Bayesian Analysis (2023) **TBA**, Number TBA, pp. 1–27

A Conditional Bayesian Approach with Valid Inference for High Dimensional Logistic Regression*

Abhishek Ojha† and Naveen N. Narisetty†

Abstract. We consider the problem of performing inference for a continuous treatment effect on a binary outcome variable while controlling for high dimensional baseline covariates. We propose a novel Bayesian framework for performing inference for the desired low-dimensional parameter in a high-dimensional logistic regression model. While it is relatively easier to address this problem in linear regression, the nonlinearity of the logistic regression poses additional challenges that make it difficult to orthogonalize the effect of the treatment variable from the nuisance variables. Our proposed approach provides the first Bayesian alternative to the recent frequentist developments and can incorporate available prior information on the parameters of interest, which plays a crucial role in practical applications. In addition, the proposed approach incorporates uncertainty in orthogonalization in high dimensions instead of relying on a single instance of orthogonalization as done by frequentist methods. We provide uniform convergence results that show the validity of credible intervals resulting from the posterior. Our method has competitive empirical performance when compared with state-of-the-art methods.

1 Introduction

In the modern era of big data, the analysis of large-scale data sets has become increasingly common across various scientific domains. High-dimensional problems, characterized by data with a large number of variables or features are particularly prevalent in many applications such as healthcare, genomics, finance, and econometrics. In genomics, for instance, researchers analyze large genetic data to identify associations between genetic variations and disease susceptibility. In finance, high-dimensional data are employed to model and predict market trends, and asset prices, and use them to devise investment strategies. While there has been substantial progress in modeling and estimation of high dimensional models both in the contexts of linear regression models and binary regression models (see Section 1.2 for a review of relevant literature), performing inference for certain parameters of importance in the model is still a challenging problem. For instance, in analyzing the efficacy of a drug in treating a medical condition, it is important to provide quantitative inference for the effect of the dosage of the drug while controlling for a large number of covariates in the high dimensional model. Motivated by this, we consider the problem of performing Bayesian inference for low-dimensional parameters of interest in high dimensional logistic regression model.

*The authors gratefully acknowledge funding support from NSF grant CAREER-1943500.
†Department of Statistics, University of Illinois at Urbana-Champaign, ojha4@illinois.edu
†Department of Statistics, University of Illinois at Urbana-Champaign, naveen@illinois.edu

 https://doi.org/10.1214/23-BA1408

1.1 Problem Formulation

Our data consist of a sample of n observations containing binary responses denoted by Y_i, $i = 1, \ldots, n$, q-dimensional target covariates of interest denoted by X_i, $i = 1, \ldots, n$, and d-dimensional nuisance variables (including the intercept) denoted by Z_i, $i = 1, \ldots, n$. We assume that each of the binary response variables is independently generated from the following logistic regression model:

$$P[Y_i = 1 \mid X_i, Z_i, \theta, \beta] = \frac{\exp(X_i\theta + Z_i^T\beta)}{1 + \exp(X_i\theta + Z_i^T\beta)}, \tag{1.1}$$

for $i = 1, \ldots, n$, where θ and β are random parameters having dimensions q and d, respectively.

We consider the problem of providing valid inference for the low dimensional parameter θ in the above high dimensional logistic regression model where the dimension d of the nuisance parameter β can be much larger than the sample size. Our goal is to construct a posterior distribution for the parameter θ given the data (Y, X, Z) that can be used for obtaining interval estimators having valid frequentist coverage. We would like the validity of our inferential results to hold for all values of θ.

1.2 Literature Review

For high dimensional data, the number of covariates d can be much larger than the sample size n. Statistical estimation in such problems is ill-posed unless further assumptions about sparsity or smoothness are made on the high dimensional parameters. It has led to extensive development of various high dimensional regularization and variable selection methods both from frequentist and Bayesian paradigms. Instead of searching for the optimal parameter in the whole $\mathbb{R}^d$ space, these methods restrict the parameter search to smaller subspaces through the use of an additional penalty or regularization over the parameters.

Some notable examples of penalty-based methods in the case of linear models include LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), and MCP (Zhang, 2010). Penalized variable selection methods along with their theoretical properties have been extended to the case of Generalized Linear Models (GLMs) as well (Van de Geer, 2008; Friedman et al., 2010; Breheny and Huang, 2011; Huang and Zhang, 2012). Approaches have been developed for logistic regression in particular (Bach, 2010; Bunea, 2008; Kwemou, 2016). In the Bayesian paradigm, several methods have been developed for high dimensional estimation and variable selection using prior distributions that induce sparsity or shrinkage (George and McCulloch, 1993; Ishwaran and Rao, 2005; Park and Casella, 2008; Liang et al., 2008; Carvalho et al., 2009; Johnson and Rossell, 2012; Bondell and Reich, 2012; Armagan et al., 2013; Ročková and George, 2014; Narisetty and He, 2014; Bhattacharya et al., 2015). Bayesian approaches have also been developed for generalized linear models and logistic regression in particular (O'brien and Dunson, 2004; Genkin et al., 2007; Ghosh et al., 2018; Narisetty et al., 2019).

The aforementioned penalization methods have been shown to possess oracle properties under certain regularity conditions and sparsity assumptions. These results imply that the estimation error converges to zero and the true set of covariates can be recovered with high probability under regularity conditions. Selection consistency and Strong Selection consistency are well-studied in the Bayesian literature (Johnson and Rossell, 2012; Ročková and George, 2014; Narisetty and He, 2014). The asymptotic normality of posterior distributions in one-dimensional parameter settings and low-dimensional multivariate settings have been established (Walker, 1969; Schervish, 2012). The asymptotic normality of posteriors in generalized linear models has been proven when the number of covariates is much smaller than the sample size n (Ghosal, 1997, 2000; Dasgupta et al., 2014).

Variable selection methods are not directly useful for inference tasks where we need to construct valid credible or confidence intervals for the model parameters. Theoretical consistency for selection requires that the minimum signal is sufficiently bounded away from zero. This assumption can be suitable for selection tasks but the inferential results eventually break down when the true signal for some covariates is close to zero and it becomes difficult to differentiate such covariates from noise. Moreover, the bias due to high dimensional estimation is in the order of $(\log d/n)^{1/2}$ (where d is the number of covariates), which is much larger than the standard $n^{-1/2}$ rate of convergence in low dimensions. The impact of model selection on inference has been discussed in the literature (Pötscher, 1991; Kabaila, 1995). Examples have been investigated where the signal does not meet the minimum signal strength condition and consequently, the corresponding estimates are no longer $\sqrt{n}$-consistent and asymptotically normal (Leeb and Pötscher, 2005, 2008; Pötscher and Leeb, 2009). Furthermore, it has been shown that the confidence sets are necessarily larger for the methods with oracle properties when the signal strength assumption is violated (Pötscher, 2009).

In the past decade, multiple methods have proposed estimators with the desired $\sqrt{n}$-consistency which is essential to provide valid inference in different high-dimensional settings. Broadly, these approaches can be classified into two categories: exploratory and selective. In the first line of research, users first perform the variable selection and then aim to provide asymptotically valid confidence regions for the selected set of covariates. There are some approaches that aim at providing valid inference for all the model parameters irrespective of the model selection criteria that the user utilizes (Berk et al., 2013; Kuchibhotla et al., 2020). These methods possess the desired asymptotic consistency and normality but their practical utility is often suboptimal in high dimensions. The volume of confidence regions for these methods becomes too large in practice even for a slightly large number of covariates even when $d < n$.

In the second line of research, the goal is to perform inference on a given set of variables of interest (such as, treatment variables) and we want to conduct valid inference controlling the effect of other high dimensional covariates. These methods do not assume any minimum signal strength for the true signal and have been developed for both the linear models and the generalized linear models (Taylor and Tibshirani, 2015; Panigrahi and Taylor, 2018; Panigrahi et al., 2021; Van de Geer et al., 2014; Javanmard and Montanari, 2014; Zhang and Zhang, 2014; Belloni et al., 2010, 2012, 2014; Chernozhukov et al., 2018; Wang et al., 2020; Belloni et al., 2013). A Bayesian method based on conditional posteriors has been recently proposed for valid inference in linear models which forms the methodological motivation for our paper (Wu et al., 2023).

Our specific goal of constructing a posterior distribution for θ with valid frequentist properties has not been previously studied in the context of high dimensional logistic regression. The conditional posterior method developed for linear models by Wu et al. (2023) is demonstrated to have asymptotic normality at an optimal $n^{-1/2}$ rate. However, the method heavily relies on the linearity of the regression model. Bayesian model averaging approaches have been developed (Torrens-i Dinarès et al., 2021; Antonelli et al., 2022) for inference of average treatment effect which is different from the parameter θ of our interest. In the frequentist paradigm, the problem of constructing valid confidence intervals for low-dimensional parameters in a high-dimensional model has received more attention. Double selection and double machine learning which are based on the idea of Neyman orthogonality (Neyman, 1959, 1979) have been explored for the construction of confidence intervals in linear regression (Belloni et al., 2010, 2012, 2014; Chernozhukov et al., 2018). Belloni et al. (2013) extended the double selection approach to the logistic regression setting. Shi et al. (2021) proposed an inference procedure for a low-dimensional parameter in a high-dimensional logistic regression setting by devising a recursive online-score estimation approach. Ma et al. (2021) proposed a generalized low dimensional projection method for logistic regression which can be used to construct global test statistics and confidence intervals for model parameters. Cai et al. (2021) proposed a two-step bias correction method for generalized linear models for constructing confidence intervals and performing simultaneous hypothesis testing for each component of the regression vector. All these recent approaches for inference in high dimensional logistic regression are based on the frequentist framework whereas our proposed approach aims to devise a Bayesian paradigm that has valid frequentist properties.

1.3 Our Contributions

In this paper, we construct a conditional posterior for θ which is motivated by Neyman orthogonality principle (Neyman, 1959, 1979). As opposed to the double selection-based method (Belloni et al., 2013) which only uses a single instance of the nuisance parameter selection, the proposed Bayesian approach has the advantage of utilizing the complete posterior distributions of the nuisance parameters. We shall show that the posterior distribution of the parameter of interest achieves asymptotic normality (uniformly at $n^{-1/2}$ rate) with valid inference properties. Moreover, our proposed approach inherits the qualities of a standard Bayesian approach for the low dimensional parameter of interest where prior information or prior belief about the parameter can be incorporated. The ideas of constructing orthogonal scores (Belloni et al., 2013) and projection-based de-biasing (Ma et al., 2021; Cai et al., 2021) have been used in the frequentist paradigm for obtaining valid inference. Therefore, the novel component of our paper is not the proposal of orthogonalization but rather providing a novel Bayesian framework for this problem that is carefully devised to induce orthogonalization and is demonstrated to have conceptual and empirical advantages over existing methods. The primary contribution of the research is to generalize the conditional posterior approach discussed in linear regression (Wu et al., 2023) to the more challenging situation of logistic regression which demonstrates that the scope of Bayesian inference for high dimensional models is not just restricted to linear model settings. This is particularly a challenging task since developing conditional posteriors based on orthogonality is much more convenient in the linear model setting as opposed to the non-linear logistic regression case.

In Section 2, we introduce the proposed method based on the conditional posteriors and provide motivation behind their construction. In Section 3, we discuss the regularity assumptions and the theoretical properties of our estimator. In Section 4, we perform simulation studies to evaluate the proposed method and compare its performance with existing Bayesian and frequentist methods. In Section 5, we perform real data analysis on an RNA-seq data set. In Section 6, we provide a final conclusion.

2 Bayesian Conditional Posterior for Inference

We first briefly discuss the concept of conditional posterior (Wu et al., 2023) before providing a detailed explanation of the proposed method. In a standard Bayesian procedure, a model assumption is made, which leads to a likelihood for the model parameters. Given a prior distribution for the model parameters and the likelihood based on the observed data, Bayes' formula is used to derive a posterior distribution for the model parameters.

The intuition behind the conditional posterior, as discussed in this paper, is slightly different. We begin with a prior distribution for the model parameters and given the observed data, we directly propose a distribution for the model parameters. This proposal is referred to as the "conditional posterior". The conditional posterior represents our updated beliefs about the parameters based on the observed data and the conditioning parameters. It is important to note that the data component of this conditional posterior may correspond to a specific working model. The proposal for the conditional posterior and the choice of the working model are driven by theoretical and practical considerations such as the attainment of valid inferential properties and the facilitation of efficient sampling. Our proposed method utilizes the Neyman orthogonality principle for the construction of our conditional posterior that yields valid inference for the parameters of interest.

Notations: Recall that θ is the parameter of interest and β is the high-dimensional nuisance parameter. We use the notations θ_0 and β_0 to denote the corresponding oracle quantities. We denote the diagonal covariance matrix of the binary output Y (given X, Z) using W_0. The diagonal entries of W_0 are defined in terms of (θ_0, β_0) as:

$$W_{0_{i,i}} = \frac{\exp(X_i\theta_0 + Z_i^T\beta_0)}{(1 + \exp(X_i\theta_0 + Z_i^T\beta_0))^2}, \tag{2.1}$$

for $i = 1, 2, \ldots, n$. As W_0 depends on the unknown parameters, we use a sample of the parameters (θ, β) to obtain an estimator for W_0 which we denote by W. However, it is important to note that the sample of θ used in W is not the final sample used for inference on the parameter of interest. To emphasize this distinction, we denote this sample with an additional tilde symbol overhead, representing it as $\tilde{\theta}$. Therefore, the diagonal entries of W are defined in terms of $(\tilde{\theta}, \beta)$ as:

$$W_{i,i} = \frac{\exp(X_i\tilde{\theta} + Z_i^T\beta)}{(1 + \exp(X_i\tilde{\theta} + Z_i^T\beta))^2}, \tag{2.2}$$

for $i = 1, 2, \ldots, n$. We introduce a high-dimensional parameter γ which captures the dependence of the target covariate X on the nuisance covariates Z through a working regression model. Based on $(\tilde{\theta}, \beta)$ and γ, we define a re-parameterized high-dimensional parameter ϕ as $\phi = \tilde{\theta}\gamma + \beta$. Moreover, we denote the oracle values of γ and ϕ using γ_0 and ϕ_0 respectively. We use the symbol "**WM**" to denote the working models and the symbol "**CP**" to denote the conditional posteriors for the parameters.

2.1 Formulation of Proposed Conditional Posteriors

In this paper, for simplicity, we have provided theory and simulations for one-dimensional X with $q = 1$. It is straightforward to generalize these to a finite-dimensional X where q is finite. Our goal is to devise a conditional posterior for θ given (Y, X, Z) which can be used for valid inference. In order to capture the dependence between X and Z and to use it later for inference on θ, we introduce the following working linear regression model which is governed by the regression parameter γ. This approach is consistent with the existing literature (Belloni et al., 2013; Zhang and Zhang, 2014; Ma et al., 2021; Cai et al., 2021; Wu et al., 2023).

Working Model for γ:

The working model for the parameter γ is given by the following regression model with X as the response and Z as the covariates:

$$X \mid (Z, \gamma, \sigma^2, W_0) \sim N(Z\gamma, \sigma^2 W_0^{-1}), \tag{γ-\textbf{WM}}$$

where W_0 is a diagonal matrix with entries $W_{0_{i,i}} = Var(Y_i \mid X_i, Z_i)$ as defined in (2.1) and the parameter σ^2 denotes the variance of the residuals $W_{0_{i,i}}^{1/2}(X_i - Z_i^T\gamma_0)$. Belloni et al. (2013) argued that the inclusion of such a parameter is crucial to establish an orthogonal score for the parameter of interest θ. The role of this additional parameter γ is akin to the direction of projections used in de-biasing methods in the literature (Zhang and Zhang, 2014; Ma et al., 2021; Cai et al., 2021; Wu et al., 2023).

With an appropriate prior distribution over γ and the working model provided by (γ-**WM**), we can construct a conditional posterior for γ. However, given the observed data (X, Y, Z), we do not have the knowledge of W_0. Therefore, we use a sample of the parameters $(\tilde{\theta}, \beta)$ to obtain an estimator for W_0 denoted as W having diagonal elements as defined in (2.2). The tilde notation in $\tilde{\theta}$ emphasizes that this sample is different from the final sample used for inference on the parameter of interest defined later by (θ-**CP**).

The working model for $(\tilde{\theta}, \beta)$ is the same as the original logistic model given by (1.1).

Consequently, we obtain samples of $(\tilde{\theta}, \beta)$ based on the following conditional posterior:

$$\boxed{f((\tilde{\theta}, \beta) \mid Y, X, Z) \propto \pi(\tilde{\theta})\pi(\beta) \prod_{i=1}^{n} \frac{\exp(X_i\tilde{\theta} + Z_i^T\beta)^{Y_i}}{1 + \exp(X_i\tilde{\theta} + Z_i^T\beta)},} \qquad ((\tilde{\theta}, \beta)\textbf{-CP})$$

where $\pi(\tilde{\theta})$ and $\pi(\beta)$ are the priors on the model parameters θ and β, respectively.

Conditional Posterior for γ:

A sample of $(\tilde{\theta}, \beta)$ based on ($(\tilde{\theta}, \beta)$-**CP**) leads to a sample of W defined based on (2.2). Motivated by the working model (γ-**WM**), we sample γ and σ^2 from the following conditional posterior:

$$\boxed{f((\gamma, \sigma^2) \mid Y, X, Z, \tilde{\theta}, \beta) \propto \pi(\gamma)\pi(\sigma^2) \exp\left(-\frac{1}{2\sigma^2}(X - Z\gamma)^T W (X - Z\gamma)\right),} \quad (\gamma\textbf{-CP})$$

where $\pi(\gamma)$ is the prior on the parameter γ and W is a function of $(\tilde{\theta}, \beta)$ as given in (2.2). Furthermore, we place an inverse-gamma prior ($\pi(\sigma^2)$) on σ^2, which results in an inverse-gamma posterior for σ^2 after integrating out γ from (γ-**CP**).

Motivation Behind the Choice of the Working Model for γ:

We shall now describe the main motivation for the choice of the working model (γ-**WM**). The score function corresponding to our logistic regression model (1.1) takes the following form:

$$\varphi_{\text{usual}}(\theta; X, Y, Z, \beta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log P[Y_i = 1 \mid X_i, Z_i, \theta, \beta] = X^T(Y - \mu), \tag{2.3}$$

where μ is the $n \times 1$ mean vector with $\mu_i = \exp(X_i\theta + Z_i^T\beta)/(1 + \exp(X_i\theta + Z_i^T\beta))$. The expectation of the derivative of this score (2.3) with respect to the nuisance parameter β, when evaluated at the oracle quantities θ_0 and β_0 leads to:

$$E\left[\frac{\partial}{\partial\beta}(\varphi_{\text{usual}}(\theta; X, Y, Z, \beta))\big|_{\theta_0, \beta_0}\right] = -E[X^T W_0 Z] \neq 0. \tag{2.4}$$

The derivative of the usual score for θ with respect to the nuisance parameters (2.4) involves a dot product between $W_0^{1/2}X$ and $W_0^{1/2}Z$, which need not be small. This results in a bias for estimating θ that is usually in the order of $(\log d/n)^{1/2}$ and slower than the desired $n^{-1/2}$ rate (Belloni et al., 2013). On the other hand, the product between $W_0^{1/2}Z$ and the residual obtained after performing a linear regression of $W_0^{1/2}X$ on $W_0^{1/2}Z$ can be expected to be small due to the orthogonality of the residuals and the covariates. Therefore, the working model (γ-**WM**) is based on weighted regression of the target covariate on the nuisance covariates.

Working Model for the Parameter of Interest θ:

We first define a re-parameterized nuisance parameter based on the samples $(\tilde{\theta}, \beta)$ and γ given by:

$$\phi = \tilde{\theta}\gamma + \beta. \tag{2.5}$$

Note that we use $\tilde{\theta}$ in the estimation of this parameter ϕ as it is sampled based on ($(\tilde{\theta}, \beta)$-**CP**) and is different from the conditional posterior for θ used for final inference. This makes the posterior concentration for ϕ and θ not depend on each other. We assume the following working model for the parameter of interest θ conditional on γ and ϕ:

$$P(Y_i = 1 \mid X_i, Z_i, \theta, \gamma, \phi) = \frac{\exp\{(X_i - Z_i^T\gamma)\theta + Z_i^T\phi\}}{1 + \exp\{(X_i - Z_i^T\gamma)\theta + Z_i^T\phi\}}, \tag{θ-WM}$$

for $i = 1, \ldots, n$.

Conditional Posterior for θ:

Based on the working model (θ-**WM**), we obtain a sample for the parameter of interest θ based on the following conditional posterior:

$$\boxed{f(\theta \mid Y, X, Z, \gamma, \phi) \propto \pi(\theta) \prod_{i=1}^{n} \frac{\exp((X_i - Z_i^T\gamma)\theta + Z_i^T\phi)^{Y_i}}{1 + \exp((X_i - Z_i^T\gamma)\theta + Z_i^T\phi)},} \tag{θ-CP}$$

where $\pi(\theta)$ is the prior distribution on the parameter of interest θ.

In conclusion, γ is sampled from its conditional posterior (γ-**CP**), ϕ is sampled based on (γ-**CP**) and ($(\tilde{\theta}, \beta)$-**CP**) together, and θ is sampled based on (θ-**CP**). Algorithm 1 provides an outline of the main steps involved in defining the conditional posterior distributions for all the parameters. These three conditional posteriors together provide a joint posterior for all the parameters involved:

$$f(\theta, \gamma, \phi \mid Y, X, Z) \propto f(\theta \mid Y, X, Z, \gamma, \phi) \times f(\phi \mid Y, X, Z, \gamma) \times f(\gamma \mid Y, X, Z). \tag{2.6}$$

Here, $f(\theta \mid Y, X, Z, \gamma, \phi)$ is given by (θ-**CP**) and $f(\phi \mid Y, X, Z, \gamma)$ is obtained from ($(\tilde{\theta}, \beta)$-**CP**) as ϕ is a linear function of $\tilde{\theta}$ and β given γ. We obtain $f(\gamma \mid Y, X, Z)$ by integrating out $\tilde{\theta}, \beta$, and σ^2 from the product of (γ-**CP**) and ($(\tilde{\theta}, \beta)$-**CP**). It is worth noting that this joint posterior is not based on a single working model for all the parameters together but rather utilizes different working models for each parameter separately as described earlier.

Motivation Behind the Choice of the Working Model of θ:

Given the oracle quantities θ_0 and β_0, the logistic regression model (1.1) takes the form:

$$P(Y_i = 1 \mid X_i, Z_i, \theta_0, \beta_0) = \frac{\exp(X_i\theta_0 + Z_i^T\beta_0)}{1 + \exp(X_i\theta_0 + Z_i^T\beta_0)}. \tag{2.7}$$

An important observation is that the working model (θ-**WM**) given the oracle quantities θ_0, γ_0, and ϕ_0 ($= \gamma_0\theta_0 + \beta_0$) is exactly same as the true model (2.7). To see this:

$$\begin{aligned}
P(Y_i = 1 \mid X_i, Z_i, \theta_0, \gamma_0, \phi_0) &= \frac{\exp\{(X_i - Z_i^T\gamma_0)\theta_0 + Z_i^T\phi_0\}}{1 + \exp\{(X_i - Z_i^T\gamma_0)\theta_0 + Z_i^T\phi_0\}} \\
&= \frac{\exp\{X_i\theta_0 + Z_i^T(\phi_0 - \theta_0\gamma_0)\}}{1 + \exp\{X_i\theta_0 + Z_i^T(\phi_0 - \theta_0\gamma_0)\}} \\
&= \frac{\exp(X_i\theta_0 + Z_i^T\beta_0)}{1 + \exp(X_i\theta_0 + Z_i^T\beta_0)}.
\end{aligned}$$

This implies that the working model (θ-**WM**) retains the same form as the true model under the oracle values of the parameters. This correspondence is essential to capture the essential features of the original model and to yield valid inference.

We shall now discuss how the working model for θ yields an orthogonal score. For given values of γ and ϕ, the score for the working model for θ in (θ-**WM**) is:

$$\varphi_{\text{orth}}(\theta; X, Y, Z, \gamma, \phi) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}[\log P(Y_i = 1 \mid X_i, Z_i, \gamma, \phi]) = (X - Z\gamma)^T(Y - \mu),$$

where μ is n-dimensional vector with each component being $\mu_i = \exp\{(X_i - Z_i^T\gamma)\theta + Z_i^T\phi\}/(1 + \exp\{(X_i - Z_i^T\gamma)\theta + Z_i^T\phi\})$ for $i = 1, \dots, n$.

Next, we will verify the orthogonal property for this score. The derivative of the score with respect to the nuisance parameter ϕ is given by:

$$E\Big[\frac{\partial}{\partial\phi}\varphi_{\text{orth}}(\theta; X, Y, Z, \gamma, \phi)\Big|_{(\theta_0,\gamma_0,\phi_0)}\Big] = -E[(X - Z\gamma_0)^T W_0 Z] = 0, \tag{2.8}$$

where the last equality holds because of Assumption A6. Furthermore, the derivative of the score w.r.t the nuisance parameter γ is given by:

$$E\Big[\frac{\partial}{\partial\gamma}\varphi_{\text{orth}}(\theta; X, Y, Z, \gamma, \phi)\Big|_{(\theta_0,\gamma_0,\phi_0)}\Big] = -E[(X - Z\gamma_0)^T W_0 Z] - E[Z^T(Y - \mu_0)] = 0, \tag{2.9}$$

where μ_0 is the true mean vector $\mu_0 = E[Y \mid X, Z]$. Equations (2.8) and (2.9) assure that the working model (θ-**WM**) gives an orthogonal score for estimating θ. This orthogonal property facilitates obtaining inference for θ at an $n^{-1/2}$ rate, as expected in low-dimensional settings, even though γ and ϕ are high-dimensional.

Further Discussion on the Use of $\tilde{\theta}$:

One can argue that the sample of θ, obtained from equation (θ-**CP**), can be utilized to obtain W in equation (2.2) for estimating W_0. However, this will make the posterior sampling of γ to be conditional on the samples of θ which could interfere with the

Algorithm 1 The Proposed Conditional Posterior Formulation.

1. To get a sample of the nuisance parameter γ, we first need to estimate the unknown diagonal matrix W_0 (2.1), $W_{0_{i,i}} = Var(Y_i \mid X_i, Z_i)$.
2. We estimate W_0 using W based on a sample of $(\tilde{\theta}, \beta)$ as defined in (2.2). Samples of $(\tilde{\theta}, \beta)$ are obtained from the conditional posterior given by (**$(\tilde{\theta}, \beta)$-CP**):
$$f((\tilde{\theta}, \beta) \mid Y, X, Z) \propto \pi(\tilde{\theta})\pi(\beta) \prod_{i=1}^{n} \frac{\exp(X_i\tilde{\theta} + Z_i^T\beta)^{Y_i}}{1 + \exp(X_i\tilde{\theta} + Z_i^T\beta)}.$$
3. Given a sample of W, we obtain a sample of γ from the conditional posterior given by (**γ-CP**):
$$f((\gamma, \sigma^2) \mid Y, X, Z, (\tilde{\theta}, \beta)) \propto \pi(\gamma)\pi(\sigma^2) \exp\left(- \frac{1}{2\sigma^2}(X - Z\gamma)^T W (X - Z\gamma)\right).$$
4. Define a re-parameterization $\phi = \tilde{\theta}\gamma + \beta$. Given a sample of γ and ϕ, we obtain a sample for the parameter of interest θ based on its conditional posterior given by (**θ-CP**):
$$f(\theta \mid Y, X, Z, \gamma, \phi) \propto \pi(\theta) \prod_{i=1}^{n} \frac{\exp((X_i - Z_i^T\gamma)\theta + Z_i^T\phi)^{Y_i}}{1 + \exp((X_i - Z_i^T\gamma)\theta + Z_i^T\phi)}.$$

implicit orthogonalization induced by the conditional posterior (**θ-CP**). Moreover, it could lead to intractability of the final posterior distributions of all the parameters involved which makes the corresponding posterior concentrations difficult to achieve. We use samples $\tilde{\theta}$ while estimating the weights W so that the posterior concentration for γ is achieved separately without relying on the posterior concentration of θ.

Conceptual Advantage of the Proposed Method:

Based on the orthogonal properties in (2.8) and (2.9) one can observe that the assumption about weight matrix W_0 in (3.6) is crucial for ensuring that the orthogonal property holds. In the double selection algorithm (Belloni et al., 2013), this weight matrix is estimated once based on a post-LASSO logistic regression using all the covariates. In the projection-based methods, the debiasing is based on an estimate of the nuisance parameters and a single projection direction (Ma et al., 2021; Cai et al., 2021). In the proposed Bayesian framework, the weight estimates are adaptive to the posterior of nuisance parameters. Moreover, the adaptive sampling of the extra nuisance parameter γ allows extensive exploration of the orthogonality. The samples can depend on different instances of the orthogonality and therefore provide a more complete picture. This intuition is similar to the advantages observed in the use of integrated likelihood methods over profile likelihood methods for eliminating nuisance parameters (Berger et al., 1999; Severini, 2007, 2011). In frequentist approaches, the estimation of nuisance parameters is equivalent to maximizing a penalized likelihood function. As discussed in the aforementioned references, this approach may not fully capture the uncertainty in nuisance parameters, resulting in inadequate uncertainty quantification for the parameter of interest. In contrast, integrated likelihood methods average over multiple conditional likelihoods and do a better job in capturing the uncertainty associated with nuisance parameters. The adaptive sampling employed in the proposed method resembles the averaging behavior of integrated likelihood methods and shares similar advantages.

3 Theoretical Guarantees

3.1 Notations and Definitions

Notations: We use X_i to denote the covariate of interest and $(Z_i)_{d\times 1}$ to denote the high dimensional nuisance covariates for the observation i. We use the term $\text{logit}(x)$ to represent the logistic link function (1.1). $L_n(\theta \mid \gamma, \phi)$ denotes the conditional log-likelihood function of the parameter of interest θ for a given γ and ϕ. $L_n(\theta \mid \gamma_0, \phi_0)$ denotes the conditional log-likelihood function given the oracle nuisance parameters γ_0, ϕ_0. For sequences a_n and b_n, $a_n = O(b_n)$ means $a_n/b_n \leq M$ for some $M > 0$ and $a_n = o(b_n)$ means $a_n/b_n \xrightarrow{n\to\infty} 0$. Convergence in probability is denoted by the symbol $\xrightarrow{\text{P}}$. $o_P(1)$ stands for convergence in probability to zero. $O_P(1)$ stands for stochastically bounded. $\|\cdot\|_0, \|\cdot\|_1$, and $\|\cdot\|_2$ represent the l_0, l_1 and l_2 norms of a vector, respectively.

We first explicitly write the log-likelihoods. Based on (**θ-CP**), the conditional log-likelihood is:

$$L_n(\theta \mid \gamma, \phi) = \sum_{i=1}^{n} Y_i\{(X_i - Z_i^T\gamma)\theta + Z_i^T\phi\} - \log\{1 + \exp((X_i - Z_i^T\gamma)\theta + Z_i^T\phi)\}. \tag{3.1}$$

Similarly, under γ_0 and ϕ_0, the conditional log-likelihood can be written as:

$$L_n(\theta \mid \gamma_0, \phi_0) = \sum_{i=1}^{n} Y_i\{(X_i - Z_i^T\gamma_0)\theta + Z_i^T\phi_0\} - \log\{1 + \exp((X_i - Z_i^T\gamma_0)\theta + Z_i^T\phi_0)\}. \tag{3.2}$$

3.2 Assumptions/Regularity Conditions

A1 *On dimension of nuisance parameters:* $\log d = o(n)$ as $n \to \infty$.

A2 *On the regularity of design:* The nuisance variables are bounded, that is,

$$\max\{|Z_{ij}|, 1 \leq i \leq n, 1 \leq j \leq d\} \leq C, \text{ for some } 0 < C < \infty.$$

A3 *On sparsity of true nuisance parameter:* Suppose the number of non-zero elements in γ_0 is s_1 and the number of non-zero elements in ϕ_0 is s_2. Let $s := \max\{s_1, s_2\}$, then

$$s^2 \log(d \vee n) = o(\sqrt{n}). \tag{3.3}$$

A4 *Concentration of the marginal posterior of nuisance parameters* γ, ϕ*:* Let γ_0 and ϕ_0 be the oracle nuisance parameters from the frequentist perspective. We assume the following properties:

$$P\left(\max\left\{\frac{\|\gamma-\gamma_0\|_1}{s}, \frac{\|\gamma-\gamma_0\|_2}{s^{1/2}}\right\} \geq M\sqrt{\frac{\log(d\vee n)}{n}} \mid Y, X, Z\right) \leq C(d\vee n)^{-c_1}, \tag{3.4}$$

on a set $\mathcal{E}_1$, with $P(\mathcal{E}_1) \geq 1-(d\vee n)^{-c_1}$, and

$$P\left(\max\left\{\frac{\|\phi-\phi_0\|_1}{s}, \frac{\|\phi-\phi_0\|_2}{s^{1/2}}\right\} \geq M\sqrt{\frac{\log(d\vee n)}{n}} \mid Y, X, Z\right) \leq C(d\vee n)^{-c_2}, \tag{3.5}$$

on a set $\mathcal{E}_2$, with $P(\mathcal{E}_2) \geq 1-(d\vee n)^{-c_2}$, where M, C, c_1 and c_2 are some positive constants.

A5 *On prior of the parameter of interest* $\pi(\theta)$*:* The prior density of the parameter of interest $\pi(\theta)$ is continuous at $\theta=\theta_0$, and $\pi(\theta_0)>0$.

A6 *On the oracle quantity* γ_0*:* We assume that γ_0 satisfies the following moment condition:

$$E[(X-Z\gamma_0)^T W_0 Z]=0, \tag{3.6}$$

where W_0 is the diagonal matrix of the variances of the binary response Y (given X, Z) given by (2.1).

Assumption A1 is the growth condition for the number of high dimensional covariates in the model. We consider the setting where the total number of covariates can increase as the sample size increases at a sub-exponential rate. Assumption A2 requires that each component of the nuisance covariate vector is bounded. Assumption A3 presents a bound on the level of sparsity for the nuisance parameters. Assumption A4 intuitively means that with high probability, the posterior of the nuisance parameter concentrates around their oracle values at $\sqrt{\log(d\vee n)/n}$ rate. This is the standard concentration rate in high-dimensional settings and is commonly satisfied by most regularization and selection methods both from the frequentist or Bayesian paradigms (Ishwaran and Rao, 2005; Fan and Li, 2001; Van de Geer, 2008; Liang et al., 2008; Castillo et al., 2015; Song and Liang, 2022). In this paper, we use the Skinny Gibbs (Narisetty et al., 2019) method for the nuisance estimation which satisfies this concentration rate as well. However, it is important to note that this concentration rate does not guarantee valid inference, and the resulting estimates may exhibit significant bias. As mentioned in Section 1, this concentration rate is slower than the desired $1/\sqrt{n}$ rate of convergence. Assumption A5 guarantees the positive support of prior at θ_0 which can be verified in the case of a Gaussian prior for example. Assumption A6 is made to achieve orthogonality. This assumption has been previously considered in the existing literature (Belloni et al., 2013). This assumption can be relaxed further to $E[(X-Z\gamma_0)^T W_0 Z]=o(n^{1/2})$ to achieve similar theoretical properties.

Remark 1 (On the use of eigenvalue condition)**.** *In our theoretical analysis, we assumed that the nuisance covariates are bounded, but we did not impose further regularity conditions on the maximum eigenvalue of the covariate matrix. Although our sparsity requirement $s^2 \log(d \vee n) = o(\sqrt{n})$ is stronger than some other results in the literature due to this, we can also relax it with an additional assumption. Define the maximum s-sparse empirical eigenvalues of the nuisance design matrix as:*

$$\phi_{\max}(s) = \max_{1 \le \|v\|_0 \le s} \frac{\frac{1}{n}\sum_{i=1}^{n} \|Z_i^T v\|_2^2}{\|v\|_2^2}. \tag{3.7}$$

Consider some fixed sequence of constants $l_n \to \infty$ and $\delta_n \to 0$ and a constant $0 < C < \infty$. If we impose further regularity conditions over nuisance design matrix Z and assume that the sparse maximal eigenvalues are bounded with high probability, i.e. $P(\phi_{\max}(sl_n) \le C) \ge 1 - \delta_n$, then Assumption A3 can be relaxed to

$$s \log(d \vee n) = o(\sqrt{n}). \tag{3.8}$$

The bounded maximum eigenvalue condition is commonly used in the existing variable selection literature (Van de Geer, 2008; Bondell and Reich, 2012; Belloni et al., 2013). The sparsity assumption in (3.8) is commonly assumed for the methods in the high-dimensional literature that can achieve $\sqrt{n}$-consistency and valid inferential results (Zhang and Zhang, 2014; Van de Geer et al., 2014; Belloni et al., 2013; Wu et al., 2023).

3.3 Main Results

When the oracle values of the nuisance parameters are known, the conditional posterior of θ is given by:

$$f(\theta \mid Y, X, Z, \gamma_0, \phi_0) \propto \pi(\theta) \prod_{i=1}^{n} \frac{\exp\{(X_i - Z_i^T\gamma_0)\theta + Z_i^T\phi_0\}^{Y_i}}{1 + \exp\{(X_i - Z_i^T\gamma_0)\theta + Z_i^T\phi_0\}}. \tag{3.9}$$

This posterior is for one-dimensional parameter θ and there is no variable selection involved. This corresponds to a logistic model with $(X_i - Z_i^T\gamma_0)$ as the input variable and $Z_i^T\phi_0$ as the random effect. Moreover, the log-likelihood in this case is given by:

$$L_n(\theta \mid \gamma_0, \phi_0) = \sum_{i=1}^{n} Y_i\{(X_i - Z_i^T\gamma_0)\theta + Z_i^T\phi_0\} - \log\{1 + \exp((X_i - Z_i^T\gamma_0)\theta + Z_i^T\phi_0)\}. \tag{3.10}$$

The corresponding score has an expectation zero at the oracle value θ_0. Therefore, the posterior concentration theorem (Walker, 1969; Schervish, 2012) applies here, and the posterior concentrates around θ_0 at the desired $n^{-1/2}$ rate. We denote the Maximum Likelihood Estimator for the likelihood in (3.10) by $\hat{\theta}_0$.

$$\hat{\theta}_0 = \arg\max_{\theta \in \Theta} \frac{1}{n} L_n(\theta \mid \gamma_0, \phi_0). \tag{3.11}$$

This is the Maximum Likelihood Estimate of a one-dimensional parameter in a logistic model which does not have a closed-form expression. However, based on the theory of the Maximum Likelihood Estimates (Fahrmeir and Kaufmann, 1985), we know that it concentrates around θ_0 at the desired $n^{-1/2}$ rate. The standard deviation of $\hat{\theta}_0$ will be denoted by σ_n and is defined as:

$$\sigma_n = \sqrt{-\frac{\partial^2}{\partial\theta^2} L_n(\theta \mid \gamma_0, \phi_0)\Big|_{\theta=\hat{\theta}_0}}. \tag{3.12}$$

We now provide the main theoretical results. The conditional posterior for θ as defined in (θ**-CP**) is:

$$f(\theta \mid Y, X, Z, \gamma, \phi) = \frac{\pi(\theta)\exp\{L_n(\theta \mid \gamma, \phi)\}}{\int_\Theta \pi(\theta)\exp(L_n(\theta \mid \gamma, \phi))d\theta}, \tag{3.13}$$

where L_n is the log likelihood function based on the (θ**-CP**) as defined in (3.1). Let $f(\theta \mid Y, X, Z)$ be the marginal posterior for θ obtained after integrating out γ and ϕ from the joint posterior $f(\theta, \gamma, \phi \mid Y, X, Z)$ given by (2.6). Our theorem below will guarantee the concentration of this posterior for θ at the optimal rate.

Theorem 1 (Posterior Concentration)**.** *Suppose Assumptions A1–A6 presented above hold. Consider $\hat{\theta}_0$ and σ_n as defined in (3.11) and (3.12) respectively. If a and b are constants, where $a < b$, then the marginal posterior probability $P(\hat{\theta}_0 + a\sigma_n < \theta < \hat{\theta}_0 + b\sigma_n \mid Y, X, Z)$ given by*

$$\int_{\hat{\theta}_0 + a\sigma_n}^{\hat{\theta}_0 + b\sigma_n} f(\theta \mid Y, X, Z)d\theta$$

converges in probability to

$$\frac{1}{\sqrt{2\pi}}\int_a^b \exp\left(-\frac{1}{2}z^2\right)dz,$$

as $n \to \infty$.

Therefore, the theorem states that the posterior concentrates around $\hat{\theta}_0$ which is the oracle estimator (3.11) at $n^{-1/2}$ rate because $\sigma_n = O(\sqrt{n})$. Concentration of $\hat{\theta}_0$ around θ_0 implies that the conditional posterior concentrates around the true parameter θ_0 at $n^{-1/2}$ rate. The conditional posterior attains asymptotic normality and hence the interval estimates would be valid as confirmed by the following corollary:

Corollary 1.1. *The $(1-\alpha)$-credible interval has the correct coverage in the frequentist sense. Let $\hat{q}^{CB}_{\alpha/2}$ and $\hat{q}^{CB}_{1-\alpha/2}$ be the $\alpha/2^{th}$ and $(1-\alpha/2)^{th}$ percentiles of the marginal posterior distribution $f(\theta \mid Y, X, Z)$ respectively. Then, we have*

$$\left| P\Big[\theta_0 \in \big(\hat{q}^{CB}_{\alpha/2}, \hat{q}^{CB}_{1-\alpha/2}\big) \mid Y, X, Z\Big] - (1-\alpha)\right| \xrightarrow{\text{P}} 0.$$

The proofs of the theorem and the corollary have been presented in Ojha and Narisetty (2023).

4 Simulation Studies

4.1 Simulation Setup

We investigate the performance of our proposed method under different simulation settings. Through simulation results, we see that the proposed Bayesian method achieves valid coverage in the frequentist sense. Even with the default hyper-parameter values without tuning, we observe that the coverage, bias, and interval lengths are more robust to the simulation settings than the existing frequentist methods. Our simulations are based on the following models:

$$P[Y_i = 1 \mid X_i, \mathbf{Z}_i] = \text{logit}(X_i\theta_0 + \mathbf{Z}_i^T\beta), \quad X_i = \mathbf{Z}_i^T\gamma + \epsilon_i,$$

where $\epsilon_i \overset{i.i.d.}{\sim} N(0,1)$. $\mathbf{Z_i}$ is the high-dimensional nuisance variable with intercept as the first component and the last $d-1$ components are generated according to a multivariate normal, $N(0, H_{d-1\times d-1})$ with $H_{ij} = 0.5^{|i-j|}$. The nuisance parameters β and γ are chosen to be sparse vectors according to

$$\beta_{d\times 1} = (1, 1/2, 1/3, 1/4, 1/5, 0, 0, 0, 0, 0, 1, 1/2, 1/3, 1/4, 1/5, 0, 0, \dots, 0)^T,$$
$$\gamma_{d\times 1} = (1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9, 1/10, 0, 0, \dots, 0)^T.$$

The data-generating process for the target covariate X_i does not involve weights which were part of the assumption (3.6) for both our Bayesian approach and the frequentist approach of Belloni et al. (2013).

We now provide our prior specifications which we use for our empirical investigations.

P1 **Prior for** β: Let I_{1j} denote the binary latent variables that decide the active and inactive state of the component β_j. The priors are:

$$\beta_j \mid I_{1j} = 0 \sim N(0, \tau_{0n}^2), \beta_j \mid I_{1j} = 1 \sim N(0, \tau_{1n}^2),$$
$$P(I_{1j} = 1) = 1 - P(I_{1j} = 0) = q_n,$$

for $j = 1, \dots, d$. The parameters τ_{0n}, τ_{1n}, and q_n are chosen exactly the same way as in Narisetty et al. (2019). To be specific, we have chosen $\tau_{0n}^2 = 1/n$, $n\tau_{1n}^2 = \max\{n, 0.01d^{2.1}\}$, and q_n such that $P[\sum_{j=1}^d I_{1j} > K] = 0.1$ for $K = \max\{10, \log n\}$.

P2 **Prior for** γ**:** Let I_{2j} denote the binary latent variable that decides the active and inactive state of the component γ_j. The priors are:

$$\gamma_j \mid I_{2j} = 0 \sim N(0, \tau_{0n}^2), \gamma_j \mid I_{2j} = 1 \sim N(0, \tau_{1n}^2),$$
$$P(I_{2j} = 1) = 1 - P(I_{2j} = 0) = q_n,$$

for $j = 1, \dots, d$. The choice of the prior hyper-parameters is the same as that discussed in the prior for β.

P3 **Prior for σ^2:** This is the variance parameter in defining the working model between X and Z (**γ-WM**). We take the inverse-Gamma prior for this variance parameter. $\sigma^2 \sim \mathrm{IG}(a, b)$. We have chosen $a = b = 1$ in our simulations.

P4 **Prior for θ:** This is a finite-dimensional parameter that does not require any variable selection step. We assume a Gaussian prior for this parameter with a large variance $\lambda > 0$: $\theta \sim N(0, \lambda)$. We have chosen $\lambda = 10$ in our simulations.

Remark 2 (On the choice of priors and high dimensional posterior samplers)**.** *In this paper, we use continuous spike-and-slab priors with the Skinny Gibbs algorithm (Narisetty et al., 2019) for posterior sampling. However, this is just an example choice but any other method that facilitates computationally feasible sampling from the conditional posteriors of the high-dimensional nuisance parameters can be used alternatively. For the validity of posterior based inference, we require that the samples of the high-dimensional nuisance parameters satisfy the concentration rates as presented in Assumption 4. The theoretical results of Narisetty et al. (2019) guarantee concentration of the stationary distribution corresponding to the algorithm although this stationary distribution is not the same as the exact posterior distribution. For continuous spike and slab priors, Biswas et al. (2022) introduced an exact and efficient sampling method for high dimensional posteriors that can also be used.*

We consider two pairs of sample sizes and number of covariates such that $(n, d) \in \{(200, 300), (300, 400)\}$ and select the true signal θ_0 from a set of varying signal strengths, $\theta_0 \in \{0, 0.1, 0.2, \ldots, 1\}$. We calculate the empirical 95% credible (confidence) intervals for different values of θ_0 based on 1000 Monte Carlo simulations. We report the frequentist coverage, interval length, and bias for each method under consideration. The methods that we compare in this study are:

- **DS**: *Double Selection* – To implement the Double Selection algorithm proposed by Belloni et al. (2013), we used the pseudo-code given in Table 2 of their paper using the exact values of the penalty parameters they described.

- **WLP**: *Global and Simultaneous Hypothesis Testing for High-Dimensional Logistic Regression Models* (Ma et al., 2021) – This method uses a generalized low-dimensional projection technique for the bias correction of the parameter estimates which are obtained after employing logistic LASSO. They construct global test statistics and construct confidence intervals for all the parameters using the debiased estimates. We use the intervals and estimates for the parameter of interest θ in our results. This method has been implemented based on the code provided by the authors.

- **LSW**: *Statistical Inference for High-Dimensional Generalized Linear Models With Binary Outcomes* (Cai et al., 2021) – This is a two-step bias correction method that employs a novel weighting strategy for the bias correction and consequently construct confidence intervals. This method has been implemented based on the code provided by the authors.

- **NAIVE**: *Post LASSO Logistic Regression* – In the first step of this method, we perform LASSO using the `glmnet` package to select the significant nuisance covariates. In the second step, we use the covariate of interest X along with the selected nuisance covariates from the first step to perform a low-dimensional logistic regression and use the estimated parameter and its variance for inference using Wald intervals.

- **BMA**: *Bayesian Model Averaging* (Tüchler, 2008) – This method utilizes spike and slab priors over the high-dimensional nuisance parameter for model selection and subsequently averages over selected models to obtain estimates and standard deviation for the parameter of interest. The point estimator and the estimated standard deviation are used to obtain intervals. We implement this method using the package `BoomSpikeSlab`.

- **BLASSO**: *Bayesian LASSO* (Park and Casella, 2008) – This is the Bayesian logistic regression method with Laplace priors on the model parameters. We use the `bayesreg` package to implement this method.

- **CB-SG**: This is our proposed Bayesian approach based on conditional posteriors using the *Skinny Gibbs sampler*.

- **ORACLE** – We regress the output Y on the target covariate X and the nuisance covariates Z which truly affect Y. The selection of nuisance covariates is based on the indices where the components of β are non-zero. As this corresponds to a low-dimensional logistic regression scenario, we employ the `glm` package without any form of regularization. We include this approach as a benchmark but it cannot be implemented in practice as it uses the knowledge of the unknown sparsity structure of the true data generating model.

4.2 Results and Discussion

In this section, we compare the results for all the methods considered. Table 1 and Table 2 contain the coverage and the interval lengths of methods under comparison for the $(n = 200, d = 300)$ and $(n = 300, d = 400)$ cases, respectively. Table 3 contains the biases of the estimates produced by the methods under comparison based on 1000 Monte Carlo simulations.

Based on the results in Tables 1–3, we find that the coverage of the proposed method (CB-SG) is close to the desired coverage in most settings while keeping the length of the interval and the bias similar to the optimal ones obtained by ORACLE. While the coverage of BLASSO is at least as much as the nominal coverage, it has very large interval lengths especially for large signal values. On the other hand, DS, WLP, NAIVE, and BMA methods suffer from severe under-coverage and large bias for most signal values. Among the frequentist methods, LSW has the least bias and better coverage for most of the signal values. When compared to our CB-SG method, LSW has lower coverage even though the interval lengths are similar. In summary, the proposed method demonstrates

		Frequentist				Bayesian			
Quantities	θ_0	DS	WLP	LSW	NAIVE	BMA	BLASSO	CB-SG	ORACLE
Coverage	0.0	0.127	0.007	0.882	0.906	0.495	1.000	0.931	0.940
	0.1	0.206	0.020	0.906	0.892	0.482	0.998	0.922	0.942
	0.2	0.297	0.028	0.924	0.877	0.431	0.981	0.920	0.954
	0.3	0.370	0.045	0.951	0.874	0.416	0.947	0.932	0.955
	0.4	0.425	0.113	0.953	0.846	0.403	0.956	0.939	0.942
	0.5	0.510	0.240	0.967	0.833	0.379	0.937	0.962	0.959
	0.6	0.549	0.443	0.963	0.806	0.396	0.939	0.977	0.957
	0.7	0.667	0.706	0.962	0.824	0.394	0.955	0.978	0.937
	0.8	0.713	0.889	0.955	0.797	0.371	0.966	0.982	0.944
	0.9	0.767	0.968	0.950	0.820	0.377	0.953	0.985	0.941
	1.0	0.828	0.971	0.930	0.830	0.432	0.960	0.976	0.938
Interval Length	0.0	0.440	0.403	0.749	1.119	0.795	0.600	0.730	0.738
	0.1	0.460	1.380	0.760	1.188	0.823	0.758	0.752	0.760
	0.2	0.476	0.600	0.796	1.209	0.856	0.922	0.771	0.781
	0.3	0.496	0.695	0.820	1.309	0.882	1.069	0.798	0.801
	0.4	0.521	0.638	0.860	1.333	0.939	1.262	0.823	0.823
	0.5	0.538	0.773	0.885	1.560	0.981	1.469	0.861	0.867
	0.6	0.564	0.669	0.915	1.645	1.051	1.763	0.913	0.908
	0.7	0.596	1.642	0.967	1.736	1.119	1.994	0.966	0.956
	0.8	0.622	1.693	1.041	1.971	1.186	2.235	1.023	1.001
	0.9	0.652	1.250	1.143	1.962	1.265	2.405	1.080	1.054
	1.0	0.680	2.369	1.154	2.091	1.345	2.658	1.174	1.113

Table 1: Coverage and interval length for each method considered for $n = 200, d = 300$ under signal strengths $\theta_0 \in \{0, 0.1, 0.2, \dots, 1\}$.

a competitive overall performance in terms of achieving the desired coverage, smaller interval length and bias.

Recent studies have highlighted that double selection methods often run into problems related to variance inflation (Antonelli et al., 2022; Torrens-i Dinarès et al., 2021). In Ojha and Narisetty (2023), we present an additional simulation study where the confounding between the covariate of interest X and the nuisance covariates Z is larger. We choose larger non-zero coefficients for the nuisance parameter γ and study its impact on coverage, interval length, and bias. We observe that our proposed method has relatively less impact due to confounding compared to other methods.

Furthermore, in Ojha and Narisetty (2023), we provide simulation results where d is much larger than the sample size.

5 Analysis of a Single-Cell RNA-seq Data

We revisit the single-cell RNA-seq data (Shalek et al., 2014), which includes expression estimates in transcripts per million for mouse genes that are annotated and calculated by Li and Dewey (2011). The data consist of 1861 cells, specifically dendritic cells derived from mouse bone marrow. The main goal of the study is to understand how gene

		Frequentist				Bayesian			
Quantities	θ_0	DS	WLP	LSW	NAIVE	BMA	BLASSO	CB-SG	ORACLE
Coverage	0.0	0.177	0.000	0.843	0.907	0.454	0.999	0.946	0.936
	0.1	0.266	0.004	0.883	0.870	0.442	1.000	0.929	0.948
	0.2	0.368	0.006	0.904	0.846	0.390	0.990	0.935	0.941
	0.3	0.381	0.007	0.928	0.796	0.375	0.976	0.951	0.948
	0.4	0.419	0.026	0.950	0.775	0.378	0.969	0.960	0.949
	0.5	0.386	0.082	0.934	0.733	0.341	0.974	0.962	0.950
	0.6	0.420	0.205	0.957	0.719	0.348	0.973	0.983	0.946
	0.7	0.464	0.425	0.953	0.682	0.342	0.966	0.976	0.945
	0.8	0.567	0.741	0.935	0.685	0.337	0.968	0.966	0.947
	0.9	0.608	0.908	0.926	0.668	0.320	0.967	0.952	0.939
	1.0	0.652	0.976	0.924	0.665	0.337	0.964	0.955	0.952
Interval Length	0.0	0.365	0.333	0.568	0.824	0.627	0.483	0.569	0.585
	0.1	0.379	0.363	0.589	0.876	0.650	0.595	0.582	0.600
	0.2	0.394	0.400	0.605	0.999	0.665	0.732	0.590	0.612
	0.3	0.409	0.440	0.631	1.010	0.686	0.898	0.607	0.632
	0.4	0.431	0.501	0.651	1.038	0.713	1.082	0.628	0.656
	0.5	0.455	0.592	0.689	1.203	0.738	1.276	0.651	0.683
	0.6	0.474	0.631	0.707	1.357	0.775	1.445	0.684	0.710
	0.7	0.501	0.746	0.771	1.413	0.812	1.774	0.733	0.749
	0.8	0.523	0.736	0.778	1.468	0.857	1.850	0.779	0.775
	0.9	0.552	1.018	0.814	1.634	0.900	2.143	0.833	0.823
	1.0	0.575	1.004	0.867	1.820	0.952	2.363	0.897	0.863

Table 2: Coverage and interval length for each method considered for $n = 300, d = 400$ under signal strengths $\theta_0 \in \{0, 0.1, 0.2, \dots, 1\}$.

expressions vary under specific stimuli. We focus on Particle In Cell (PIC) stimulations, which are viral-like double-stranded RNAs. Following the analysis of Cai et al. (2021) who previously analyzed this data, we examine the expression profiles of cells after 6 hours of stimulation, including 96 control cells and 96 cells stimulated with PIC. To ensure reliable analysis, we preprocess the data by removing genes that are not expressed in at least 80% of the cells and keep genes that have variance within the top 10 percentile for our analysis. This results in a total of 697 genes. We apply log transformation and normalize the data set to have mean zero and unit variance in each cell.

5.1 Goal of the Study

Based on previous experimental findings (Jang et al., 2018), which suggest that RSAD2 plays an important in the immune response against viruses mediated by mature dendritic cells, we focus on RSAD2 as our target covariate. Our aim is to quantify the association between RSAD2 expression and PIC stimulations while controlling for the influence of all other genes in the analysis. Formally, we assume a logistic model for the association between gene expressions and the stimulations. For $i = 1, 2, \dots, 192,$:

$$P(\text{stimulation}_i = 1 \mid RSAD2, \text{Gene}_1, \dots, \text{Gene}_{696}) =$$

Bias Table		Frequentist				Bayesian			
(n,d)	θ_0	DS	WLP	LSW	NAIVE	BMA	BLASSO	CB-SG	ORACLE
n=300 d=400	0.0	0.363	0.483	0.181	0.009	0.365	0.113	0.073	0.000
	0.1	0.340	0.473	0.166	0.196	0.409	0.079	0.091	0.022
	0.2	0.306	0.445	0.146	0.391	0.451	0.045	0.080	0.023
	0.3	0.290	0.421	0.122	0.496	0.486	0.028	0.083	0.025
	0.4	0.281	0.387	0.091	0.595	0.521	0.021	0.092	0.020
	0.5	0.254	0.367	0.076	0.811	0.539	0.050	0.106	0.049
	0.6	0.250	0.333	0.058	1.081	0.586	0.123	0.103	0.076
	0.7	0.216	0.310	0.042	1.090	0.630	0.180	0.095	0.092
	0.8	0.205	0.285	0.016	1.181	0.684	0.227	0.081	0.093
	0.9	0.183	0.239	0.002	1.117	0.718	0.281	0.056	0.114
	1.0	0.136	0.215	0.011	1.098	0.749	0.372	0.038	0.135
n=400 d=500	0.0	0.308	0.495	0.154	0.001	0.327	0.106	0.052	0.008
	0.1	0.284	0.470	0.138	0.163	0.340	0.062	0.051	0.004
	0.2	0.248	0.448	0.119	0.272	0.376	0.041	0.037	0.017
	0.3	0.248	0.426	0.105	0.388	0.402	0.043	0.048	0.019
	0.4	0.244	0.403	0.079	0.424	0.417	0.074	0.036	0.019
	0.5	0.253	0.384	0.076	0.614	0.446	0.132	0.028	0.041
	0.6	0.248	0.357	0.054	0.681	0.467	0.186	0.013	0.043
	0.7	0.243	0.347	0.062	0.766	0.488	0.336	0.012	0.069
	0.8	0.223	0.292	0.025	0.793	0.526	0.335	0.038	0.053
	0.9	0.216	0.284	0.002	0.894	0.549	0.475	0.073	0.075
	1.0	0.207	0.243	0.015	1.058	0.573	0.569	0.094	0.088

Table 3: Bias for different methods under signal strengths $\theta_0 \in \{0, 0.1, 0.2, \ldots, 1\}$.

$$\left(1 + \exp -\left(\alpha_0 + \theta * RSAD2 + \sum_{j=1}^{696} \alpha_j * \text{Gene}_j\right)\right)^{-1}, \tag{5.1}$$

where stimulation_i is marked as "1" if that cell is stimulated by PIC and "0" otherwise. Moreover, $RSAD2, \text{Gene}_1, \ldots,$ and Gene_{696} are the normalized gene expressions. Our aim is to estimate θ in (5.1) and provide a valid interval estimator for θ.

5.2 Results and Discussion

Table 4 presents the results of our analysis on the association between RSAD2 and PIC stimulations. An additional row labeled as "MARGINAL on RSAD2" is included, which reports the results obtained when fitting a binary output (PIC stimulations $\in \{0, 1\}$) against RSAD2 and an intercept term.

We observe that the inferences obtained by CB-SG and DS show a strong positive association between RSAD2 expression and the PIC stimulations, similar to the results of the "MARGINAL on RSAD2" approach. However, BLASSO and BMA fail to detect this association. On the other hand, WLP and LSW methods also indicate a positive effect, but the effect estimated by these methods is not as strong as CB-SG and DS.

METHODS	Estimate for θ	SE of the Estimate	LOWER	UPPER
DS	6.44	1.27	3.94	8.93
WLP	2.81	0.34	2.13	3.49
LSW	2.12	0.43	1.27	2.97
BMA	0.00	0.01	−0.02	0.02
BLASSO	0.33	0.55	−0.74	1.41
CB-SG	5.72	1.41	2.96	8.49
MARGINAL on RSAD2	4.75	0.77	3.25	6.26

Table 4: Estimates, standard errors (SE), lower and upper ends of interval estimators for PIC stimulation data.

Furthermore, the variance estimates of WLP and LSE are smaller than the variance obtained from "MARGINAL on RSAD2". This is surprising as we would expect the variance to inflate in comparison with the marginal model as these methods consider a potentially larger model with high dimensional nuisance covariates.

6 Conclusion

There are several common applications in various fields including economics and biostatistics where it is important to understand the impact of a specific covariate of interest on a binary output. We propose a new conditional Bayesian approach for constructing Bayesian posterior for a finite-dimensional parameter of interest in the high dimensional logistic regression model. Our approach leads to credible intervals with $O(n^{-1/2})$ length and have valid frequentist coverage properties. The proposed posterior is based on the idea of Neyman orthogonality to construct a conditional Bayesian model. Moreover, the proposed Bayesian approach has multiple conceptual and practical advantages compared to the frequentist approach based on double selection. The proposed Bayesian framework can be used to incorporate prior information on the parameters of interest. Moreover, the proposed Bayesian approach provides a more extensive exploration of the orthogonalization as opposed to a fixed orthogonalization used by the frequentist approach. Similarly, the underlying weights required for the orthogonalization are updated adaptively in our Bayesian approach leading to better empirical performance.

Supplementary Material

A Conditional Bayesian Approach with Valid Inference for High Dimensional Logistic Regression (DOI: 10.1214/23-BA1408SUPP; .pdf). In the supplementary file, we discuss the computational aspects of the proposed method, provide all the proofs related to the theoretical results, and present further empirical studies.

References

Antonelli, J., Papadogeorgou, G., and Dominici, F. (2022). "Causal inference in high dimensions: a marriage between Bayesian modeling and good frequentist properties." *Biometrics*, 78(1): 100–114. MR4408573. doi: https://doi.org/10.1111/biom.13417. 4, 18

Armagan, A., Dunson, D. B., and Lee, J. (2013). "Generalized double Pareto shrinkage." *Statistica Sinica*, 23(1): 119. MR3076161. 2

Bach, F. (2010). "Self-concordant analysis for logistic regression." *Electronic Journal of Statistics*, 4: 384–414. MR2645490. doi: https://doi.org/10.1214/09-EJS521. 2

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). "Sparse models and methods for optimal instruments with an application to eminent domain." *Econometrica*, 80(6): 2369–2429. MR3001131. doi: https://doi.org/10.3982/ECTA9626. 3, 4

Belloni, A., Chernozhukov, V., and Hansen, C. (2010). "Lasso methods for Gaussian instrumental variables models." *arXiv preprint arXiv:1012.1297*. 3, 4

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). "Inference on treatment effects after selection among high-dimensional controls." *The Review of Economic Studies*, 81(2): 608–650. MR3207983. doi: https://doi.org/10.1093/restud/rdt044. 3, 4

Belloni, A., Chernozhukov, V., and Wei, Y. (2013). "Honest confidence regions for a regression parameter in logistic regression with a large number of controls." CeMMAP working papers CWP67/13, Centre for Microdata Methods and Practice, Institute for Fiscal Studies. URL https://ideas.repec.org/p/ifs/cemmap/67-13.html MR1015135. doi: https://doi.org/10.1214/aos/1176347253. 3, 4, 6, 7, 10, 12, 13, 15, 16

Berger, J. O., Liseo, B., and Wolpert, R. L. (1999). "Integrated likelihood methods for eliminating nuisance parameters." *Statistical Science*, 1–22. MR1702200. doi: https://doi.org/10.1214/ss/1009211803. 10

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). "Valid post-selection inference." *The Annals of Statistics*, 802–837. MR3099122. doi: https://doi.org/10.1214/12-AOS1077. 3

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). "Dirichlet–Laplace priors for optimal shrinkage." *Journal of the American Statistical Association*, 110(512): 1479–1490. MR3449048. doi: https://doi.org/10.1080/01621459.2014.960967. 2

Biswas, N., Mackey, L., and Meng, X.-L. (2022). "Scalable spike-and-slab." In *International Conference on Machine Learning*, 2021–2040. PMLR. 16

Bondell, H. D. and Reich, B. J. (2012). "Consistent high-dimensional Bayesian variable selection via penalized credible regions." *Journal of the American Statistical Association*, 107(500): 1610–1624. MR3036420. doi: https://doi.org/10.1080/01621459.2012.716344. 2, 13

Breheny, P. and Huang, J. (2011). "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection." *The Annals of Applied Statistics*, 5(1): 232. MR2810396. doi: https://doi.org/10.1214/10-AOAS388. 2

Bunea, F. (2008). "Honest variable selection in linear and logistic regression models via l_1 and l_1+l_2 penalization." *Electronic Journal of Statistics*, 2: 1153–1194. MR2461898. doi: https://doi.org/10.1214/08-EJS287. 2

Cai, T. T., Guo, Z., and Ma, R. (2021). "Statistical inference for high-dimensional generalized linear models with binary outcomes." *Journal of the American Statistical Association*, 1–14. MR4595497. doi: https://doi.org/10.1080/01621459.2021.1990769. 4, 6, 10, 16, 19

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). "Handling sparsity via the horseshoe." In *Artificial Intelligence and Statistics*, 73–80. PMLR. 2

Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. (2015). "Bayesian linear regression with sparse priors." *The Annals of Statistics*, 43(5): 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 12

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal*, 21(1): C1–C68. MR3769544. doi: https://doi.org/10.1111/ectj.12097. 3, 4

Dasgupta, S., Khare, K., and Ghosh, M. (2014). "Asymptotic expansion of the posterior density in high dimensional generalized linear models." *Journal of Multivariate Analysis*, 131: 126–148. MR3252640. doi: https://doi.org/10.1016/j.jmva.2014.06.013. 3

Fahrmeir, L. and Kaufmann, H. (1985). "Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models." *The Annals of Statistics*, 13(1): 342–368. MR0773172. doi: https://doi.org/10.1214/aos/1176346597. 14

Fan, J. and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American statistical Association*, 96(456): 1348–1360. MR1946581. doi: https://doi.org/10.1198/016214501753382273. 2, 12

Friedman, J., Hastie, T., and Tibshirani, R. (2010). "Regularization paths for generalized linear models via coordinate descent." *Journal of Statistical Software*, 33(1): 1. 2

Genkin, A., Lewis, D. D., and Madigan, D. (2007). "Large-scale Bayesian logistic regression for text categorization." *Technometrics*, 49(3): 291–304. MR2408634. doi: https://doi.org/10.1198/004017007000000245. 2

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 2

Ghosal, S. (1997). "Normal approximation to the posterior distribution for generalized linear models with many covariates." *Mathematical Methods of Statistics*, 6(3): 332–348. MR1475901. 3

Ghosal, S. (2000). "Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity." *Journal of Multivariate Analysis*, 74(1): 49–68. MR1790613. doi: https://doi.org/10.1006/jmva.1999.1874. 3

Ghosh, J., Li, Y., and Mitra, R. (2018). "On the use of Cauchy prior distributions for Bayesian logistic regression." *Bayesian Analysis*, 13(2): 359–383. MR3780427. doi: https://doi.org/10.1214/17-BA1051. 2

Huang, J. and Zhang, C.-H. (2012). "Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications." *The Journal of Machine Learning Research*, 13(1): 1839–1864. MR2956344. 2

Ishwaran, H. and Rao, J. S. (2005). "Spike and slab variable selection: Frequentist and Bayesian strategies." *The Annals of Statistics*, 33(2): 730–773. MR2163158. doi: https://doi.org/10.1214/009053604000001147. 2, 12

Jang, J.-S., Lee, J.-H., Jung, N.-C., Choi, S.-Y., Park, S.-Y., Yoo, J.-Y., Song, J.-Y., Seo, H. G., Lee, H. S., and Lim, D.-S. (2018). "Rsad2 is necessary for mouse dendritic cell maturation via the IRF7-mediated signaling pathway." *Cell Death & Disease*, 9(8): 823. 19

Javanmard, A. and Montanari, A. (2014). "Confidence intervals and hypothesis testing for high-dimensional regression." *The Journal of Machine Learning Research*, 15(1): 2869–2909. MR3277152. 3

Johnson, V. E. and Rossell, D. (2012). "Bayesian model selection in high-dimensional settings." *Journal of the American Statistical Association*, 107(498): 649–660. MR2980074. doi: https://doi.org/10.1080/01621459.2012.682536. 2, 3

Kabaila, P. (1995). "The effect of model selection on confidence regions and prediction regions." *Econometric Theory*, 11(3): 537–549. MR1349934. doi: https://doi.org/10.1017/S0266466600009403. 3

Kuchibhotla, A. K., Brown, L. D., Buja, A., Cai, J., George, E. I., and Zhao, L. H. (2020). "Valid post-selection inference in model-free linear regression." *The Annals of Statistics*, 48(5): 2953–2981. MR4152630. doi: https://doi.org/10.1214/19-AOS1917. 3

Kwemou, M. (2016). "Non-asymptotic oracle inequalities for the Lasso and group Lasso in high dimensional logistic model." *ESAIM: Probability and Statistics*, 20: 309–331. MR3533711. doi: https://doi.org/10.1051/ps/2015020. 2

Leeb, H. and Pötscher, B. M. (2005). "Model selection and inference: Facts and fiction." *Econometric Theory*, 21(1): 21–59. MR2153856. doi: https://doi.org/10.1017/S0266466605050036. 3

Leeb, H. and Pötscher, B. M. (2008). "Sparse estimators and the oracle property, or the return of Hodges' estimator." *Journal of Econometrics*, 142(1): 201–211. MR2394290. doi: https://doi.org/10.1016/j.jeconom.2007.05.017. 3

Li, B. and Dewey, C. N. (2011). "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." *BMC Bioinformatics*, 12: 1–16. 18

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g-priors for Bayesian variable selection." *Journal of the American Statistical Association*, 103(481): 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 2, 12

Ma, R., Tony Cai, T., and Li, H. (2021). "Global and simultaneous hypothesis testing for high-dimensional logistic regression models." *Journal of the American Statistical Association*, 116(534): 984–998. MR4270038. doi: https://doi.org/10.1080/01621459.2019.1699421. 4, 6, 10, 16

Narisetty, N. N. and He, X. (2014). "Bayesian variable selection with shrinking and diffusing priors." *The Annals of Statistics*, 42(2): 789–817. MR3210987. doi: https://doi.org/10.1214/14-AOS1207. 2, 3

Narisetty, N. N., Shen, J., and He, X. (2019). "Skinny Gibbs: A consistent and scalable Gibbs sampler for model selection." *Journal of the American Statistical Association*, 114(527): 1205–1217. MR4011773. doi: https://doi.org/10.1080/01621459.2018.1482754. 2, 12, 15, 16

Neyman, J. (1959). "Optimal asymptotic tests of composite hypotheses." *Probability and Statistics*, 213–234. MR0112201. 4

Neyman, J. (1979). "C (α) tests and their use." *Sankhyā: The Indian Journal of Statistics, Series A*, 1–21. 4

O'brien, S. M. and Dunson, D. B. (2004). "Bayesian multivariate logistic regression." *Biometrics*, 60(3): 739–746. MR2089450. doi: https://doi.org/10.1111/j.0006-341X.2004.00224.x. 2

Ojha, A. and Narisetty, N. N. (2023). "Supplementary Material for "A conditional Bayesian approach with valid inference for high dimensional logistic regression"." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1408SUPP. 14, 18

Panigrahi, S. and Taylor, J. (2018). "Scalable methods for Bayesian selective inference." *Electronic Journal of Statistics*, 12(2): 2355–2400. MR3832095. doi: https://doi.org/10.1214/18-EJS1452. 3

Panigrahi, S., Taylor, J., and Weinstein, A. (2021). "Integrative methods for post-selection inference under convex constraints." *The Annals of Statistics*, 49(5): 2803–2824. MR4338384. doi: https://doi.org/10.1214/21-aos2057. 3

Park, T. and Casella, G. (2008). "The Bayesian lasso." *Journal of the American Statistical Association*, 103(482): 681–686. MR2524001. doi: https://doi.org/10.1198/016214508000000337. 2, 17

Pötscher, B. M. (1991). "Effects of model selection on inference." *Econometric Theory*, 7(2): 163–185. MR1128410. doi: `https://doi.org/10.1017/S0266466600004382`. 3

Pötscher, B. M. (2009). "Confidence sets based on sparse estimators are necessarily large." *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, 1–18. 3

Pötscher, B. M. and Leeb, H. (2009). "On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding." *Journal of Multivariate Analysis*, 100(9): 2065–2082. MR2543087. doi: `https://doi.org/10.1016/j.jmva.2009.06.010`. 3

Ročková, V. and George, E. I. (2014). "EMVS: The EM approach to Bayesian variable selection." *Journal of the American Statistical Association*, 109(506): 828–846. MR3223753. doi: `https://doi.org/10.1080/01621459.2013.869223`. 2, 3

Schervish, M. J. (2012). *Theory of Statistics*. Springer Science & Business Media. MR1354146. doi: `https://doi.org/10.1007/978-1-4612-4250-5`. 3, 13

Severini, T. A. (2007). "Integrated likelihood functions for non-Bayesian inference." *Biometrika*, 94(3): 529–542. MR2410006. doi: `https://doi.org/10.1093/biomet/asm040`. 10

Severini, T. A. (2011). "Frequency properties of inferences based on an integrated likelihood function." *Statistica Sinica*, 433–447. 10

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N., et al. (2014). "Single-cell RNA-seq reveals dynamic paracrine control of cellular variation." *Nature*, 510(7505): 363–369. 18

Shi, C., Song, R., Lu, W., and Li, R. (2021). "Statistical inference for high-dimensional models via recursive online-score estimation." *Journal of the American Statistical Association*, 116(535): 1307–1318. MR4309274. doi: `https://doi.org/10.1080/01621459.2019.1710154`. 4

Song, Q. and Liang, F. (2022). "Nearly optimal Bayesian shrinkage for high-dimensional regression." *Science China Mathematics*, 1–34. MR4535982. doi: `https://doi.org/10.1007/s11425-020-1912-6`. 12

Taylor, J. and Tibshirani, R. J. (2015). "Statistical learning and selective inference." *Proceedings of the National Academy of Sciences*, 112(25): 7629–7634. MR3371123. doi: `https://doi.org/10.1073/pnas.1507583112`. 3

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288. MR1379242. 2

Torrens-i Dinarès, M., Papaspiliopoulos, O., and Rossell, D. (2021). "Confounder importance learning for treatment effect inference." *arXiv preprint arXiv:2110.00314*. 4, 18

Tüchler, R. (2008). "Bayesian variable selection for logistic models using auxiliary mixture sampling." *Journal of Computational and Graphical Statistics*, 17(1): 76–94. MR2424796. doi: `https://doi.org/10.1198/106186008X289849`. 17

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). “On asymptotically optimal confidence regions and tests for high-dimensional models.” *The Annals of Statistics*, 42(3): 1166–1202. MR3224285. doi: https://doi.org/10.1214/14-AOS1221. 3, 13

Van de Geer, S. A. (2008). “High-dimensional generalized linear models and the lasso.” *The Annals of Statistics*, 36(2): 614–645. MR2396809. doi: https://doi.org/10.1214/009053607000000929. 2, 12, 13

Walker, A. M. (1969). “On the asymptotic behaviour of posterior distributions.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 31(1): 80–88. URL http://www.jstor.org/stable/2984328. MR0269000. 3, 13

Wang, J., He, X., and Xu, G. (2020). “Debiased inference on treatment effect in a high-dimensional model.” *Journal of the American Statistical Association*, 115(529): 442–454. MR4078474. doi: https://doi.org/10.1080/01621459.2018.1558062. 3

Wu, T., N. Narisetty, N., and Yang, Y. (2023). “Statistical inference via conditional Bayesian posteriors in high-dimensional linear regression.” *Electronic Journal of Statistics*, 17(1): 769–797. MR4551566. doi: https://doi.org/10.1214/23-ejs2113. 3, 4, 5, 6, 13

Zhang, C.-H. (2010). “Nearly unbiased variable selection under minimax concave penalty.” *The Annals of Statistics*, 38(2): 894–942. MR2604701. doi: https://doi.org/10.1214/09-AOS729. 2

Zhang, C.-H. and Zhang, S. S. (2014). “Confidence intervals for low dimensional parameters in high dimensional linear models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1): 217–242. MR3153940. doi: https://doi.org/10.1111/rssb.12026. 3, 6, 13