On high-dimensional Poisson models with measurement error: Hypothesis testing for nonlinear nonconvex optimization

Fei Jiang[1], **Yeqing Zhou**[2,*], **Jianxuan Liu**[3,†], **Yanyuan Ma**[4,‡]

[1]Department of Epidemiology and Biostatistics, The University of California, San Francisco

[2]School of Mathematical Sciences, Tongji University

[3]Department of Mathematics, Syracuse University

[4]Department of Statistics, Pennsylvania State University

Abstract

We study estimation and testing in the Poisson regression model with noisy high dimensional covariates, which has wide applications in analyzing noisy big data. Correcting for the estimation bias due to the covariate noise leads to a non-convex target function to minimize. Treating the high dimensional issue further leads us to augment an amenable penalty term to the target function. We propose to estimate the regression parameter through minimizing the penalized target function. We derive the L_1 and L_2 convergence rates of the estimator and prove the variable selection consistency. We further establish the asymptotic normality of any subset of the parameters, where the subset can have infinitely many components as long as its cardinality grows sufficiently slow. We develop Wald and score tests based on the asymptotic normality of the estimator, which permits testing of linear functions of the members if the subset. We examine the finite sample performance of the proposed tests by extensive simulation. Finally, the proposed method is successfully applied to the Alzheimer's Disease Neuroimaging Initiative study, which motivated this work initially.

Keywords and phrases:

High dimension Inference; Measurement Error; Non-convex optimization; Poisson model

MSC2020 subject classifications:

Primary 00X00; 00X00; secondary 00X00

1. Introduction.

Count data are routinely encountered in practice. For example, cognitive scores in a neuroscience study, the number of deaths in an infectious disease study, and the number of clicks on a particular product on an e-commerce platform, are all count data. Because

*zhouyeqing@tongji.edu.cn . †jliu193@syr.edu . ‡yzm63@psu.edu .

most of the count data are concentrated on a few small discrete values rather than expanded on the entire real line and because the distribution of count variables is often skewed, the familiar linear model becomes less ideal to capture these features. In the literature, Poisson regression (McCullagh & Nelder 2019) is arguably the most popular model to describe count outcomes, because it naturally models the skewed distribution for positive outcomes. On the other hand, together with the count data, a large number of covariates are often collected thanks to the ever advancing capability of modern technologies. However, these covariates are often contaminated with errors due to imperfect data acquisition and processing procedures. Ignoring these errors can produce biased results, which can finally lead to misleading statistical inference on the model parameters (Carroll et al. 2006) that explain the association between covariates and outcomes. Our goal is to develop rigorous statistical inference procedures to test linear hypotheses in the high dimensional Poisson model with noisy covariates. Such inference tools will enable explaining the association between the count outcome and the individual covariate or combination of covariate, quantifying the uncertainties of the estimated association, and controlling the false discovery rate when testing scientifically important hypotheses.

Let Y be the count outcome and $\mathbf{X}$ be its associated covariate vector. In the Poisson model, Y is related to $\mathbf{X}$ as

$$\mathrm{pr}(Y \mid \mathbf{X}) = e^{-\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})}\left\{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\right\}^{Y}/Y!. \tag{1}$$

We study the testing problem in (1) under the situation that the covariate vector $\mathbf{X}$ is both high dimensional and contaminated with noise. When $\mathbf{X}$ is accurately observed, the testing problem has been extensively discussed in the literature (Ning & Liu 2017, Zhang & Cheng 2017, Van de Geer et al. 2014, Shi et al. 2019). However, when $\mathbf{X}$ is not accurately observed, it is unclear that any of the existing proposed tests are applicable, and testing in the high dimensional noisy Poisson regression model has not been explored. The major obstacles in constructing valid hypothesis testing procedures are as follows. 1) The existing lasso-type penalized Poisson estimator (Jiang & Ma 2021) does not enjoy the variable selection consistency when the number of parameters is much larger than the sample size. 2) The asymptotic normality of the estimator has not been established. We develop Wald and score tests targeting at linear hypothesis on the parameters of interest in (1). To overcome obstacle 1), we improve the penalized Poisson estimator proposed in Jiang & Ma (2021) by using a class of "amenable" penalty functions first defined in Loh & Wainwright (2015, 2017) in combination with a modified log-likelihood function to construct estimators. We establish the estimation consistency and variable selection consistency of the resulting estimators. To bypass obstacle 2), we derive the asymptotic linear form of the estimators, and establish the asymptotic normality. The asymptotic normal estimator has a wider range of applications than the lasso type estimator does, because it facilitates subsequent inference procedures such as constructing hypothesis testing procedures.

Even after establishing the asymptotic normal properties, it is still challenging to generalize Wald and score tests to the high dimensional setting for Poisson regression with noisy data. This is because under the amenable penalties (Loh & Wainwright 2015, 2017), the

asymptotic normality of the estimators is built on a minimal signal condition, which requires the nonzero elements in $\boldsymbol{\beta}$ to be at least of order λ. Here λ is the penalty parameter which goes to zero when sample size increases. Now consider testing the null hypothesis $\beta_1 = 0$ versus the alternative $\beta_1 = h_n$, where β_1 is the first element of $\boldsymbol{\beta}$. The minimal signal condition implies that the test will have no power in testing the local alternative when $|h_n|_2 \ll \lambda$. To resolve this issue, we remove the penalties on the subvector of the parameters involved in the test. However, it is still unclear how fast the dimension of the subvector can grow while still ensuring sufficient power. To this end, we derive the convergence property of the estimators, which provides the explicit rate at which the dimension of the subvector is allowed to grow with the sample size in order to achieve consistency, asymptotic normality, and sufficient power in testing. Furthermore, to implement the score test, we need to estimate the regression parameters under the null hypothesis, which involves optimization under linear equality constraints. This type of constrained parameter estimation for noisy Poisson model has not yet been developed. To fill this gap, we develop a general procedure for parameter estimation under linear constraints. The constraints include inequality constraints for the parameter estimation under general Poisson model and an additional equality constraint imposed by the null hypothesis, which leads to great challenge in establishing the convexity. Incorporating inequality constraints is practically important because it allows to incorporate additional parameter information, which will reduce the estimation variation and in turn the sample size needed to achieve satisfactory estimation accuracy.

We briefly summarize our contributions as follows. First, we develop a new estimation procedure of the Poisson model with amenable regularization for noisy data. Second, we show the variable selection consistency and the consistency of the resulting estimator. We provide explicit convergence rate of the estimator. Third, we derive the asymptotic normality of the estimator for the nonzero parameters and the parameters to test. Fourth, we propose the Wald and score test procedures by constructing the corresponding test statistics. Fifth, we derive the asymptotic distributions of the Wald and score test statistics. These five essential elements combined together finally allow us to perform hypothesis testing for Poisson model with high dimensional noisy covariates, which allows us to answer important questions such as "if the left inferior temporal gyrus has a significant impact on the development of Alzheimer's disease". These estimation and inference tasks are not straightforward to achieve, they require building up a series of theoretical properties first, which involves techniques related to analyzing conditional sub-Gaussian distribution tails, utilizing and modifying various concentration inequalities, constructing the prime-dual equivalence, carefully bounding various quantities, linking different vector and matrix norms, and establishing a Lyapunov-type bound (Bentkus 2005) on the probability distribution to derive the asymptotic distribution of proposed test statistics. All these analyses are performed under the unusual constraints involving both linear equality constraints and parameter restrictions. We also modify the alternating direction method of multipliers (ADMM) algorithm to solve a regularized optimization problem under linear constraint in constrained parameter space. Although each individual technique in its basic form has been used in the literatures of mathematical analysis, statistics, combinatorics,

operations research and computer science, a seamless combination of all these into a general tool to solve the problem under study is very challenging and difficult.

Count data occur frequently in practice, and it is a rule rather than exception that the covariates can be contaminated. In modern data collection mechanism, covariates are almost always high dimensional. Hence, estimation and inference in Poisson regression with high dimensional noisy covariates is a general problem with wide applications. A direct motivation of this work is the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, which is a multi-site longitudinal study investigating early detection of Alzheimer's disease (AD) and tracking disease progression biomarkers (Weiner et al. 2017). Recently, the advent of tau-targeted positron emission tomography (PET) tracers such as flortaucipir (^{18}F-AV-1451) has made it possible to investigate the relative (to patient's body weight) tissue radioactivity concentration of the tracers, quantified as standardized uptake value ratio (SUVR), in relationship to the cognitive function. Therefore, we aim to study the association between cognitive scores and SUVRs from PEG image data. We extract Montreal Cognitive Assessment (MoCa) scores (Y) and SUVRs ($\mathbf{X}$) from the PET image in the ADNI study taken within 14 days of the cognitive tests from 196 subjects in the ADNI phase 3 study. We first perform a linear lasso regression between the logarithm of MoCa score and the 218 covariates including age, gender, SUVRs, and volumes of whole brain ROIs. Figure 1 shows the density of the residuals from the lasso regression, which suggests that the residuals are skewed and hence the linear lasso regression does not provide a satisfactory fit for the data. This motivates us to consider Poisson regression. We utilize the Poisson high dimensional hypothesis testing procedure developed in Shi et al. (2019) to examine which SUVRs are significantly associated with the MoCa scores. For each covariate of interest, we test the hypothesis that the corresponding coefficient is greater than zero. We plot the logarithm of the p-values from the score and Wald tests proposed in Shi et al. (2019) for the coefficients of the SUVRs at cortical ROIs. Figure 1 shows that if using 0.05/218 as a cut off for the p-value, both the Wald and score test identify the SUVRs at all cortical ROIs as significant predictors, which contradicts the fact that the cognitive functions are controlled by a subset of brain ROIs (Leisman et al. 2016). This unsatisfactory result likely attributes to the fact the Shi et al. (2019)'s method relies on the assumption that the expectation of the exponential of the distance between outcome and regression function is bounded (Condition (A3) in (Shi et al. 2019)) while neuroimage data are often subject to data acquisition and processing errors, which likely leads to violation of this assumption. This motivating example demonstrates the necessity of developing novel statistical inference procedure to test linear hypothesis in the high dimensional Poisson model with noisy data.

The rest of the paper is organized as follows. Section 2 discusses related work. In Section 3, we describe our model assumptions and the overall estimation strategy. We further detail the estimation with and without the null constraint, and the construction of the test statistics. The fundamental theoretical developments are provided in Section 4, where we establish convergence rates, the asymptotic normal results, and the properties of the test procedures. We study the practical implementation and the numerical performances in Section 5, where a detailed algorithm is provided, extensive simulations are carried out, and a ADNI data set

is analyzed. We conclude the paper in Section 6. The main mathematical proofs are provided in an Appendix given in a Supplementary Document.

2. Related Works and Notations.

Nonlinear models with high dimensional noisy data are in general hard problems to work with, partly because existing treatments usually lead to non-convex optimization, which violates standard requirements in the high dimensional data analysis literature. Thus, only linear models, which are the simplest in all noisy data problems, have received relatively thorough investigation (Loh & Wainwright 2012, Belloni & Rosenbaum 2016, Datta & Zou 2017, Belloni, Rosenbaum & Tsybakov 2017, Belloni, Chernozhukov & Kaul 2017, Li et al. 2021). Expanding the research framework to the Poisson regression context is difficult because the link function in the Poisson model is nonlinear. Subsequently, it is not easy to construct noise adjusted quantities such as a noise adjusted Hessian matrix like in the linear case. In addition, the Hessian matrix involves heavy tailed random variables due to the exponential link, even if all the covariates are sub-Gaussian in their original scale. These difficulties require additional restrictions on the moments of the covariate distribution as well as on the parameter searching space, which complicates all the subsequent computation and analysis. Indeed, the only works we are aware of in the high dimensional Poisson model with noisy data are Jiang & Ma (2021), Sørensen et al. (2015, 2018), Brown et al. (2019), while only the estimator in Jiang & Ma (2021) has been shown to be consistent. However, because all these methods use lasso-type L_1 penalty in the estimation, the resulting estimators do not enjoy variable selection consistency and their asymptotic distribution results are not established.

There is extensive literature on the linear hypothesis testing under high dimensional noise free setting. (Ning & Liu 2017) introduced a decorrelated score function to construct confidence regions for low dimensional components in high dimensional models. Zhang & Cheng (2017) used the desparsifying lasso estimator (Van de Geer et al. 2014) to propose a maximal-type statistic allowing the number of parameters that are involved in the test to grow with the sample size. Moreover, Shi et al. (2019) proposed a partial penalized likelihood ratio test, a score test, and a Wald test for testing the linear hypothesis of the parameters in high dimensional generalized linear models.

Notations.

We introduce some general notation that will be used throughout the text. For a matrix $\mathbf{M}$, let $\|\mathbf{M}\|_{\max}$ be the matrix maximum norm, $\|\mathbf{M}\|_{\infty}$ be the L_{∞} norm and $\|\mathbf{M}\|_p$ be the L_p norm. Let $\mathcal{F}(\boldsymbol{\beta})$ be the σ-field generated by $\mathbf{X}_i \boldsymbol{\beta}^{\mathrm{T}} \mathbf{W}_i$, $i = 1, \ldots, n$. Further, let $\mathcal{F}_X$ be the sigma-field generated by $\mathbf{X}_i$, $i = 1, \ldots, n$. For a general vector $\mathbf{a}$, let $\|\mathbf{a}\|_{\infty}$ be the vector sup-norm, $\|\mathbf{a}\|_p$ be the vector l_p-norm. Let $\mathbf{e}_j$ be the unit vector with 1 on its jth entry. For a vector $\mathbf{v} = (v_1, \ldots, v_p)^{\mathrm{T}}$, let sup($\mathbf{v}$) be the set of indices with $v_i \neq 0$ and $\|\mathbf{v}\|_0 = |\mathrm{supp}(\mathbf{v})|$, where $|\mathcal{U}|$ stands for the cardinality of the set $\mathcal{U}$. For a vector $\mathbf{v} \in \mathbb{R}^p$ and a subset $S \subseteq (1, \ldots, p)$, we use $\mathbf{v}_S \in \mathbb{R}^S$ to denote the vector obtained by restricting $\mathbf{v}$ on the set S.

Following Fletcher & Watson (1980), for an arbitrary norm $\|\cdot\|_A$ and its dual normal $\|\cdot\|_D$, we define $\|\mathbf{x}\|_A$ as the set $(\mathbf{v} : \|\mathbf{x}\|_A = \mathbf{v}^{\mathrm{T}}\mathbf{x}, \|\mathbf{v}\|_D \leq 1)$. Thus, for an arbitrary vector $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$, $\|\mathbf{x}\|_1 = \{\mathbf{v} = (v_1, \ldots, v_p)^{\mathrm{T}} : v_j = \mathrm{sign}(x_j)$ if $x_j \neq 0$, and $|v_j| \leq 1$ if $x_j = 0\}$, and $\|\mathbf{x}\|_2 = \{\mathbf{v} = (v_1, \ldots, v_p)^{\mathrm{T}} : v_j = x_j/\|\mathbf{x}\|_2\}$.

3. Model, Estimation and Test Statistics.

3.1 Problem Formulation: High dimensional Poisson model with noisy data.

Let $\mathbf{X}_i$ be a p-dimensional covariate, for example the image features, and let Y_i be a count random response variable, for example the MoCa score from the ADNI data. We model the relationship between Y_i and $\mathbf{X}_i$ $(i = 1, \ldots, n)$ through a Poisson model $\mathrm{pr}(Y_i = y \mid \mathbf{X}_i = \mathbf{x}) = e^{-\exp(\beta_t^{\mathrm{T}}\mathbf{x})}\{\exp(\beta_t^{\mathrm{T}}\mathbf{x})\}^y/y!$. Here, β_t is a p-dimensional sparse parameter vector. We allow the number of nonzero entries in β_t to grow with the sample size. We consider Poisson model here because our response is a count, and Poisson model is arguably the most standard model for count data. Indeed, Poisson model has been widely used to model the distribution of cognitive scores (Katz et al. 2021, Fallah et al. 2011, Mitnitski et al. 2014). We use $e^{\beta_t^{\mathrm{T}}\mathbf{x}}$ to model the conditional mean of the Poisson model to ensure the positiveness of the mean, and to allow possible skewness in the distribution (McCullagh & Nelder 2019). We assume β_t to be sparse because it often happens that only a few covariates have effect on the outcome. For example, in the ADNI data, because the cognitive functions are controlled by a subset of brain ROIs (Leisman et al. 2016), only a subset of brain features contributes to the cognitive function.

Furthermore, we assume the covariate $\mathbf{X}_i$ is not precisely observed and instead, a contaminated version of $\mathbf{X}_i$, denoted $\mathbf{W}_i$, is observed, where $\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i$, and $\mathbf{U}_i$ is the noise that is independent of both $\mathbf{X}_i$ and Y_i. For example, in the ADNI data, $\mathbf{X}_i$ can be the true image features, while $\mathbf{W}_i$ represent the observed image features which can deviate from the truth due to imperfect data collection and processing procedure. Without loss of generality, assume that $E(\mathbf{X}_i) = \mathbf{0}$, which can always be achieved by centering the observed covariates in practice. Furthermore, we assume $\mathbf{U}_i$ is a normally distributed random noise vector with mean zero and known covariance matrix $\mathbf{\Omega}$. The normal assumption for $\mathbf{U}_i$ is the common assumption at the state of the art in the Poisson measurement error literature and allows to derive analytic form of the loss function, which is the only setting that we can directly examine the convexity of the loss function. The known $\mathbf{\Omega}$ assumption is only for convenience of presentation. In practice, it is often replaced by an estimated version based on multiple observations, validation data or other standard instruments under both low and high dimensional settings (Carroll et al. 2006, Loh & Wainwright 2012), and the corresponding analysis is routine. Let $(\mathbf{X}_i, \mathbf{W}_i, Y_i, \mathbf{U}_i)$ be independent and identically distributed (iid) and assume $(\mathbf{W}_i, Y_i)$, $i = 1, \ldots, n$ are the iid observations. In this work, we devise estimation procedures for β and establishing theoretical properties of the estimator, we further aim at performing inference, such as conduct hypothesis testing. Throughout, we allow the covariate dimension to be much higher than the number of observations, i.e. $p \gg n$. We assume β_t is in the feasible set: $\{\beta : \|\beta\|_0 \leq k, \|\beta\|_2 \leq b_0\}$, which is practically sensible. A vector β in the feasible set automatically satisfies $\|\beta\|_1 \leq b_0\sqrt{k}$.

3.2. General Estimation Strategy.

If the true covariates $\mathbf{X}_i$ can be observed and the dimension p is fixed, this is a standard regression model and we routinely estimate $\boldsymbol{\beta}$ by minimizing the negative loglikelihood, which is proportional to

$$-n^{-1} \sum_{i=1}^{n} \left\{ Y_i \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta} - \exp\left(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_i\right) \right\}.$$

Here we use $\exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_i)$ to model the mean of Y_i because it preserves the positiveness of the mean estimate, and it is a standard choice in the generalized linear model (McCullagh & Nelder 2019). It is useful to note that for normal noise $\mathbf{U}_i$, we have the relation

$$E\left\{ \exp\left(\boldsymbol{\beta}_t^{\mathrm{T}} \mathbf{W}_i - \boldsymbol{\beta}_t^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{\beta}_t / 2\right) \mid \mathbf{X}_i \right\} = \exp\left(\boldsymbol{\beta}_t^{\mathrm{T}} \mathbf{X}_i\right), \tag{2}$$

$$E\left\{ \exp\left(\boldsymbol{\beta}_i^{\mathrm{T}} \mathbf{W}_i - \boldsymbol{\beta}_i^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{\beta}_i / 2\right)(\mathbf{W}_i - \boldsymbol{\Omega} \boldsymbol{\beta}_i) \mid \mathbf{X}_i \right\} = \exp\left(\boldsymbol{\beta}_i^{\mathrm{T}} \mathbf{X}_i\right) \mathbf{X}_i, \tag{3}$$

$$E\left[\exp\left(\boldsymbol{\beta}_i^{\mathrm{T}} \mathbf{W}_i - \boldsymbol{\beta}_i^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{\beta}_i / 2\right)\left\{ (\mathbf{W}_i - \boldsymbol{\Omega} \boldsymbol{\beta}_i)^{\otimes 2} - \boldsymbol{\Omega} \right\} \mid \mathbf{X}_i \right] = \exp\left(\boldsymbol{\beta}_i^{\mathrm{T}} \mathbf{X}_i\right) \mathbf{X}_i^{\otimes 2}. \tag{4}$$

Due to the conditional independence of $\mathbf{W}_i$ and Y_i given $\mathbf{X}_i$, (2) leads to

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

Consequently, it is a reasonable practice to estimate $\boldsymbol{\beta}$ by minimizing the loss function

$$\mathscr{L}(\boldsymbol{\beta}) = -n^{-1} \sum_{i=1}^{n} \left\{ Y_i \mathbf{W}_i^{\mathrm{T}} \boldsymbol{\beta} - \exp\left(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{W}_i - \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{\beta} / 2\right) \right\}, \tag{5}$$

which has the same mean as the negative log-likelihood function when $\mathbf{X}_i$ is accurately observed. When $n > p$, the estimator for $\boldsymbol{\beta}$ can be obtained by minimizing $\mathscr{L}(\boldsymbol{\beta})$ using the standard gradient descent method. However, when $n < p$, without addition regularization, optimizing (5) is an ill-posed mathematical problem because it does not have a unique solution. To take into account the ultra-high dimension nature of the model, using the fact that $\boldsymbol{\beta}$ is sparse, we propose to estimate $\boldsymbol{\beta}$ through solving the following constrained minimization problem

$$\min_{\|\boldsymbol{\beta}\|_1 \leqslant R_1, \|\boldsymbol{\beta}\|_2 \leqslant R_2} \left\{ \mathscr{L}(\boldsymbol{\beta}) + \rho_\lambda(\boldsymbol{\beta}) \right\} \tag{6}$$

at suitable R_1, R_2, where $\rho_\lambda(\boldsymbol{\beta})$ is a suitable penalty function. For convenience, define the set $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_1 \leqslant R_1, \|\boldsymbol{\beta}\|_2 \leqslant R_2\}$ as the feasible set (Fletcher & Watson 1980). Here R_1, R_2 can be any constants that are greater than the true $\|\boldsymbol{\beta}\|_1$ and $\|\boldsymbol{\beta}\|_2$, respectively. The condition $\|\boldsymbol{\beta}\|_1 \leqslant R_1$ is imposed to guarantee that the objective function satisfies the restricted eigenvalue condition discussed in Loh & Wainwright (2012) and therefore the objective function is convex in the feasible set, while the condition $\|\boldsymbol{\beta}\|_2 \leqslant R_2$ is imposed to avoid the explosion

of the mean function $\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\beta}/2)$. In practice, we often set R_1, R_2 to be a constant times the L_1, L_2 norms of the initial estimators of $\boldsymbol{\beta}$. Here, with a slight abuse of notation, we use the same symbol ρ_λ to denote both multivariate and univariate penalty functions and let $\rho_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{\|\beta\|_0} \rho_\lambda(\beta_j)$, where β_j is the jth element of $\boldsymbol{\beta}$ and $\|\boldsymbol{\beta}\|_0$ is the number of nonzero elements in $\boldsymbol{\beta}$.

3.3. Estimation under Hypotheses.

Consider testing the hypothesis that $\mathbf{C}\boldsymbol{\beta}_{\mathcal{M}} = \mathbf{t} + \mathbf{h}_n$ for some $\mathbf{h}_n \in \mathbb{R}^r$, where $\mathbf{C}$ is a $r \times m$ matrix with $r \leqslant m$, $\boldsymbol{\beta}_{\mathcal{M}}$ is a m-dimensional subvector of $\boldsymbol{\beta}$ with index set $\mathcal{M}$. The null hypothesis holds when $\mathbf{h}_n = \mathbf{0}$, while the alternative hypothesis holds when $\mathbf{h}_n \neq \mathbf{0}$. For example if $\mathbf{t} = 0$, $\mathbf{h}_n = 1$, $\mathbf{C} = (1, 0)$, $\mathcal{M}$ contains the index of the first element in $\boldsymbol{\beta}_t$, then testing $\mathbf{C}\boldsymbol{\beta}_{\mathcal{M}} = \mathbf{t} + \mathbf{h}_n$ is testing the null hypothesis that $\beta_{t1} = 0$ versus the alterative that $\beta_{t1} = 1$. Similarly, we can test $\beta_{t1} - \beta_{t2} = 0$ versus $\beta_{t1} - \beta_{t2} \neq 0$ by choosing $\mathbf{C} = (1, -1)$, $\mathbf{t} = 0$, $\mathbf{h}_n = 0$ or nonzero, and $\mathcal{M} = \{1, 2\}$. In summary, by varying $\mathbf{C}$, $\mathbf{t}$, $\mathbf{h}_n$, and $\mathcal{M}$, we can generate different linear hypotheses to test. Without loss of generality, we assume $\boldsymbol{\beta}_{\mathcal{M}}$ contains the first m elements of $\boldsymbol{\beta}$. Further, let $\boldsymbol{\beta}_{\mathcal{M}}^c$ be the vector containing elements that are not in $\mathcal{M}$, i.e. the last $p - m$ components of $\boldsymbol{\beta}$. Let $S \subseteq \mathcal{M}^c$ be the index set of the nonzero elements of $\beta_{t,\mathcal{M}^c}$. We assume $\beta_{t,\mathcal{M}^c}$ to be k sparse, i.e. $|S| = k$. Note that k is allowed to diverge with n. Without loss of generality, we assume the first k elements in $\beta_{t,\mathcal{M}^c}$ are none zero.

Suppose we are interested in testing whether $\mathbf{C}\boldsymbol{\beta}_{\mathcal{M}} = \mathbf{t}$ or not. Under the null hypothesis that $H_0 : \mathbf{C}\boldsymbol{\beta}_{\mathcal{M}} = \mathbf{t}$, we modify the general estimation strategy slightly and consider the estimator resulting from the equality and inequality constrained minimization:

$$\widehat{\boldsymbol{\beta}} = \mathrm{argmin}_{\|\beta\|_1 \leqslant R_1, \|\beta\|_2 \leqslant R_2}\{\mathscr{L}(\boldsymbol{\beta}) + \rho_\lambda(\boldsymbol{\beta}_{\mathcal{M}^c})\}, \text{ s.t. } \mathbf{C}\boldsymbol{\beta}_{\mathcal{M}} = \mathbf{t} \tag{7}$$

for suitable R_1, R_2. Without assuming the null hypothesis, we consider a similar estimator resulting from the inequality constrained minimization:

$$\widehat{\boldsymbol{\beta}}_a = \mathrm{argmin}_{\|\beta\|_1 \leqslant R_1, \|\beta\|_2 \leqslant R_2}\{\mathscr{L}(\boldsymbol{\beta}) + \rho_\lambda(\boldsymbol{\beta}_{\mathcal{M}^c})\}. \tag{8}$$

Note that here, both (7) and (8) are slightly different from the general strategy in (6), in that we do not place the penalty ρ_λ on the parameters in $\mathcal{M}$, which are to be tested for the linear relation $\mathbf{C}\boldsymbol{\beta}_{\mathcal{M}} = \mathbf{t}$. This special treatment is to avoid the situation that the penalty forces some components in $\boldsymbol{\beta}_{\mathcal{M}}$ to be zero, and therefore the null hypothesis $\mathbf{C}\boldsymbol{\beta}_{\mathcal{M}} = \mathbf{t}$ is affected not only by the data but also by our penalization.

3.4. Test statistics.

We define

$$\mathbf{Q}(\beta) \equiv E\{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\mathbf{X}\mathbf{X}^{\mathrm{T}}\}, \tag{9}$$

define the covariance of the residuals

$$\boldsymbol{\Sigma}(\beta) \equiv E\left[\left\{Y_i \mathbf{W}_i - \exp\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}_i - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\beta}/2\right)(\mathbf{W}_i - \boldsymbol{\Omega}\boldsymbol{\beta})\right\}^{\otimes 2}\right],$$

and define

$$\boldsymbol{\Psi}(\boldsymbol{\Sigma}, \mathbf{Q}, \boldsymbol{\beta}) \equiv \left(\mathbf{C}[\mathbf{I}_{m \times m}, \mathbf{0}_{m \times k}]\mathbf{Q}^{-1}_{\mathcal{M} \cup S, \mathcal{M} \cup S}(\boldsymbol{\beta})\boldsymbol{\Sigma}_{\mathcal{M} \cup S, \mathcal{M} \cup S}(\boldsymbol{\beta})\mathbf{Q}^{-1}_{\mathcal{M} \cup S, \mathcal{M} \cup S}(\boldsymbol{\beta})[\mathbf{I}_{m \times m}, \mathbf{0}_{m \times k}]^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\right).$$

Furthermore, let $\widehat{\boldsymbol{\Sigma}}(\beta)$ and $\widehat{\mathbf{Q}}(\beta)$ be a sample estimator of $\boldsymbol{\Sigma}(\beta)$ and $\mathbf{Q}(\beta)$, respectively. To test $\mathbf{C}\boldsymbol{\beta}_{\mathcal{M}} = \mathbf{t}$, we introduce two statistics, the Wald statistic

$$T_W = n\left(\mathbf{C}\widehat{\boldsymbol{\beta}}_{a.\mathcal{M}} - \mathbf{t}\right)^{\mathrm{T}}\boldsymbol{\Psi}\left(\widehat{\boldsymbol{\Sigma}}, \widehat{\mathbf{Q}}, \widehat{\boldsymbol{\beta}}_a\right)^{-1}\left(\mathbf{C}\widehat{\boldsymbol{\beta}}_{a.\mathcal{M}} - \mathbf{t}\right), \tag{10}$$

and the score statistic

$$T_S = n\left\{\frac{\partial\mathscr{L}(\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\right\}_{.\mathcal{M} \cup S}\left(\mathbf{C}[\mathbf{I}_{m \times m}, \mathbf{0}_{m \times k}]\widehat{\mathbf{Q}}^{-1}_{\mathcal{M} \cup S, \mathcal{M} \cup S}(\widehat{\boldsymbol{\beta}})\right)^{\mathrm{T}}$$

$$\times \boldsymbol{\Psi}^{-1}(\widehat{\boldsymbol{\Sigma}}, \widehat{\mathbf{Q}}, \widehat{\boldsymbol{\beta}})\mathbf{C}[\mathbf{I}_{m \times m}, \mathbf{0}_{m \times k}]\widehat{\mathbf{Q}}^{-1}_{\mathcal{M} \cup S, \mathcal{M} \cup S}(\widehat{\boldsymbol{\beta}})\left\{\frac{\partial\mathscr{L}(\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}}\right\}_{.\mathcal{M} \cup S}. \tag{11}$$

As we will show later in Section 4.4 that T_W and T_S are both asymptotically chi-square distributed with r degrees of freedom under the null hypothesis. Therefore, to control the false discovery rate at level α, we reject the null hypothesis if $T_W > \chi^2_{1-\alpha}(r)$ when we perform Wald test, or if $T_S > \chi^2_{1-\alpha}(r)$ when we perform score test. Here $\chi^2_{1-\alpha}(r)$ is the $1 - \alpha$ quantile of the chi-square distribution.

4. Theoretical Properties.

Define

$$\breve{\boldsymbol{\beta}}_{\mathcal{M}} \equiv \boldsymbol{\beta}_{t.\mathcal{M}} - \mathbf{C}^{\mathrm{T}}\left(\mathbf{C}\mathbf{C}^{\mathrm{T}}\right)^{-1}\mathbf{h}_n,$$

and let $\breve{\boldsymbol{\beta}} = \left(\breve{\boldsymbol{\beta}}_{t.\mathcal{M}}^{\mathrm{T}}, \boldsymbol{\beta}_{t.\mathcal{M}^c}^{\mathrm{T}}\right)^{\mathrm{T}}$. Thus, the last $p - m$ components of $\breve{\boldsymbol{\beta}}$, i.e. $\breve{\boldsymbol{\beta}}_{.\mathcal{M}^c}$, and the last $p - m$ components of $\boldsymbol{\beta}_0$, i.e. $\boldsymbol{\beta}_{0.\mathcal{M}^c}$, are identical. However, the first m components of $\breve{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ are different, in that $\mathbf{C}\breve{\boldsymbol{\beta}}_{\mathcal{M}} = \mathbf{t}$ under both null and alternative, while $\mathbf{C}\boldsymbol{\beta}_{\mathcal{M}} = \mathbf{t}$ under the null alone. Under some conditions, we first show that the inequality and equality constrained estimator $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\breve{\boldsymbol{\beta}}$ regardless the null or the alternative holds, and when $\|\mathbf{h}_n\|_2$ vanishes, $\widehat{\boldsymbol{\beta}}$ is also consistent as an estimator of the true parameter $\boldsymbol{\beta}_t$. Furthermore, we show that $\widehat{\boldsymbol{\beta}}_a$ is a consistent estimator of $\boldsymbol{\beta}_t$ regardless the null or the alternative holds. We then establish the asymptotic linear form of the estimators of a subvector $\widehat{\boldsymbol{\beta}}$ and a subvector of $\widehat{\boldsymbol{\beta}}_a$, which are formed by components of $\boldsymbol{\beta}_t$ that are either to

be tested or nonzero. Finally, using the asymptotic linear forms, we construct test statistics and prove the convergence properties of these test statistics under both null and alternative.

4.1. Conditions.

Before we proceed with the specific results, we first list a set of assumptions on the univariate penalty function ρ_λ which are similar to those in Loh & Wainwright (2015) and Loh & Wainwright (2017).

(A1) The function $\rho_\lambda(t)$ satisfies $\rho_\lambda(0) = 0$ and is symmetric around zero.

(A2) (A2) On the nonnegative real line $t \geq 0$, the function $\rho_\lambda(t)$ is nondecreasing. Furthermore, $\rho_\lambda(t)$ is subadditive, i.e. $\rho_\lambda(t_1 + t_2) \leq \rho_\lambda(t_1) + \rho_\lambda(t_2)$ for all $t_1, t_2 \geq 0$.

(A3) For $t > 0$, the function $\rho_\lambda(t)/t$ is non-increasing in t.

(A4) The function $\rho_\lambda(t)$ is differentiable at all $t \neq 0$ and sub-differentiable at $t = 0$, with $\lim_{t \to 0+} \rho'_\lambda(t) = \lambda$, where $\rho'(t)$ denotes the derivative of $\rho(t)$. Together with the symmetric Condition in (A1), this leads to $\lim_{t \to 0-} \rho'_\lambda(t) = -\lambda$.

(A5) There exists $\mu > 0$ so that $\rho_\lambda(t) + \mu t^2/2$ is convex.

(A6) There exists a $\gamma \in (0, +\infty)$ such that $\rho'_\lambda(t) = 0$ for all $t \geq \gamma\lambda$.

Conditions (A1)–(A3) are some general requirements as discussed in Zhang et al. (2012). Condition (A4) restricts the class of penalties by excluding regularizers that are not differentiable at 0, for example, the lasso penalty is excluded. Condition (A5) is known as weak convexity (Vial 1982, Chen & Gu 2014) and is a type of curvature constraint that controls the level of nonconvexity of ρ_λ. Condition (A6) is imposed to allow penalty to be zero if the estimator is $\gamma\lambda$ away from zero, which removes the estimation bias for the nonzero parameters. We say ρ_λ is μ-amenable if Conditions (A1)–(A5) hold, and we name ρ_λ (μ, γ)-amenable if Conditions (A1)–(A6) hold. The (μ, γ)-amenable penalty includes the smoothly clipped absolute deviation (SCAD) and the minimax concave penalty (Loh & Wainwright 2017).

We need some additional regularity conditions to support the theoretical development. These conditions impose upper and lower bounds on various quantities to ensure that the upper bounds are finite and the lower bounds are positive. They also restrict the relation between the sample size and parameter number so that $\log(p)/n \to 0$ in a slow rate of $1/\{\log(n)\}^2$. To save space, we only provide a discussion of these conditions here, while provide the details in the supplementary material. Specifically, Condition (C1) (a) is a standard assumption used in noisy data problem such as that used in Sentürk & Müller (2005) and is usually satisfied in practice. Condition (C1) (b) guarantees the boundedness and the invertibility of the Hessian matrix (4), i.e. the second derivative of the noise free log likelihood. Conditions (C2) and (C3) bound the total variability of both the response Y and the noise **U** marginally and conditionally on the covariates **X**. Similar requirement is also assumed in Loh & Wainwright (2012). Condition (C4) shows that the dimension of the covariate can grow exponentially faster than the sample size. Finally, Jiang & Ma (2021) have discussed

the Conditions (C5)–(C7) and provided examples showing that the conditions are usually satisfied in practice.

4.2. Consistency.

We first show that the equality and inequality constrained estimator $\widehat{\beta}$ is a consistent estimator of $\breve{\beta}$ in Theorems 1 and 2, which is the same as the true parameter β_t, except that the first m components are adjusted to ensure that H_0 holds for $\breve{\beta}$.

Theorem 1. Define

$$\alpha_1 \equiv \min_{\|\beta\|_1 \leqslant R_1, \|\beta\|_2 \leqslant R_2} \alpha_{\min}\left[E\left\{\exp\left(\beta^\mathrm{T}\mathbf{X}_i\right)\mathbf{X}_i\mathbf{X}_i^\mathrm{T}\right\}\right]/2.$$

Assume $\|\mathbf{C}_r^{-1}\mathbf{C}_{m-r}\|_2 = O(1)$, ρ_λ satisfies Conditions (A1) – (A6) and Conditions (C1) – (C6) in the supplementary material hold. Assume $\alpha_1 > 3/4\mu$, and β is in the feasible set. Let λ satisfy

$$4\max\left\{\left\|\partial\mathscr{L}(\breve{\beta})/\partial\beta\right\|_\infty, \alpha_1(\log(p)/n)^{1/4}\right\} \leqslant \lambda \leqslant \frac{\alpha_1}{6R_1}$$

and $n \geqslant \log(p)\max\left(16R_1^4\tau_1^4/\alpha_1^4, 64R_1^4\tau_1^2/\alpha_1^2\right)$. Write $t_1 \equiv \sqrt{r}\|\mathbf{C}_r^{-1}\mathbf{C}_{m-r}\|_2 + \sqrt{m-r}$ and $t \equiv \left(6\lambda\sqrt{k} + 2\lambda t_1\right)(4\alpha_1 - 3\mu)^{-1}$. Then the local minimum of (7) satisfies the error bounds

$$\left\|\widehat{\beta} - \breve{\beta}\right\|_2 \leqslant t.$$

and

$$\left\|\widehat{\beta} - \breve{\beta}\right\|_1 \leqslant \left(4\sqrt{k} + t_1\right)t.$$

Following the similar argument, we also show that the inequality constrained estimator $\widehat{\beta}_a$ is a consistent estimator of the true parameter β.

Theorem 2. Let

$$\alpha_1 = \min_{\|\beta\|_1 \leqslant R_1, \|\beta\|_2 \leqslant R_2} \alpha_{\min}\left[E\left\{\exp\left(\beta^\mathrm{T}\mathbf{X}_i\right)\mathbf{X}_i\mathbf{X}_i^\mathrm{T}\right\}\right]/2$$

and let ρ_λ satisfy Conditions (A1) – (A6) and Conditions (C1) – (C6) in the supplementary material hold. Assume $\alpha_1 > 3/4\mu$, and β is in the feasible set. Let λ satisfy

$$4\max\left\{\left\|\partial\mathscr{L}(\beta_t)/\partial\beta\right\|_\infty, \alpha_1(\log(p)/n)^{1/4}\right\} \leqslant \lambda \leqslant \frac{\alpha_1}{6R_1}$$

and $n \geqslant \log(p) \max\left(16 R_1^4 \tau_1^4 / \alpha_1^4, 64 R_1^4 \tau_1^2 / \alpha_1^2\right)$. Then the local minimum of (8) satisfies the error bounds

$$\left\| \widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_t \right\|_2 \leqslant \frac{6\lambda\sqrt{k} + 2\lambda\sqrt{m}}{4\alpha_1 - 3\mu}.$$

and

$$\left\| \widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_t \right\|_1 \leqslant (4\sqrt{k} + \sqrt{m})\frac{6\lambda\sqrt{k} + 2\lambda\sqrt{m}}{4\alpha_1 - 3\mu}.$$

Theorems 1 and 2 suggest that when $\log(p)/n \to 0$, and when λ is suitably chosen, for example, λ is at least no smaller than $O[\{\log(p)/n\}^{1/4}]$, both $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_a$ converge to their corresponding true values in terms of both l_1 and l_2 norms, as long as k and m grow slower than $\{n/\log(p)\}^{1/2}$. These theoretical results suggest that the dimension of $\boldsymbol{\beta}_{\mathcal{M}}$, i.e., the number of parameters involved in the tests, and the number of nonzero entries in $\boldsymbol{\beta}_t$ can grow at a slower rate of $\{n/\log(p)\}^{1/2}$ under noisy Poisson model. These results also assist us to find reasonable ranges for λ in practice to obtain consistent estimators.

4.3. Asymptotic linear forms.

We denote $\tilde{\boldsymbol{\beta}}$ as a stationary point of (7), which satisfies the first order condition that

$$\left\{ \partial \mathscr{L}(\tilde{\boldsymbol{\beta}}) / \partial \boldsymbol{\beta}^{\mathrm{T}} + \partial \rho_\lambda(\tilde{\boldsymbol{\beta}}_{\mathcal{M}^c}) / \partial \boldsymbol{\beta}_{\mathcal{M}^c}^{\mathrm{T}} \mathbf{A} \right\}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \geqslant 0, \tag{12}$$

for all $\boldsymbol{\beta} \in \mathbb{R}^p$ in the feasible set and satisfies $\mathbf{C}\boldsymbol{\beta}_{\mathcal{M}} = \mathbf{t}$. Here $\mathbf{A} = (\mathbf{0}_{p-m,m}, \mathbf{I}_{p-m,p-m})$ is a matrix that satisfies $\|\mathbf{A}\|_\infty = \|\mathbf{A}\|_1 = 1$. Likewise, we denote $\tilde{\boldsymbol{\beta}}_a$ as a stationary point of (8), which satisfies the first order condition that

$$\left\{ \partial \mathscr{L}(\tilde{\boldsymbol{\beta}}_a) / \partial \boldsymbol{\beta}^{\mathrm{T}} + \partial \rho_\lambda(\tilde{\boldsymbol{\beta}}_{a.\mathcal{M}^c}) / \partial \boldsymbol{\beta}_{a.\mathcal{M}^c}^{\mathrm{T}} \mathbf{A} \right\}(\boldsymbol{\beta}_a - \tilde{\boldsymbol{\beta}}_a) \geqslant 0, \tag{13}$$

for all $\boldsymbol{\beta}_a \in \mathbb{R}^p$ in the feasible set.

To show the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_a$, our first step is to establish that the local minimizers $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}_a$ achieve variable selection consistency. To do this, we follow the prime-dual construction introduced in Wainwright (2009). We first show that both

$$\min_{\|\boldsymbol{\beta}\|_1 \leqslant R_1, \|\boldsymbol{\beta}\|_2 \leqslant R_2, \boldsymbol{\beta} \in \mathbb{R}^{\mathcal{M} \cup S}} \left\{ \mathscr{L}(\boldsymbol{\beta}) + \rho_\lambda(\boldsymbol{\beta}_{\mathcal{M}^c}) \right\}, \text{ such that } \mathbf{C}\boldsymbol{\beta}_{\mathcal{M}} = \mathbf{t} \tag{14}$$

and

$$\min_{\|\boldsymbol{\beta}\|_1 \leqslant R_1, \|\boldsymbol{\beta}\|_2 \leqslant R_2, \boldsymbol{\beta} \in \mathbb{R}^{\mathcal{M} \cup S}} \left\{ \mathscr{L}(\boldsymbol{\beta}) + \rho_\lambda(\boldsymbol{\beta}_{\mathcal{M}^c}) \right\} \tag{15}$$

have unique local minimizer in the interior of the feasible set. Then we show that all stationary points of (7) and (8) must have support in $\mathcal{M} \cup S$. Since the local minimizers

of (7) and (8) are automatically stationary points of (7) and (8) respectively, the local minimizers of (7) and (8) must also have support in $\mathcal{M} \cup S$. Therefore, the local minimizers of (7) and (8) are actually the local minimizers of (14) and (15) respectively, so are also unique. In other words, $\widehat{\beta}$ and $\widehat{\beta}_a$ are respectively the unique solution of (14) and (15) hence achieve the variable selection consistency. The details of the above analysis are presented in Theorem A.1 and Theorem A.2 in the Appendix A in the supplementary material.

In our second step to establish the asymptotic distribution properties of $\widehat{\beta}$ and $\widehat{\beta}_a$, we define

$$\widehat{\mathbf{Q}}(\beta) = \frac{\partial^2 \mathscr{L}(\beta)}{\partial \beta \partial \beta^{\mathrm{T}}},$$

and define

$$\mathbf{A}_2 = [\mathbf{I}_{m \times m}, \mathbf{0}_{m \times k}]^{\mathrm{T}} \mathbf{C}^{\mathrm{T}} \left(\mathbf{C}[\mathbf{I}_{m \times m}, \mathbf{0}_{m \times k}] \left\{ \widehat{\mathbf{Q}}_{\mathcal{M} \cup S, \mathcal{M} \cup S}(\beta^*) \right\}^{-1} \times [\mathbf{I}_{m \times m}, \mathbf{0}_{m \times k}]^{\mathrm{T}} \mathbf{C}^{\mathrm{T}} \right)^{-1} \mathbf{C}[\mathbf{I}_{m \times m}, \mathbf{0}_{m \times k}],$$

where β^* is the point in between $\widehat{\beta}$ and β_t and

$$\mathbf{A}_2^* = [\mathbf{I}_{m \times m}, \mathbf{0}_{m \times k}]^{\mathrm{T}} \mathbf{C}^{\mathrm{T}} \left(\mathbf{C}[\mathbf{I}_{m \times m}, \mathbf{0}_{m \times k}] \{ \mathbf{Q}_{\mathcal{M} \cup S, \mathcal{M} \cup S}(\beta) \}^{-1} \times [\mathbf{I}_{m \times m}, \mathbf{0}_{m \times k}]^{\mathrm{T}} \mathbf{C}^{\mathrm{T}} \right)^{-1} \mathbf{C}[\mathbf{I}_{m \times m}, \mathbf{0}_{m \times k}],$$

where $\mathbf{Q}(\beta) = E\{\exp(\beta^{\mathrm{T}}\mathbf{X})\mathbf{X}\mathbf{X}^{\mathrm{T}}\}$ is defined in (9). Based on the variable selection consistency established in the first step, we derive the asymptotic linear form of $\widehat{\beta}_{\mathcal{M} \cup S}$ and $\widehat{\beta}_{a.\mathcal{M} \cup S}$ under null and alternative hypothesis in Theorems 3 and 4, respectively.

Theorem 3. Assume ρ_λ satisfies Conditions (A1) – (A6) and Conditions (C1) – (C7) in the supplementary material hold, $\lambda = O_p[\{\log(p)/n\}^{1/4}]$, $\|\mathbf{C}_r^{-1}\mathbf{C}_{m-r}\|_2 = O(1)$, and $\lambda \leqslant a_1/(8R_1)$. Further we assume the boundedness $\left\| \{\mathbf{Q}_{(\mathcal{M} \cup S), \mathcal{M} \cup S}^{-1}(\beta_t)\} \right\|_\infty \leqslant c_\infty$, and $\left\| \{\mathbf{Q}_{(\mathcal{M} \cup S), \mathcal{M} \cup S}(\beta_t)\}^{-1} \mathbf{A}_2 \mathbf{Q}_{(\mathcal{M} \cup S), \mathcal{M} \cup S}^{-1}(\beta_t) \right\|_\infty \leqslant c_\infty$. In addition assume $\|\mathbf{h}_n\|_2 = O\{\sqrt{\max(m + k - r, r)/n}\}$, $\min(|\beta_j|) \geqslant \lambda(\gamma + 5c_\infty)$ for $j \in S$ and $n \geqslant c_\infty(m + k)^4\log(p)$. Then we have

$$\widehat{\beta}_{\mathcal{M} \cup S} - \beta_{t.\mathcal{M} \cup S} = -\left(\{\mathbf{Q}_{\mathcal{M} \cup S, <\mathcal{M} \cup S}(\beta_t)\}^{-1} - \{\mathbf{Q}_{\mathcal{M} \cup S, \mathcal{M} \cup S}(\beta_t)\}^{-1}\mathbf{A}_2^* \times \{\mathbf{Q}_{\mathcal{M} \cup S, \mathcal{M} \cup S}(\beta_t)\}^{-1} \right) \left\{ \frac{\partial \mathscr{L}(\beta_t)}{\partial \beta} \right\}_{\mathcal{M} \cup S}$$
$$\{1 + o_p(1)\} + \{\mathbf{Q}_{\mathcal{M} \cup S, \mathcal{M} \cup S}(\beta_t)\}^{-1}\mathbf{A}_2^* \left[\left\{ \left(\mathbf{C}\mathbf{C}^{\mathrm{T}}\right)^{-1}\mathbf{C}, \mathbf{0}_{r \times k} \right\}^{\mathrm{T}} \mathbf{h}_n \right] \{1 + o_p(1)\}$$

and $\widehat{\beta}_{(\mathcal{M} \cup S)^c} = \mathbf{0}$.

Theorem 4. Assume ρ_λ satisfies Conditions (A1) – (A6) and Conditions (C1) – (C7) in the supplementary material hold, $\lambda = O_p[\{\log(p)/n\}^{1/4}]$, and $\lambda \leqslant a_1/(8R_1)$. Further we assume $\|\{\mathbf{Q}_{(\mathcal{M} \cup S), \mathcal{M} \cup S}(\beta_t)\}^{-1}\|_\infty$, $\leqslant c_\infty$, $\min(|\beta_j|) \geqslant \lambda(\gamma + 5c_\infty)$ for $j \in S$ and $n \geqslant c_\infty(m + k)^4\log(p)$. Then we have

$$\widehat{\boldsymbol{\beta}}_{a,\mathcal{M}\cup S} - \boldsymbol{\beta}_{t,\mathcal{M}\cup S} = -\{\mathbf{Q}_{\mathcal{M}\cup S, \mathcal{M}\cup S}(\boldsymbol{\beta}_t)\}^{-1}\left\{\frac{\partial \mathcal{L}(\boldsymbol{\beta}_t)}{\partial \boldsymbol{\beta}}\right\}_{\mathcal{M}\cup S}\{1 + o_p(1)\}$$

and $\widehat{\boldsymbol{\beta}}_{(\mathcal{M}\cup S)^c} = \mathbf{0}$.

Theorems 3 and 4 suggest that the asymptotic linear forms of $\widehat{\boldsymbol{\beta}}_{\mathcal{M}\cup S}$ and $\widehat{\boldsymbol{\beta}}_{a,\mathcal{M}\cup S}$ are the usual product of the inverse of Hessian matrix and the score function. Furthermore, only the first $(m+k) \times (m+k)$ block in the Hessian matrix and the first $m+k$ elements in the score function contribute to the asymptotic distribution. Therefore, when $m+k$ grows slower than $\{n/\log(p)\}^{1/4}$ and $\|\mathbf{h}_n\| \to 0$, it is easy to see that the asymptotic linear forms converge in distribution to Gaussian random vectors. It is worth mentioning that the minimal signal condition $\min(|\beta_j|) \geqslant \lambda(\gamma + 5c_\infty)$ for $j \in S$ is a standard requirement for the optimization using nonconvex penalty such as SCAD (Fan & Li 2001). This condition is also very weak because $\lambda \to 0$, which allows the minimal signal vanishing to zero.

4.4. Asymptotic distribution of the test statistics.

To study the asymptotic behavior of T_S and T_W, we first investigate the distribution of their asymptotic form T_0 defined by

$$T_0 \equiv \left(\boldsymbol{\omega}_n + \sqrt{n}\mathbf{h}_n\right)^{\mathrm{T}} \boldsymbol{\Psi}^{-1}(\boldsymbol{\Sigma}, \mathbf{Q}, \boldsymbol{\beta}_t)\left(\boldsymbol{\omega}_n + \sqrt{n}\mathbf{h}_n\right),$$

where

$$\boldsymbol{\omega}_n = -\sqrt{n}\mathbf{C}[\mathbf{I}_{m\times m}, \mathbf{0}_{m\times k}]\mathbf{Q}_{\mathcal{M}\cup S, \mathcal{M}\cup S}^{-1}(\boldsymbol{\beta}_t)\left\{\frac{\partial \mathcal{L}(\boldsymbol{\beta}_t)}{\partial \boldsymbol{\beta}}\right\}_{\mathcal{M}\cup S}.$$

As shown in Lemma 1, T_0 is asymptotically noncentral chi-square distributed with the noncentral parameter approaches $n\mathbf{h}_n^{\mathrm{T}}\boldsymbol{\Psi}^{-1}(\boldsymbol{\Sigma}, \mathbf{Q}, \boldsymbol{\beta}_t)\mathbf{h}_n$.

__Lemma__ 1. Assume ρ_λ satisfies Conditions (A1) – (A6) and Conditions (C1) and (D1) in the supplementary material hold and $n \geqslant c_\infty(m+k)^4\log(p)$, then

$$\lim_{n\to\infty}\sup_{\mathcal{C}}\left|\Pr(T_0 \leqslant x) - \Pr\left\{\chi^2\left(r, n\mathbf{h}_n^{\mathrm{T}}\boldsymbol{\Psi}^{-1}(\boldsymbol{\Sigma}, \mathbf{Q}, \boldsymbol{\beta}_t)\mathbf{h}_n\right) \leqslant x\right\}\right| = 0,$$

where $\chi^2(r, \gamma)$ is a non-central chi-square random variable, with non-centrality parameter γ.

Here Condition (D1) provides upper bound of the third moment of each summand in $\boldsymbol{\omega}$ (note that $\mathcal{L}(\boldsymbol{\beta}_t)/\partial\boldsymbol{\beta}$ is the summation of the derivatives of the negative log-likelihood from n samples), which is a necessary condition to establish convergence in distribution. See Theorem 3.1 in Shi et al. (2019) for example. To establish the asymptotic distribution of T_W and T_S, in Theorems 5 and 6 respectively, we show that T_W and T_S are close to T_0, hence has the same testing property asymptotically when r is finite.

Theorem 5. Assume the conditions in Theorem 4 and Conditions (D1) and (D2) in the Section B.4.2 in the supplementary material hold, we have $T_W - T_0 = o_p(r)$. Therefore,

$$\lim_{n \to \infty} \sup_{\mathscr{C}} \left| \Pr(T_W \leqslant x) - \Pr\left\{ \chi^2\left(r, n\mathbf{h}_n^{\mathrm{T}} \mathbf{\Psi}^{-1}(\mathbf{\Sigma}, \mathbf{Q}, \boldsymbol{\beta}_t)\mathbf{h}_n\right) \leqslant x \right\} \right| = 0,$$

where $\chi^2(r, \gamma)$ is a non-central chi-square random variable, with non-centrality parameter γ.

Theorem 6. Assume the conditions in Theorem 3, Conditions (D1) and (D2) in the Section B.4.2 in the supplementary material hold, we have $T_S - T_0 = o_p(r)$. Therefore,

$$\lim_{n \to \infty} \sup_{\mathscr{C}} \left| \Pr(T_S \leqslant x) - \Pr\left\{ \chi^2\left(r, n\mathbf{h}_n^{\mathrm{T}} \mathbf{\Psi}^{-1}(\mathbf{\Sigma}, \mathbf{Q}, \boldsymbol{\beta}_t)\mathbf{h}_n\right) \leqslant x \right\} \right| = 0,$$

where $\chi^2(r, \gamma)$ is a non-central chi-square random variable, with non-centrality parameter γ.

Here Condition (D2) in the Section B.4.2 is a regularity condition ensures $\mathbf{\Psi}(\mathbf{\Sigma}, \mathbf{Q}, \boldsymbol{\beta}_t)$ to be positive definite. Theorems 5 and 6 show that the two test statistics T_W and T_S indeed have the same $\chi^2(r, \gamma)$ distribution as T_0 in large samples, hence can be used to perform the standard chi-square test. A curious question is whether or not a likelihood ratio type of test can also be constructed. We feel it is hard in this context because it is almost impossible to obtain a likelihood function in the functional measurement error context. Much work is needed to overcome this obstacles.

5. Numerical Implementation.

5.1. Computational algorithms.

We compute the estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_a$ using the popular ADMM. In what follows, we only detail the algorithm to estimate $\widehat{\boldsymbol{\beta}}$. The estimator $\widehat{\boldsymbol{\beta}}_a$ can be computed in a similar way. For a given λ, we consider

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmin}_{\|\boldsymbol{\beta}\|_1 \leqslant R_1, \|\boldsymbol{\beta}\|_2 \leqslant R_2} \{ \mathscr{L}(\boldsymbol{\beta}) + \rho_\lambda(\boldsymbol{\beta}_{\mathscr{M}^c}) \}, \text{ s.t. } \mathbf{C}\boldsymbol{\beta}_{\mathscr{M}} = \mathbf{t}$$

for constants R_1, R_2. Similar to Shi et al. (2019), this optimization problem is equivalent to

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}) = \operatorname{argmin}_{\|\boldsymbol{\beta}\|_1 \leqslant R_1, \|\boldsymbol{\beta}\|_2 \leqslant R_2} \{ \mathscr{L}(\boldsymbol{\beta}) + \rho_\lambda(\boldsymbol{\beta}_{\mathscr{M}^c}) \}, \text{ s.t. } \mathbf{C}\boldsymbol{\beta}_{\mathscr{M}} = \mathbf{t}, \boldsymbol{\beta}_{\mathscr{M}^c} = \boldsymbol{\theta}.$$

By the augmented Lagrangian method, the estimators can be obtained by minimizing

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{v}) = \mathscr{L}(\boldsymbol{\beta}) + \rho_\lambda(\boldsymbol{\beta}_{\mathscr{M}^c}) + \boldsymbol{v}^{\mathrm{T}} \begin{pmatrix} \mathbf{C}\boldsymbol{\beta}_{\mathscr{M}} - \mathbf{t} \\ \boldsymbol{\beta}_{\mathscr{M}^c} - \boldsymbol{\theta} \end{pmatrix} + \frac{\rho}{2} \left\| \begin{matrix} \mathbf{C}\boldsymbol{\beta}_{\mathscr{M}} - \mathbf{t} \\ \boldsymbol{\beta}_{\mathscr{M}^c} - \boldsymbol{\theta} \end{matrix} \right\|_2^2,$$

with $\|\boldsymbol{\beta}\|_1 \leqslant R_1$, $\|\boldsymbol{\beta}\|_2 \leqslant R_2$, where the dual variables $\boldsymbol{v}$ are Lagrange multipliers and $\rho > 0$ is a given penalty parameter. We compute the estimators of $(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{v})$ through iterations. Let

the sup-script (t) indicate the t-th iteration, we describe the main steps of ADMM methods in Algorithm 1.

Algorithm 1 ADMM Algorithm for estimating $\hat{\beta}$.

For $t = 0, 1, \ldots, t_{max}$, perform:

Step 1. Use the Newton-Raphson algorithm to solve (17) to obtain $\tilde{\beta}^{(t+1)}$.

$$(17) \qquad \tilde{\beta}^{(t+1)} = \arg\min_\beta \left\{ \mathcal{L}(\beta) + v^{(t)\mathrm{T}} \begin{pmatrix} C\beta_\mathcal{M} - t \\ \beta_{\mathcal{M}^c} - \theta^{(t)} \end{pmatrix} + \frac{\rho}{2} \left\| \begin{matrix} C\beta_\mathcal{M} - t \\ \beta_{\mathcal{M}^c} - \theta^{(t)} \end{matrix} \right\|_2^2 \right\}.$$

Step 2. Project $\tilde{\beta}^{(t+1)}$ to a L_1 ball with radius R_1 to obtain $\hat{\beta}^{(t+1)}$ by the simplex projection method (Duchi et al. 2008). If $\|\hat{\beta}^{(t+1)}\|_2 > R_2$, we shrink it to get $\beta^{(t+1)} = \hat{\beta}^{(t+1)} R_2 / \|\hat{\beta}^{(t+1)}\|_2$. Otherwise, $\beta^{(t+1)} = \hat{\beta}^{(t+1)}$.

Step 3. Obtain $\theta^{(t+1)}$ by solving (17), where the penalty term we use is SCAD with $a = 3.7$.

$$(18) \qquad \theta^{(t+1)} = \arg\min_\theta \left\{ \rho_\lambda(\theta) + \frac{\rho}{2} \left\| \beta_{\mathcal{M}^c}^{(t+1)} - \theta \right\|_2^2 + v^{(t)\mathrm{T}} \begin{pmatrix} C\beta_\mathcal{M}^{(t+1)} - t \\ \beta_{\mathcal{M}^c}^{(t+1)} - \theta \end{pmatrix} \right\}.$$

Step 4. Update the dual variables v by

$$v^{(t+1)} \leftarrow v^{(t)} + \rho \begin{pmatrix} C\beta_\mathcal{M}^{(t+1)} - t \\ \beta_{\mathcal{M}^c}^{(t+1)} - \theta^{(t+1)} \end{pmatrix}.$$

Step 5. If stopping rule $\|\beta^{(t+1)} - \beta^{(t)}\|_2 \leq \delta_{tol}$ or $\|\theta^{(t+1)} - \theta^{(t)}\|_2 \leq \delta_{tol}$ is satisfied, where δ_{tol} denotes the tolerance of error, then terminate the algorithm.

End of the main loop.

In the implementation, the initial value $\beta^{(0)}$ can be computed by a penalized Poisson regression following Jiang & Ma (2021). For the radii R_1 and R_2, we consider $R_1 = \sqrt{2} R_2$ and $R_2 = 1.5 \|\beta\|_2^{(0)}$. In the implementation, if the algorithm converges to the boundary, we can increase the corresponding norm R_1 or R_2 slightly. In contrast, if multiple minimum problems are encountered, we can decrease R_1 and when the estimation procedure leads to a very large $\exp(\beta^\mathrm{T} X)$, we can decrease R_2, gradually. The tuning parameter λ is selected by minimizing

$$\mathrm{BIC}(\lambda) = n \mathscr{L}(\hat{\beta}) + c_n \|\hat{\beta}\|_0 \qquad (16)$$

with respect to λ, where c_n is a positive number that may depend on n. In our analysis, we follow Shi et al. (2019) to adopt $c_n = \max\{\log n, \log(\log(n)) \log p\}$. For simplicity, we set $\rho = 1$.

5.2. Simulation Experiments.

We generate the outcome Y_i from the Poisson model

$$\Pr(Y_i = y \mid \mathbf{X}_i) = \exp\left\{ -\exp\left(\beta^\mathrm{T} \mathbf{X}_i\right) \right\} \exp\left(y \beta^\mathrm{T} \mathbf{X}_i\right) / y!,$$

where the covariates $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,p})^\mathrm{T}$ are generated from two distributions: (I) the multivariate normal distribution with mean zero and covariance matrix Σ. (II) the uniform distribution in the interval $(-\sqrt{6}/2, \sqrt{6}/2)$. To generate correlated uniform distribution, we first draw covariates independently from $\mathcal{U}(-\sqrt{6}/2, \sqrt{6}/2)$, and then transform these covariates by multiplying the Choleski factorization of covariance Σ. We consider two forms of the covariance matrix: uncorrelated structure $\Sigma = 0.5 \mathbf{I}_p$ and correlated with autoregressive AR(1) structure $\Sigma = (0.5^{|i-j|+1})_{p \times p}$ for $i, j = 1, \ldots, p$. Furthermore, the noise $\mathbf{U}_i$ is drawn from the multivariate normal distribution with mean zero and covariance matrix $\Omega = 0.1\Sigma$. The true coefficient $\beta = (\beta_1, \ldots, \beta_p)^\mathrm{T} = (0.75, -0.75 + h_2, h_3, 0, \ldots, 0, h_p)^\mathrm{T}$. Here $h_j, j = 2, 3, p$ are assigned various values to check the empirical powers of the tests. We set $h_j = 0$ when j 2, 3 or p. For simplicity, the initial $\beta^{(0)}$ is set to be a p-dimensional zero vector. We select

parameter λ as described in Section 5.1. The candidate list for λ is $\{e^{-2.5}, e^{-2.245}, \ldots, e^{0.5}\}$ of length 41. We consider sample size $n = 300, 500$ and covariate dimension $p = 50, 350, 600$. The tolerance of error $\delta_{tol} = 10^{-4}$. We repeat each setting 500 times, and report the size and power of the proposed tests under different hypotheses. We perform the tests at type I error $\alpha = 0.05$ in the following scenarios.

5.2.1. Univariate parameter testing.—We first consider the following three hypotheses on a single element in $\boldsymbol{\beta}$.

$$H_{0,1}: \beta_2 = -0.75, \ v.s. \ H_{a,1}: \beta_2 \neq -0.75.$$

$$H_{0,2}: \beta_3 = 0, \ v.s. \ H_{a,2}: \beta_3 \neq 0.$$

$$H_{0,3}: \beta_p = 0, \ v.s. \ H_{a,3}: \beta_p \neq 0.$$

To test a hypothesis set regarding β_j, we simulate data with $h_j = 0, 0.1, 0.2, 0.4$, while set $h_k = 0$ for $k \neq j$. For example, to test $H_{0,1}$ and $H_{a,1}$, we simulate data with $h_3 = 0$, $h_p = 0$, and $h_2 = 0, 0.1, 0.2, 0.4$. When $h_2 = 0$, the null hypothesis $H_{0,1}$ holds, we study the type I error of the test. On the other hand, when $h_2 = 0.1$ to 0.4, the alternative hypothesis is true, which allows us to examine the power of the test. Tables 1 and 2 summarize the empirical type I error and powers of the Wald and score tests. It is clear that the empirical type I errors are controlled at the nominal level 0.05 in all scenarios, indicating that the proposed tests are consistent. The powers of the Wald and score tests increase gradually when the magnitude of $|h_j|$'s increases, and have satisfactory powers in general. The Wald and score tests yield similar performances in all scenarios. This finding is in accordance with theoretical analysis.

5.2.2. Linear hypothesis testing.—We also consider the hypotheses that contain the linear combinations of two coefficient parameters:

$$H_{0,4}: \beta_1 + \beta_2 = 0, \ v.s. \ H_{a,4}: \beta_1 + \beta_2 \neq 0.$$

$$H_{0,5}: \beta_3 + \beta_4 = 0, \ v.s. \ H_{a,5}: \beta_3 + \beta_4 \neq 0.$$

$$H_{0,6}: \beta_1 + \beta_p = 0.75, \ v.s. \ H_{a,6}: \beta_1 + \beta_p \neq 0.75.$$

$$H_{0,7}: \beta_2 + \beta_3 = -0.75, \ v.s. \ H_{a,7}: \beta_2 + \beta_3 \neq -0.75.$$

For the first three sets of hypotheses, we still set $h_j = 0, 0.1, 0.2, 0.4$ if the hypothesis involves β_j for $j = 2, 3, p$, and set $h_k = 0$ if the corresponding β_k is not involved in the

hypotheses. For the last hypothesis $H_{0,7}$, we set $h_2 = 0$, $h_p = 0$ and vary h_3 from 0 to 0.4. Tables 3 and 4 show that the Wald and score tests control the type I error at nominal level, and their powers improve when h_j increases.

5.2.3. Performance regarding m.—We further investigate how the testing performance changes as m changes. We consider three sets of hypotheses:

$$H_{0,8}: \sum_{j=1}^{4} \beta_j = 0, \ v.s. \ H_{a,8}: \sum_{j=1}^{4} \beta_j \neq 0.$$

$$H_{0,9}: \sum_{j=1}^{8} \beta_j = 0, \ v.s. \ H_{a,9}: \sum_{j=1}^{8} \beta_j \neq 0.$$

$$H_{0,10}: \sum_{j=1}^{12} \beta_j = 0, \ v.s. \ H_{a,10}: \sum_{j=1}^{12} \beta_j \neq 0,$$

corresponding to $m = 4$, 8 and 12. We set $h_2 = 0$, $h_p = 0$, and $h_3 = 0$, 0.2, 0.4, 0.8. The empirical sizes and powers are displayed in Table 5. These results suggest that under different m, the empirical sizes remain close to the nominal significance level for both the Wald and score tests. On the other hand, the empirical power decreases in general when m increases. For instance, as shown in Table 5, when $\mathbf{X}$ follows the multivariate normal distribution with mean zero and covariance $\Sigma = 0.5\mathbf{I}_p$, $p = 350$ and $h_3 = 0.8$, the powers of the Wald test are 1.000, 0.950 and 0.854 for $m = 4$, 8 and 12, respectively. This is intuitively sensible, and suggests that larger sample size is needed to reach a desired power when the hypothesis concerns more parameters.

5.2.4. Comparison with naive test.—We further compare the performances of our proposed tests with the naive Wald and score tests developed under the noise free framework. We consider the covariates $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,p})^{\mathrm{T}}$ generated from the multivariate normal distribution with mean zero and covariance matrix $0.7\mathbf{I}_p$. The noise $\mathbf{U}_i$ follows the multivariate normal distribution with mean zero and covariance matrix $0.3\mathbf{I}_p$. Other settings remain unchanged. We consider the hypotheses on a single element in $\boldsymbol{\beta}$: $H_{0,2}$, and the linear combinations of two coefficient parameters: $H_{0,5}$ and $H_{0,7}$ as described previously. We report the empirical sizes and powers of the Wald and score tests with/without noises for $p = 50$ in Table 6. It is clear that while the proposed tests achieve Type I errors reasonably close to the nominal level under different null hypotheses, the naive tests lead to precarious performance. For instance, the Type I errors of Wald and Score tests for $H_{0,5}$ are as large as 0.474 and 0.554, respectively. These Type I errors are far beyond the significance level. Because they cannot control the significance level, we do not recommend consider using them in practice.

5.3. Neuroimage application.

We apply our proposed testing procedures to study how the SUVRs from PET image data affect the MoCa score. We download the preprocessed ^{18}F-AV-1451 PET image features, and demographic and cognitive assessments from the ADNI database. The image features include ^{18}F-AV-1451 SUVRs and volumes of the cortical, sub-cortical regions, brainstem, ventricles and sub-divisions of corpus callosum. Furthermore, the demographic variables include gender and standardized age (divided by the standard deviation) at the image examining time. For each subject, we obtain his/her MoCa score within 14 days of his/her image examining time as the outcome, which ranges from 9 to 30. Furthermore, we remove the covariates with more than 100 missing values. We standardize the volumes of ROIs by subtracting the means and dividing by the standard deviations. We use the SUVR from inferior cerebellum as a reference and divide the rest of SUVRs by this reference as suggested in (Landau et al. 2016). Finally, we have $n = 196$ complete samples with $p = 218$ covariates in the analysis.

Since the neuroimage data are longitudinally collected, we estimate the covariance matrix of $\mathbf{U}$ using repeatedly measured image features, while assuming that age and gender are recorded precisely. More specifically, let $\widetilde{\mathbf{W}}_{ij}$ denote the observed image features at the jth examining time. We first perform the regression between $\widetilde{\mathbf{W}}_{ij}$ and age of the ith patient at the jth examining time, and obtain $\widetilde{\mathbf{U}}_{ij}$ as the residual of the regression. Then we obtain the estimator for the covariance matrix

$$\widetilde{\mathbf{\Omega}} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(\widetilde{\mathbf{U}}_{ij} - \overline{\mathbf{U}}_i\right)\left(\widetilde{\mathbf{U}}_{ij} - \overline{\mathbf{U}}_i\right)^{\mathrm{T}}}{\sum_{i=1}^{n} (n_i - 1)},$$

where n_i is the number of repeated measurements of $\widetilde{\mathbf{W}}_{ij}$, and $\overline{\mathbf{U}}_i = \sum_{j=1}^{n_i} \widetilde{\mathbf{U}}_{ij}/n_i$. Finally, because the first two covariates, age and gender, are measured precisely, the first two columns and rows of the estimated $\mathbf{\Omega}$, denoted by $\widehat{\mathbf{\Omega}}$, are zeros. We set the rest $(p-2) \times (p-2)$ sub-matrix of $\widehat{\mathbf{\Omega}}$ to be $\widetilde{\mathbf{\Omega}}$.

We test p hypotheses, each of the form

$$H_0 : \beta_j = 0 \quad \text{v.s.} \quad H_a : \beta_j \neq 0, \tag{19}$$

for $j = 1, \ldots, p$ at 0.05 nominal level. To implement the hypothesis testing procedure, in each test, we first fit a standard penalized Poisson regression model to obtain the initial values of the coefficient estimators. Then we construct the score test and Wald test statistics based on (11) and (10), respectively. The tuning parameter λ is selected by minimizing (16). We obtain the p-value as the probability of a $\chi^2(1)$ random variable that is greater than the resulting score and Wald test statistics. There are 33 and 69 covariate coefficients with significant p-values at 0.05 nominal level based on the score and Wald tests, respectively. Furthermore, we plot the boxplot of the resulting p-values in Figure 2. It is clear that the distribution of the p-values are similar for the score and the Wald test. For each covariate j, we obtain the estimated jth coefficient based on (8) under the corresponding alternative hypothesis, and plot the estimated coefficients of the SUVRs at the cortical regions on a

template brain in Figure 3. The results show that the SUVRs have negative effects on the cognitive score, suggesting that the higher the SUVR values, the lower the MoCa score and in turn the worse the cognitive function, which is consistent with the scientific evidences (Braak & Braak 1991, Schöll et al. 2016, Baker et al. 2017). Furthermore, the score test is more stringent and gives less number of significant SUVRs. Among 33 significant predictors from the score test, 27 of them are also significant in the Wald test. Based on this high agreement between the score and Wald tests, we believe the difference between the two tests is a small sample phenomenon.

To adjust for the multiple testing, we further performed an analysis to control false discovery rate (FDR) (Benjamini & Hochberg 1995) within 0.05 by treating the p-values as independent. Since the score test is too stringent, no significant covariate has been identified at 0.05 FDR by using the score test. Therefore, we only present the results from the Wald test. We plot the p-values versus $0.05j/218$ in Figure 4 in an increasing order, which suggests 36 covariates are selected as the important predictors. There are 13 cortical SUVRs among the 36 important predictors that are significant. We present their estimated coefficient, p-values from the Wald test in Table 7. The results show that the majority of the significant cortical SUVRs are in the temporal lobe, which consists of structures that are vital for declarative or long-term memory (Smith & Kosslyn 2008).

In addition, we perform a 5-fold cross validation and compare the prediction errors among the four methods: (a) We select the important predictors as those with p-value less than 0.05 in the test (19) based on the score statistics and then use formula $\exp\left(\widehat{\boldsymbol{\beta}}_s^{\mathrm{T}}\mathbf{W}_{Si} - \widehat{\boldsymbol{\beta}}_s^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}_s\widehat{\boldsymbol{\beta}}_s\right)$ to predict the outcome in the test sample, where $\mathbf{W}_{Si}$ is the selected covariates, $\widehat{\boldsymbol{\beta}}_s$ is estimator from (6) using selected covariates, $\widehat{\boldsymbol{\Omega}}_s$ is the subset of $\widehat{\boldsymbol{\Omega}}$ corresponding to the selected covariates. (b) We select the important predictors as those with p-value less than 0.05 in the test (19) based on the Wald statistics and then use formula $\exp\left(\widehat{\boldsymbol{\beta}}_w^{\mathrm{T}}\mathbf{W}_{Wi} - \widehat{\boldsymbol{\beta}}_w^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}_w\widehat{\boldsymbol{\beta}}_w\right)$ to predict the outcome in the test sample, where $\mathbf{W}_{Wi}$ is the selected covariates, $\widehat{\boldsymbol{\beta}}_w$ is estimator from (6) using selected covariates, $\widehat{\boldsymbol{\Omega}}_w$ is the subset of $\widehat{\boldsymbol{\Omega}}$ corresponding to the selected covariates. (c) We select the important predictors using the standard lasso regression between the logarithm of the MoCa score and all covariates and then use formula $\exp\left(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{W}_i\right)$ to predict the outcome, where $\widehat{\boldsymbol{\beta}}$ is the estimator from the lasso regression. (d) We select the important predictors using the penalized Poisson regression between the logarithm of the MoCap score and all covariates and then use formula $\exp\left(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{W}_i\right)$ to predict the outcome, where $\widehat{\boldsymbol{\beta}}$ is the estimator from the penalized Poisson regression. The penalty parameters in the lasso and penalized Poisson regression are selected using a sub-routine of 10-folder cross-validation. Method (d) breaks down because the algorithm does not converge for any selections of the penalty parameters. Therefore, in Figure 5, we show the distributions of the prediction errors, defined as $\sum_{i=1}^{n}\left|Y_i - \widehat{Y}_i\right|/|Y_i|$, only for the methods (a), (b) and (c) after 100 runs of the 5-fold cross-validation. The results shows that Method (a) and (b) have similar performance and both outperform Method (c) with much smaller prediction errors on average.

Finally, we perform the score and the Wald tests to test whether any SUVRs from any composite regions may have significant association with the MoCa score, where the composite regions, namely BRAAK12, BRAAK34, BRAAK56, are defined in (Braak & Braak 1991) and used in Landau et al. (2016) and Schöll et al. (2016). We provide the list of ROIs in each composite regions in Appendix B.5. Let β_{s_k} be the coefficients of the SUVRs from the ROIs that belong to the composite region k. We test the null hypothesis that $\beta_{s_k} = 0$. The results in Table 8 show that all the tests are significant, suggesting that at least one ROI in each of the composite region has significant association with the cognitive function. This result partially agrees with results in (Schöll et al. 2016) that the SUVRs from the composite regions are significantly different in healthy subjects and patients with a diagnosis of probable Alzheimer's disease.

6. Conclusion and discussion.

We have proposed an amenably penalized noise corrected Poisson model to study the relationship between the cognitive score and high dimensional noisy neuroimage data. Under the sparsity assumption, we established the parameter convergence rates in both l_1 and l_2 norms, the variable selection consistency property and the asymptotic normality of a subvector with possibly infinitely many components. Inference tools are subsequently developed. The neuroimage application shows that the inference tools generate scientifically meaningful results, which have potential to be used to study the cognitive function and cognitive changes for neurodegenerative diseases. Further research along this line is ongoing in our group. The neuroimage dataset and computational code are available at Jiang et al. (2021).

Thanks to an anonymous referee, we would like to point out one important extension. Instead of a constant matrix Ω, we can further allow Ω to depend on both the covariate $\mathbf{X}$ and the response Y, hence $\Omega(Y, \mathbf{X})$, and assume $E(\mathbf{U} | Y, \mathbf{X}) = \mathbf{0}$. This would include heteroscedastic measurement error and to allow dependent relation between $\mathbf{W}$ and Y given $\mathbf{X}$. All the estimation and inference results will still hold and the regularity conditions and proofs in the Suppement also do not need to be further modified to accomodate this extension.

Establishing similar results in generalized linear models beyond Poisson or general regression models with non-Gaussian noise turns out to be surprisingly difficult due to various technical obstacles. The main difficult lies in being unable to construct a loss function that is positive-definite at the true parameter value. In the case when an estimating equation is available, although one may be tempted to treat the l_2 norm square of the estimating equation as a loss function, we find other technical issues arise partially because the Hessian of the loss function may involve the response, hence some of the techniques used here cannot be directly applied. Likewise, extending the Poisson model to allow overdispersion also turns out challenging, regardless if we use a negative binomial model, or incorporate random effects, or use extra observed covariates. All these will lead to models different from Poisson. The biggest hurdle of considering general regression model and/or non-Gaussian noise is to rigorously establish that the loss function is locally convex. More investigation and dedicated effort are needed in this aspect.

The assumption that the covariance of the measurement is known is widely adopted in the low and dimensional noisy data literature (Stefanski 1989, Cook & Stefanski 1995, Loh & Wainwright 2012, Sørensen et al. 2015), because the parameter estimation in the noisy model with unknown noise covariance is a challenging, especially in high dimensional setting where the covariance is a high dimensional unknown parameter to be estimated. Thresholding techniques as those proposed in Bickel & Levina (2008), Cai & Liu (2011), Fan et al. (2011) can be used for the covariance estimation, but the theoretical properties of the resulting estimators are involved, requiring careful treatment of the additional error from the covariance estimator. In a relatively simple situation when the error variance can be estimated through estimating a parameter γ via solving $\mathbf{f}_{\gamma}(\gamma) = \mathbf{0}$, then writing $\mathscr{L}(\beta)$ as $\mathscr{L}(\beta, \gamma)$, we can accomodate the additional parameter by concatenating β with γ and carrying out the subsequent analysis. For example, in this case the result in Theorem 4 will be updated to

$$
\begin{pmatrix} \widehat{\beta}_{a,\mathcal{M} \cup S} - \beta_{t,\mathcal{M} \cup S} \\ \widehat{\gamma} - \gamma \end{pmatrix} = - \begin{bmatrix} \mathbf{Q}_{\mathcal{M} \cup S, \mathcal{M} \cup S}(\beta_t, \gamma) & \partial^2 \mathscr{L}(\beta, \gamma)/\partial \beta_{\mathcal{M} \cup S} \partial \gamma^{\mathrm{T}} \\ \partial \mathbf{f}_{\gamma}(\gamma)/\partial \beta_{\mathcal{M} \cup S}^{\mathrm{T}} & \partial \mathbf{f}_{\gamma}(\gamma)/\partial \gamma^{\mathrm{T}} \end{bmatrix}^{-1} \times \begin{bmatrix} \left\{ \dfrac{\partial \mathscr{L}(\beta_t, \gamma)}{\partial \beta} \right\}_{\mathcal{M} \cup S} \\ \mathbf{f}_{\gamma}(\gamma) \end{bmatrix} \{1 + o_p(1)\}.
$$

Letting $\mathbf{M} \equiv \mathbf{Q}_{\mathcal{M} \cup S, \mathcal{M} \cup S}(\beta_t) - \left\{ \partial^2 \mathscr{L}(\beta_t, \gamma)/\partial \beta_{\mathcal{M} \cup S} \partial \gamma^{\mathrm{T}} \right\} \left\{ \partial \mathbf{f}_{\gamma}(\gamma)/\partial \gamma^{\mathrm{T}} \right\}^{-1} \left\{ \partial \mathbf{f}_{\gamma}(\gamma)/\partial \beta_{\mathcal{M} \cup S}^{\mathrm{T}} \right\}$, then this leads to

$$
\widehat{\beta}_{a,\mathcal{M} \cup S} - \beta_{t,\mathcal{M} \cup S} = - \mathbf{M}^{-1} \left[\left\{ \dfrac{\partial \mathscr{L}(\beta_t, \gamma)}{\partial \beta} \right\}_{\mathcal{M} \cup S} + \left\{ \partial^2 \mathscr{L}(\beta_t, \gamma)/\partial \beta_{\mathcal{M} \cup S} \partial \gamma^{\mathrm{T}} \right\} \left\{ \partial \mathbf{f}_{\gamma}(\gamma)/\partial \gamma^{\mathrm{T}} \right\}^{-1} \times \mathbf{f}_{\gamma}(\gamma) \right] \{1 + o_p(1)\}.
$$

We further conduct simulations to evaluate the proposed adjustment in Section B.6 of the supplementary document. The results suggest that the proposed adjustment controls type I error rate when Ω contain small number of unknown parameters. Estimation and testing when Ω has a large number of unknown parameters are challenging problems and deserve much more extensive investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

REFERENCES

Baker SL, Lockhart SN, Price JC, He M, Huesman RH, Schonhaut D, Faria J, Rabinovici G & Jagust WJ (2017), 'Reference tissue–based kinetic evaluation of 18f-av-1451 for tau imaging', Journal of Nuclear Medicine 58(2), 332–338. [PubMed: 27587706]

Belloni A, Chernozhukov V & Kaul A (2017), 'Confidence bands for coefficients in high dimensional linear models with error-in-variables', arXiv preprint arXiv:1703.00469.

Belloni A & Rosenbaum M (2016), 'An {l_1, l_2, l_∞}-regularization approach to high-dimensional errors-in-variables models', Electronic Journal of Statistics 10, 1935–7524.

Belloni A, Rosenbaum M & Tsybakov AB (2017), 'Linear and conic programming estimators in high dimensional errors-in-variables models', Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79(3), 939–956.

Benjamini Y & Hochberg Y (1995), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', Journal of the Royal statistical society: series B (Methodological) 57(1), 289–300.

Bentkus V (2005), 'A lyapunov-type bound in r^d', Theory of Probability & Its Applications 49(2), 311–323.

Bickel PJ & Levina E (2008), 'Covariance regularization by thresholding', The Annals of Statistics 36(6), 2577–2604.

Braak H & Braak E (1991), 'Neuropathological stageing of alzheimer-related changes', Acta neuropathologica 82(4), 239–259. [PubMed: 1759558]

Brown B, Weaver T & Wolfson J (2019), 'Meboost: Variable selection in the presence of measurement error', Statistics in medicine 38(15), 2705–2718. [PubMed: 30856279]

Cai T & Liu W (2011), 'Adaptive thresholding for sparse covariance matrix estimation', Journal of the American Statistical Association 106(494), 672–684.

Carroll RJ, Ruppert D, Stefanski LA & Crainiceanu CM (2006), Measurement error in nonlinear models: a modern perspective, Chapman and Hall/CRC, New Tork.

Chen L & Gu Y (2014), 'The convergence guarantees of a non-convex approach for sparse recovery', IEEE Transactions on Signal Processing 62(15), 3754–3767.

Cook J & Stefanski LA (1995), 'A simulation extrapolation method for parametric measurement error models', Journal of the American Statistical Association 89, 1314–1328.

Datta A & Zou H (2017), 'Cocolasso for high-dimensional error-in-variables regression', The Annals of Statistics 45(6), 2400–2426.

Duchi J, Shalev-Shwartz S, Singer Y & Chandra T (2008), Efficient projections onto the l 1-ball for learning in high dimensions, in 'Proceedings of the 25th international conference on Machine learning', ACM, pp. 272–279.

Fallah N, Mitnitski A & Rockwood K (2011), 'Applying neural network poisson regression to predict cognitive score changes', Journal of Applied Statistics 38(9), 2051–2062.

Fan J & Li R (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', Journal of the American statistical Association 96(456), 1348–1360.

Fan J, Liao Y & Mincheva M (2011), 'High dimensional covariance matrix estimation in approximate factor models', Annals of statistics 39(6), 3320. [PubMed: 22661790]

Fletcher R & Watson GA (1980), 'First and second order conditions for a class of nondifferentiable optimization problems', Mathematical Programming 18(1), 291–307.

Jiang F & Ma Y (2021), 'Poisson regression with error corrupted high dimensional features', Statistica Sinica. In press.

Jiang F, Zhou Y & Ma Y (2021), 'Code and data for high dimensional poisson models with noisy data: hypothesis testing for nonlinear nonconvex optimization'. https://github.com/feigroup/high-dimenisional-inference-for-count-data-with-errors.

Katz MJ, Wang C, Nester CO, Derby CA, Zimmerman ME, Lipton RB, Sliwinski MJ & Rabin LA (2021), 'T-moca: A valid phone screen for cognitive impairment in diverse community samples', Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 13(1), e12144.

Landau S, Ward TJ, Murphy A & Jagust W (2016), 'Flortaucipir (av-1451) processing methods'. URL: https://adni.bitbucket.io/reference/docs/UCBERKELEYAV1451/UCBERKELEY_AV1451_Methods_Aug2018.pdf

Leisman G, Moustafa AA & Shafir T (2016), 'Thinking, walking, talking: integratory motor and cognitive brain function', Frontiers in public health 4, 94. [PubMed: 27252937]

Li M, Li R & Ma Y (2021), 'Inference in high dimensional linear measurement error models', Journal of Multivariate Analysis in press.

Loh P-L & Wainwright MJ (2012), 'High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity', The Annals of Statistics pp. 1637–1664.

Loh P-L & Wainwright MJ (2015), 'Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima', Journal of Machine Learning Research 16, 559–616.

Loh P-L & Wainwright MJ (2017), 'Support recovery without incoherence: A case for nonconvex regularization', The Annals of Statistics 45(6), 2455–2482.

McCullagh P & Nelder JA (2019), Generalized linear models, Routledge.

Mitnitski AB, Fallah N, Dean CB & Rockwood K (2014), 'A multi-state model for the analysis of changes in cognitive scores over a fixed time interval', Statistical methods in medical research 23(3), 244–256. [PubMed: 21937474]

Ning Y & Liu H (2017), 'A general theory of hypothesis tests and confidence regions for sparse high dimensional models', The Annals of Statistics 45, 158–195.

Schöll M, Lockhart SN, Schonhaut DR, O'Neil JP, Janabi M, Ossenkoppele R, Baker SL, Vogel JW, Faria J, Schwimmer HD et al. (2016), 'Pet imaging of tau deposition in the aging human brain', Neuron 89(5), 971–982. [PubMed: 26938442]

Sentürk D & Müller H-G (2005), 'Covariate-adjusted regression', Biometrika 92(1), 75–89.

Shi C, Song R, Chen Z, Li R et al. (2019), 'Linear hypothesis testing for high dimensional generalized linear models', The Annals of statistics 47(5), 2671–2703. [PubMed: 31534282]

Smith E & Kosslyn S (2008), Cognitive Psychology: Mind and Brain, Pearson Education, Pearson Prentice Hall. URL: https://books.google.com/books?id=YIPSNwAACAAJ

Sørensen Ø, Frigessi A & Thoresen M (2015), 'Measurement error in lasso: Impact and likelihood bias correction', Statistica sinica pp. 809–829.

Sørensen Ø, Hellton KH, Frigessi A & Thoresen M (2018), 'Covariate selection in high-dimensional generalized linear models with measurement error', Journal of Computational and Graphical Statistics 27(4), 739–749.

Stefanski LA (1989), 'Unbiased estimation of a nonlinear function a normal mean with application to measurement err oorf models', Communications in Statistics-Theory and Methods 18(12), 4335–4358.

Van de Geer S, Bühlmann P, Ritov Y & Dezeure R (2014), 'On asymptotically optimal confidence regions and tests for high-dimensional models', Annals of Statistics 42(3), 1166–1202.

Vial J-P (1982), 'Strong convexity of sets and functions', Journal of Mathematical Economics 9(1–2), 187–205.

Wainwright MJ (2009), 'Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (lasso)', IEEE transactions on information theory 55(5), 2183–2202.

Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, Harvey D, Jack CR Jr, Jagust W, Morris JC et al. (2017), 'The alzheimer's disease neuroimaging initiative 3: Continued innovation for clinical trial improvement', Alzheimer's & Dementia 13(5), 561–571.

Zhang C-H, Zhang T et al. (2012), 'A general theory of concave regularization for high-dimensional sparse estimation problems', Statistical Science 27(4), 576–593.

Zhang X & Cheng G (2017), 'Simultaneous inference for high-dimensional linear models', Journal of the American Statistical Association 112(518), 757–768.

Fig 1.

Left: The density of the residuals from lasso regression. The lasso regression does not provide a satisfactory fit for the data. Middle and Right: The logarithm of the p values from the Wald and score tests proposed by Shi et al. (2019) for testing whether the SUVR from each cortical regions is significant predictor for the cognitive score. The Wald and score tests suggest that the SUVRs at all the cortical regions have significant association with the cognitive score.

Distributions p-value
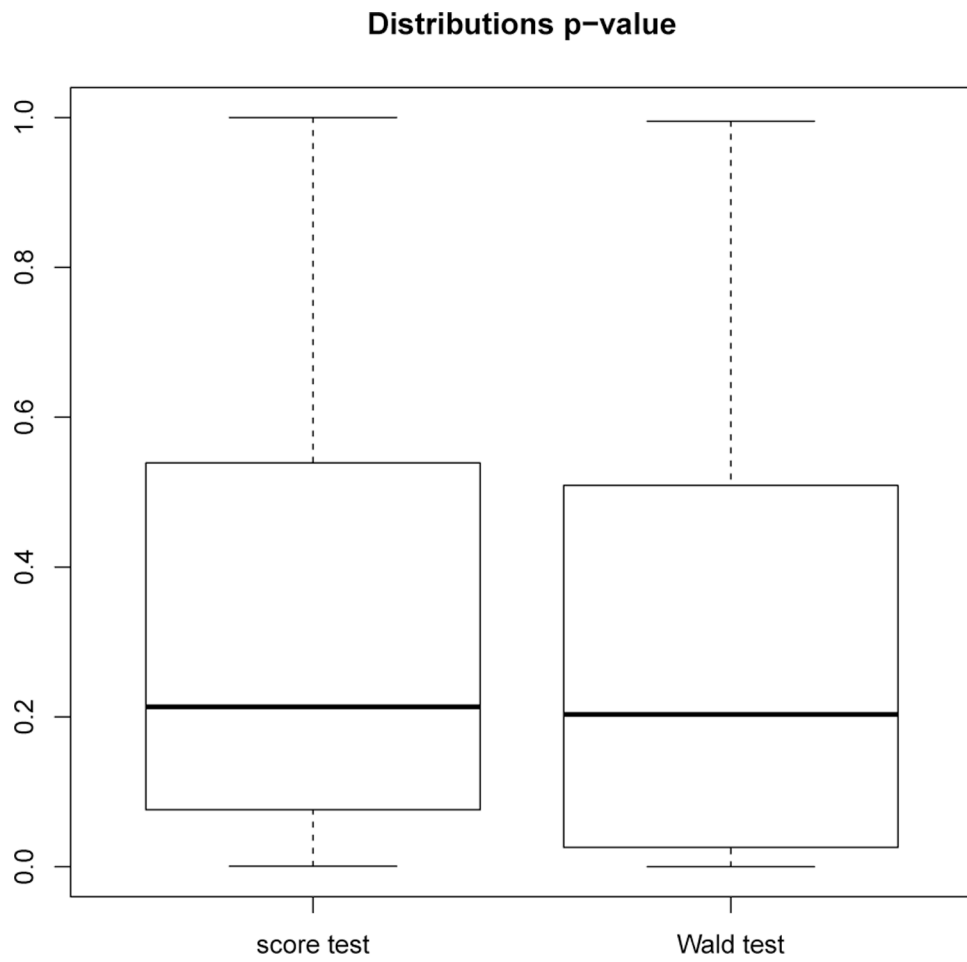
Fig 2.
The boxplot of the p-values based on the score and Wald tests. The distributions of the p-values are similar from both methods.
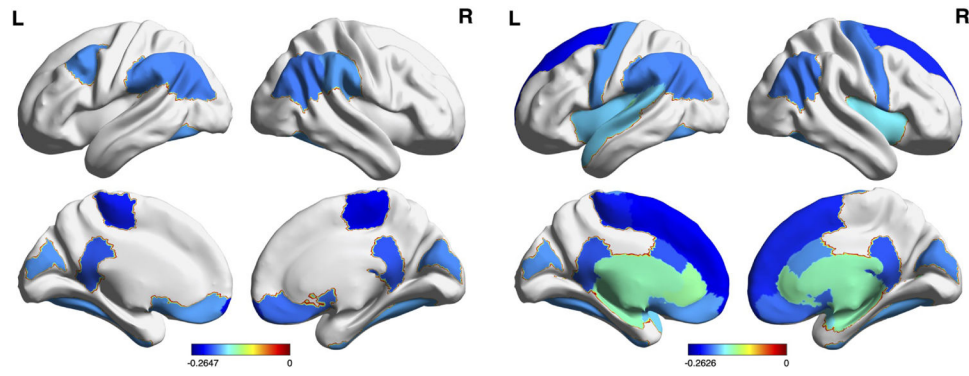
Fig 3.
The effects of SUVRs at the cortical regions. The colors represent the values (indicated by the color bars) of the estimated coefficients of the SUVRs. We only plot the coefficient values corresponding to the significant brain regions with p-value less than 0.05 from score test (left) and Wald test (right). The white areas are the non-significant brain regions. The L and R letters in the plot represent the left and right hemispheres.
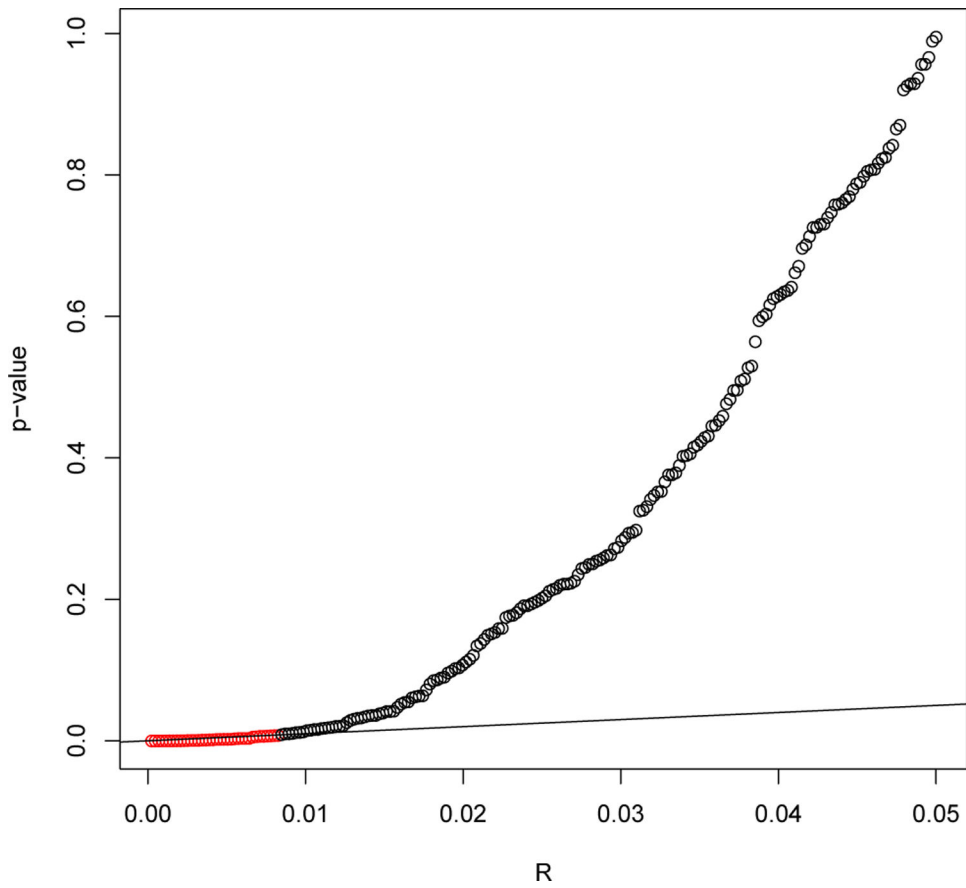
Fig 4.
Sorted p-value versus R = 0.05j/218, j = 1, …, 218. There are 36 important predictors corresponding to the p-values (in red) that below the line.
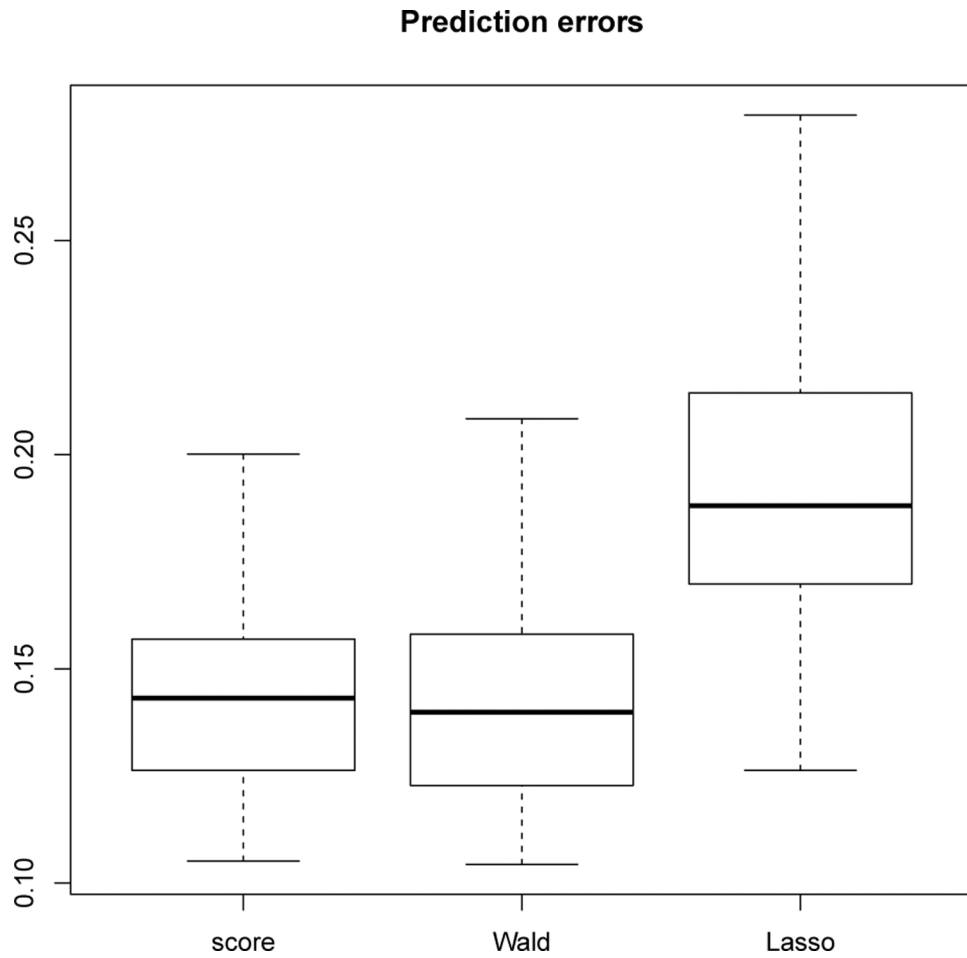
Prediction errors

Fig 5.
The distribution of the prediction errors from 100 runs of the 5-fold cross-validation based on Methods (a), (b) and (c). Method (d) breaks down because the algorithm does not converge.

Table 1

The empirical sizes and powers of Wald and score tests for univariate parameter testing with n = 300.

|		X ~ Normal				X ~ Uniform			
		$\Sigma=0.5I_p$		$\Sigma=0.5^{|i-j|+1}$		$\Sigma=0.5I_p$		$\Sigma=0.5^{|i-j|+1}$	
		T_W	T_S	T_W	T_S	T_W	T_S	T_W	T_S
	β_2	$H_{0,1}:\beta_2=-0.75$, v.s. $H_{a,1}:\beta_2\neq-0.75$							
	−0.75	0.068	0.054	0.056	0.050	0.054	0.044	0.066	0.062
	−0.65	0.352	0.288	0.222	0.176	0.292	0.276	0.232	0.208
	−0.55	0.826	0.778	0.680	0.592	0.736	0.726	0.632	0.598
	−0.35	1.000	0.996	0.996	0.976	1.000	1.000	0.990	0.972
	β_3	$H_{0,2}:\beta_3=0$, v.s. $H_{a,2}:\beta_3\neq0$							
	0.0	0.056	0.046	0.058	0.038	0.056	0.046	0.068	0.060
$p=50$	0.1	0.302	0.272	0.204	0.172	0.250	0.240	0.214	0.182
	0.2	0.752	0.724	0.554	0.524	0.692	0.682	0.530	0.508
	0.4	0.996	0.996	0.984	0.960	1.000	1.000	0.976	0.942
	β_p	$H_{0,3}:\beta_p=0$, v.s. $H_{a,3}:\beta_p\neq0$							
	0.0	0.060	0.044	0.056	0.042	0.062	0.054	0.052	0.050
	0.1	0.246	0.222	0.234	0.210	0.276	0.258	0.240	0.226
	0.2	0.708	0.672	0.666	0.634	0.714	0.698	0.708	0.682
	0.4	0.998	0.998	0.998	0.996	0.998	0.998	0.998	0.994
	β_2	$H_{0,1}:\beta_2=-0.75$, v.s. $H_{a,1}:\beta_2\neq-0.75$							
	−0.75	0.050	0.036	0.054	0.034	0.064	0.068	0.066	0.056
	−0.65	0.312	0.322	0.260	0.242	0.268	0.266	0.230	0.202
	−0.55	0.750	0.766	0.650	0.644	0.752	0.750	0.612	0.598
	−0.35	0.998	0.998	0.980	0.878	0.998	0.998	0.978	0.892
	β_3	$H_{0,3}:\beta_3=0$, v.s. $H_{a,2}:\beta_3\neq0$							
	0.0	0.064	0.048	0.066	0.066	0.066	0.058	0.068	0.054
$p=350$	0.1	0.328	0.330	0.224	0.200	0.270	0.262	0.198	0.164
	0.2	0.770	0.770	0.590	0.546	0.708	0.706	0.568	0.504
	0.4	1.000	1.000	0.950	0.846	1.000	1.000	0.942	0.830
	β_p	$H_{0,3}:\beta_p=0$, v.s. $H_{a,3}:\beta_p\neq0$							
	0.0	0.072	0.066	0.050	0.046	0.058	0.050	0.066	0.056
	0.1	0.346	0.342	0.208	0.198	0.250	0.250	0.220	0.206
	0.2	0.736	0.742	0.662	0.646	0.782	0.768	0.654	0.638
	0.4	1.000	1.000	0.996	0.994	1.000	1.000	0.994	0.992

Table 2

The empirical sizes and powers of Wald and score tests for univariate parameter testing with n = 500.

		X ~ Normal				X ~ Uniform			
		$= 0.5\mathbf{I}p$		$= 0.5^{\|i-j\|+1}$		$= 0.5\mathbf{I}p$		$= 0.5^{\|i-j\|+1}$	
		T_W	T_S	T_W	T_S	T_W	T_S	T_W	T_S
	β_2	$H_{0,1}: \beta_2 = -0.75$, v.s. $H_{a,1}: \beta_2 \neq -0.75$							
	−0.75	0.066	0.054	0.044	0.040	0.064	0.060	0.060	0.056
	−0.65	0.488	0.450	0.346	0.316	0.442	0.422	0.324	0.304
	−0.55	0.954	0.950	0.864	0.838	0.910	0.906	0.800	0.792
	−0.35	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	β_3	$H_{0,2}: \beta_3 = 0$, v.s. $H_{a,2}: \beta_3 \neq 0$							
	0.0	0.048	0.046	0.066	0.058	0.052	0.050	0.056	0.056
p = 50	0.1	0.402	0.382	0.324	0.300	0.390	0.386	0.302	0.296
	0.2	0.890	0.888	0.780	0.770	0.892	0.884	0.778	0.766
	0.4	1.000	1.000	1.000	0.998	1.000	1.000	1.000	1.000
	β_p	$H_{0,3}: \beta_p = 0$, v.s. $H_{a,3}: \beta_p \neq 0$							
	0.0	0.066	0.060	0.064	0.058	0.050	0.048	0.050	0.048
	0.1	0.400	0.368	0.350	0.338	0.390	0.380	0.336	0.320
	0.2	0.922	0.914	0.896	0.878	0.892	0.890	0.884	0.878
	0.4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	β_2	$H_{0,1}: \beta_2 = -0.75$, v.s. $H_{a,1}: \beta_2 \neq -0.75$							
	−0.75	0.052	0.056	0.046	0.046	0.062	0.056	0.040	0.038
	−0.65	0.458	0.478	0.328	0.330	0.390	0.392	0.396	0.402
	−0.55	0.920	0.930	0.864	0.868	0.926	0.928	0.902	0.902
	−0.35	0.842	0.918	0.988	0.990	1.000	1.000	1.000	1.000
	β_3	$H_{0,2}: \beta_3 = 0$, v.s. $H_{a,2}: \beta_3 \neq 0$							
	0.0	0.076	0.066	0.062	0.058	0.070	0.066	0.066	0.056
p = 600	0.1	0.454	0.452	0.344	0.342	0.392	0.392	0.454	0.454
	0.2	0.904	0.904	0.832	0.822	0.902	0.894	0.870	0.862
	0.4	0.974	0.974	0.998	0.980	1.000	1.000	0.998	0.894
	β_p	$H_{0,3}: \beta_p = 0$, v.s. $H_{a,3}: \beta_p \neq 0$							
	0.0	0.048	0.046	0.050	0.046	0.060	0.058	0.058	0.054
	0.1	0.444	0.444	0.404	0.390	0.414	0.416	0.450	0.448
	0.2	0.930	0.928	0.870	0.860	0.912	0.912	0.876	0.874
	0.4	0.980	0.980	1.000	1.000	1.000	1.000	1.000	1.000

Table 3

The empirical size and power of Wald and score tests for linear hypothesis testing with n = 300.

| | | X ~ Normal | | | | X ~ Uniform | | | |
| | | $\Sigma = 0.5\mathbf{I}p$ | | $\Sigma = 0.5^{\|i-j\|+1}$ | | $\Sigma = 0.5\mathbf{I}p$ | | $\Sigma = 0.5^{\|i-j\|+1}$ | |
		T_W	T_S	T_W	T_S	T_W	T_S	T_W	T_S
	$\beta_1 + \beta_2$	$H_{0,4} : \beta_1 + \beta_2 = 0,\ v.s.\ H_{a,4} : \beta_1 + \beta_2 \neq 0$							
	0.0	0.056	0.048	0.040	0.042	0.056	0.050	0.052	0.042
	0.1	0.154	0.132	0.174	0.162	0.138	0.136	0.208	0.184
	0.2	0.420	0.386	0.574	0.532	0.374	0.354	0.574	0.530
	0.4	0.930	0.906	0.990	0.984	0.908	0.902	0.994	0.990
	$\beta_3 + \beta_4$	$H_{0,5} : \beta_3 + \beta_4 = 0,\ v.s.\ H_{a,5} : \beta_3 + \beta_4 \neq 0.$							
	0.0	0.058	0.050	0.066	0.056	0.056	0.050	0.068	0.066
	0.1	0.130	0.116	0.154	0.138	0.152	0.136	0.190	0.174
	0.2	0.450	0.428	0.470	0.450	0.412	0.388	0.450	0.424
	0.4	0.930	0.920	0.946	0.930	0.932	0.920	0.958	0.932
$p = 50$	$\beta_1 + \beta_p$	$H_{0,6} : \beta_1 + \beta_p = 0.75,\ v.s.\ H_{a,6} : \beta_1 + \beta_p \neq 0.75$							
	0.75	0.050	0.044	0.060	0.046	0.058	0.048	0.044	0.036
	0.85	0.172	0.146	0.138	0.116	0.162	0.154	0.154	0.130
	0.95	0.490	0.444	0.408	0.354	0.462	0.436	0.418	0.376
	1.15	0.970	0.950	0.936	0.916	0.970	0.958	0.942	0.914
	$\beta_2 + \beta_3$	$H_{0,7} : \beta_2 + \beta_3 = -0.75,\ v.s.\ H_{a,7} : \beta_2 + \beta_3 \neq -0.75$							
	−0.75	0.056	0.050	0.054	0.044	0.060	0.044	0.062	0.062
	−0.65	0.200	0.176	0.182	0.172	0.148	0.138	0.180	0.174
	−0.55	0.484	0.444	0.516	0.486	0.422	0.408	0.468	0.466
	−0.35	0.922	0.910	0.966	0.962	0.920	0.910	0.960	0.962
	$\beta_1 + \beta_2$	$H_{0,4} : \beta_1 + \beta_2 = 0,\ v.s.\ H_{a,4} : \beta_1 + \beta_2 \neq 0$							
	0.0	0.062	0.056	0.062	0.056	0.050	0.046	0.048	0.046
	0.1	0.164	0.160	0.216	0.202	0.106	0.096	0.206	0.184
	0.2	0.472	0.438	0.612	0.572	0.402	0.378	0.536	0.510
	0.4	0.940	0.934	0.988	0.988	0.910	0.900	0.982	0.980
	$\beta_3 + \beta_4$	$H_{0,5} : \beta_3 + \beta_4 = 0,\ v.s.\ H_{a,5} : \beta_3 + \beta_4 \neq 0.$							
	0.0	0.058	0.046	0.070	0.040	0.038	0.040	0.068	0.048
$p = 350$	0.1	0.192	0.188	0.174	0.138	0.126	0.124	0.172	0.136
	0.2	0.454	0.442	0.462	0.410	0.392	0.378	0.404	0.356
	0.4	0.952	0.952	0.912	0.814	0.944	0.942	0.916	0.828
	$\beta_1 + \beta_p$	$H_{0,6} : \beta_1 + \beta_p = 0.75,\ v.s.\ H_{a,6} : \beta_1 + \beta_p \neq 0.75$							
	0.75	0.046	0.044	0.058	0.038	0.056	0.042	0.060	0.054
	0.85	0.148	0.142	0.280	0.292	0.110	0.110	0.240	0.242
	0.95	0.466	0.472	0.562	0.566	0.394	0.390	0.496	0.512
	1.15	0.942	0.944	0.960	0.940	0.938	0.940	0.954	0.932

		X ~ Normal				X ~ Uniform			
		$= 0.5\mathbf{I}_p$		$= 0.5^{|i-j|+1}$		$= 0.5\mathbf{I}_p$		$= 0.5^{|i-j|+1}$	
		T_W	T_S	T_W	T_S	T_W	T_S	T_W	T_S
$\beta_2 + \beta_3$		$H_{0,7} : \beta_2 + \beta_3 = -0.75,$ *v.s.* $H_{a,7} : \beta_2 + \beta_3$ 0.75							
−0.75		0.052	0.038	0.052	0.046	0.052	0.038	0.062	0.038
−0.65		0.154	0.138	0.174	0.138	0.134	0.120	0.142	0.130
−0.55		0.450	0.420	0.478	0.442	0.400	0.374	0.488	0.456
−0.35		0.932	0.916	0.968	0.966	0.914	0.912	0.960	0.960

Table 4

The empirical size and power of Wald and score tests for linear hypothesis testing with n = 500.

| | | X ~ Normal | | | | X ~ Uniform | | | |
| | | $= 0.5\mathbf{I}p$ | | $= 0.5^{|i-j|+1}$ | | $= 0.5\mathbf{I}p$ | | $= 0.5^{|i-j|+1}$ | |
		T_W	T_S	T_W	T_S	T_W	T_S	T_W	T_S
	$\beta_1 + \beta_2$	$H_{0,4} : \beta_1 + \beta_2 = 0$, v.s. $H_{a,4} : \beta_1 + \beta_2 \neq 0$							
	0.0	0.058	0.052	0.048	0.046	0.062	0.056	0.046	0.044
	0.1	0.240	0.220	0.286	0.280	0.202	0.196	0.296	0.274
	0.2	0.670	0.656	0.810	0.794	0.586	0.580	0.758	0.754
	0.4	0.996	0.996	1.000	1.000	0.988	0.988	0.998	0.998
	$\beta_3 + \beta_4$	$H_{0,5} : \beta_3 + \beta_4 = 0$, v.s. $H_{a,5} : \beta_3 + \beta_4 \neq 0.$							
	0.0	0.046	0.040	0.064	0.058	0.060	0.048	0.060	0.062
	0.1	0.244	0.216	0.254	0.244	0.242	0.238	0.260	0.244
	0.2	0.646	0.616	0.696	0.692	0.666	0.648	0.686	0.678
	0.4	0.996	0.996	0.998	0.996	0.996	0.998	0.998	0.998
$p = 50$	$\beta_1 + \beta_p$	$H_{0,6} : \beta_1 + \beta_p = 0.75$, v.s. $H_{a,6} : \beta_1 + \beta_p \neq 0.75$							
	0.75	0.068	0.062	0.058	0.056	0.050	0.046	0.054	0.048
	0.85	0.242	0.222	0.190	0.162	0.226	0.212	0.176	0.160
	0.95	0.676	0.640	0.620	0.580	0.646	0.626	0.646	0.614
	1.15	0.996	0.994	0.986	0.980	0.996	0.994	0.998	0.990
	$\beta_2 + \beta_3$	$H_{0,7} : \beta_2 + \beta_3 = -0.75$, v.s. $H_{a,7} : \beta_2 + \beta_3 \neq -0.75$							
	−0.75	0.060	0.058	0.056	0.044	0.056	0.052	0.050	0.048
	−0.65	0.238	0.234	0.302	0.302	0.214	0.206	0.226	0.218
	−0.55	0.640	0.622	0.730	0.720	0.608	0.598	0.664	0.660
	−0.35	0.992	0.992	1.000	1.000	0.994	0.992	0.998	0.998
	$\beta_1 + \beta_2$	$H_{0,4} : \beta_1 + \beta_2 = 0$, v.s. $H_{a,4} : \beta_1 + \beta_2 \neq 0$							
	0.0	0.054	0.044	0.056	0.050	0.046	0.042	0.046	0.042
	0.1	0.192	0.180	0.286	0.268	0.190	0.182	0.292	0.288
	0.2	0.602	0.594	0.790	0.786	0.578	0.558	0.824	0.816
	0.4	0.688	0.700	0.998	1.000	0.986	0.984	1.000	1.000
	$\beta_3 + \beta_4$	$H_{0,5} : \beta_3 + \beta_4 = 0$, v.s. $H_{a,5} : \beta_3 + \beta_4 \neq 0.$							
	0.0	0.072	0.064	0.066	0.060	0.044	0.046	0.060	0.056
$p = 600$	0.1	0.222	0.218	0.244	0.238	0.206	0.194	0.264	0.264
	0.2	0.654	0.650	0.678	0.670	0.614	0.608	0.712	0.714
	0.4	0.974	0.974	0.994	0.980	0.992	0.990	0.994	0.978
	$\beta_1 + \beta_p$	$H_{0,6} : \beta_1 + \beta_p = 0.75$, v.s. $H_{a,6} : \beta_1 + \beta_p \neq 0.75$							
	0.75	0.044	0.040	0.056	0.048	0.058	0.056	0.052	0.050
	0.85	0.238	0.242	0.316	0.316	0.196	0.200	0.386	0.390
	0.95	0.670	0.672	0.716	0.722	0.604	0.616	0.790	0.788
	1.15	0.986	0.994	1.000	1.000	0.994	0.994	0.9996	0.994

		X ~ Normal				X ~ Uniform			
		$= 0.5 \mathbf{I}p$		$= 0.5^{\|i-j\|+1}$		$= 0.5 \mathbf{I}p$		$= 0.5^{\|i-j\|+1}$	
		T_W	T_S	T_W	T_S	T_W	T_S	T_W	T_S
$\beta_2 + \beta_3$		$H_{0,7} : \beta_2 + \beta_3 = -0.75$, *v.s.* $H_{a,7} : \beta_2 + \beta_3$ -0.75							
−0.75		0.058	0.052	0.052	0.044	0.056	0.052	0.044	0.046
−0.65		0.222	0.200	0.230	0.208	0.162	0.146	0.264	0.246
−0.55		0.630	0.608	0.692	0.680	0.544	0.516	0.686	0.678
−0.35		0.974	0.990	1.000	1.000	0.994	0.992	1.000	1.000

Table 5

The empirical size and power of Wald and score tests under different m.

		X ~ Normal				X ~ Uniform			
		$\Sigma = 0.5\mathbf{I}p$		$\Sigma = 0.5^{\lvert i-j\rvert+1}$		$\Sigma = 0.5\mathbf{I}p$		$\Sigma = 0.5^{\lvert i-j\rvert+1}$	
		T_W	T_S	T_W	T_S	T_W	T_S	T_W	T_S
	$\sum_{j=1}^{4}\beta_j$	$H_{0,8}: \sum_{j=1}^{4}\beta_j = 0,\ v.s.\ H_{a,8}: \sum_{j=1}^{4}\beta_j \neq 0$							
	0.0	0.062	0.052	0.046	0.038	0.062	0.060	0.060	0.052
	0.2	0.238	0.220	0.434	0.406	0.244	0.228	0.404	0.386
	0.4	0.754	0.724	0.914	0.904	0.660	0.646	0.914	0.900
	0.8	1.000	1.000	1.000	1.000	0.998	0.998	1.000	1.000
	$\sum_{j=1}^{8}\beta_j$	$H_{0,9}: \sum_{j=1}^{8}\beta_j = 0,\ v.s.\ H_{a,9}: \sum_{j=1}^{8}\beta_j \neq 0$							
	0.0	0.054	0.052	0.062	0.058	0.052	0.054	0.062	0.052
$p=50$	0.2	0.162	0.142	0.316	0.308	0.178	0.172	0.296	0.282
	0.4	0.410	0.388	0.764	0.732	0.424	0.406	0.764	0.742
	0.8	0.966	0.952	1.000	1.000	0.920	0.906	1.000	1.000
	$\sum_{j=1}^{12}\beta_j$	$H_{0,10}: \sum_{j=1}^{12}\beta_j = 0,\ v.s.\ H_{a,10}: \sum_{j=1}^{12}\beta_j \neq 0$							
	0.0	0.046	0.044	0.062	0.062	0.046	0.052	0.052	0.050
	0.2	0.116	0.124	0.250	0.240	0.084	0.096	0.210	0.210
	0.4	0.288	0.298	0.610	0.604	0.318	0.330	0.642	0.642
	0.8	0.854	0.796	0.996	0.994	0.802	0.756	0.994	0.992
	$\sum_{j=1}^{4}\beta_j$	$H_{0,8}: \sum_{j=1}^{4}\beta_j = 0\ v.s.\ H_{a,8}: \sum_{j=1}^{4}\beta_j \neq 0$							
	0.0	0.062	0.052	0.062	0.052	0.060	0.054	0.036	0.036
	0.2	0.260	0.244	0.420	0.400	0.230	0.222	0.402	0.382
	0.4	0.718	0.684	0.918	0.916	0.710	0.672	0.924	0.912
	0.8	1.000	0.998	1.000	1.000	1.000	1.000	1.000	1.000
	$\sum_{j=1}^{8}\beta_j$	$H_{0,9}: \sum_{j=1}^{8}\beta_j = 0,\ v.s.\ H_{a,9}: \sum_{j=1}^{8}\beta_j \neq 0$							
	0.0	0.062	0.060	0.068	0.060	0.066	0.058	0.042	0.040
$p=350$	0.2	0.132	0.130	0.266	0.258	0.180	0.166	0.262	0.238
	0.4	0.452	0.420	0.770	0.746	0.380	0.360	0.764	0.752
	0.8	0.950	0.936	1.000	1.000	0.936	0.926	1.000	1.000
	$\sum_{j=1}^{12}\beta_j$	$H_{0,10}: \sum_{j=1}^{12}\beta_j = 0,\ v.s.\ H_{a,10}: \sum_{j=1}^{12}\beta_j \neq 0$							
	0.0	0.056	0.056	0.054	0.054	0.052	0.050	0.038	0.038
	0.2	0.098	0.106	0.170	0.174	0.100	0.100	0.178	0.172
	0.4	0.296	0.296	0.616	0.604	0.276	0.274	0.584	0.574
	0.8	0.854	0.792	0.998	0.996	0.838	0.818	0.992	0.990

Table 6

The empirical size and power of Wald and score tests with/without noise considered.

	With noise		Without noise		With noise		Without noise		With noise		Without noise	
	T_W	T_S	T_W	T_S	T_W	T_S	T_W	T_S	T_W	T_S	T_W	T_S
β_3	$H_{0,2} : \beta_3 = 0$				$H_{0,5} : \beta_3 + \beta_4 = 0$				$H_{0,7} : \beta_2 + \beta_3 = -0.75$			
0.0	0.084	0.078	0.774	0.824	0.064	0.066	0.474	0.554	0.056	0.054	0.538	0.604
0.1	0.340	0.316	0.332	0.410	0.106	0.094	0.784	0.838	0.166	0.188	0.234	0.288
0.2	0.692	0.646	0.104	0.092	0.270	0.194	0.930	0.956	0.388	0.378	0.096	0.090
0.4	0.914	0.954	0.690	0.362	0.700	0.540	0.996	1.000	0.882	0.880	0.406	0.196

Table 7

The estimated coefficients, p-values from score and Wald tests for the significant SUVRs at 27 cortical regions. We also include the specific brain lobe that contains each cortical region. BANKSSTS stands for banks of the superior temporal sulcus.

Cortical regions	Brain lobes	Estimated coefficient	Wald test p-value
left middle temporal gyrus	Temporal lobe	−0.214	0.0002
left inferior parietal cortex	Parietal lobe	−0.214	0.0003
left inferior temporal gyrus	Temporal lobe	−0.223	0.0007
right inferior parietal cortex	Temporal lobe	−0.211	0.0007
left BANKSSTS	Temporal lobe	−0.174	0.0015
left fusiform gyrus	Temporal lobe	−0.262	0.0016
right middle temporal gyrus	Temporal lobe	−0.229	0.0024
left caudal middle frontal gyrus	Frontal lobe	−0.236	0.0030
left precuneus cortex	Parietal lobe	−0.215	0.0034
left entorhinal cortex	Temporal lobe	−0.217	0.0036
right inferior temporal	Temporal lobe	−0.225	0.0059
right left entorhinal cortex	Temporal lobe	−0.221	0.0065
right BANKSSTS	Temporal lobe	−0.168	0.0076

Table 8

The score and Wald test results of hypothesis that $\beta_{s_k} = 0$. TS and TW are the score and Wald test statistics, DF is the degree of freedom in the asymptotic distribution of the score and Wald statistics, which equals the number of the ROIs in the composite regions.

Composite regions	TS	score test p-value	TW	Wald test p-value	DF
BRAAK12	10.10	0.039	15.41	0.0039	4
BRAAK34	42.51	0.0113	89.45	1.774e-09	24
BRAAK56	66.29	0.0165	71.44	0.0055	44