Sunder, and J. Zhang. Pilot CEOs and corporate innovation. *Journal of Financial Economics*, 123:209–224, 2017.

A. T. K. Wan, X. Zhang, and G. Zou. Least squares model averaging by Mallows criterion. *Journal of Econometrics*, 156:277–283, 2010.

X. Wei and Y. Yang. Robust combination of model selection methods for prediction. *Statistica Sinica*, 22:1021–1040, 2012.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25, 1982.

T. Xie. Prediction model averaging estimator. *Economics Letters*, 131:5–8, 2015.

Y. Yang. Adaptive regression by mixing. *Journal of the American Statistical Association*, 96:574–588, 2001.

Z. Yuan and Y. Yang. Combining linear regression models: When and how? *Journal of the American Statistical Association*, 100:1202–1214, 2005.

X. Zhang. Consistency of model averaging estimators. *Economics Letters*, 130:120–123, 2015.

X. Zhang, D. Yu, G. Zou, and H. Liang. Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association*, 111:1775–1790, 2016.

X. Zhang, J. Chiou, and Y. Ma. Functional prediction through averaging estimated functional linear regression models. *Biometrika*, 105:945–962, 2018.

X. Zhang, G. Zou, H. Liang, and R. J. Carroll. Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association*, 115:972–984, 2020.

Table 1: Relative KL divergence in Design 1.1 ($\times 10^{-3}$)

n		KL-MA	AIC	BIC	Lasso	Post-Lasso	SAIC	SBIC	OPT$_1$	OPT$_2$	Bagging	Full
50	Mean	**19.836**	35.881	41.728	39.554	50.581	23.530	25.216	19.897	22.278	38.143	41.852
	Std	25.242	40.814	43.785	39.720	40.398	25.374	26.614	25.107	28.134	23.831	38.093
100	Mean	**8.929**	13.894	17.965	21.313	24.365	9.902	11.692	8.949	11.436	38.847	17.077
	Std	12.856	16.626	21.724	18.854	18.413	12.758	14.643	12.872	16.385	16.313	15.021
200	Mean	4.785	7.173	8.377	13.814	14.335	**4.618**	5.198	4.799	6.552	43.320	8.153
	Std	5.990	7.600	8.762	9.079	8.581	5.708	6.302	6.001	8.012	10.323	7.206
400	Mean	**1.798**	3.073	4.294	8.791	9.264	2.063	2.587	1.806	2.913	43.622	3.349
	Std	3.003	3.769	4.743	4.003	3.927	2.832	3.076	3.004	4.042	7.547	3.428
800	Mean	**0.884**	1.401	1.820	6.755	6.957	0.937	1.289	0.900	1.494	47.393	1.499
	Std	1.462	1.696	2.838	2.080	2.081	1.380	1.785	1.470	2.075	5.885	1.625

Notes: This table presents the relative KL divergence, defined by $\mathrm{RKL} = \mathrm{KL} - \mathrm{KL}_{\mathrm{bs}}$, where KL is the KL divergence produced by each method, and $\mathrm{KL}_{\mathrm{bs}}$ is the minimum KL divergence produced by the best single candidate model. The results here are for Design 1.1, where the candidate models are finite-dimensional and all misspecified. The mean and standard deviation are obtained from 1000 replications.

Table 2: The MSE of KL-MA coefficient estimates when candidate models include correct models

	n	50	100	200	400	800
Finite dimension	p	7	7	7	7	7
(Design 1.2)	$\sqrt{p/n}$	0.374	0.265	0.187	0.132	0.094
	MSE	0.444	0.320	0.254	0.180	0.143
Divergent dimension	p	3	3	4	6	7
(Design 2.2)	$\sqrt{p/n}$	0.245	0.173	0.141	0.122	0.094
	MSE	0.178	0.129	0.176	0.161	0.143

Notes: n is the sample size for estimation. p is the dimension of covariates of the full model. Design 1.2 and 2.2 consider finite- and divergent-dimensional candidate models, respectively, with at least one correct model in the model space.

Table 3: Relative KL divergence in Design 2.1 ($\times 10^{-3}$)

n		KL-MA	AIC	BIC	Lasso	post-Lasso	SAIC	SBIC	OPT$_1$	OPT$_2$	Bagging	Full
50	Mean	3.175	9.139	8.572	13.289	13.773	4.218	3.475	3.150	**2.693**	3.779	8.294
	Std	11.740	15.907	15.025	22.082	22.643	13.290	12.815	11.793	11.062	7.960	16.683
100	Mean	0.854	6.119	6.636	8.291	8.237	2.381	2.362	0.884	**0.688**	1.439	2.739
	Std	4.902	7.779	8.083	6.875	6.715	5.268	5.386	4.938	4.551	5.704	6.202
200	Mean	1.839	3.105	6.837	10.468	10.151	2.387	4.407	**1.830**	3.990	15.149	1.800
	Std	3.658	4.500	7.306	6.874	5.970	3.592	5.080	3.645	5.611	5.777	3.517
400	Mean	**1.318**	2.305	4.467	7.620	7.693	1.575	2.726	1.319	2.934	45.731	1.871
	Std	2.406	2.966	5.003	3.032	2.947	2.292	3.065	2.407	3.884	7.902	2.512
800	Mean	**0.884**	1.401	1.820	6.755	6.957	0.937	1.289	0.900	1.494	47.393	1.499
	Std	1.462	1.696	2.838	2.080	2.081	1.380	1.785	1.470	2.075	5.885	1.625

Notes: This table presents the relative KL divergence, defined by RKL $= \mathrm{KL} - \mathrm{KL}_{\mathrm{bs}}$, where KL is the KL divergence produced by each method, and $\mathrm{KL}_{\mathrm{bs}}$ is the minimum KL divergence produced by the best single candidate model. The results here are for Design 2.1, where the candidate models are divergent-dimensional and all misspecified. The mean and standard deviation are obtained from 1000 replications.

Table 4: Coefficient estimates of innovation regression

	Full model		KL-MA			
	R^2=0.024		Screening 1		Screening 2	
	Coef.	p-value	Coef.	p-value	Coef.	p-value
Pilot CEO	0.204	0.017	0.194	0.042	0.192	0.042
log(assets)	0.027	0.236	0.024	0.210	0.024	0.210
log(PPE/EMP)	0.069	0.000	0.065	0.000	0.065	0.000
Stock return	−0.0004	0.355	0.000	0.406	0.000	0.404
Tobin's Q	0.041	0.006	0.041	0.000	0.041	0.000
Inst. holdings	0.180	0.033	0.174	0.042	0.174	0.042
log(1+tenure)	−0.050	0.158	0.000	0.154	0.000	0.158
log(1+delta)	0.015	0.427	0.012	0.600	0.012	0.600
log(1+vega)	0.014	0.580	0.011	0.560	0.011	0.560
log(CEO age)	0.536	0.017	0.465	0.010	0.466	0.010
Top university	0.176	0.012	0.172	0.020	0.172	0.020
Finance education	−2.14	0.001	−2.025	0.000	−2.024	0.000
Technical education	0.212	0.165	0.219	0.210	0.219	0.210
PhD in tech. edu	−0.065	0.465	0.000	0.414	0.000	0.426
No school info	0.064	0.296	0.000	0.256	0.000	0.244
Military	−0.134	0.402	−0.000	0.000	−0.000	0.000
Overconfidence	0.017	0.774	0.000	0.810	0.000	0.820
constant	−1.071	0.221	−0.880	0.226	−0.882	0.226

Notes: This table presents the coefficient estimates and associated p-values of the full model and KL-MA using three different pre-screening methods. The p-values of KL-MA are obtained from bootstrapping with 1000 resamplings. Screening 1 includes pilot CEO and the intercept in all candidate models, and adds one extra covariate at each time based on the absolute value of their bivariate correlations with the outcome variable. Screening 2 resembles Screening 1 but does not fix pilot CEO in all candidate models.

Table 5: Prediction of innovation outcomes using Screening 1: SRKL and RMSPE

n_0		KL-MA	AIC	BIC	Lasso	Post-Lasso	SAIC	SBIC	OPT_1	OPT_2	Bagging	Full
							SRKL					
$\lfloor 0.7n \rfloor = 3759$	Mean	**0.100**	0.190	0.174	0.133	0.128	0.189	0.170	0.164	0.113	1.069	0.190
	Std	0.121	0.195	0.170	0.034	0.035	0.192	0.161	0.156	0.104	0.123	0.195
$\lfloor 0.8n \rfloor = 4296$	Mean	**0.082**	0.156	0.150	0.140	0.135	0.156	0.148	0.140	0.097	1.073	0.156
	Std	0.134	0.176	0.172	0.039	0.039	0.176	0.170	0.158	0.109	0.141	0.176
$\lfloor 0.9n \rfloor = 4833$	Mean	**0.057**	0.122	0.122	0.154	0.149	0.122	0.122	0.110	0.081	1.087	0.122
	Std	0.107	0.200	0.197	0.055	0.055	0.199	0.197	0.173	0.107	0.184	0.200
							RMSPE					
$\lfloor 0.7n \rfloor = 3759$	Mean	10.607	64.924	46.345	**0.925**	0.981	63.541	41.259	43.591	24.000	17.871	64.982
	Std	21.058	198.863	128.045	0.659	0.656	192.195	109.358	112.565	60.516	2.930	198.378
$\lfloor 0.8n \rfloor = 4296$	Mean	5.386	16.980	13.904	1.105	**1.154**	16.935	13.204	13.202	7.853	17.551	16.979
	Std	13.126	37.604	29.980	0.678	0.677	37.532	27.684	26.277	15.434	2.713	37.578
$\lfloor 0.9n \rfloor = 4833$	Mean	**0.821**	7.199	6.912	1.426	1.475	7.198	6.805	6.111	2.889	17.602	7.204
	Std	3.956	21.995	21.449	0.810	0.811	21.985	21.303	19.285	10.627	3.168	22.011

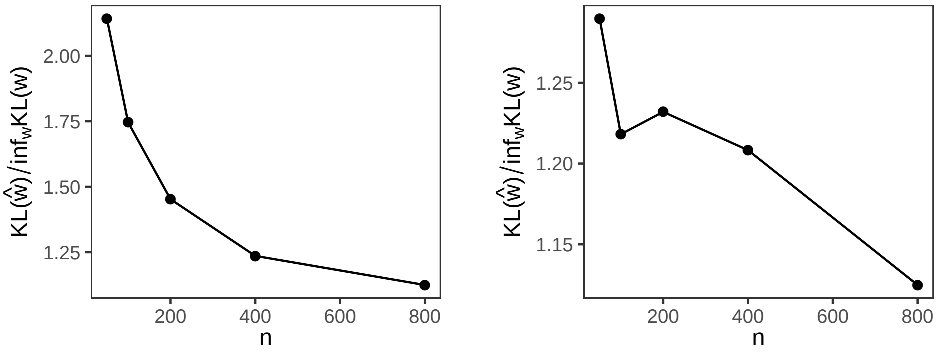
Notes: This table presents the sample version of the relative KL divergence (SRKL) and the relative mean squared prediction error (RMSPE). The estimation sample n_0 ranges from $\lfloor 0.7n \rfloor$ to $\lfloor 0.9n \rfloor$. The mean and standard deviation are obtained from 1000 replications. Screening 1 includes pilot CEO and the intercept in all candidate models, and adds one extra covariate at each time based on the absolute value of their bivariate correlations with the outcome variable.

Table 6: Prediction of innovation outcomes using Screening 2: SRKL and RMSPE

n_0		KL-MA	AIC	BIC	Lasso	Post-Lasso	SAIC	SBIC	OPT_1	OPT_2	Bagging	Full
						SRKL						
$\lfloor 0.7n \rfloor = 3759$	Mean	**0.099**	0.189	0.174	0.133	0.128	0.188	0.170	0.163	0.112	1.067	0.190
	Std	0.120	0.194	0.169	0.034	0.035	0.191	0.160	0.155	0.104	0.122	0.195
$\lfloor 0.8n \rfloor = 4296$	Mean	**0.082**	0.155	0.154	0.140	0.135	0.155	0.148	0.139	0.097	1.072	0.156
	Std	0.134	0.176	0.172	0.039	0.039	0.176	0.170	0.158	0.109	0.141	0.176
$\lfloor 0.9n \rfloor = 4833$	Mean	**0.057**	0.123	0.123	0.154	0.149	0.123	0.122	0.111	0.081	1.087	0.122
	Std	0.107	0.199	0.197	0.055	0.055	0.199	0.196	0.173	0.107	0.183	0.200
						RMSPE						
$\lfloor 0.7n \rfloor = 3759$	Mean	10.630	64.909	46.330	**0.925**	0.981	63.525	41.245	43.585	24.035	17.751	64.982
	Std	21.077	198.851	128.034	0.659	0.656	192.184	109.347	112.575	60.661	2.785	198.378
$\lfloor 0.8n \rfloor = 4296$	Mean	5.381	16.974	13.898	**1.105**	1.154	16.929	13.199	13.197	7.854	17.480	16.979
	Std	13.088	37.602	29.968	0.678	0.677	37.531	27.673	26.270	15.424	2.624	37.578
$\lfloor 0.9n \rfloor = 4833$	Mean	**0.838**	7.216	6.929	1.426	1.475	7.215	6.822	6.122	2.903	17.586	7.204
	Std	3.974	21.997	21.452	0.810	0.811	21.988	21.307	19.275	10.626	3.101	22.011

Notes: This table presents the sample version of the relative KL divergence (SRKL) and the relative mean squared prediction error (RMSPE). The estimation sample n_0 ranges from $\lfloor 0.7n \rfloor$ to $\lfloor 0.9n \rfloor$. The mean and standard deviation are obtained from 200 replications. Screening 2 sorts *all* covariates and constructs candidate models by including one extra covariate at each time based on the descending order of the absolute value of bivariate correlation.

Figure 1: The ratio of KL divergence of KL-MA over the infeasible best possible model averaging



(a) Finite dimension

(b) Divergent dimension

Notes: This figure plots the ratio of KL divergence produced by KL-MA over that given by the infeasible optimal averaging, i.e., $\mathrm{KL}(\widehat{\mathbf{w}})/\inf_{\mathbf{w}} \mathrm{KL}(\mathbf{w})$, when all candidate models are misspecified. The left subfigure considers the finite-dimensional case, and the right subfigure considers the divergent-dimensional case.

Appendix

This appendix includes the proofs of theorems and additional results of the empirical application. First, Section A.1 presents some lemmas and their proofs. Then Sections A.2 and A.3 provide the proofs of Theorems 1 and 2, respectively. Finally, Section A.4 provides additional empirical results.

A.1 Lemmas

To prove the theorems, we first establish Lemmas 1–5. All limiting results below are obtained by letting n go to infinity unless stated otherwise.

Proof of Lemma 1. To show the unbiasedness of our weight-choice criterion, first note that for any $i = 1, 2, ..., n$,

$$\begin{aligned}
\mathrm{E}_{f(\mathbf{y})}\{\mu_i \log \widehat{\mu}_i(\mathbf{w},\mathbf{y})\} &= \sum_{j=1,j\neq i}^{n}\sum_{y_j=0}^{\infty}\sum_{y_i=0}^{\infty}\left[\log\{\widehat{\mu}_i(\mathbf{w},\mathbf{y})\}\mu_i\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\prod_{j\neq i}\frac{e^{-\mu_j}\mu_j^{y_j}}{y_j!}\right]\\
&= \sum_{j=1,j\neq i}^{n}\sum_{y_j=0}^{\infty}\sum_{y_i=0}^{\infty}\left[(y_i+1)\log\{\widehat{\mu}_i(\mathbf{w},\mathbf{y})\}\frac{e^{-\mu_i}\mu_i^{y_i+1}}{(y_i+1)!}\prod_{j\neq i}\frac{e^{-\mu_j}\mu_j^{y_j}}{y_j!}\right]\\
&= \sum_{j=1,j\neq i}^{n}\sum_{y_j=0}^{\infty}\sum_{y_i=1}^{\infty}\left[y_i\log\{\widehat{\mu}_i(\mathbf{w},\mathbf{y}^{(y_i-1)})\}\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\prod_{j\neq i}\frac{e^{-\mu_j}\mu_j^{y_j}}{y_j!}\right]\\
&= \sum_{j=1,j\neq i}^{n}\sum_{y_j=0}^{\infty}\sum_{y_i=0}^{\infty}\left[y_i\log\{\widehat{\mu}_i(\mathbf{w},\mathbf{y}^{(y_i-1)})\}\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\prod_{j\neq i}\frac{e^{-\mu_j}\mu_j^{y_j}}{y_j!}\right]\\
&= \mathrm{E}_{f(\mathbf{y})}[y_i\log\{\widehat{\mu}_i(\mathbf{w},\mathbf{y}^{(y_i-1)})\}]. \qquad (A.1)
\end{aligned}$$

Combining (A.1) with (5) and (6), we obtain

$$\begin{aligned}
\mathrm{E}_{f(\mathbf{y})}\{\mathcal{C}(\mathbf{w})\} &= \mathrm{E}_{f(\mathbf{y})}\log f(\mathbf{y}) + \mathrm{E}_{f(\mathbf{y})}\sum_{i=1}^{n}[\widehat{\mu}_i(\mathbf{w},\mathbf{y}) + \log(y_i!) - y_i\log\{\widehat{\mu}_i(\mathbf{w},\mathbf{y}^{(y_i-1)})\}]\\
&= \mathrm{E}_{f(\mathbf{y})}\log f(\mathbf{y}) + \mathrm{E}_{f(\mathbf{y})}\sum_{i=1}^{n}\{\widehat{\mu}_i(\mathbf{w},\mathbf{y}) + \log(y_i!) - \mu_i\log\widehat{\mu}_i(\mathbf{w},\mathbf{y})\}\\
&= \mathrm{E}_{f(\mathbf{y})}\{\mathrm{KL}(\mathbf{w})\}.
\end{aligned}$$

□

Lemma 2 *Under Conditions 2 and 3, we have*

$$\left\|\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i\right\| = O_P(\sqrt{pn}), \quad \textit{and} \quad \left\|\sum_{i=1}^{n}y_i\mathbf{x}_i\right\| = O_P(\sqrt{pn}),$$

where $\varepsilon_i = y_i - \mu_i$*.*

Proof of Lemma 2. Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^{\mathrm{T}}$. We first consider $\|\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i\|$.

$$\begin{aligned}
\mathrm{E}\left(\frac{1}{\sqrt{pn}}\left\|\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i\right\|\right)^2 &= \frac{1}{pn}\mathrm{E}(\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}) \\
&= \frac{1}{pn}\mathrm{tr}\{\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathrm{Cov}(\boldsymbol{\varepsilon})\} \\
&= \frac{1}{pn}\sum_{i=1}^{n}\|\mathbf{x}_i\|^2\mathrm{Var}(y_i) \\
&\le \max_{1\le i\le n}\frac{\|\mathbf{x}_i\|^2}{p}\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(y_i) \\
&\le \max_{1\le i\le n}\frac{\|\mathbf{x}_i\|^2}{p}\sqrt{\frac{\|\boldsymbol{\mu}\|^2}{n}} \\
&\le \sqrt{C_2}C_3^2 < \infty,
\end{aligned} \tag{A.2}$$

where the last inequality is due to Lemma 2 and (12) in Lemma 3. So we have

$$\frac{1}{\sqrt{pn}}\left\|\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i\right\| = O_P(1). \tag{A.3}$$

Next, we consider $\|\sum_{i=1}^{n}y_i\mathbf{x}_i\|$. Under Condition 2 and (12) in Condition 3, we have

$$\begin{aligned}
\left\|\sum_{i=1}^{n}\mu_i\mathbf{x}_i\right\| &\le \sum_{i=1}^{n}|\mu_i|\|\mathbf{x}_i\| \\
&\le \sqrt{\sum_{i=1}^{n}|\mu_i|^2}\sqrt{\sum_{i=1}^{n}\|\mathbf{x}_i\|^2} \\
&\le \sqrt{C_2 n}\sqrt{n\max_{1\le i\le n}\|\mathbf{x}_i\|^2} \\
&\le \sqrt{C_2 n}\sqrt{C_3^2 np} \\
&\le \sqrt{C_2}C_3\sqrt{p}n.
\end{aligned} \tag{A.4}$$

Then combining (A.3) and (A.4), we obtain

$$\left\|\sum_{i=1}^{n}y_i\mathbf{x}_i\right\| \le \left\|\sum_{i=1}^{n}\mu_i\mathbf{x}_i\right\| + \left\|\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i\right\|$$

$$
\begin{aligned}
&= O(\sqrt{p}n) + O_P(\sqrt{p}n) \\
&= O_P(\sqrt{p}n).
\end{aligned}
$$

□

Lemma 3 *Under Conditions 1, 3 and 4, we have*

$$\max_{1\le s\le S}\left\|\sum_{i=1}^{n}\varepsilon_{(s),i}\mathbf{x}_{(s),i}\right\| = O_P(\sqrt{Spn}), \tag{A.5}$$

where $\varepsilon_{(s),i} = y_i - \exp(\mathbf{x}_{(s).i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^{*})$.

Proof of Lemma 3. First, recall that

$$
\begin{aligned}
f_{(s)}(\mathbf{y}|\boldsymbol{\beta}_{(s)}) &= \prod_{i=1}^{n}\frac{\mu_{(s),i}^{y_i}e^{-\mu_{(s),i}}}{y_i!}, \quad \text{and} \quad \mu_{(s),i} = \exp(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}), \\
i &= 1,\ldots,n, \quad y_i = 0,1,\ldots
\end{aligned}
$$

Then the KL divergence is written as

$$
\begin{aligned}
\mathrm{KL}(\boldsymbol{\beta}_{(s)}) &= \mathrm{E}_{\mathbf{y}}\log\frac{f(\mathbf{y})}{f_{(s)}(\mathbf{y}|\boldsymbol{\beta}_{(s)})} \\
&= \mathrm{E}_{\mathbf{y}}\log f(\mathbf{y}) + \sum_{i=1}^{n}\mathrm{E}_{\mathbf{y}}\log(y_i!) + \sum_{i=1}^{n}\left\{\exp(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}) - \mu_i\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}\right\}.
\end{aligned}
$$

Using (15) in Condition 4, we have

$$\frac{\partial^2\mathrm{KL}(\boldsymbol{\beta}_{(s)}^{*})}{\partial\boldsymbol{\beta}_{(s)}\partial\boldsymbol{\beta}_{(s)}^{\mathrm{T}}} = \sum_{i=1}^{n}\exp(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^{*})\mathbf{x}_{(s),i}\mathbf{x}_{(s),i}^{\mathrm{T}} > 0.$$

With the second-order derivative of $\mathrm{KL}(\boldsymbol{\beta}_{(s)}^{*})$ larger than zero, $\boldsymbol{\beta}_{(s)}^{*}$ that leads to the minimum $\mathrm{KL}(\boldsymbol{\beta}_{(s)})$ satisfies the first-order condition, i.e.,

$$0 = \frac{\partial\mathrm{KL}(\boldsymbol{\beta}_{(s)}^{*})}{\partial\boldsymbol{\beta}_{(s)}} = \sum_{i=1}^{n}\left\{\exp(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^{*}) - \mu_i\right\}\mathbf{x}_{(s),i},$$

which, according to Condition 1, further implies that

$$
\begin{aligned}
&\mathrm{E}\left[\sum_{i=1}^{n}\left\{y_i - \exp(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^{*})\right\}\mathbf{x}_{(s),i}\right] \\
&= \mathrm{E}\left\{\sum_{i=1}^{n}(y_i-\mu_i)\mathbf{x}_{(s),i}\right\} + \sum_{i=1}^{n}\left\{\mu_i - \exp(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}^{*})\right\}\mathbf{x}_{(s),i}
\end{aligned}
$$

$$= \quad 0. \tag{A.6}$$

Next, based on Condition 2,

$$\begin{aligned}
& \frac{1}{Spn}\sum_{s=1}^{S} \mathrm{E}\left\|\boldsymbol{\varepsilon}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)}\right\|^2 \\
= & \frac{1}{Spn}\sum_{s=1}^{S} \mathrm{E}\left(\boldsymbol{\varepsilon}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)}\mathbf{X}_{(s)}^{\mathrm{T}}\boldsymbol{\varepsilon}_{(s)}\right) \\
= & \frac{1}{Spn}\sum_{s=1}^{S} \mathrm{tr}\left\{\mathbf{X}_{(s)}\mathbf{X}_{(s)}^{\mathrm{T}}\mathrm{Cov}(\boldsymbol{\varepsilon}_{(s)})\right\} \\
= & \frac{1}{Spn}\sum_{s=1}^{S}\sum_{i=1}^{n} \|\Pi_s^{\mathrm{T}}\mathbf{x}_i\|^2 \mathrm{Var}(\varepsilon_{(s),i}) \\
\leq & \frac{1}{Spn}\sum_{s=1}^{S}\sum_{i=1}^{n} \|\mathbf{x}_i\|^2 \mathrm{Var}(y_i) \\
\leq & \max_{1\leq i\leq n}\frac{\|\mathbf{x}_i\|^2}{p}\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(y_i) \\
\leq & \sqrt{C_2}C_3^2 < \infty,
\end{aligned} \tag{A.7}$$

where Π_s is a selection matrix picking the covariates included in the s^{th} model, i.e., $\mathbf{X}\Pi_s = \mathbf{X}_{(s)}$ and $\mathbf{x}_i^{\mathrm{T}}\Pi_s = \mathbf{x}_{(s),i}^{\mathrm{T}}$. The second step in (A.7) is based on (A.6), and the last inequality is from Condition 2 and (12) in Condition 3. Hence, we have

$$\frac{1}{\sqrt{Spn}}\max_{1\leq s\leq S}\left\|\sum_{i=1}^{n}\varepsilon_{(s),i}\mathbf{x}_{(s),i}\right\| = O_P(1), \tag{A.8}$$

and this completes the proof.

□

Lemma 4 *Under Conditions 1, 3 and 4,*

$$\max_{1\leq s\leq S}\left\|\widehat{\boldsymbol{\beta}}_{(s)} - \boldsymbol{\beta}_{(s)}^{*}\right\| = O_P(S^{1/2}p^{1/2}n^{-1/2}), \tag{A.9}$$

where $\widehat{\boldsymbol{\beta}}_{(s)}$ is the ML estimator of the s^{th} candidate model.

Proof of Lemma 4. The log-likelihood function of the s^{th} model is written as

$$l_n(\boldsymbol{\beta}_{(s)}) = \sum_{i=1}^{n}\left\{y_i\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)} - \exp(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}) - \log(y_i!)\right\}, \tag{A.10}$$

and the ML estimator $\widehat{\boldsymbol{\beta}}_{(s)}$ satisfies

$$\frac{\partial l_n(\widehat{\boldsymbol{\beta}}_{(s)})}{\partial \boldsymbol{\beta}_{(s)}} = \sum_{i=1}^{n} \left\{ y_i - \exp(\mathbf{x}_{(s),i}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}) \right\} \mathbf{x}_{(s),i} = 0. \tag{A.11}$$

Let $\mathcal{A}_n(\boldsymbol{\beta}^*_{(s)}|\delta) = \{\boldsymbol{\gamma} \in \mathcal{R}^{p_s} : \sqrt{n}\|\boldsymbol{\gamma} - \boldsymbol{\beta}^*_{(s)}\|/\sqrt{Sp} \leq \delta\}$ and $\partial\mathcal{A}_n(\boldsymbol{\beta}^*_{(s)}|\delta)$ be the boundary of $\mathcal{A}_n(\boldsymbol{\beta}^*_{(s)}|\delta)$. By the second-order Taylor expansion of (A.11) at $\boldsymbol{\beta}^*_{(s)}$, there exists some $\delta > 0$, such that when n is large enough and $\boldsymbol{\gamma}_s \in \partial\mathcal{A}_n(\boldsymbol{\beta}^*_{(s)}|\delta)$,

$$\begin{aligned}
&\max_{1\leq s\leq S} \left\{ l_n(\boldsymbol{\gamma}_s) - l_n(\boldsymbol{\beta}^*_{(s)}) \right\} \\
= \;& \max_{1\leq s\leq S} \left\{ \sum_{i=1}^{n} \varepsilon_{(s),i} \mathbf{x}_{(s),i}^{\mathrm{T}} (\boldsymbol{\gamma}_s - \boldsymbol{\beta}^*_{(s)}) - \frac{1}{2} (\boldsymbol{\gamma}_s - \boldsymbol{\beta}^*_{(s)})^{\mathrm{T}} \sum_{i=1}^{n} \exp(\mathbf{x}_{(s),i}^{\mathrm{T}} \widetilde{\boldsymbol{\beta}}_{(s)}) \mathbf{x}_{(s),i} \mathbf{x}_{(s),i}^{\mathrm{T}} (\boldsymbol{\gamma}_s - \boldsymbol{\beta}^*_{(s)}) \right\} \\
\leq \;& Sp \left[\frac{1}{\sqrt{Spn}} \max_{1\leq s\leq S} \left\| \sum_{i=1}^{n} \varepsilon_{(s),i} \mathbf{x}_{(s),i} \right\| \sqrt{\frac{n}{Sp}} \left\| \boldsymbol{\gamma}_s - \boldsymbol{\beta}^*_{(s)} \right\| \right. \\
& \left. - \frac{1}{2} \frac{n}{Sp} \left\| \boldsymbol{\gamma}_s - \boldsymbol{\beta}^*_{(s)} \right\|^2 \min_{1\leq s\leq S} \lambda_{\min} \left\{ I_{(s)}(\widetilde{\boldsymbol{\beta}}_{(s)}) \right\} \right] \\
\leq \;& Sp \left\{ \frac{1}{\sqrt{Spn}} \max_{1\leq s\leq S} \left\| \sum_{i=1}^{n} \varepsilon_{(s),i} \mathbf{x}_{(s),i} \right\| \delta \|\nu_s\| - \frac{1}{2} C_{\min} \delta^2 \|\nu_s\|^2 \right\} \\
= \;& Sp \left\{ \delta O_P(1) - \frac{1}{2} C_{\min} \delta^2 \right\},
\end{aligned} \tag{A.12}$$

where $\nu_s = \sqrt{n}\left(\boldsymbol{\gamma} - \boldsymbol{\beta}^*_{(s)}\right)/\delta\sqrt{Sp}$ and $\widetilde{\boldsymbol{\beta}}_{(s)}$ lies between $\boldsymbol{\gamma}_s$ and $\boldsymbol{\beta}^*_{(s)}$. The last inequality in (A.12) is due to (15) in Condition 4, and the last equality holds because

$$\max_{1\leq s\leq S} \left\| \sum_{i=1}^{n} \varepsilon_{(s),i} \mathbf{x}_{(s),i} \right\| / \sqrt{Spn} = O_P(1),$$

according to Lemma 3.

Finally, because of the nonnegativity of exponential function $\exp(\cdot)$ and (15) in Condition 4, we know

$$\frac{\partial^2 l_n(\boldsymbol{\beta}_{(s)})}{\partial \boldsymbol{\beta}_{(s)} \partial \boldsymbol{\beta}_{(s)}^{\mathrm{T}}} = -\sum_{i=1}^{n} \exp(\mathbf{x}_{(s),i}^{\mathrm{T}} \boldsymbol{\beta}_{(s)}) \mathbf{x}_{(s),i} \mathbf{x}_{(s),i}^{\mathrm{T}} < 0.$$

Thus, the log-likelihood function is concave, and we obtain the desired result (A.9) based on (A.12).

□

Lemma 5 *Under Conditions 1, 3 and 4, we have*

$$\max_{1\leq i\leq n} \max_{1\leq s\leq S} \left\| \widehat{\boldsymbol{\beta}}_{(s)}^{(y_i-1)} - \widehat{\boldsymbol{\beta}}_{(s)} \right\| \;=\; O_P(p^{1/2}n^{-1}) \tag{A.13}$$

and

$$\max_{1\le i\le n}\max_{1\le s\le S}\left\|\widehat{\boldsymbol{\beta}}_{(s)}^{(y_i-1)}-\boldsymbol{\beta}^*_{(s)}\right\| \;=\; O_P(S^{1/2}p^{1/2}n^{-1/2}). \tag{A.14}$$

Proof of Lemma 5. First, denote $l_n^{(y_i-1)}(\boldsymbol{\beta}_{(s)})$ as the log-likelihood function of the s^{th} model but using y_i-1 instead of y_i, then we have

$$l_n^{(y_i-1)}(\boldsymbol{\beta}_{(s)}) \;=\; \sum_{j=1}^{n}\left\{y_j\mathbf{x}_{(s),j}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}-\exp(\mathbf{x}_{(s),j}^{\mathrm{T}}\boldsymbol{\beta}_{(s)})-\log(y_j!)\right\}-\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}+\log(y_i).$$

Let $\mathcal{B}_n(\widehat{\boldsymbol{\beta}}_{(s)}|\delta)=\left\{\boldsymbol{\gamma}\in\mathcal{R}^{p_s}: n\|\boldsymbol{\gamma}-\widehat{\boldsymbol{\beta}}_{(s)}\|/\sqrt{p}\le\delta\right\}$ and $\partial\mathcal{B}_n(\widehat{\boldsymbol{\beta}}_{(s)}|\delta)$ be the boundary of $\mathcal{B}_n(\widehat{\boldsymbol{\beta}}_{(s)}|\delta)$. There exists some $\delta>0$, such that when n is large enough and $\boldsymbol{\gamma}_s\in\partial\mathcal{B}_n(\widehat{\boldsymbol{\beta}}_{(s)}|\delta)$, Taylor expansion gives that

$$\begin{aligned}
&\max_{1\le i\le n}\max_{1\le s\le S}\left\{l_n^{(y_i-1)}(\boldsymbol{\gamma}_s)-l_n^{(y_i-1)}(\widehat{\boldsymbol{\beta}}_{(s)})\right\}\\
=\;&\max_{1\le i\le n}\max_{1\le s\le S}\left\{\frac{\partial l_n^{(y_i-1)}(\widehat{\boldsymbol{\beta}}_{(s)})}{\partial\boldsymbol{\beta}_s}\left(\boldsymbol{\gamma}_s-\widehat{\boldsymbol{\beta}}_{(s)}\right)+\frac{1}{2}\left(\boldsymbol{\gamma}_s-\widehat{\boldsymbol{\beta}}_{(s)}\right)^{\mathrm{T}}\frac{\partial^2 l_n^{(y_i-1)}(\widetilde{\boldsymbol{\gamma}}_s)}{\partial\boldsymbol{\beta}_s\partial\boldsymbol{\beta}_s^{\mathrm{T}}}\left(\boldsymbol{\gamma}_s-\widehat{\boldsymbol{\beta}}_{(s)}\right)\right\}\\
=\;&\max_{1\le i\le n}\max_{1\le s\le S}\left\{-\mathbf{x}_{(s),i}^{\mathrm{T}}\left(\boldsymbol{\gamma}_s-\widehat{\boldsymbol{\beta}}_{(s)}\right)-\frac{1}{2}\left(\boldsymbol{\gamma}_s-\widehat{\boldsymbol{\beta}}_{(s)}\right)^{\mathrm{T}}\sum_{j=1}^{n}\exp\left(\mathbf{x}_{(s),j}^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}_s\right)\mathbf{x}_{(s),j}\mathbf{x}_{(s),j}^{\mathrm{T}}\left(\boldsymbol{\gamma}_s-\widehat{\boldsymbol{\beta}}_{(s)}\right)\right\}\\
=\;&\frac{p}{n}\max_{1\le i\le n}\max_{1\le s\le S}\left\{-\frac{\mathbf{x}_{(s),i}^{\mathrm{T}}}{\sqrt{p}}\frac{n}{\sqrt{p}}\left(\boldsymbol{\gamma}_s-\widehat{\boldsymbol{\beta}}_{(s)}\right)-\frac{1}{2}\frac{n}{\sqrt{p}}\left(\boldsymbol{\gamma}_s-\widehat{\boldsymbol{\beta}}_{(s)}\right)^{\mathrm{T}}I_{(s)}(\widetilde{\boldsymbol{\gamma}}_s)\frac{n}{\sqrt{p}}\left(\boldsymbol{\gamma}_s-\widehat{\boldsymbol{\beta}}_{(s)}\right)\right\}\\
\le\;&\frac{p}{n}\left\{-\frac{1}{2}(C_{\min}+o_P(1))\delta^2\|\nu_s\|^2+\max_{1\le i\le n}\max_{1\le s\le S}\frac{\|\mathbf{x}_{(s),i}\|}{\sqrt{p}}\delta\|\nu_s\|\right\}\\
\le\;&\frac{p}{n}\left\{-\frac{1}{2}\delta^2(C_{\min}+o_P(1))+\delta C_3\right\},
\end{aligned}\tag{A.15}$$

where $\nu_s=n(\boldsymbol{\gamma}_s-\widehat{\boldsymbol{\beta}}_{(s)})/\delta\sqrt{p}$ and $\widetilde{\boldsymbol{\gamma}}_s$ lies between $\boldsymbol{\gamma}_s$ and $\widehat{\boldsymbol{\beta}}_{(s)}$. The penultimate inequality is due to (15) of Condition 4 and the fact that $\|\gamma_s-\boldsymbol{\beta}^*_{(s)}\|\le\|\gamma_s-\widehat{\boldsymbol{\beta}}_{(s)}\|+\|\widehat{\boldsymbol{\beta}}_{(s)}-\boldsymbol{\beta}^*_{(s)}\|\le\delta\sqrt{p}/n+O_P(S^{1/2}p^{1/2}n^{-1/2})=o_P(1)$ when n is large enough. The last inequality is based on (12) in Condition 3. Considering that $l_n^{(y_i-1)}(\boldsymbol{\gamma}_s)$ is concave and reaches its maximum value at $\widehat{\boldsymbol{\beta}}_{(s)}^{(y_i-1)}$, (A.15) further implies that

$$\max_{1\le i\le n}\max_{1\le s\le S}\left\|\widehat{\boldsymbol{\beta}}_{(s)}^{(y_i-1)}-\widehat{\boldsymbol{\beta}}_{(s)}\right\|=O_P(p^{1/2}n^{-1}). \tag{A.16}$$

Combining (A.9) and (A.16), we obtain that

$$\begin{aligned}
\max_{1\le i\le n}\max_{1\le s\le S}\left\|\widehat{\boldsymbol{\beta}}_{(s)}^{(y_i-1)}-\boldsymbol{\beta}^*_{(s)}\right\| &\le \max_{1\le i\le n}\max_{1\le s\le S}\left\|\widehat{\boldsymbol{\beta}}_{(s)}^{(y_i-1)}-\widehat{\boldsymbol{\beta}}_{(s)}\right\|+\max_{1\le i\le n}\max_{1\le s\le S}\left\|\widehat{\boldsymbol{\beta}}_{(s)}-\boldsymbol{\beta}^*_{(s)}\right\|\\
&= O_P(p^{1/2}n^{-1})+O_P(S^{1/2}p^{1/2}n^{-1/2})\\
&= O_P(S^{1/2}p^{1/2}n^{-1/2}).
\end{aligned}$$

□

A.2 Proof of Theorem 1

From the proof of Theorem 1' of Wan et al. (2010), to prove (18), it is sufficient to verify that

$$\sup_{\mathbf{w}\in\mathcal{W}_n}\frac{|\mathrm{KL}(\mathbf{w})-\mathrm{KL}^*(\mathbf{w})|}{\mathrm{KL}^*(\mathbf{w})}=o_p(1) \tag{A.17}$$

and

$$\sup_{\mathbf{w}\in\mathcal{W}_n}\frac{|\mathrm{KL}(\mathbf{w})-\mathcal{C}(\mathbf{w})|}{\mathrm{KL}^*(\mathbf{w})}=o_p(1). \tag{A.18}$$

First, from (12) and (13) of Condition 3, we have

$$\max_{1\leq s\leq S}\sup_{\boldsymbol{\beta}_{(s)}\in O(\boldsymbol{\beta}^*_{(s)},\rho)}\frac{1}{\sqrt{p_s}n}\sum_{i=1}^{n}\exp\left(\mathbf{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\beta}_{(s)}\right)\|\mathbf{x}_{(s),i}\|\leq C_3C_4. \tag{A.19}$$

Next, by the definition of KL divergence in (5) and the differential mean value theorem, we obtain that, when n is large enough,

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{W}_n}|\mathrm{KL}(\mathbf{w})-\mathrm{KL}^*(\mathbf{w})|\\
=\;&\sup_{\mathbf{w}\in\mathcal{W}_n}\left|\sum_{i=1}^{n}[\widehat{\mu}_i(\mathbf{w},\mathbf{y})-\mu_i\log\{\widehat{\mu}_i(\mathbf{w},\mathbf{y})\}]-\sum_{i=1}^{n}[\mu_i^*(\mathbf{w})-\mu_i\log\{\mu_i^*(\mathbf{w})\}]\right|\\
\leq\;&\sup_{\mathbf{w}\in\mathcal{W}_n}\sum_{i=1}^{n}|\widehat{\mu}_i(\mathbf{w},\mathbf{y})-\mu_i^*(\mathbf{w})|+\sup_{\mathbf{w}\in\mathcal{W}_n}\sum_{i=1}^{n}\mu_i\left|\log\left\{\widehat{\mu}_i(\mathbf{w},\mathbf{y})\right\}-\log\left\{\mu_i^*(\mathbf{w})\right\}\right|\\
\leq\;&\sup_{\mathbf{w}\in\mathcal{W}_n}\sum_{i=1}^{n}\left|\exp\left(\sum_{s=1}^{S}w_s\mathbf{x}_{(s),i}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_{(s)}\right)\sum_{s=1}^{S}w_s\mathbf{x}_i^{\mathrm{T}}\Pi_s\left(\widehat{\boldsymbol{\beta}}_{(s)}-\boldsymbol{\beta}^*_{(s)}\right)\right|\\
&+\sum_{i=1}^{n}\mu_i\|\mathbf{x}_i\|\sup_{\mathbf{w}\in\mathcal{W}_n}\sum_{s=1}^{S}w_s\left\|\Pi_s\widehat{\boldsymbol{\beta}}_{(s)}-\Pi_s\boldsymbol{\beta}^*_{(s)}\right\|\\
\leq\;&\sup_{\mathbf{w}\in\mathcal{W}_n}\sum_{s=1}^{S}\sum_{i=1}^{n}w_s\exp\left(\mathbf{x}_{(s),i}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_{(s)}\right)\|\mathbf{x}_i\|\max_{1\leq s\leq S}\left\|\widehat{\boldsymbol{\beta}}_{(s)}-\boldsymbol{\beta}^*_{(s)}\right\|\\
&+\sum_{i=1}^{n}\mu_i\|\mathbf{x}_i\|\max_{1\leq s\leq S}\left\|\widehat{\boldsymbol{\beta}}_{(s)}-\boldsymbol{\beta}^*_{(s)}\right\|\\
\leq\;&\max_{1\leq s\leq S}\sum_{i=1}^{n}\exp\left(\mathbf{x}_{(s),i}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_{(s)}\right)\|\mathbf{x}_i\|\max_{1\leq s\leq S}\left\|\widehat{\boldsymbol{\beta}}_{(s)}-\boldsymbol{\beta}^*_{(s)}\right\|\\
&+\sum_{i=1}^{n}\mu_i\|\mathbf{x}_i\|\max_{1\leq s\leq S}\left\|\widehat{\boldsymbol{\beta}}_{(s)}-\boldsymbol{\beta}^*_{(s)}\right\|\\
=\;&O_P(p^{1/2}n)O_P\left(S^{1/2}p^{1/2}n^{-1/2}\right)+O(p^{1/2}n)O_P(S^{1/2}p^{1/2}n^{-1/2})
\end{aligned}$$

$$= \quad O_P(S^{1/2}pn^{1/2}), \tag{A.20}$$

where $\widetilde{\boldsymbol{\beta}}_{(s)}$ lies between $\widehat{\boldsymbol{\beta}}_{(s)}$ and $\boldsymbol{\beta}^*_{(s)}$, and the last second equality is due to (A.19) and Lemmas 2 and 4.

Finally, using Conditions 2–4 and the triangle inequality, we have

$$\begin{aligned}
& \sup_{\mathbf{w}\in\mathcal{W}_n} |\mathrm{KL}(\mathbf{w}) - \mathcal{C}(\mathbf{w})| \\
= \quad & \sup_{\mathbf{w}\in\mathcal{W}_n} \left| \sum_{i=1}^n y_i \log\left\{\widehat{\mu}_i(\mathbf{w}, \mathbf{y}^{(y_i-1)})\right\} - \mu_i \log\{\widehat{\mu}_i(\mathbf{w}, \mathbf{y})\} \right| \\
\leq \quad & \sup_{\mathbf{w}\in\mathcal{W}_n} \left| \sum_{i=1}^n (y_i - \mu_i) \log\{\mu_i^*(\mathbf{w})\} \right| \\
& + \sup_{\mathbf{w}\in\mathcal{W}_n} \left| \sum_{i=1}^n y_i \left[\log\left\{\widehat{\mu}_i(\mathbf{w}, \mathbf{y}^{(y_i-1)})\right\} - \log\{\mu_i^*(\mathbf{w})\} \right] \right| \\
& + \sup_{\mathbf{w}\in\mathcal{W}_n} \left| \sum_{i=1}^n \mu_i \left[\log\{\widehat{\mu}_i(\mathbf{w}, \mathbf{y})\} - \log\{\mu_i^*(\mathbf{w})\} \right] \right| \\
\leq \quad & \left\| \sum_{i=1}^n \varepsilon_i \mathbf{x}_i \right\| \sup_{\mathbf{w}\in\mathcal{W}_n} \left\| \sum_{s=1}^S w_s \Pi_s \boldsymbol{\beta}^*_{(s)} \right\| \\
& + \sup_{\mathbf{w}\in\mathcal{W}_n} \left| \sum_{i=1}^n y_i \mathbf{x}_i^{\mathrm{T}} \sum_{s=1}^S w_s \left(\Pi_s \widehat{\boldsymbol{\beta}}^{(y_i-1)}_{(s)} - \Pi_s \boldsymbol{\beta}^*_{(s)} \right) \right| \\
& + \sum_{i=1}^n \mu_i \|\mathbf{x}_i\| \sup_{\mathbf{w}\in\mathcal{W}_n} \sum_{s=1}^S w_s \left\| \Pi_s \widehat{\boldsymbol{\beta}}_{(s)} - \Pi_s \boldsymbol{\beta}^*_{(s)} \right\| \\
\leq \quad & \left\| \sum_{i=1}^n \varepsilon_i \mathbf{x}_i \right\| \max_{1\leq s\leq S} \|\boldsymbol{\beta}^*_{(s)}\| \\
& + \left| \sum_{i=1}^n y_i \mathbf{x}_i \right| \max_{1\leq i\leq n} \max_{1\leq s\leq S} \left\| \widehat{\boldsymbol{\beta}}^{(y_i-1)}_{(s)} - \boldsymbol{\beta}^*_{(s)} \right\| \\
& + \sum_{i=1}^n \mu_i \|\mathbf{x}_i\| \max_{1\leq s\leq S} \left\| \widehat{\boldsymbol{\beta}}_{(s)} - \boldsymbol{\beta}^*_{(s)} \right\| \\
= \quad & O_P(pn^{1/2}) + \left\{ O(p^{1/2}n) + O_P(p^{1/2}n^{1/2}) \right\} O_P(S^{1/2}p^{1/2}n^{-1/2}) \\
& + O(p^{1/2}n) O_P(S^{1/2}p^{1/2}n^{-1/2}) \\
= \quad & O_P(S^{1/2}pn^{1/2}),
\end{aligned} \tag{A.21}$$

where $\mu_i^*(\mathbf{w}) = \exp(\sum_{s=1}^S w_s \mathbf{x}^{\mathrm{T}}_{(s),i} \boldsymbol{\beta}^*_{(s)})$, and the last second equality is due to Lemmas 2, 4 and 5 and (14) in Condition 3.

Assuming that Condition 5 is satisfied, the above two results (A.20) and (A.21) imply (A.17) and (A.18), respectively. This completes the proof.

A.3 Proof of Theorem 2

Assume that the s_0^{th} model is a correct candidate model, we know that $\|\widehat{\boldsymbol{\beta}}_{s_0}-\boldsymbol{\beta}_{\text{true}}\| = O_P(p^{1/2}n^{-1/2})$. Let $\mathbf{w}_{s_0}^{o}$ denotes the weight vector whose s_0^{th} element is 1 and others are 0. We note

$$
\begin{aligned}
\mathcal{C}(\mathbf{w}_{s_0}^{o}) &= \log f(\mathbf{y}) + \sum_{i=1}^{n}\left\{\exp(\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{s_0}) + \log(y_i!) - y_i\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{s_0}^{(y_i-1)}\right\} \\
&= \log f(\mathbf{y}) + \sum_{i=1}^{n}\left\{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\text{true}}) + \log(y_i!) - y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\text{true}}\right\} \\
&\quad + \sum_{i=1}^{n}\Big\{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\text{true}})\mathbf{x}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_{s_0}-\boldsymbol{\beta}_{\text{true}}) + \frac{1}{2}(\widehat{\boldsymbol{\beta}}_{s_0}-\boldsymbol{\beta}_{\text{true}})^{\mathrm{T}}\exp(\mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}})\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_{s_0}-\boldsymbol{\beta}_{\text{true}}) \\
&\quad - y_i\mathbf{x}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_{s_0}^{(y_i-1)}-\boldsymbol{\beta}_{\text{true}})\Big\} \\
&= \log f(\mathbf{y}) + \sum_{i=1}^{n}\left\{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\text{true}}) + \log(y_i!) - y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\text{true}}\right\} \\
&\quad + p\Bigg[-\frac{1}{\sqrt{pn}}\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i^{\mathrm{T}}\sqrt{\frac{n}{p}}\left(\widehat{\boldsymbol{\beta}}_{s_0}-\boldsymbol{\beta}_{\text{true}}\right) \\
&\quad + \frac{1}{2}\sqrt{\frac{n}{p}}\left(\widehat{\boldsymbol{\beta}}_{s_0}-\boldsymbol{\beta}_{\text{true}}\right)^{\mathrm{T}} I_n(\widetilde{\boldsymbol{\beta}})\sqrt{\frac{n}{p}}\left(\widehat{\boldsymbol{\beta}}_{s_0}-\boldsymbol{\beta}_{\text{true}}\right) \\
&\quad - \sum_{i=1}^{n}\left(\frac{1}{\sqrt{pn}}y_i\mathbf{x}_i^{\mathrm{T}}\right)\frac{n}{\sqrt{p}}\left(\widehat{\boldsymbol{\beta}}_{s_0}^{(y_i-1)}-\widehat{\boldsymbol{\beta}}_{s_0}\right)\Bigg],
\end{aligned}
\tag{A.22}
$$

where $I_n(\widetilde{\boldsymbol{\beta}}) = n^{-1}\sum_{i=1}^{n}\exp(\mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}})\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$, $\varepsilon_i = y_i - \exp\left(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\text{true}}\right)$, and $\widetilde{\boldsymbol{\beta}}$ lies between $\widehat{\boldsymbol{\beta}}_{s_0}$ and $\boldsymbol{\beta}_{\text{true}}$.

For notation convenience, we define

$$
\begin{aligned}
\eta_n &= -\frac{1}{\sqrt{pn}}\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i^{\mathrm{T}}\sqrt{\frac{n}{p}}\left(\widehat{\boldsymbol{\beta}}_{s_0}-\boldsymbol{\beta}_{\text{true}}\right) \\
&\quad + \frac{1}{2}\sqrt{\frac{n}{p}}\left(\widehat{\boldsymbol{\beta}}_{s_0}-\boldsymbol{\beta}_{\text{true}}\right)^{\mathrm{T}} I_n(\widetilde{\boldsymbol{\beta}})\sqrt{\frac{n}{p}}\left(\widehat{\boldsymbol{\beta}}_{s_0}-\boldsymbol{\beta}_{\text{true}}\right) \\
&\quad - \sum_{i=1}^{n}\left(\frac{1}{\sqrt{pn}}y_i\mathbf{x}_i^{\mathrm{T}}\right)\frac{n}{\sqrt{p}}\left(\widehat{\boldsymbol{\beta}}_{s_0}^{(y_i-1)}-\widehat{\boldsymbol{\beta}}_{s_0}\right),
\end{aligned}
$$

and under Conditions 2–4, we have

$$
|\eta_n| \leq \sqrt{\frac{n}{p}}\left\|\frac{1}{\sqrt{pn}}\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i\right\|\left\|\widehat{\boldsymbol{\beta}}_{s_0}-\boldsymbol{\beta}_{\text{true}}\right\| + \frac{n}{p}\lambda_{\max}\{I_n(\widetilde{\boldsymbol{\beta}})\}\left\|\widehat{\boldsymbol{\beta}}_{s_0}-\boldsymbol{\beta}_{\text{true}}\right\|^2
$$

$$
\begin{aligned}
&+\frac{n}{2\sqrt{p}}\left\|\frac{1}{\sqrt{p}n}\sum_{i=1}^{n} y_i\mathbf{x}_i^{\mathrm{T}}\right\| \max_{1\le i\le n}\left\|\widehat{\boldsymbol{\beta}}_{s_0}^{(y_i-1)}-\widehat{\boldsymbol{\beta}}_{s_0}\right\| \\
= \;& O_P(1)+(C_{\max}+o_P(1))O_P(1)+O_P(1) \\
= \;& O_P(1), \qquad \text{(A.23)}
\end{aligned}
$$

where the last second equality is implied by (16) in Condition 4, Lemmas 2 and 5, and the fact that $\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_{s_0}^*\| \le \|\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{s_0}\|+\|\widehat{\boldsymbol{\beta}}_{s_0}-\boldsymbol{\beta}_{s_0}^*\| \le \|\boldsymbol{\beta}_{\text{true}}-\widehat{\boldsymbol{\beta}}_{s_0}\|+\|\widehat{\boldsymbol{\beta}}_{s_0}-\boldsymbol{\beta}_{s_0}^*\| = O_P(p^{1/2}n^{-1/2})+O_P(S^{1/2}p^{1/2}n^{-1/2}) = o_P(1)$ when n is large enough.

Note that from Lemma 2 and Conditions 3 and 4, when n is large enough,

$$
\begin{aligned}
\max_{1\le i\le n}\left\|\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})^{(y_i-1)}-\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})\right\| &= \max_{1\le i\le n}\left\|\sum_{s=1}^{S}\widehat{w}_s\Pi_s\left(\widehat{\boldsymbol{\beta}}_{(s)}^{(y_i-1)}-\widehat{\boldsymbol{\beta}}_{(s)}\right)\right\| \\
&\le \max_{1\le i\le n}\max_{1\le s\le S}\left\|\widehat{\boldsymbol{\beta}}_{(s)}^{(y_i-1)}-\widehat{\boldsymbol{\beta}}_{(s)}\right\| \\
&= O_P\left(p^{1/2}n^{-1}\right). \qquad \text{(A.24)}
\end{aligned}
$$

Furthermore, under Conditions 6-7, we have

$$
\begin{aligned}
\max_{1\le i\le n}|\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})| &\le \max_{1\le i\le n}\left|\sum_{s=1}^{S}\widehat{w}_s\mathbf{x}_i^{\mathrm{T}}\left(\widehat{\boldsymbol{\beta}}_s-\boldsymbol{\beta}_s^*\right)\right|+\max_{1\le i\le n}\left|\sum_{s=1}^{S}\widehat{w}_s\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_s^*\right| \\
&\le \max_{1\le i\le n}\max_{1\le s\le S}\|\mathbf{x}_i\|\left\|\widehat{\boldsymbol{\beta}}_s-\boldsymbol{\beta}_s^*\right\|+\max_{1\le i\le n}\max_{1\le s\le S}|\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_s^*| \\
&= O_P(S^{1/2}pn^{-1/2})+C_6 \\
&= C_6+o_P(1). \qquad \text{(A.25)}
\end{aligned}
$$

Let $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})=\sum_{s=1}^{S}\widehat{w}_s\Pi_s\widehat{\boldsymbol{\beta}}_{(s)}$, where $\widehat{\mathbf{w}}$ minimizes $\mathcal{C}(\mathbf{w})$. Combining with (A.22) and using Taylor expansion, when n is large enough we have

$$
\begin{aligned}
&\mathcal{C}(\widehat{\mathbf{w}})-\mathcal{C}(\mathbf{w}_{s_0}^o) \\
= \;& p\Bigg[-\frac{1}{\sqrt{pn}}\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i^{\mathrm{T}}\sqrt{\frac{n}{p}}\left\{\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})-\boldsymbol{\beta}_{\text{true}}\right\} \\
&+\frac{1}{2}\sqrt{\frac{n}{p}}\left\{\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})-\boldsymbol{\beta}_{\text{true}}\right\}^{\mathrm{T}} I_n(\widetilde{\boldsymbol{\beta}})\sqrt{\frac{n}{p}}\left\{\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})-\boldsymbol{\beta}_{\text{true}}\right\} \\
&-\sum_{i=1}^{n}\left(\frac{1}{\sqrt{p}n}y_i\mathbf{x}_i^{\mathrm{T}}\right)\frac{n}{\sqrt{p}}\left\{\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})^{(y_i-1)}-\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})\right\}\Bigg]-p\eta_n \\
= \;& p\Bigg[-\frac{1}{\sqrt{pn}}\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i^{\mathrm{T}}\nu+\frac{1}{2}\nu^{\mathrm{T}}I_n(\widetilde{\boldsymbol{\beta}})\nu \\
&-\sum_{i=1}^{n}\left(\frac{1}{\sqrt{p}n}y_i\mathbf{x}_i^{\mathrm{T}}\right)\frac{n}{\sqrt{p}}\left\{\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})^{(y_i-1)}-\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})\right\}\Bigg]-p\eta_n
\end{aligned}
$$

$$
\begin{aligned}
&\geq p\left[\frac{1}{2}\{c_0\exp(-C_6)+o_P(1)\}\|\nu\|^2-\left\|\frac{1}{\sqrt{pn}}\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i^{\mathrm{T}}\right\|\|\nu\|\right.\\
&\quad\left.-\left\|\frac{1}{\sqrt{pn}}\sum_{i=1}^{n}y_i\mathbf{x}_i^{\mathrm{T}}\right\|\frac{n}{\sqrt{p}}\max_{1\leq i\leq n}\left\|\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})^{(y_i-1)}-\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})\right\|-|\eta_n|\right]\\
&= p\left[2^{-1}\{c_0\exp(-C_6)+o_P(1)\}\|\nu\|^2-\|\nu\|O_P(1)-O_P(1)-O_P(1)\right], \qquad \text{(A.26)}
\end{aligned}
$$

where $\nu=\sqrt{n}\{\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})-\boldsymbol{\beta}_{\text{true}}\}/\sqrt{p}$, and $\widetilde{\boldsymbol{\beta}}$ lies between $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$ and $\boldsymbol{\beta}_{\text{true}}$. The last inequality is due to

$$
\begin{aligned}
\lambda_{\min}\{I_n(\widetilde{\boldsymbol{\beta}})\}&=\lambda_{\min}\left\{\frac{1}{n}\sum_{i=1}^{n}\exp(\mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}})\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right\}\\
&\geq\min_{1\leq i\leq n}\inf_{t\in(0,1)}\exp\left\{t\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})+(1-t)\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\text{true}}\right\}\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\\
&\geq\exp\left\{-\max_{1\leq i\leq n}\sup_{t\in(0,1)}t|\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})|+(1-t)|\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\text{true}}|\right\}c_0\\
&\geq c_0\exp\{-C_6+o_P(1)\}\\
&=c_0\exp(-C_6)+o_P(1), \qquad \text{(A.27)}
\end{aligned}
$$

where $\max\left\{\max_{1\leq i\leq n}|\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})|,\max_{1\leq i\leq n}|\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\text{true}}|\right\}\leq C_6+o_P(1)$ according to (A.25) and Condition 6, and the last equality is based on (A.23), (A.24) and Lemma 2.

According to the definition of $\widehat{\mathbf{w}}$, we see that $\Pr\{\mathcal{C}(\widehat{\mathbf{w}})-\mathcal{C}(\mathbf{w}_s^o)\geq 0\}=0$. This means that the probability of the right-hand side of (A.26) being nonnegative is also zero. Thus, $\|\nu\|$ must be bounded in probability. Therefore, we conclude that

$$
\left\|\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})-\boldsymbol{\beta}_{\text{true}}\right\|=O_P(p^{1/2}n^{-1/2}).
$$

This completes the proof.

A.4 Additional empirical results

Table A.1: Definition of covariates

Pilot CEO	A dummy variable that equals 1 for CEOs with a pilot license and zero otherwise
Log(Assets)	The logarithm of total assets in millions
Log(PPE/EMP)	The logarithm of the ratio of net property, plant, and equipment over the number of employees
Stock return	Firm buy-and-hold return over the fiscal year
Tobin's q	The market value of assets divided by the book value of assets
Inst. holdings	Percentage of shares held by financial institutions
Tenure	The CEO tenure in months
Delta	The dollar change in CEO stock and option portfolio for a 1% change in stock price
Vega	The dollar change in CEO option holdings for a 1% change in stock return volatility
Log(CEO age)	The logarithm of CEO age in years.
Top university	A dummy variable that equals one if the CEOs undergraduate institution is listed as one of the top 50 schools ranked by *U.S. News & World Report* in any year during the period 1983 through 2007 and zero otherwise
Finance education	A dummy variable that equals one if the CEO received a degree in accounting, finance, business (including MBA), or economics and zero otherwise
Technical education	A dummy variable that equals one for CEOs with undergraduate or graduate degrees in engineering, physics, operations research, chemistry, mathematics, biology, pharmacy, or other applied science and zero otherwise
PhD in technical edu.	A dummy variable that equals one for CEOs with a PhD in engineering, technology, science, or mathematics and zero otherwise
No school info.	A dummy variable that equals one if we cannot identify the CEOs undergraduate school and zero otherwise.
Military	A dummy variable that equals one for CEOs with military background and zero otherwise
Overconfidence	A dummy variable that equals one for all years after the CEOs options exceed 67% moneyness and zero otherwise

Table A.2: Descriptive statistics of the outcome variable and covariates

	Mean	Std	Max	Min
R&D	5.504	10.739	50.000	0.000
Pilot CEO	0.085	0.279	1.000	0.000
log(Assets)	6.808	1.459	12.048	1.233
log(PPE/EMP)	2.385	73.202	5316.6	0.023
Stock return	3.828	1.323	9.491	-1.292
Tobin's q	2.169	2.503	78.565	0.404
Inst. holdings	0.442	0.319	1.331	0.000
log(1+Tenure)	3.507	0.762	4.970	0.000
log(1+Delta)	4.537	1.993	11.267	0.000
log(1+Vega)	3.450	1.639	9.192	0.000
log(CEO age)	3.971	0.135	4.382	3.367
Top university	0.143	0.350	1.000	0.000
Finance education	0.001	0.039	1.000	0.000
Technical education	0.014	0.118	1.000	0.000
PhD in technical edu.	0.093	0.290	1.000	0.000
No school info.	0.315	0.464	1.000	0.000
Military	0.021	0.851	1.000	0.000
Overconfidence	0.626	0.484	1.000	0.000

Table A.3: Coefficient estimates of innovation regression with different sets of covariates

	Full model		Model (1)		Model (2)		Model (3)		Model (4)	
	Coef.	p-val.	Coef.	p-val.	Coef.	p-val.	Coef.	p-val.	Coef.	p-val.
Pilot CEO	0.204	0.017	0.219	0.000	0.193	0.000	0.214	0.000	0.218	0.000
log(assets)	0.027	0.236			0.055	0.000				
log(PPE/EMP)	0.069	0.000								
Stock return	−0.0004	0.355								
Tobin’s Q	0.041	0.006								
Inst. holdings	0.180	0.033								
log(1+tenure)	−0.050	0.158								
log(1+delta)	0.015	0.427								
log(1+vega)	0.014	0.580								
log(CEO age)	0.536	0.017	0.409	0.000						
Top university	0.176	0.012	0.218	0.000						
Finance education	−2.14	0.001	−2.093	0.000						
Technical education	0.212	0.165	0.269	0.000						
PhD in tech. edu	−0.065	0.465	−0.074	0.000						
No school info	0.064	0.296	0.013	0.315						
Military	−0.134	0.402	−0.129	0.045			−0.122	0.038		
Overconfidence	0.017	0.774	0.041	0.001					0.0399	0.000
constant	−1.071	0.221	0.000	0.998	1.309	0.000	1.686	0.000	1.660	0.000

*Notes:*All models are estimated using standard Poisson regressions.

Table A.4: The pre-screening procedures for empirical application

Screening 1	Step 1: Compute the absolute value of the bivariate correlation between each covariate and the outcome variable. Step 2: The first candidate model only includes pilot CEO and the intercept. Step 3: The second candidate model contains pilot CEO and the intercept, as well as an additional covariate with the largest absolute value of the bivariate correlation. Step 4: Continue Step 3 by adding one extra covariate at each time based on their bivariate correlations to construct the remaining candidate models, until all covariates are included.
Screening 2	Step 1: Compute the absolute value of the bivariate correlation between each covariate and the outcome variable. Step 2: The first candidate model includes the intercept and a covariate with the largest absolute value of the bivariate correlation. Step 3: Continue to add one extra covariate at each time based on their bivariate correlations to construct the remaining candidate models, until all covariates are included.