

Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: www.tandfonline.com/journals/tstf20

Inference on the significance of modalities in multimodal generalized linear models

Wanting Jin , Guorong Wu & Quefeng Li

To cite this article: Wanting Jin , Guorong Wu & Quefeng Li (20 Nov 2026): Inference on the significance of modalities in multimodal generalized linear models, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2026.2765038

To link to this article: https://doi.org/10.1080/24754269.2026.2765038

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

View supplementary material

Published online: 20 Nov 2026.

Submit your article to this journal

Article views: 10

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tstf20

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2026.2765038

Taylor & Francis
by informa

OPEN ACCESS Check for updates

Inference on the significance of modalities in multimodal generalized linear models

Wanting Jin[a], Guorong Wu[b] and Quefeng Li [a]

[a]Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [b]Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

ABSTRACT

Despite the popularity of multimodal statistical models, there is a lack of rigorous statistical inference tools for inferring the significance of a single modality within a multimodal model, especially in high-dimensional models. For high-dimensional multimodal generalized linear models, we propose a novel entropy-based metric, called the expected relative entropy, to quantify the information gain of one modality in addition to all other modalities in the model. We propose a deviance-based statistic to estimate the expected relative entropy, prove that it is consistent, and show that its asymptotic distribution can be approximated by a non-central chi-squared distribution. That enables the calculation of confidence intervals and *p*-values to assess the significance of the expected relative entropy for a given modality. We numerically evaluate the empirical performance of our proposed inference tool by simulations and apply it to a multimodal neuroimaging dataset to demonstrate its good performance on various high-dimensional multimodal generalized linear models.

ARTICLE HISTORY
Received 1 February 2026
Revised 26 May 2026
Accepted 12 November 2026

KEYWORDS
High-dimensional inference; multimodal data; relative entropy; sure independence screening

1. Introduction

Multimodal data, which are collected from multiple sources or modalities of the same subjects, are rapidly emerging in many areas, including artificial intelligence, biomedical research, finance, social sciences, etc. For example, multimodal neuroimaging employs various imaging techniques, such as magnetic resonance imaging (MRI), positron emission tomography (PET) and computed tomography (CT) to study structural and functional changes in the brain. As different imaging modalities characterize different aspects of the brain, recent neuroimaging studies (Uludag & Roebroeck, 2014) have demonstrated that fused multimodal models often outperform single-modal models in predicting disease phenotypes and understanding disease progression mechanisms. Multi-omics data are another notable example of multimodal data in biomedical research. Richardson et al. (2016) illustrated that the integrative analysis of multi-omics data can enhance our understanding of the genomic mechanisms underlying various diseases. In recent statistics literature, many new integrative models have been proposed for multimodal data analysis, including unsupervised (Gaynanova & Li, 2019; Li & Jung, 2017; Lock et al., 2013) and supervised learning methods

CONTACT Quefeng Li quefeng@email.unc.edu

Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/24754269.2026.2765038.

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

(Li & Li, 2022; Xue & Qu, 2021). There are also theoretical studies proving that, under certain conditions, multimodal models are guaranteed to have better statistical performance than unimodal models (Li & Li, 2018).

In multimodal data analysis, a critical yet challenging problem is how to quantify the information contributed by individual modalities within a multimodal model. This problem is particularly relevant in various scientific domains. For instance, in genetics, heritability is a crucial concept that aims to determine the proportion of phenotypic variability attributable to genetic variation, as opposed to environmental, behavioural, or other factors. In neuroscience, when treating patients with dementia, physicians often employ multiple imaging modalities for diagnosis based on the patient's specific condition. Therefore, assessing the informational value of various imaging modalities in the disease phenotype is crucial for devising accurate treatment plans and preventing unnecessary diagnostic procedures. These practical problems raise the great need for rigorous statistical methods to quantify the information contributed by individual modalities and assess its statistical significance in multimodal statistical models.

In the classical statistical literature, there are many metrics quantifying the goodness-of-fit of a statistical model. These metrics can be regarded as the quantification of information contributed by all variables in a unimodal statistical model. For instance, R^2 and adjusted-R^2 were used to quantify the proportion of variation explained by all variables in a linear regression. These metrics were further extended to pseudo-R^2 metrics in the context of the generalized linear model (Zheng & Agresti, 2000). Hu et al. (2006) studied the property of the pseudo-R^2 metric under multinomial regression and proved that its limit can be interpreted as the relative entropy between the full and the null models. Unfortunately, these classic metrics are inadequate to measure and infer the contribution of individual modalities in multimodal models. The reason is that multimodal models are often high-dimensional, as each modality can contain a large number of variables. Consequently, inferring the significance of modalities constitutes a high-dimensional inference problem. In recent years, many high-dimensional inference methods have been developed. For example, Javanmard and Montanari (2014), Ning and Liu (2017), van de Geer et al. (2014), and Zhang and Zhang (2014) have developed de-biased or de-sparsifying least absolute shrinkage and selection operator (LASSO) methods and de-correlated score tests for testing the significance of a subvector of the covariate coefficients in unimodal regression models. Zhang and Cheng (2017) further developed a test for the whole covariate coefficients using the max norm of the de-sparsifying estimators. Shi et al. (2019) and Zhu and Bradic (2018) developed partially penalized methods for testing an arbitrary linear combination of covariate coefficients in a sparse generalized linear model and a dense high-dimensional linear regression model. Despite these advancements, existing inference methods are limited to the variable level and cannot be directly applied to our specific problem of evaluating and comparing high-dimensional modalities. This is because the contribution of individual modalities depends not only on the magnitude of covariate coefficients, but also on other population parameters such as the correlation between different modalities. Therefore, inferring such a quantity needs to infer the significance of a complicated function of covariate coefficients and some other population parameters, making the task more challenging than inferring covariate coefficients alone.

Recent years have also witnessed advancements in the inference of covariate coefficient functions. For example, Cai and Guo (2020) developed methods to infer the explained variance in a high-dimensional linear model, which is a quadratic function of the covariate coefficients. As will be discussed in Section 2, our proposed metric is more complicated than

a quadratic function, which poses a more challenging high-dimensional inference problem. In addition, we consider the inference problem in the context of multimodal models, instead of unimodal models.

In this article, we propose an entropy-based metric to quantify the contribution of individual modalities in multimodal generalized linear models (MGLM). We justify such a metric by studying its mathematical properties and establishing connections with conventional goodness-of-fit metrics for specific GLMs. We show that those conventional metrics can be treated as special cases of our proposed metric. Then, we develop a deviance-based estimator for this metric and study its statistical properties. We show that the estimator is consistent and provide an error bound. Moreover, we derive the limiting distribution of the estimator, enabling us to construct confidence intervals and calculate p-values to assess the significance of individual modalities. We show that such an asymptotic distributional result does not rely on variable selection consistency, which makes it more flexible for high-dimensional models. To the best of our knowledge, these inference tools are among the first to perform rigorous statistical inference at the modality level in the context of MGLM. They address the critical need of inferring the significance of modalities in multimodal models and can be applied to a wide range of multimodal data analysis problems.

The rest of the article is organized as follows. In Section 2, we propose to use the expected relative entropy (ERE) as a metric for quantifying the information contributed by individual modalities. We study its mathematical properties and relate it to conventional goodness-of-fit metrics for unimodal regression models. We also propose a two-step, deviance-based estimator for this metric. In Section 3, we give the error bound of the estimator and derive its asymptotic distribution, upon which the confidence interval and the p-value for the significance of a modality can be calculated. We use simulation studies to assess the numerical performance of the proposed metric and the corresponding estimator in Section 4. In Section 5, we apply our proposed method to analyze a multimodal neuroimaging dataset to demonstrate how it can be applied to a real-world problem. We conclude that the sure screeninge our article with a discussion of how our methods can be extended to other multimodal regression models in Section 6.

2. Expected relative entropy

2.1. Model setup

Suppose that there are M modalities of variables. Let $\boldsymbol{x}_m \in \mathbb{R}^{p_m}$ denote the vector of p_m variables from the m-th modality, $\boldsymbol{x} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_M^\top)^\top \in \mathbb{R}^p$ and $p = \sum_{m=1}^M p_m$. Let y denote the response variable. We assume that y relates to $\boldsymbol{x}$ through a multimodal generalized linear model such that the conditional distribution of y given $\boldsymbol{x}$ belongs to the canonical exponential family, which has the following density function

$$p(y \mid \boldsymbol{x}) = \exp\{\phi^{-1}[y\boldsymbol{x}^\top \boldsymbol{\beta}^* - b(\boldsymbol{x}^\top \boldsymbol{\beta}^*)] + c(y)\}, \tag{1}$$

where $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*\top}, \ldots \boldsymbol{\beta}_M^{*\top})^\top \in \mathbb{R}^p$ is the vector of true regression coefficients, $\boldsymbol{\beta}_m^* \in \mathbb{R}^{p_m}$ is the vector of true regression coefficients in the m-th modality, $b(\cdot)$ and $c(\cdot)$ are known functions, and $b(\cdot)$ is a twice continuously differentiable function with a positive second derivative. We assume the dispersion parameter $\phi = 1$ for simplicity. This assumption has also been used in other high-dimensional GLM inference studies (Li et al., 2024; Ning & Liu, 2017; Ouyang et al., 2023). When ϕ is unknown in practice, we can replace it with

some consistent estimator. As proved in some previous work (Li & Li, 2022), the asymptotic distribution of the resulting test statistic remains the same when such a consistent estimator is used.

Next, we propose a metric to quantify the information gain of the m-th modality in the MGLM. Let $\mathcal{M} = \{j : \beta_j^* \neq 0\}$ be the set of indices for non-zero elements of $\boldsymbol{\beta}^*$ and $\mathcal{M}_{-m}$ be the subset of $\mathcal{M}$ with indices outside the m-th modality. Excluding the m-th modality, we define an oracle working model that regresses y on $\boldsymbol{x}_{\mathcal{M}_{-m}}$, the subvector of $\boldsymbol{x}$ outside the m-th modality with nonzero coefficients, and let

$$\boldsymbol{\beta}_{0_{\mathcal{M}_{-m}}} = \text{argmax}_{\boldsymbol{\beta}_{\mathcal{M}_{-m}}} \, \mathrm{E}[\exp\{\phi^{-1}[y\boldsymbol{x}_{\mathcal{M}_{-m}}^\top \boldsymbol{\beta}_{\mathcal{M}_{-m}} - b(\boldsymbol{x}_{\mathcal{M}_{-m}}^\top \boldsymbol{\beta}_{\mathcal{M}_{-m}})] + c(y)\}]. \tag{2}$$

In other words, $\boldsymbol{\beta}_{0_{\mathcal{M}_{-m}}}$ can be treated as the pseudo true coefficients for the oracle working model that utilizes the true predictors outside the m-th modality. Finally, we expand $\boldsymbol{\beta}_{0_{\mathcal{M}_{-m}}}$ to $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0_{\mathcal{M}_{-m}}}^\top, \mathbf{0}^\top)^\top \in \mathbb{R}^p$ and define

$$p(y \mid \boldsymbol{x}_{-m}) = \exp\{\phi^{-1}[y\boldsymbol{x}^\top \boldsymbol{\beta}_0 - b(\boldsymbol{x}^\top \boldsymbol{\beta}_0)] + c(y)\},$$

where $\boldsymbol{x}_{-m}$ is the $(p - p_m)$-dim subvector of $\boldsymbol{x}$ without variables in the m-th modality. The Kullback–Leibler (KL) divergence, also known as the relative entropy, measuring the disparity between $p(y \mid \boldsymbol{x})$ and $p(y \mid \boldsymbol{x}_{-m})$, is defined as

$$D_{\mathrm{KL}}(p(y \mid \boldsymbol{x}) \| p(y \mid \boldsymbol{x}_{-m})) = \int_{y \in \mathcal{Y}} p(y \mid \boldsymbol{x}) \log\{p(y \mid \boldsymbol{x}) / p(y \mid \boldsymbol{x}_{-m})\} \, \mathrm{d}y,$$

where $\mathcal{Y}$ is the sample space of y. The KL divergence has been widely used in statistics and information theory (Amari, 2016). Our proposed KL divergence quantifies the entropy difference between the full and reduced models, and is a random quantity depending on $\boldsymbol{x}$. To quantify the information gain of the m-th modality, we propose to use

$$H_m = 2\mathrm{E}_{\boldsymbol{x}}[D_{\mathrm{KL}}(p(y \mid \boldsymbol{x}) \parallel p(y \mid \boldsymbol{x}_{-m}))], \tag{3}$$

where the expectation is taken with respect to the distribution of $\boldsymbol{x}$. In other words, H_m, up to a factor of 2, represents the expected relative entropy (ERE) of models with and without the m-th modality. The factor of 2 is introduced here to simplify the writing of the asymptotic results in Theorem 3.2. Our paper's main goal is to develop valid tools to perform statistical inference on H_m. As will be demonstrated in several concrete examples below, H_m depends not only on the magnitude of coefficients, but also on other population parameters. Therefore, existing high-dimensional inference methods (Ning & Liu, 2017; Shi et al., 2019; van de Geer et al., 2014; Zhang & Zhang, 2014) cannot be directly applied to perform inference on H_m, which motivates us to develop new inference tools. In the following, we will first justify why H_m is an appropriate metric for measuring the information gain of the m-th modality and connect it with existing goodness-of-fit measures for conventional unimodal GLMs. In this way, our proposed metric can be viewed as a natural extension of those existing measurements. Next, we study some mathematical properties of H_m.

First, H_m is non-negative, which is due to the non-negativity of the KL divergence. Second, Proposition 2.1 shows that H_m is monotone, which implies that the ERE contributed by multiple modalities is always no less than that of one modality. The non-negativity and monotonicity properties demonstrate that the ERE is an appropriate metric for quantifying

the information gain of an individual modality in MGLM. Next, we calculate H_m for some popular generalized linear models and show how it relates to existing goodness-of-fit metrics.

Proposition 2.1: *If* $\mathrm{E}|\log\{p(y\,|\,\boldsymbol{x}_{-m})/p(y\,|\,\boldsymbol{x}_{-(m,l)})\}| < \infty$, *then it holds that* $H_{(m,l)} \geq H_m$, *where* $H_{(m,l)} = 2\mathrm{E}_{\boldsymbol{x}}[D_{\mathrm{KL}}(p(y\,|\,\boldsymbol{x}) \parallel p(y\,|\,\boldsymbol{x}_{-(m,l)}))]$ *and* $\boldsymbol{x}_{-(m,l)}$ *is the subvector for* $\boldsymbol{x}$ *without variables in the m- and l-th modalities.*

Example 2.2 (Linear regression): Suppose that $y = \sum_{m=1}^{M} \boldsymbol{x}_m^\top \boldsymbol{\beta}_m^* + \epsilon$, where ϵ is the random error with $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma_\epsilon^2$. If we further assume $\boldsymbol{x} \sim N_p(0, \boldsymbol{\Sigma})$, then it follows that

$$H_m = \log \frac{\sigma_{m\,|\,-m}^2 + \sigma_\epsilon^2}{\sigma_\epsilon^2} + \frac{\boldsymbol{\beta}_m^{*\top} \boldsymbol{\Sigma}_m \boldsymbol{\beta}_m^* - \sigma_{m\,|\,-m}^2}{\sigma_{m\,|\,-m}^2 + \sigma_\epsilon^2}, \tag{4}$$

where $\boldsymbol{\Sigma}_m = \mathrm{Cov}(\boldsymbol{x}_m)$ and $\sigma_{m\,|\,-m}^2 = \mathrm{Var}(\boldsymbol{x}_m^\top \boldsymbol{\beta}_m^* | \boldsymbol{x}_{-m})$. The proof of (4) can be found in the Appendix.

As shown in (4), H_m depends not only on the coefficient $\boldsymbol{\beta}_m^*$, but also on the covariance matrix $\boldsymbol{\Sigma}_m$, the conditional variance $\sigma_{m\,|\,-m}^2$, and σ_ϵ^2. Thus, performing inference on H_m presents greater challenges than inference on $\boldsymbol{\beta}_m^*$ alone. Li and Li (2022) considered the multimodal linear regression model and proposed to use the conditional variance $\sigma_{m\,|\,-m}^2$ to quantify the contribution of the m-th modality in that model. They showed that $\sigma_{m\,|\,-m}^2$ can be interpreted as the improvement of the prediction error or the increment of the adjusted R^2 that is attributed to the inclusion of the m-th modality in addition to other modalities; see Section 6 of Li and Li (2022). Equation (4) gives a closed-form expression describing the relationship between $\sigma_{m\,|\,-m}^2$ and our proposed metric H_m.

For an MGLM with the density function shown in (1), the KL divergence is

$$D_{\mathrm{KL}}(p(y\,|\,\boldsymbol{x}) \parallel p(y\,|\,\boldsymbol{x}_{-m})) = b'(\boldsymbol{x}^\top \boldsymbol{\beta}^*)\boldsymbol{x}^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) + b(\boldsymbol{x}^\top \boldsymbol{\beta}_0) - b(\boldsymbol{x}^\top \boldsymbol{\beta}^*). \tag{5}$$

Using (5), we give more concrete forms of H_m for logistic, exponential and Poisson regressions.

Example 2.3 (Logistic regression): For logistic regression, the conditional density of y given $\boldsymbol{x}$ is $p(y\,|\,\boldsymbol{x}) = \exp[y\boldsymbol{x}^\top \boldsymbol{\beta}^* - \log\{1 + \exp(\boldsymbol{x}^\top \boldsymbol{\beta}^*)\}]$ so that $b(\theta) = \log\{1 + \exp(\theta)\}$. It follows from (5) that

$$H_m = 2\mathrm{E}_{\boldsymbol{x}} \left\{ \boldsymbol{x}^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) \frac{\exp(\boldsymbol{x}^\top \boldsymbol{\beta}^*)}{1 + \exp(\boldsymbol{x}^\top \boldsymbol{\beta}^*)} - \log \frac{1 + \exp(\boldsymbol{x}^\top \boldsymbol{\beta}_0)}{1 + \exp(\boldsymbol{x}^\top \boldsymbol{\beta}^*)} \right\}. \tag{6}$$

Next, we discuss how H_m relates to some conventional goodness-of-fit metric of logistic regression. We observe that $p(y\,|\,\boldsymbol{x})$ can also be expressed as $p(y\,|\,\boldsymbol{x}) = \exp\{I(y = 1)\log P_1(\boldsymbol{x}) + I(y = 0)\log P_0(\boldsymbol{x})\}$, where $P_1(\boldsymbol{x}) = P(y = 1\,|\,\boldsymbol{x}) = \exp(\boldsymbol{x}^\top \boldsymbol{\beta}^*)/\{1 + \exp(\boldsymbol{x}^\top \boldsymbol{\beta}^*)\}$, $P_0(\boldsymbol{x}) = P(y = 0|\boldsymbol{x}) = 1/\{1 + \exp(\boldsymbol{x}^\top \boldsymbol{\beta}^*)\}$. If we view all covariates as one single modality, the reduced model without any covariate follows the Bernoulli distribution with $P(y = 1) = \mathrm{E}_{\boldsymbol{x}}\{P_1(\boldsymbol{x})\}$, which is the marginal distribution of y denoted by $p(y)$. Then,

we have $\log p(y) = I(y=1)\log P(y=1) + I(y=0)\log P(y=0)$. Following our definition in (3), the ERE of all variables is given by

$$H_{\text{all}} = 2\mathrm{E}_{\boldsymbol{x}}[D_{\text{KL}}(p(y\,|\,\boldsymbol{x}) \parallel p(y))] = 2[\mathrm{E}_{\boldsymbol{x}}\{\log p(y\,|\,\boldsymbol{x})\} - \mathrm{E}_{\boldsymbol{x}}\{\log p(y)\}]$$
$$= 2\left[\sum_{k=0}^{1}\mathrm{E}_{\boldsymbol{x}}\{P_k(\boldsymbol{x})\log P_k(\boldsymbol{x})\} - \sum_{k=0}^{1}\mathrm{E}_{\boldsymbol{x}}\{P_k(\boldsymbol{x})\}\log E_{\boldsymbol{x}}\{P_k(\boldsymbol{x})\}\right]. \tag{7}$$

In (7), the term $\sum_{k=0}^{1}\mathrm{E}_{\boldsymbol{x}}\{P_k(\boldsymbol{x})\log P_k(\boldsymbol{x})\}$ can be viewed as the entropy of the conditional distribution of y given $\boldsymbol{x}$, while the term $\sum_{k=0}^{1}\mathrm{E}_{\boldsymbol{x}}\{P_k(\boldsymbol{x})\}\log \mathrm{E}_{\boldsymbol{x}}\{P_k(\boldsymbol{x})\}$ represents the entropy of the marginal distribution of y. In this case, we can verify that H_{all} is the same as the entropy-based metric proposed by Hu et al. (2006) to quantify the goodness-of-fit of the conventional logistic regression model. Consequently, Hu's entropy-based metric can be considered as a special case of our ERE metric in a classic unimodal setting that regards all covariates as a single modality.

Example 2.4 (Exponential regression): For exponential regression, the conditional density takes the form $p(y\,|\,\boldsymbol{x}) = \exp\{-y\boldsymbol{x}^\top\boldsymbol{\beta}^* + \log(\boldsymbol{x}^\top\boldsymbol{\beta}^*)\}$ so that $b(\theta) = -\log(\theta)$. Therefore,

$$H_m = 2\mathrm{E}_{\boldsymbol{x}}\left\{\log\frac{\boldsymbol{x}^\top\boldsymbol{\beta}^*}{\boldsymbol{x}^\top\boldsymbol{\beta}_0} - \left(1 - \frac{\boldsymbol{x}^\top\boldsymbol{\beta}_0}{\boldsymbol{x}^\top\boldsymbol{\beta}^*}\right)\right\}.$$

Example 2.5 (Poisson regression): For the Poisson regression, the conditional density has the form of $p(y\,|\,\boldsymbol{x}) = \exp\{y\boldsymbol{x}^\top\boldsymbol{\beta}^* - \exp(\boldsymbol{x}^\top\boldsymbol{\beta}^*) - \log(y!)\}$ so that $b(\theta) = \exp(\theta)$. Therefore,

$$H_m = 2\mathrm{E}_{\boldsymbol{x}}\{\exp(\boldsymbol{x}^\top\boldsymbol{\beta}^*)\boldsymbol{x}^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) + \exp(\boldsymbol{x}^\top\boldsymbol{\beta}_0) - \exp(\boldsymbol{x}^\top\boldsymbol{\beta}^*)\}.$$

Unlike linear regression, the closed forms of H_m for logistic, exponential and Poisson regressions are hard to derive due to their non-linear link functions, even when assuming $\boldsymbol{x}$ is multivariate normal. However, in simulation studies, when the distribution of $\boldsymbol{x}$ is known, we can use the Markov chain Monte Carlo method to approximate H_m. More details of such an approximation will be given in Section 4.

2.2. Estimation of the expected relative entropy

In practice, suppose that we have independent and identically distributed (i.i.d.) samples $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ from the distribution of $(y, \boldsymbol{x})$. We propose a deviance-based estimator of H_m, which is motivated by a seminal work (Hastie, 1987) that explored the relationship between the deviance and the KL divergence. The deviance of the full model using all modalities is $D(\boldsymbol{\beta}) = 2\sum_{i=1}^n\{\log p(Y_i) - \log p(Y_i\,|\,\boldsymbol{X}_i;\boldsymbol{\beta})\}$, where $p(Y_i)$ denotes the marginal density of y and $p(Y_i\,|\,\boldsymbol{X}_i;\boldsymbol{\beta})$ denotes the conditional density of $y\,|\,\boldsymbol{x}$ with the parameter $\boldsymbol{\beta} \in \mathbb{R}^p$. The deviance of the reduced model using all but the m-th modality is $D_{-m}(\boldsymbol{\beta}_0) = 2\sum_{i=1}^n\{\log p(Y_i) - \log p(Y_i\,|\,\boldsymbol{X}_i;\boldsymbol{\beta}_0)\}$, where $p(Y_i\,|\,\boldsymbol{X}_i;\boldsymbol{\beta}_0)$ denotes the conditional density of $y|\boldsymbol{x}_{-m}$ with the parameter $\boldsymbol{\beta}_{0_{-m}} \in \mathbb{R}^{p-p_m}$ in the reduced model and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0_{-m}}^\top, \mathbf{0}^\top)^\top \in \mathbb{R}^p$ being its expansion to a p-dim vector. Here, we slightly abuse the notation $\boldsymbol{\beta}_0$ by denoting it as an arbitrary parameter vector in the reduced model. Its true value is defined in (2) using

the same notation. Hastie (1987) showed that the average deviance difference $n^{-1}\{D(\boldsymbol{\beta}) - D_{-m}(\boldsymbol{\beta}_0)\}$ can be treated as a sample analog of the KL divergence $D_{\mathrm{KL}}(p(y \mid \boldsymbol{x}) \parallel p(y \mid \boldsymbol{x}_{-m}))$. This motivates us to estimate H_m by

$$\hat{H}_m = \frac{2}{n}\left\{\log L_n(\hat{\boldsymbol{\beta}}) - \log L_n(\hat{\boldsymbol{\beta}}_0)\right\}, \tag{8}$$

where $\log L_n(\boldsymbol{\beta}) = \sum_{i=1}^n Y_i \boldsymbol{X}_i^\top \boldsymbol{\beta} - b(\boldsymbol{X}_i^\top \boldsymbol{\beta})$ is the log-likelihood function and $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_0$ are estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$, respectively.

In the high-dimensional MGLM, we propose a two-step procedure to obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_0$. In the first step, we employ a sure independence screening (SIS, Fan & Song, 2010) method to screen the marginal effects of variables on the outcome and threshold those effects to reduce the ultra-high dimensional problem to a moderate dimensional one. In the second step, we solve a partially penalized maximum likelihood problem to obtain the final estimators.

In the screening step, define the maximum marginal likelihood estimator (MMLE) as

$$\hat{\boldsymbol{\beta}}_j^M = \mathrm{argmin}_{\boldsymbol{\beta}_j} -\frac{1}{n}\sum_{i=1}^n l(Y_i, \tilde{\boldsymbol{X}}_{ij}^\top \boldsymbol{\beta}_j), \tag{9}$$

for $j = 1, \ldots, p$, where $l(Y_i, \tilde{\boldsymbol{X}}_{ij}^\top \boldsymbol{\beta}_j) = Y_i \tilde{\boldsymbol{X}}_{ij}^\top \boldsymbol{\beta}_j - b(\tilde{\boldsymbol{X}}_{ij}^\top \boldsymbol{\beta}_j)$, $\boldsymbol{\beta}_j = (\beta_{cj}, \beta_j)^\top$, $\tilde{\boldsymbol{X}}_{ij} = (1, X_{ij})^\top$, β_{cj} is the intercept term and X_{ij} is the j-th element of $\boldsymbol{X}_i$.

The population counterpart of $\hat{\boldsymbol{\beta}}_j^M$ is defined as $\boldsymbol{\beta}_j^M = \mathrm{argmin}_{\boldsymbol{\beta}_j} \mathrm{E}\{-l(Y_1, \tilde{\boldsymbol{X}}_{1j}^\top \boldsymbol{\beta}_j)\}$. Then we threshold $\hat{\boldsymbol{\beta}}_j^M$ to obtain the set of screened variables $\tilde{\mathcal{M}} = \{j : |\hat{\beta}_j^M| \geq \gamma_n\}$, where γ_n is a user-defined threshold level. We let $\tilde{s} = |\tilde{\mathcal{M}}|$ be the number of screened variables. As will be discussed in Section 3, the purpose of screening is to reduce the problem from an ultra-high dimensional one to a moderate dimensional one, while ensuring that the screened set covers the true set, i.e., $\tilde{\mathcal{M}} \supset \mathcal{M} = \{j : \beta_j^* \neq 0\}$, with high probability. For that reason, we suggest setting a relatively small γ_n. In this paper, we use the following procedure to choose the optimal γ_n. We conduct a grid search within a user-specified interval of thresholds. For each grid point $\gamma_n^\dagger$, we regress the outcome $\boldsymbol{Y}$ on the subset of covariates $\boldsymbol{X}_{\mathcal{M}^\dagger}$, where $\mathcal{M}^\dagger = \{j : |\hat{\beta}_j^M| \geq \gamma_n^\dagger\}$, to obtain the maximum likelihood estimator under that model. Then, we calculate the corresponding Bayesian Information Criterion (BIC) value. It is important to note that the lower bound of the user-specified interval should be chosen such that $|\mathcal{M}^\dagger| < n$, ensuring that the maximum likelihood estimators can be calculated. The optimal value of γ_n is chosen as the smaller one of the two thresholds that yields the smallest BIC value plus one standard error. We remark that from a theoretical perspective, we only need to choose a threshold so that $\tilde{\mathcal{M}} \supset \mathcal{M}$ and the size of $\tilde{\mathcal{M}}$ is controllable. For more details, see our discussion before Theorem 3.1. In practice, some other screening procedures can also be employed such as the iterative sure independence screening (ISIS) method (Fan & Song, 2010).

In the second step, to estimate $\boldsymbol{\beta}$, we propose to solve the following partially penalized maximization likelihood problem:

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}_{\tilde{\mathcal{M}}^\perp} = \boldsymbol{0}} -\frac{1}{n}\sum_{i=1}^n \{Y_i \boldsymbol{X}_i^\top \boldsymbol{\beta} - b(\boldsymbol{X}_i^\top \boldsymbol{\beta})\} + \sum_{j \in \tilde{\mathcal{M}} \setminus \tilde{\mathcal{M}}_m} p_{\lambda_1}(|\beta_j|), \tag{10}$$

where $\tilde{\mathcal{M}}_m = \{j : j \in \tilde{\mathcal{M}}$ and $j \in$ the m-th modality$\}$ is the set of screened variables in the m-th modality, and $p_{\lambda_1}(|\beta_j|)$ is a folded-concave penalty function with a tuning parameter

λ_1. The folded concave penalty function $p_\lambda(|t|)$ is required to satisfy the following conditions: (i) $p_\lambda(|t|)$ is non-decreasing and concave in $t \in \mathbb{R}_+$ with $p_\lambda(0) = 0$ and $p_\lambda(t) > 0$ if $t > 0$; (ii) $p_\lambda(t)$ is differentiable for any $t \in \mathbb{R}_+$; (iii) the first derivative $p'_\lambda(t) = 0$ for any $t \geq a\lambda$; (iv) $0 \leq p'_\lambda(t) \leq \lambda$ for any $t \geq 0$. Examples of folded concave penalty functions include the SCAD penalty (Fan & Li, 2001) and the MCP penalty (Zhang, 2010), among others. In (10), we only penalize the screened variables outside the m-th modality. Variables in $\tilde{\mathcal{M}}_m$ are not penalized in order to avoid introducing bias in their estimators. This has been shown to be critical for correctly conducting the corresponding statistical inference (Shi et al., 2019). Moreover, as variables outside $\tilde{\mathcal{M}}$ have been screened out in the first step, $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{M}}^\perp} = 0$.

For the reduced model without the m-th modality, we solve another problem

$$\hat{\boldsymbol{\beta}}_0 = \text{argmin}_{\boldsymbol{\beta}_{\tilde{\mathcal{M}}^\perp}=\mathbf{0}, \boldsymbol{\beta}_{\tilde{\mathcal{M}}_m}=\mathbf{0}} -\frac{1}{n}\sum_{i=1}^{n}[Y_i \boldsymbol{X}_i^\top \boldsymbol{\beta} - b(\boldsymbol{X}_i^\top \boldsymbol{\beta})] + \sum_{j \in \tilde{\mathcal{M}} \setminus \tilde{\mathcal{M}}_m} p_{\lambda_2}(|\beta_j|), \tag{11}$$

where $p_{\lambda_2}(|\beta_j|)$ is also a folded concave penalty function with a tuning parameter λ_2. In practice, the same function can be used for p_{λ_1} and p_{λ_2}, with the two parameters tuned separately. We utilize the SCAD penalty for both of them in our paper and choose optimal values of λ_1 and λ_2 that minimize the BIC of the full and reduced models. In (11), we require all variables in the m-th modality to have zero coefficients. Therefore, $\hat{\boldsymbol{\beta}}_0 = (\hat{\boldsymbol{\beta}}_{0_{-m}}^\top, \mathbf{0}^\top)^\top \in \mathbb{R}^p$, where $\hat{\boldsymbol{\beta}}_{0_{-m}}$ can be treated as an estimator of $\boldsymbol{\beta}_{0_{-m}}$ for the reduced model. Finally, we remark that the penalties used in (10) and (11) further reduce the number of non-zero variables out of the m-th modality after the screening step. This yields sparser estimators in the final stage and controls the estimation bias. In Section 4, we will demonstrate that this extra regularization also improves the accuracy of the statistical inference on H_m.

3. Theoretical results

In this section, we present two key statistical results on $\hat{H}_m$. We first prove that $\hat{H}_m$ is consistent for H_m and give the error bound. Then, we prove that the limiting distribution of $n\hat{H}_m$ follows a non-central χ^2-distribution, with a non-centrality parameter closely related to nH_m. This distributional result enables us to construct confidence intervals for H_m based on $\hat{H}_m$, as well as computing p-values to assess the statistical significance of a given modality. These findings meet the critical need for how to perform statistical inference at the modality level for MGLM.

We first introduce some notations. For a vector $\boldsymbol{a} \in \mathbb{R}^d$, let $\|\boldsymbol{a}\|_\infty = \max_j |a_j|$, $\|\boldsymbol{a}\|_1 = \sum_{j=1}^d |a_j|$, $\|\boldsymbol{a}\|_2 = (\sum_{j=1}^d a_j^2)^{1/2}$ denote its sup-norm, L_1-norm, and Euclidean norm, respectively. For an index set S, let $\boldsymbol{a}_S$ denote the subvector of $\boldsymbol{a}$ with indices in S. In particular, let the subscripts m and $-m$ denote the sets of indices in and out of the m-th modality, respectively. Let $|S|$ denote the size of S. For a square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$, let $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ denote its minimal and maximal eigenvalues. Let $\|\boldsymbol{A}\|_\infty = \sup_{ij} |a_{ij}|$, $\|\boldsymbol{A}\|_1 = \max_{1 \leq j \leq d} \sum_{i=1}^d |a_{ij}|$, $\|\boldsymbol{A}\|_2 = \lambda_{\max}(\boldsymbol{A})$, $\|\boldsymbol{A}\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$ denote its element-wise sup-norm, L_1-norm, L_2-norm, and Frobenius norm, respectively. For index sets S_1 and S_2, let $\boldsymbol{A}_{S_1 S_2}$ denote the submatrix of $\boldsymbol{A}$ with rows in S_1 and columns in S_2. Finally, we introduce the following quantities that will be used in the theoretical results. Recalling the definitions of $\mathcal{M}$ and $\tilde{\mathcal{M}}$ in Section 2, let $s = |\mathcal{M}|$ and $\tilde{s} = |\tilde{\mathcal{M}}|$. Let $\mathcal{M}_m$ and $\tilde{\mathcal{M}}_m$ be the subsets of $\mathcal{M}$ and $\tilde{\mathcal{M}}$ with indices in the m-th modality. Define $s_m = |\mathcal{M}_m|$ and $\tilde{s}_m = |\tilde{\mathcal{M}}_m|$. To give the theoretical results, we impose the following regularity conditions.

Condition 1: $\boldsymbol{x} \in \mathbb{R}^p$ is a sub-Gaussian random vector with parameter σ^2. That is, for any $\mathbf{v} \in \mathbb{R}^p$ with $\|\mathbf{v}\|_2 = 1$, $P(|\mathbf{v}^\top \boldsymbol{x}| > t) \leq \exp\{-t^2/(2\sigma^2)\}$. $c \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq C$, where $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{x})$ and c and C are some generic positive constants.

Condition 2: There exist generic positive constants c and C such that $\mathrm{E}[\exp\{b(\boldsymbol{x}^\top \boldsymbol{\beta}^* + c)\} - \exp\{b(\boldsymbol{x}^\top \boldsymbol{\beta}^*)\}] + \mathrm{E}[\exp\{b(\boldsymbol{x}^\top \boldsymbol{\beta}^* - c)\} - \exp\{b(\boldsymbol{x}^\top \boldsymbol{\beta}^*)\}] \leq C$.

Condition 3: $\log p = O(n^a)$ for some $0 < a < 1$ and $\gamma_n = c_0 n^{-\kappa}$ for some $c_0 > 0$ and $0 < \kappa < (1-a)/2$.

Condition 4: The marginal Fisher information $\boldsymbol{I}_j(\boldsymbol{\beta}_j) = \mathrm{E}\{b''(\tilde{\boldsymbol{x}}_j \boldsymbol{\beta}_j)\tilde{\boldsymbol{x}}_j \tilde{\boldsymbol{x}}_j^\top\}$ is finite and positive definite at $\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^M$, where $\boldsymbol{\beta}_j^M$ is defined below (9) and $\tilde{\boldsymbol{x}}_j = (1, x_j)^\top$, $j = 1, \ldots, p$. Moreover, $\|\boldsymbol{I}_j(\boldsymbol{\beta}_j)\|_2 < \infty$ for all $\boldsymbol{\beta}_j \in \mathcal{B} = \{\boldsymbol{\beta}_j = (\beta_{cj}, \beta_j)^\top : |\beta_{cj}| \leq C, |\beta_j| \leq C\}$, where C is a generic positive constant.

Condition 5: For $\boldsymbol{\beta}_j \in \mathcal{B}$, $\mathrm{E}\{l(y, \tilde{\boldsymbol{x}}_j^\top \boldsymbol{\beta}_j) - l(y, \tilde{\boldsymbol{x}}_j^\top \boldsymbol{\beta}_j^M)\} \geq c\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^M\|_2^2$ for some generic positive constant c, where the function $l(\cdot)$ is defined in (9).

Condition 6: There exists a generic positive constant C, such that $|b''(\theta_1) - b''(\theta_2)| \leq C|\theta_1 - \theta_2|$ for any θ_1 and θ_2.

Condition 7: There exists a positive constant δ such that, for some $0 < q < 1/2 - \kappa$, $\sup_{\boldsymbol{\beta}_j \in \mathcal{B}, \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^M\|_2 \leq \delta} \mathrm{E}[b(\tilde{\boldsymbol{x}}_j^\top \boldsymbol{\beta}_j) I(|x_j| > Cn^q)] = o(n^{-1})$.

Condition 8: $\min_{j \in \mathcal{M}^*} |\beta_j^M| \geq cn^{-\kappa}$ for some constant $c > 0$ and $0 < \kappa < (1-a)/2$.

Condition 9: There exist some generic positive constants c and C such that $c < \mathrm{Var}(\boldsymbol{x}^\top \boldsymbol{\beta}^*) < C$. Moreover, $\mathrm{E}[b''(\boldsymbol{x}^\top \boldsymbol{\beta}^*)^4] < \infty$.

Condition 10: $\lambda_1 = c_1\sqrt{(\log \tilde{s})/n}$, $\lambda_2 = c_2\sqrt{(\log \tilde{s})/n}$ for some positive constants c_1, c_2.

Condition 1 is a typical sub-Gaussian assumption for high-dimensional statistical problems. Condition 2, which is analogous to Condition D from Fan and Song (2010), ensures that the response variable has an exponentially light tail. Condition 3 imposes assumptions on p and the threshold γ_n. Conditions 4 and 5 are assumptions about the marginal likelihood function, which are analogous to the regularity conditions A' and C' imposed in Fan and Song (2010). Condition 6 ensures the smoothness of the link function. Fan and Song (2010) proved that Conditions 4–6 are satisfied for most common generalized linear models. Conditions 7–9 are some technical conditions needed in the proof. Condition 10 gives the rate of a proper choice of tuning parameters.

Under Conditions 1–8, Fan and Song (2010) proved that the sure screening property holds with high probability, that is, $P(\mathcal{M} \subset \tilde{\mathcal{M}}) \to 1$ as $n \to \infty$. Besides, they also proved that $\tilde{s} = O_P(n^{2\kappa}) = o_P(n)$. These results show that SIS can reach the sure screening property with controllable size of screened variables. This is critical for us to establish the error bound of $\hat{H}_m - H_m$ below.

Theorem 3.1: *Under Conditions* 1–10*, it holds that*

$$\hat{H}_m - H_m = O_P(\max\{sn^{-1}\log n, 1/\sqrt{n}\}).$$

Theorem 3.1 implies that $\hat{H}_m$ is consistent for H_m under the sparsity assumption that $s = o(n/\log n)$ with a convergence rate no faster than the typical root-n rate. Next, we derive the asymptotic distribution of $\hat{H}_m$. This task is inherently a high-dimensional inference problem. The bulk of existing high-dimensional inference methods focuses on testing the significance of covariate coefficients in a high-dimensional regression model, or their linear combinations (Javanmard & Montanari, 2014; Ning & Liu, 2017; Shi et al., 2019; van de Geer et al., 2014; Zhang & Zhang, 2014; Zhu & Bradic, 2018). However, as seen in Section 2.1, $\hat{H}_m$ and H_m both depend on some complex functions of the covariate coefficients and other population parameters. Cai and Guo (2020) have developed some methods to infer the explained variance in a high-dimensional linear model, where the explained variance is a quadratic function of the covariate coefficients, which differs from our proposed metric. Moreover, their inference problem is in the context of a linear model. Hence, their results are not applicable to our problem. Furthermore, unlike the hypothesis testing problems considered in previous works, which focus more on the limiting distribution of the test statistic under the null hypothesis, we aim to derive the limiting distribution of $\hat{H}_m$ for all cases of H_m. This distinction introduces additional complexity to our inference task. Nevertheless, we find that we can still borrow ideas from existing works such as Shi et al. (2019) to obtain the limiting distribution of $\hat{H}_m$. In addition, since we adopt a two-step procedure to obtain $\hat{H}_m$, the first screening step can efficiently reduce the high-dimensional inference problem to a moderate-dimensional one, control the estimation bias introduced by high dimensionality, and simplify the statistical inference. We find that due to the sure screening property, our asymptotic distributional result no longer needs to be built upon variable selection consistency. As will be seen in conditions of Theorem 3.2, we do not require any irrepresentable condition, which is different from the conditions needed in Shi et al. (2019). To obtain the asymptotic distribution of $\hat{H}_m$, we require the following additional regularity conditions:

Condition 11: $\|\boldsymbol{\beta}^*_{\mathcal{M}_m}\|_2 = O(n^{\kappa-1/2})$.

Condition 12: $\max_j \sup_{\boldsymbol{\beta}\in\mathcal{N}_0} \lambda_{\max}\{\mathbf{X}^\top \mathrm{diag}\{|X_j| \circ b'''(\mathbf{X}\boldsymbol{\beta})\}\boldsymbol{X}\} = O_P(n)$, and $\sup_{\boldsymbol{\beta}\in\mathcal{N}_0} \lambda_{\max}(\mathrm{E}\{b''(\boldsymbol{x}^\top\boldsymbol{\beta})\boldsymbol{x}\boldsymbol{x}^\top\}) < \infty$, where $\mathcal{N}_0 = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq C\sqrt{n^{-1}s\log n}, \boldsymbol{\beta}_{\tilde{\mathcal{M}}^\perp} = \mathbf{0}\}$ for some constant $C > 0$.

Condition 13: The constant κ in Conditions 3, 7, 8, and 11 satisfies $0 < \kappa < 1/6$.

Condition 11 requires the magnitude of $\boldsymbol{\beta}^*_{\mathcal{M}_m}$ to be controllable. Condition 12 is a technical condition needed in the proof. Its first component is analogous to Condition (A1) in Shi et al. (2019) and has been shown to hold with probability tending to 1 under the sub-Gaussian assumption on $\boldsymbol{X}$ in their paper. Condition 13 imposes a more stringent requirement on κ compared to its counterpart in Theorem 3.1. This is expected, as deriving the asymptotic distribution typically requires stronger conditions than proving consistency, which has been extensively discussed in the literature, e.g., in Shi et al. (2019).

Theorem 3.2: *Under Conditions* 1–13,

$$\sup_{x>0} |\mathrm{P}(n\hat{H}_m \le x) - \mathrm{P}(\chi^2_{\tilde{s}_m,\Gamma_n} \le x)| = o_P(1), \tag{12}$$

where $\Gamma_n = n\boldsymbol{\beta}^{*\top}_{\tilde{\mathcal{M}}_m} \boldsymbol{\Omega}^{-1}_{mm} \boldsymbol{\beta}^*_{\tilde{\mathcal{M}}_m}$ *and* $\boldsymbol{\Omega}_{mm} = (\mathbf{0}, \boldsymbol{I}_{\tilde{\mathcal{M}}_m})[\mathrm{E}\{b''(\boldsymbol{x}^\top\boldsymbol{\beta}^*)\boldsymbol{x}_{\tilde{\mathcal{M}}}\boldsymbol{x}^\top_{\tilde{\mathcal{M}}}\}]^{-1}(\mathbf{0}, \boldsymbol{I}_{\tilde{\mathcal{M}}_m})^\top$.

Theorem 3.2 shows that the limiting distribution of $n\hat{H}_m$ can be approximated by the non-central χ^2-distribution with $\tilde{s}_m$ degrees of freedom and a non-centrality parameter of Γ_n. Next, we establish the relationship between H_m and Γ_n in the following proposition.

Proposition 3.3: *Assume that the parameter space* $\mathcal{B}$ *of the MGLM in (1) is compact, and that there exist positive constants* b_0 *and* C_b *such that* $|b(u)| \le C_b \exp(b_0|u|)$ *for* $u \in \mathbb{R}$. *Then, under the conditions of Theorems* 3.1 *and* 3.2, *we have* $H_m = \Gamma_n/n + o(1)$.

Proposition 3.3 shows that the non-centrality parameter of the limiting distribution derived in Theorem 3.2 is intrinsically linked to H_m under a mild condition on the parameter space and an envelope condition on $b(\cdot)$. These conditions are satisfied by standard GLM families, including Gaussian, Bernoulli, Poisson, Gamma, and inverse Gaussian distributions. Theorem 3.2 and Proposition 3.3 together enable us to construct a confidence interval for H_m. The problem essentially lies in constructing a confidence interval for the non-centrality parameter of the non-central χ^2-distribution based on a point estimator, a topic previously studied by Kent and Hainsworth (1995). Inspired by this work, we propose to use the pivotal quantity $F_{\tilde{s}_m,\theta}(n\hat{H}_m)$ to construct the confidence interval, where $F_{\tilde{s}_m,\theta}(t)$ denotes the cumulative distribution function of the non-central χ^2-distribution with $\tilde{s}_m$ degrees of freedom and a non-centrality parameter of θ. It can be verified that $F_{\tilde{s}_m,\theta}(t)$ is a strictly decreasing function of θ for any fixed t, and $F_{\tilde{s}_m,\theta}(n\hat{H}_m)$ asymptotically follows a uniform distribution in $[0, 1]$ as $n\hat{H}_m$ asymptotically follows the distribution of $F_{\tilde{s}_m,\theta}$. Suppose that we can find a real number C_l such that $F_{\tilde{s}_m,C_l}(n\hat{H}_m) = 1 - \alpha$. Then, setting $\theta = \Gamma_n + o(n)$, we have $P(C_l/n \le H_m) \ge P(C_l \le \theta) = P(F_{\tilde{s}_m,\theta}(n\hat{H}_m) \le F_{\tilde{s}_m,C_l}(n\hat{H}_m)) = 1 - \alpha$, where Proposition 3.3 is used in the first inequality. Thus, the $(1 - \alpha)$ confidence lower bound of H_m is given by $[C_l/n, +\infty)$. Similarly, the $(1 - \alpha)$ confidence interval of H_m, up to an $o(1)$ term, is given by $[C_l/n, C_u/n]$, where C_l and C_u satisfy $F_{\tilde{s}_m,C_l}(n\hat{H}_m) = 1 - \alpha/2$ and $F_{\tilde{s}_m,C_u}(n\hat{H}_m) = \alpha/2$. Since the form of the cumulative distribution function of the non-central χ^2-distribution is complicated, directly solving for C_l and C_u is computationally challenging. In practice, we propose to use a bisection method along the real line to find C_l and C_u such that $F_{\tilde{s}_m,C_l}(n\hat{H}_m) \approx 1 - \alpha/2$ and $F_{\tilde{s}_m,C_u}(n\hat{H}_m) \approx \alpha/2$. Moreover, Theorem 3.2 also enables us to test if H_m is zero. Under the null hypothesis that $H_m = 0$, it can be shown that $n\hat{H}_m$ asymptotically follows the standard $\chi^2_{\tilde{s}_m}$-distribution. Thus, the p-value is given by $P(\chi^2_{\tilde{s}_m} > n\hat{H}_m)$.

Finally, we remark that unlike some high-dimensional inference methods (Shi et al., 2019), Theorem 3.2 does not require variable selection consistency for either $\hat{\boldsymbol{\beta}}$ or $\hat{\boldsymbol{\beta}}_0$, which can be seen from the absence of the irrepresentable condition in Theorem 3.2. The reason is that the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_0$ are constructed via two-step procedures. In the first step, Sure Independence Screening (SIS) reduces the high-dimensional estimation problem to a relatively low-dimensional one; see the remark in the last paragraph preceding Theorem 3.1. The sure screening property guarantees that the screened set $\tilde{\mathcal{M}}$ contains the true active set $\mathcal{M}$ with high probability, while its cardinality is ensured to be smaller than n. This dimension reduction step prevents certain error terms arising in the proof of Theorem 3.2 from diverging. In

contrast, some high-dimensional inference methods (Shi et al., 2019) that directly address the high-dimensional problem without an initial SIS step rely on variable selection consistency to control analogous error terms. The simulation results in Section 4 further demonstrate this point as our proposed method still gives good coverage probabilities, even when the estimators contain false positives or false negatives.

In practice, quantifying the information gain of individual modalities in multimodal models as a percentage is often useful. To address this need, we propose the following pseudo-R^2 metric: $R_m^2 = 1 - \exp(-H_m)$, which can be estimated by $\hat{R}_m^2 = 1 - \exp(-\hat{H}_m)$. It can be easily shown that $R_m^2 \in [0, 1]$ as H_m is non-negative. Given that the transformation function mapping H_m to R_m^2 is strictly monotonically increasing, we can derive the confidence interval for R_m^2 by directly applying this transformation function to the confidence interval of H_m. We remark that the pseudo-R^2s introduced by Hu et al. (2006) for unimodal logistic regression can be treated as special cases of our proposed pseudo-R^2 metric.

4. Simulation

We conducted simulation studies to assess the empirical performance of the proposed metric on the evaluation of information gain of modalities. In all simulations, we generated i.i.d. samples of $\{\boldsymbol{X}_i\}_{i=1}^n$ from $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. We consider the following five models. For Models 1 to 3, we set $\boldsymbol{\Sigma} = 0.8\boldsymbol{I}_p + 0.2\boldsymbol{J}_p$, where $\boldsymbol{J}_p \in \mathbb{R}^{p\times p}$ is a matrix with all entries equal to 1.

Model 1 (Linear regression): We set $M = 3, p_1 = p_2 = p_3 = 200, \boldsymbol{\beta}_1^*=(-1.15, 1, 1.75, \mathbf{0})^\top \in \mathbb{R}^{200}, \boldsymbol{\beta}_2^* = (-0.6, -0.8, 0.45, \mathbf{0})^\top \in \mathbb{R}^{200}, \boldsymbol{\beta}_3^* = (-0.75, 0.8, -0.75, \mathbf{0})^\top \in \mathbb{R}^{200}$, and $\boldsymbol{\beta}^*= \delta \cdot (\boldsymbol{\beta}_1^{*\top}, \boldsymbol{\beta}_2^{*\top}, \boldsymbol{\beta}_3^{*\top})^\top \in \mathbb{R}^{600}$, where $\delta \in \{0.6, 0.8, 1, 1.2, 1.6, 2\}$ controls the magnitude of $\boldsymbol{\beta}^*$. We generated the response vector from the linear model that $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ is a vector of i.i.d. errors from $N(0, 1)$.

Model 2 (Logistic regression): We set $M = 2, p_1 = p_2 = 300, \boldsymbol{\beta}_1^* = (0.5, -1, -1.6, 0.9, \mathbf{0})^\top \in \mathbb{R}^{300}$, $\boldsymbol{\beta}_2^* = (0.4, 0.8, -0.7, -1.4, \mathbf{0})^\top \in \mathbb{R}^{300}$, and $\boldsymbol{\beta}^* = \delta \cdot (\boldsymbol{\beta}_1^{*\top}, \boldsymbol{\beta}_2^{*\top})^\top \in \mathbb{R}^{600}$, where $\delta \in \{1, 1.2, 1.4, 1.7, 2, 2.3\}$ controls the magnitude of $\boldsymbol{\beta}^*$. We generated the response vector $\boldsymbol{Y}$ from the logistic regression model that $P(Y_i = 1 \mid \boldsymbol{X}_i) = \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*)/\{1 + \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*)\}$.

Model 3 (Probit regression): We set $M = 2$, $p_1 = p_2 = 300$, and $\boldsymbol{\beta}_1^* = (0.5, 0.6, -0.7, -0.9, \mathbf{0})^\top \in \mathbb{R}^{300}, \boldsymbol{\beta}_2^* = (0.4, 0.5, -0.6, -0.7, \mathbf{0})^\top \in \mathbb{R}^{300}$. We let $\boldsymbol{\beta}^*=\delta \cdot (\boldsymbol{\beta}_1^{*\top}, \boldsymbol{\beta}_2^{*\top})^\top \in \mathbb{R}^{600}$, where $\delta \in \{1, 1.1, 1.2, 1.3, 1.4, 1.5\}$ controls the magnitude of $\boldsymbol{\beta}^*$. We generated the response vector $\boldsymbol{Y}$ from the Probit regression model that $P(Y_i = 1 \mid \boldsymbol{X}_i) = \Phi(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

Model 4 (Linear regression): We set $M = 3$, $p_1 = p_2 = p_3 = 400$, $n = 200$, $\boldsymbol{\beta}_1^* = (-1, 1, 1.2, \mathbf{0})^\top \in \mathbb{R}^{200}, \boldsymbol{\beta}_2^* = (-0.6, -0.8, 0.45, \mathbf{0})^\top \in \mathbb{R}^{200}, \boldsymbol{\beta}_3^* = (-0.75, 0.8, -0.75, \mathbf{0})^\top \in \mathbb{R}^{200}$, and $\boldsymbol{\beta}^* = \delta \cdot (\boldsymbol{\beta}_1^{*\top}, \boldsymbol{\beta}_2^{*\top}, \boldsymbol{\beta}_3^{*\top})^\top \in \mathbb{R}^{600}$, where $\delta \in \{0.6, 0.8, 1, 1.2, 1.6, 2\}$ controls the magnitude of $\boldsymbol{\beta}^*$. We generated the response vector from the linear model that $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ is a vector of i.i.d. errors from $N(0, 1)$. We set

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} \\ \boldsymbol{\Sigma}_{12}^\top & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} \\ \boldsymbol{\Sigma}_{13}^\top & \boldsymbol{\Sigma}_{23}^\top & \boldsymbol{\Sigma}_{33} \end{pmatrix} \in \mathbb{R}^{(p_1+p_2+p_3)\times(p_1+p_2+p_3)}. \tag{13}$$

In this construction, the intra-modality blocks $\mathbf{\Sigma}_{mm}$, for $m \in \{1, 2, 3\}$, adopt an AR(1) structure, whereby

$$(\mathbf{\Sigma}_{mm})_{ij} = 0.2^{|i-j|}, \quad \forall\, i, j \in \{1, \ldots, p_m\}$$

and the inter-modality blocks $\mathbf{\Sigma}_{ml} \in \mathbb{R}^{p_l \times p_m}$, $m \neq l$ are sparse matrices defined by

$$(\mathbf{\Sigma}_{ml})_{jk} = \begin{cases} 0.4, & \text{if } j = k \text{ and } j \leq 12, \\ 0, & \text{otherwise.} \end{cases}$$

Model 5 (Logistic regression): We set $M = 2, p_1 = p_2 = 500, n = 200, \boldsymbol{\beta}_1^* = (1, -1, -1.4, 1.2, \mathbf{0})^\top \in \mathbb{R}^{300}$, $\boldsymbol{\beta}_2^* = (0.9, 0.8, -0.8, -0.9, \mathbf{0})^\top \in \mathbb{R}^{300}$, and $\boldsymbol{\beta}^* = \delta \cdot (\boldsymbol{\beta}_1^{*\top}, \boldsymbol{\beta}_2^{*\top})^\top \in \mathbb{R}^{600}$, where $\delta \in \{0.9, 1.0, 1.2, 1.4, 1.6, 1.8\}$ controls the magnitude of $\boldsymbol{\beta}^*$. We generated the response vector $\boldsymbol{Y}$ from the logistic regression model that $P(Y_i = 1 \mid \boldsymbol{X}_i) = \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*)/\{1 + \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*)\}$. We generated i.i.d. samples of $\{\boldsymbol{X}_i\}_{i=1}^n$ from $N_p(\mathbf{0}, \mathbf{\Sigma})$, where

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{12}^\top & \mathbf{\Sigma}_{22} \end{pmatrix} \in \mathbb{R}^{(p_1+p_2)\times(p_1+p_2)},$$

where $\mathbf{\Sigma}_{11}$, $\mathbf{\Sigma}_{12}$, and $\mathbf{\Sigma}_{22}$ have the same structure as in Model 4.

We included three methods in the comparison. The oracle method computed the maximum likelihood estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_0$ by regressing the response variable on the set of variables with truly nonzero coefficients, with and without the m-th modality, respectively. Our proposed method, denoted as SIS+SCAD, first performed SIS, followed by solving (10) and (11) with a SCAD penalty. The screening threshold for SIS was chosen via a grid search procedure as described in Section 2.2. The tuning parameters λ_1 and λ_2 were chosen to minimize the BIC criteria of the full and reduced models. The third method, denoted as SIS+refit, first performed SIS, and then solved (10) and (11) with $\lambda_1 = \lambda_2 = 0$. In other words, this approach refitted the model using only those variables that survived the SIS screening procedure. For each method, we plugged $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_0$ into (8) to obtain the corresponding $\hat{H}_m$. Then, we used the procedure described after Theorem 3.2 to construct the two-sided confidence intervals. In particular, for the oracle method, we used quantiles of the non-central χ^2-distribution with the oracle s_m degrees of freedom. For the other two methods, we used quantiles of the non-central χ^2-distribution with $\tilde{s}_m$ degrees of freedom. To calculate H_m, we used (4) for Models 1 and 4. For Models 2, 3 and 5, we used the following procedure to calculate H_m. First, we independently generated another 10000 samples from these three models. We regressed the responses on variables outside the m-th modality with truly nonzero coefficients and obtained the maximum likelihood estimator as an approximation of $\boldsymbol{\beta}_0$. Then, we plugged it and $\boldsymbol{\beta}^*$ into (3) and used the sample average over these 10000 samples as an approximation of H_m. For each model, we conducted 5000 simulation replications to evaluate the empirical coverage probabilities of the two-sided confidence intervals for H_m. We also represented the sensitivity and specificity of variable selection performance given by the SIS+SCAD and SIS+refit methods.

The results are represented in Figures 1–5 for linear, logistic and probit regression models.

Each of Figures 1–5 shows that our method's confidence intervals exhibit satisfactory coverage probability when H_m is not small. In general, when H_m increases, the coverage

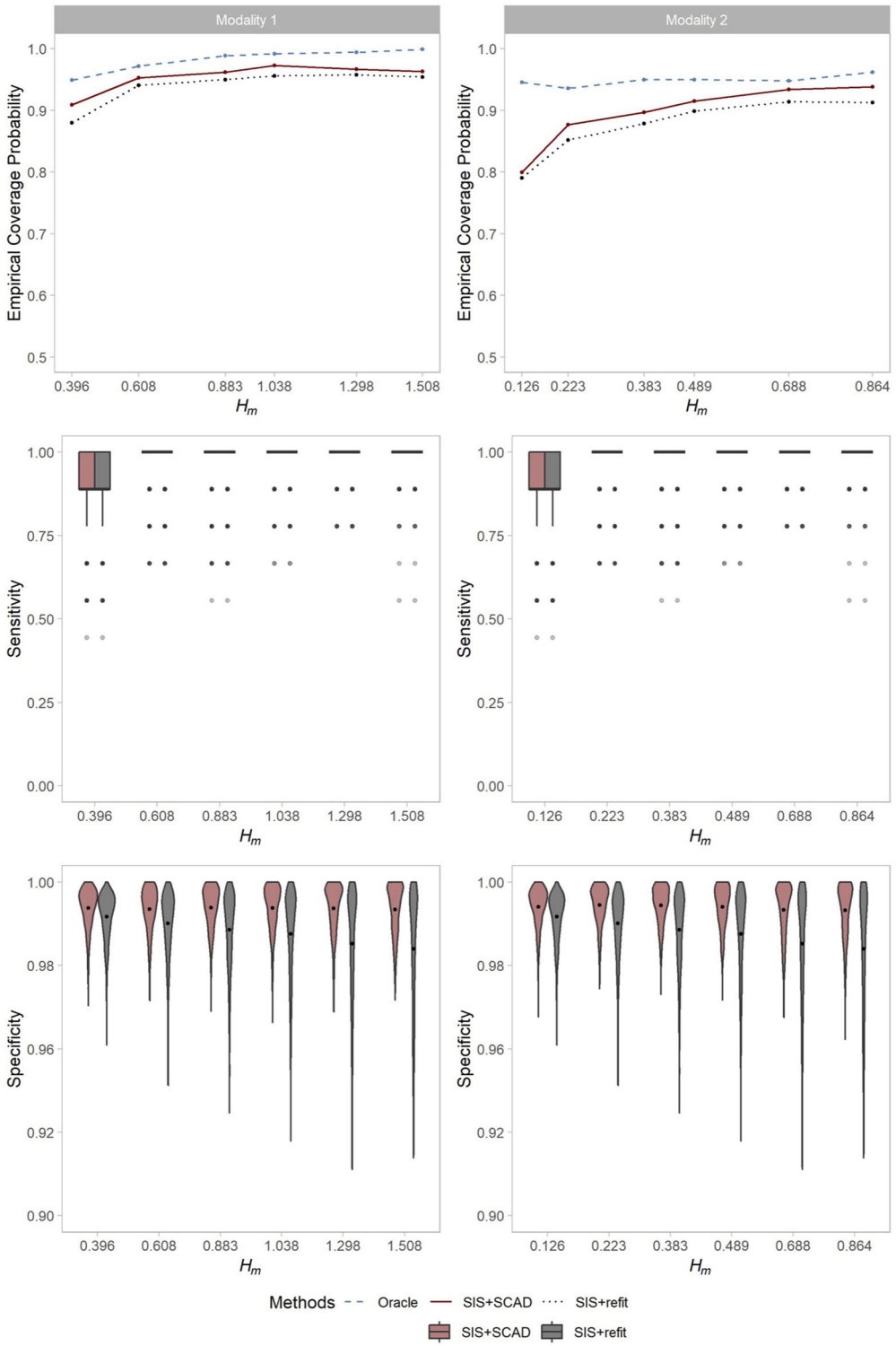


Figure 1. Empirical coverage probabilities and variable selection performance (Sensitivity and Specificity) by the three methods for Model 1. The oracle method (blue), SIS + SCAD penalized method (red), SIS + refit method (gray).

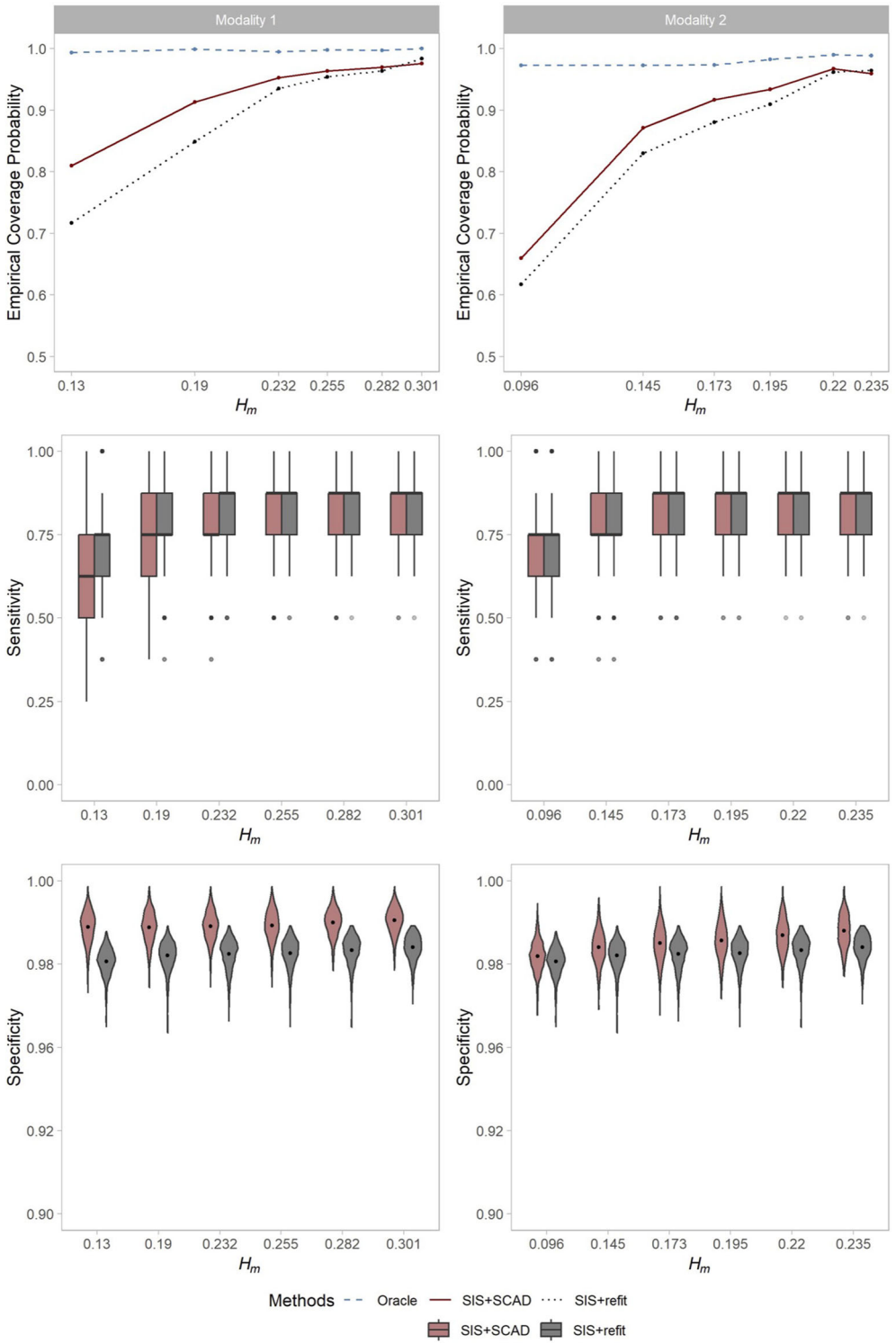


Figure 2. Empirical coverage probabilities and variable selection performance (Sensitivity and Specificity) by the three methods for Model 2. The oracle method (blue), SIS + SCAD penalized method (red), SIS + refit method (gray).

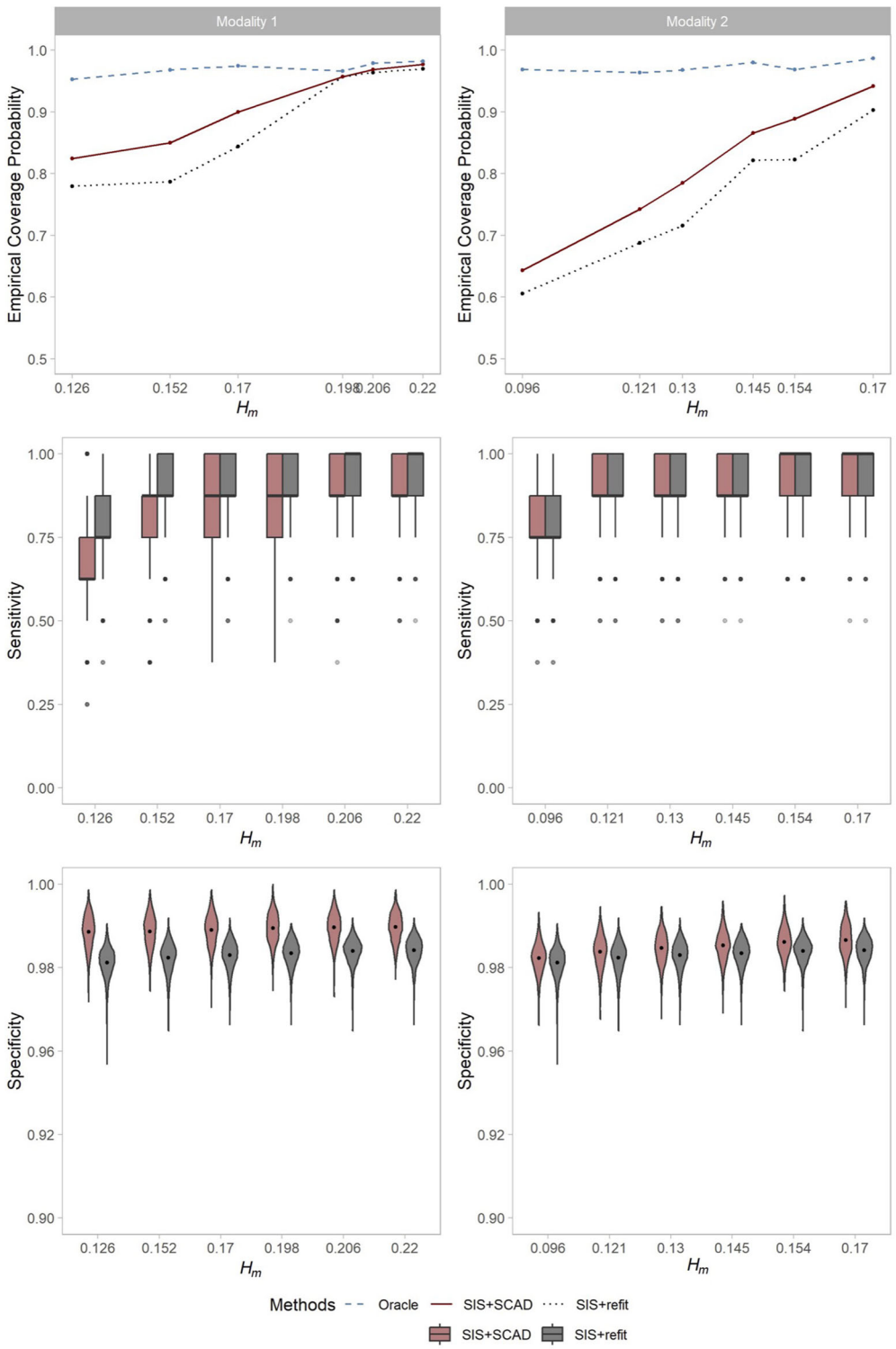


Figure 3. Empirical coverage probabilities and variable selection performance (Sensitivity and Specificity) by the three methods for Model 3. The oracle method (blue), SIS + SCAD penalized method (red), SIS + refit method (gray).

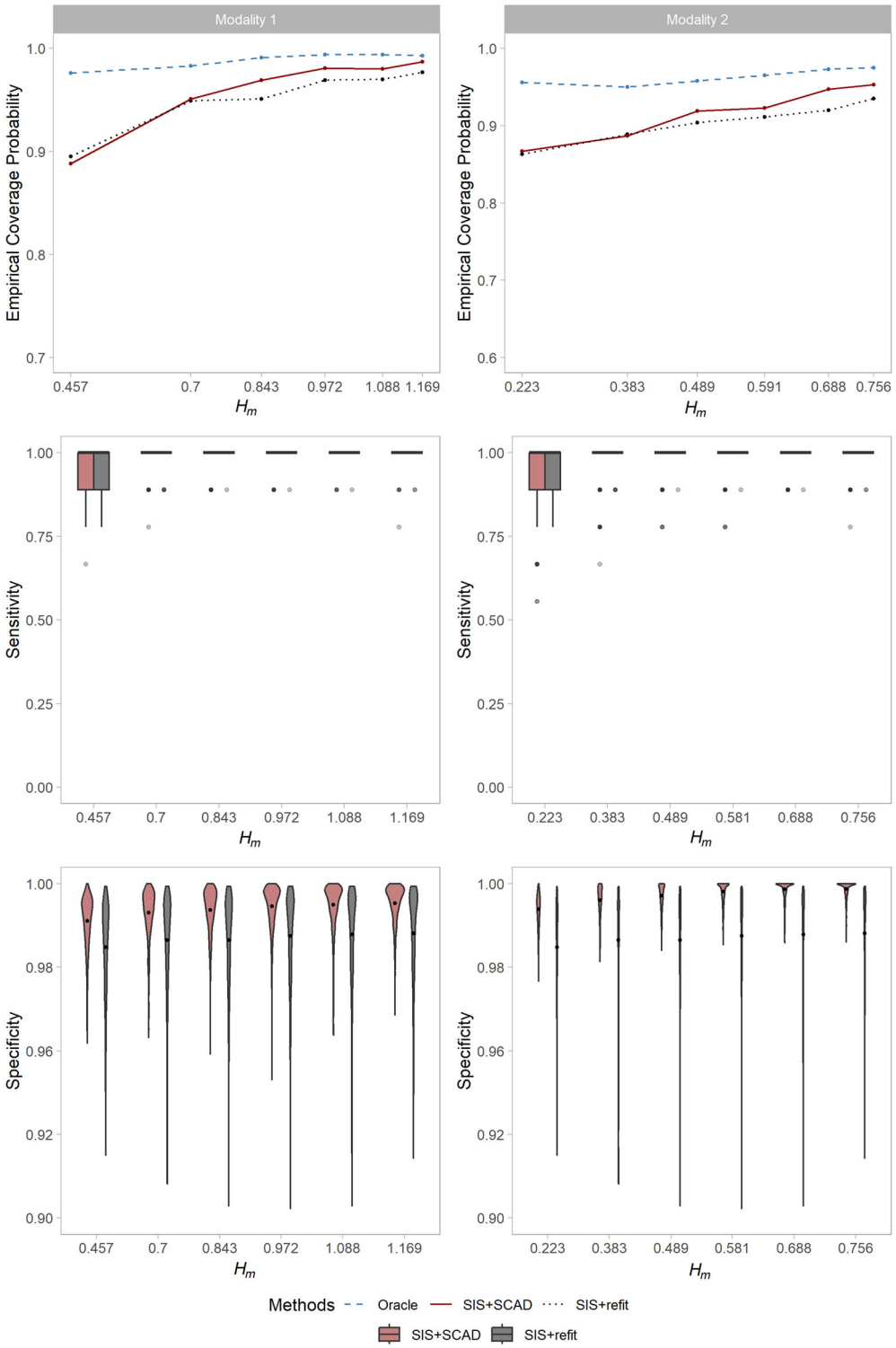


Figure 4. Empirical coverage probabilities and variable selection performance (Sensitivity and Specificity) by the three methods for Model 4. The oracle method (blue), SIS + SCAD penalized method (red), SIS + refit method (gray).

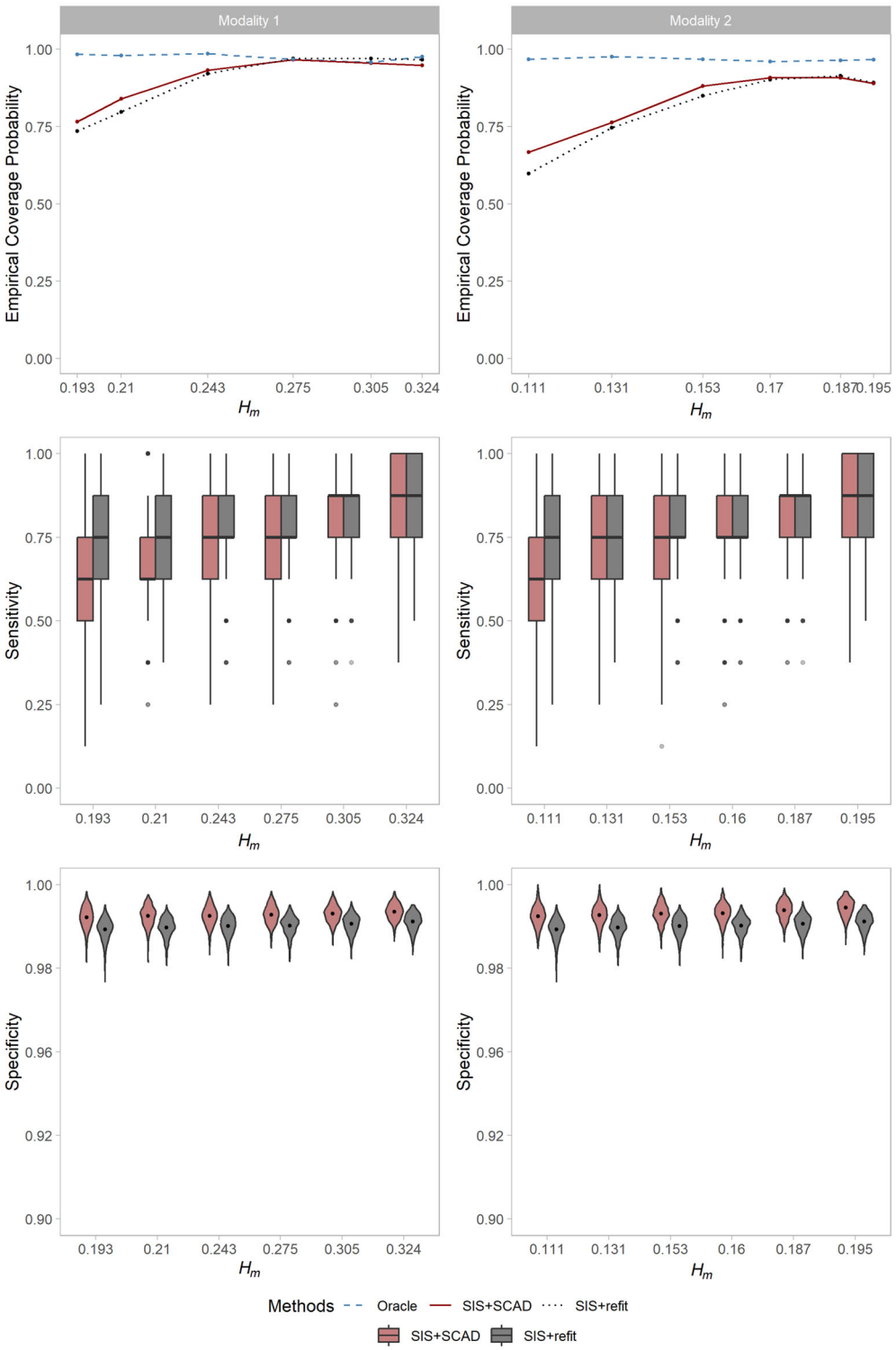


Figure 5. Empirical coverage probabilities and variable selection performance (Sensitivity and Specificity) by the three methods for Model 5. The oracle method (blue), SIS + SCAD penalized method (red), SIS + refit method (gray).

probability of our method also increases, which is expected. Comparing coverage probabilities between the two modalities in each model shows that Modality 1 consistently achieves higher coverage than Modality 2. This superior performance can be attributed to Modality 1's larger H_m value, which further validates our previous findings. We observe that when H_m is relatively large, the coverage probability of our method is close to that of the oracle method. Considering that our method deals with high-dimensional modalities, whereas the oracle method uses only the low-dimensional true method, this result is promising. Furthermore, our method achieves higher coverage probability compared to the SIS+refit approach, particularly when H_m is small. This demonstrates that the second-stage partially penalized method effectively eliminates false positives, leading to improved coverage probability. This improvement is evident in the specificity panels of Figures 1–5, where our method consistently achieves higher specificity than the SIS+refit method. Moreover, our method does not require perfect variable selection in order to yield good coverage probabilities. For example, in the logistic and probit regressions, the sensitivity panels of Figures 2, 3 and 5 reveal that our method occasionally omits some important variables. Nevertheless, it can still reach the nominal 95% coverage probability when H_m is relatively large. This observation aligns with our discussion of the theoretical results following Theorem 3.2.

5. Analysis of multimodal neuroimaging data

We applied our proposed method to analyze a multimodal neuroimaging dataset from the Alzheimer's disease neuroimaging initiative (ADNI) database. Alzheimer's disease (AD) is an irreversible neurodegenerative disorder characterized by progressive impairment of cognitive and memory functions. It is a major form of dementia among elderly subjects and is one of the leading causes of death in the United States. Neuroimaging techniques, such as positron emission tomography (PET), provide valuable tools for the diagnosis of AD. Our study included $n = 669$ subjects with Amyloid-PET and FDG-PET scans from the ADNI database. The Amyloid-PET image detects the presence and distribution of amyloid-beta proteins in the brain, which is a crucial biomarker of AD. The FDG-PET image uses fluorodeoxyglucose (FDG) to measure metabolic activity and glucose uptake in the brain, which provides insights into brain function and identifies areas of abnormal activity in neurological disorders like AD. We used the Destrieux atlas (Destrieux et al., 2010) to parcellate each brain into 160 regions of interest (ROIs), consisting of 148 cortical regions and 12 sub-cortical regions ($p_1 = p_2 = 160$). Three measurements were used to assess the cognitive decline of AD patients: the memory score (MEM), the executive function score (EF), and the AD diagnostic label (DX). MEM and EF are continuous scores, while DX is an ordinal variable with levels ranging from normal (NC) to early mild cognitive impairment (EMCI), stable MCI (SMCI), late MCI (LMCI), and AD.

In our analysis, we regressed MEM and EF on the two imaging modalities using linear regression. Additionally, we dichotomized DX, categorizing individuals with LMCI and AD as the case group and the remaining participants as the control group. We then regressed the dichotomized DX variable on the two imaging modalities using logistic regression. For each model, we first performed SIS followed by solving the partially penalized problem with the SCAD penalty in (10) to select ROIs associated with the outcome. The selected ROIs are shown in Figure 6. The results indicate that the MEM score is more sensitive to the accumulation of AD-related biomarkers than the EF score, as more ROIs are selected in the MEM-based model than the EF-based model. This is aligned with current clinical findings

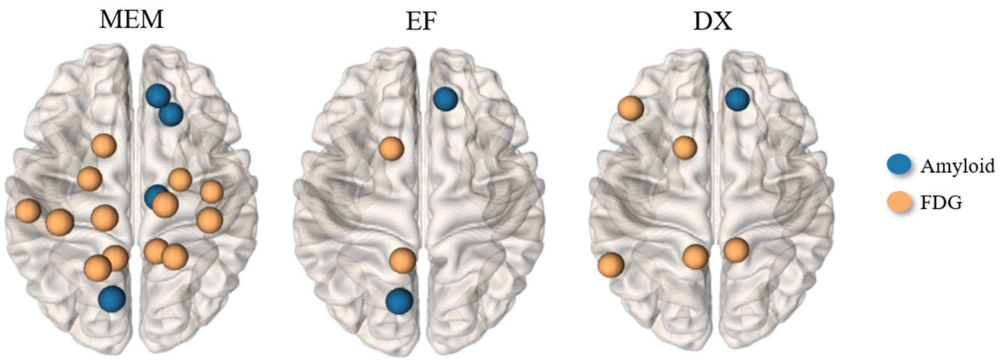


Figure 6. Selected regions of interest for MEM, EF and DX by two modalities. MEM, the memory score; EF, the executive function score; DX, the AD diagnostic label.

Table 1. Point estimates, confidence intervals and p-values of expected relative entropy and pseudo-R^2 for different neuroimaging techniques.

	FDG vs. FDG+Amyloid		
Outcome	$\hat{H}_m$ (95% CI)	R^2_m (95% CI)	p-value
MEM	0.170 (0.109, 0.233)	0.156 (0.103, 0.208)	1.19×10^{-23}
EF	0.089 (0.048, 0.139)	0.085 (0.047, 0.130)	1.10×10^{-13}
DX	0.033 (0.011, 0.066)	0.032 (0.011, 0.064)	2.77×10^{-6}
	Amyloid vs. Amyloid+FDG		
Outcome	$\hat{H}_m$(95% CI)	R^2_m (95% CI)	p-value
MEM	0.147 (0.078, 0.191)	0.137 (0.075, 0.174)	3.93×10^{-15}
EF	0.230 (0.161, 0.306)	0.205 (0.149, 0.264)	4.28×10^{-34}
DX	0.168 (0.106, 0.229)	0.155 (0.101, 0.205)	1.5×10^{-22}

Note: MEM: the memory score; EF: the executive function score; DX: the AD diagnostic label.

that memory decline is a typical clinical symptom (Jahn, 2013). Moreover, we used our proposed method to calculate the estimated ERE and pseudo-R^2 with 95% confidence intervals and corresponding p-values for each of the two modalities and regression models. The results, shown in Table 1, suggest that FDG-PET is significantly more informative than Amyloid-PET for the outcomes of EF and DX. This shows that the metabolism level is a more putative biomarker for monitoring cognitive decline than pathology biomarkers such as amyloid.

6. Discussion

We propose an ERE metric to quantify the information gain of individual modalities in high-dimensional multimodal generalized linear models. The metric is justified through some discussion of its mathematical properties and is related to conventional goodness-of-fit metrics for unimodal models, as well as recent developments in modality evaluation under more restrictive models. We demonstrate that these metrics can be viewed as special cases or are equivalent to our proposed metric. Moreover, we develop a deviance-based estimator of the ERE metric by using a two-step procedure. We derive the error bound and the asymptotic distribution of this estimator, which enables us to perform statistical inference on the ERE. This meets the growing demand for rigorous statistical inference tools at the modality level in multimodal models. Through theoretical studies and numerical experiments, we

demonstrate that our constructed confidence intervals maintain good coverage probabilities without relying on variable selection consistency, a desirable property given the difficulty of achieving perfect variable selection in high-dimensional models.

While our study focuses on generalized linear models, we note that our entropy-based metric can possibly be extended to other models, such as the Cox proportional hazards model. Further research into such extensions would be valuable, as it would provide a comprehensive tool set for assessing the significance of individual modalities across various multimodal models.

Acknowledgments

We are grateful to the editor and the referees for their insightful comments that have led to a substantial improvement to an earlier version of the paper.

Author contributions

CRediT: **Wanting Jin:** Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing; **Guorong Wu:** Conceptualization, Writing – original draft, Writing – review & editing; **Quefeng Li:** Methodology, Supervision, Writing – original draft, Writing – review & editing

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by National Institutes of Health [R01 AG073259].

ORCID

Quefeng Li http://orcid.org/0000-0003-0707-2763

References

Amari, S. I. (2016). *Information geometry and its applications* (Vol. 194). Springer.

Cai, T. T., & Guo, Z. (2020). Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(2), 391–419. https://doi.org/10.1111/rssb.12357

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, *53*(1), 1–15. https://doi.org/10.1016/j.neuroimage.2010.06.010

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. https://doi.org/10.1198/016214501753382273

Fan, J., & Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, *38*(6), 3567–3604. https://doi.org/10.1214/10-AOS798

Gaynanova, I., & Li, G. (2019). Structural learning and integrative decomposition of multi-view data. *Biometrics*, *75*(4), 1121–1132. https://doi.org/10.1111/biom.13108

Hastie, T. (1987). A closer look at the deviance. *American Statistician*, *41*(1), 16–20. https://doi.org/10.1080/00031305.1987.10475434

Hu, B., Shao, J., & Palta, M. (2006). Pseudo-R^2 in logistic regression model. *Statistica Sinica*, *16*(3), 847–860.

Jahn, H. (2013). Memory loss in Alzheimer's disease. *Dialogues in Clinical Neuroscience*, *15*(4), 445–454. https://doi.org/10.31887/DCNS.2013.15.4/hjahn

Javanmard, A., & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, *15*, 2869–2909.
Kent, J. T., & Hainsworth, T. J. (1995). Confidence intervals for the noncentral chi-squared distribution. *Journal of Statistical Planning and Inference*, *46*(2), 147–159. https://doi.org/10.1016/0378-3758(94)00104-4
Li, G., & Jung, S. (2017). Incorporating covariates into integrated factor analysis of multi-view data. *Biometrics*, *73*(4), 1433–1442. https://doi.org/10.1111/biom.12698
Li, Q., & Li, L. (2018). Integrative linear discriminant analysis with guaranteed error rate improvement. *Biometrika*, *105*(4), 917–930. https://doi.org/10.1093/biomet/asy047
Li, Q., & Li, L. (2022). Integrative factor regression and its inference for multimodal data analysis. *Journal of the American Statistical Association*, *117*(540), 2207–2221. https://doi.org/10.1080/01621459.2021.1914635
Li, S., Zhang, L., Cai, T. T., & Li, H. (2024). Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association*, *119*(546), 1274–1285. https://doi.org/10.1080/01621459.2023.2184373
Lock, E. F., Hoadley, K. A., Marron, J. S., & Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, *7*(1), 523–542. https://doi.org/10.1214/12-AOAS597
Ning, Y., & Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, *45*(1), 158–195. https://doi.org/10.1214/16-AOS1448
Ouyang, J., Tan, K. M., & Xu, G. (2023). High-dimensional inference for generalized linear models with hidden confounding. *Journal of Machine Learning Research*, *24*(296), 1–61.
Richardson, S., Tseng, G. C., & Sun, W. (2016). Statistical methods in integrative genomics. *Annual Review of Statistics and Its Application*, *3*(1), 181–209. https://doi.org/10.1146/statistics.2016.3.issue-1
van de Geer S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, *42*(3), 1166–1202. https://doi.org/10.1214/14-AOS1221
Shi, C., Song, R., Chen, Z., & Li, R. (2019). Linear hypothesis testing for high dimensional generalized linear models. *The Annals of Statistics*, *47*(5), 2671–2703. https://doi.org/10.1214/18-AOS1761
Uludag, K., & Roebroeck, A. (2014). General overview on the merits of multimodal neuroimaging data fusion. *NeuroImage*, *102*(Pt 1), 3–10. https://doi.org/10.1016/j.neuroimage.2014.05.018
Xue, F., & Qu, A. (2021). Integrating multisource block-wise missing data in model selection. *Journal of the American Statistical Association*, *116*(536), 1914–1927. https://doi.org/10.1080/01621459.2020.1751176
Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942. https://doi.org/10.1214/09-AOS729
Zhang, C. H., & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *76*(1), 217–242. https://doi.org/10.1111/rssb.12026
Zhang, X., & Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, *112*(518), 757–768. https://doi.org/10.1080/01621459.2016.1166114
Zheng, B., & Agresti, A. (2000). Summarizing the predictive power of a generalized linear model. *Statistics in Medicine*, *19*(13), 1771–1781. https://doi.org/10.1002/(ISSN)1097-0258
Zhu, Y., & Bradic, J. (2018). Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, *113*(524), 1583–1600. https://doi.org/10.1080/01621459.2017.1356319