

A Collinearity Test Statistic for Hypothesis Testing in High-Dimensional Generalized Linear Models

Zdravko Botev

botev@unsw.edu.au

UNSW Sydney

Anant Mathur

UNSW Sydney

Sarat Moka

UNSW Sydney

Spiridon Penev

UNSW Sydney

Research Article

Keywords: collinearity test, generalized linear models, high dimensions, likelihood-ratio test

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11203411/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Collinearity Test Statistic for Hypothesis Testing in High-Dimensional Generalized Linear Models

Zdravko Botev^{1*}, Anant Mathur¹, Sarat Moka¹, Spiridon Penev¹

^{1*}School of Mathematics and Statistics, UNSW Sydney, NSW 2033.

*Corresponding author(s). E-mail(s): botev@unsw.edu.au;
Contributing authors: anant.mathur@unsw.edu.au;
s.moka@unsw.edu.au; s.penev@unsw.edu.au;

Abstract

We propose a simple collinearity statistical test for nested model comparison in generalized linear models, with p predictors and sample size n . We show that the collinearity test applies without modification to a wide range of asymptotic relationships between p and n , including the difficult high-dimensional case of $p > n$. The collinearity test is simple to apply and compute and, unlike some of its competitors, requires neither the solution of a set of nonlinear equations, nor the nonconvex optimization of a partial penalized likelihood. We provide a few numerical examples comparing the test with the classical Wald and likelihood-ratio tests. The results suggest that the test enjoys high statistical power across a range of asymptotic regimes and has the highest power when the alternative hypothesis is sparse.

Keywords: collinearity test, generalized linear models, high dimensions, likelihood-ratio test

1 Introduction

Model hypothesis testing is one of the perennial problems in statistics that goes back to the pioneering work of Fisher [1], Neyman and Pearson [2], and Wilks [3]. In the context of the classical generalized linear model (GLM) with p predictors $[\mathbf{v}_1, \dots, \mathbf{v}_p] = \mathbf{X} \in \mathbb{R}^{n \times p}$, unknown regression coefficients $\boldsymbol{\beta}^* \in \mathbb{R}^p$, and response vector $\mathbf{Y} \in \mathbb{R}^n$, the hypothesis testing problem can be described as follows.

Let $\mathcal{S} \subset \{1, \dots, p\}$ with $|\mathcal{S}| < n$ represent the hypothesized model vectors $\mathbf{X}_{\mathcal{S}} := [\mathbf{v}_i, i \in \mathcal{S}]$ selected from the columns of $\mathbf{X}$. If $\mathcal{A} := \{j : \beta_j^* \neq 0, j \in \mathcal{S}^c\}$ is the set of false exclusions under $\mathcal{S}$ (the set of predictors incorrectly omitted from $\mathcal{S}$), then we wish to test the null hypothesis:

$$H_0 : \quad \mathcal{A} = \emptyset,$$

that is, the cardinality $|\mathcal{A}| = 0$, versus the alternative:

$$H_A : \quad |\mathcal{A}| > 0.$$

If the null hypothesis is true, then the conditional expected response is

$$\mathbb{E}[\mathbf{Y} \mid \mathbf{X}] = \mathbf{h}(\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}^*)$$

for some known GLM inverse-link function $\mathbf{h}$ applied componentwise, and vector $\boldsymbol{\beta}_{\mathcal{S}}^* := [\beta_i^*, i \in \mathcal{S}]$, so that $\mathbf{X}\boldsymbol{\beta}^* = \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}^* = \sum_{i \in \mathcal{S}} \mathbf{v}_i \beta_i^*$.

Which test statistic must be used to make a decision about the null hypothesis depends on the assumed asymptotic regime.

The first asymptotic regime is the classical one where $p = o(n)$ as $n \rightarrow \infty$, that is, the number of predictors is negligible relative to an abundance of observations. Here the likelihood-ratio test (LRT) and the Wald test are widely used and enjoy high statistical power [4].

The second asymptotic regime is when there exists a constant $\kappa \in (0, 1/2)$ (called the aspect ratio) such that as $n \uparrow \infty$:

$$p/n \rightarrow \kappa.$$

In other words, while p is not large enough to exceed n , compared to n it is large enough to preclude the validity of the classical LRT and Wald test [5, 6]. In this regime Sur and Candès [5, Theorem 4] prove that, at least in the case of the logistic GLM, the classical LRT statistic has an unusual rescaled χ^2 -distribution, where the rescaling factor is larger than 1 for $\kappa > 0$ (and equal to 1 for the classical case of $\kappa = 0$). To determine the correct rescaling factor, one has to solve a system of two nonlinear equations in two unknowns. With the correct rescaling of the limiting χ^2 -distribution, inference then proceeds in the usual manner and Sur and Candès [5] attest to very good empirical performance of the resulting rescaled χ^2 -distribution test.

The third asymptotic regime is when $p > n$. For example, we may have an abundance of candidate predictors, but very few observations. In this most difficult case, a popular approach is to use sophisticated partial penalized likelihood ratio test statistics. The computation of these test statistics is emphatically nontrivial and requires calls to numerical solvers for likelihood maximization with SCAD/MCP penalties [7, 8]. In general, working with log-likelihoods regularized via SCAD/MCP penalty functions results in difficult nonconvex optimization, even in cases where the negative log-likelihood is convex itself. Naturally, there are no practical guarantees that the global solution can be found.

The purpose of this article is to propose a single easy-to-compute test statistic that works in all the different asymptotic regimes described above. We will refer to it as the (GLM) collinearity test statistic, due to its similarities with how linear dependence between vectors is commonly measured.

The main advantage of the collinearity test is that, in contrast to the above-mentioned competitors, its computation requires neither the solution of a nonlinear system of equations, nor a nonconvex optimization of a partial penalized likelihood, and works without alteration even in the high-dimensional $p > n$ case. We provide numerical examples showing that the power of the collinearity test is competitive with classical alternatives.

Nevertheless, an honest assessment of the collinearity test requires emphasizing some potential limitations.

First, while the p-values under the null hypothesis can be tightly bounded from above and below via simple analytical formulas, a more precise estimation of the p-values typically calls for Monte Carlo simulation. The cost of using the collinearity test statistic combined with the Monte Carlo method with m samples is $O(nmp)$.

Second, since the test can be applied successfully across the three asymptotic regimes without any alteration, it is generally not the most statistically powerful test across the three regimes considered above. A test designed specifically for a given asymptotic regime is likely to be more powerful than the collinearity test when the corresponding asymptotic assumptions are satisfied. In particular, the LRT and Wald test can be more powerful than the collinearity test when $p = o(n)$ – the regime in which the classical asymptotic theory for the LRT and Wald test is mostly applicable.

The collinearity test is a type of max-type or L_∞ -type statistic, commonly used in econometrics [9] for robust inference. Generally, such tests have significant statistical power when the alternatives are sparse. Indeed, our numerical results and theoretical investigations are consistent with this sparse principle.

The rest of the article is organized as follows. In Section 2 we define the collinearity test statistic. In the subsequent section we derive the limiting distribution of the test statistic under the null hypothesis. In Section 3 we provide numerical experiments comparing the collinearity, Wald, and LRT tests in the setting of the logistic generalized linear model. The numerical results are consistent with our theoretical findings about the statistical power of the test for the simplest GLM — the Gaussian linear model. This is followed by the final concluding section.

2 The Collinearity Test Statistic

To lay the groundwork, define the “natural parameter” function:

$$\boldsymbol{\eta}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta},$$

with the convenient notation convention:

$$\boldsymbol{\eta}(\boldsymbol{\beta}_S) = \mathbf{X}_S\boldsymbol{\beta}_S.$$

Now, consider the canonical exponential family GLM setting, where, conditional on the selected model $\mathbf{X}_S$ and its regression parameters, each response Y_i is independent with exponential-family pdf:

$$f(y_i | \mathbf{X}_S, \boldsymbol{\beta}_S) = \exp \left(\frac{y_i \eta_i(\boldsymbol{\beta}_S) - b(\eta_i(\boldsymbol{\beta}_S))}{\phi} + c(y_i, \phi) \right), \quad (1)$$

where $\phi > 0$ is an over-dispersion parameter and $b(\cdot)$ is a given function with strictly positive and continuous second derivative over the range of all $\eta_i(\boldsymbol{\beta}_S)$. Given a coefficient $\boldsymbol{\beta}$, the conditional expectation $\boldsymbol{\mu}(\boldsymbol{\beta})$ and variance are:

$$\begin{aligned} \mu_i(\boldsymbol{\beta}) &:= \mathbb{E}[Y_i | \mathbf{X}, \boldsymbol{\beta}] = b'(\eta_i(\boldsymbol{\beta})), \quad i = 1, \dots, n \\ \sigma_i^2(\boldsymbol{\beta}) &:= \text{Var}[Y_i | \mathbf{X}, \boldsymbol{\beta}] = \phi b''(\eta_i(\boldsymbol{\beta})). \end{aligned}$$

Under the selected model S , twice the negative log-likelihood is:

$$l(\boldsymbol{\beta}_S) := -\frac{2}{\phi} \mathbf{Y}^\top \boldsymbol{\eta}(\boldsymbol{\beta}_S) + \frac{2}{\phi} \sum_{i=1}^n (b(\eta_i(\boldsymbol{\beta}_S)) - \phi c(Y_i, \phi))$$

with corresponding maximum likelihood estimator (MLE): $\hat{\boldsymbol{\beta}}_S := \text{argmin}_{\boldsymbol{\beta}_S} l(\boldsymbol{\beta}_S)$, whenever it exists. Next, define the weight diagonal matrix

$$\mathbf{D}(\boldsymbol{\beta}) := \text{diag}(b''(\eta_1(\boldsymbol{\beta})), \dots, b''(\eta_n(\boldsymbol{\beta})))^{1/2} / \phi^{1/2},$$

the standardized response:

$$\mathbf{V}(\boldsymbol{\beta}) := \frac{\mathbf{D}^{-1}(\boldsymbol{\beta})(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))}{\sqrt{n}\phi},$$

and the adjusted response:

$$\mathbf{Z}(\boldsymbol{\beta}) := \mathbf{D}(\boldsymbol{\beta})\boldsymbol{\eta}(\boldsymbol{\beta}) + \sqrt{n}\mathbf{V}(\boldsymbol{\beta}). \quad (2)$$

Further, define the projection matrix onto $\text{Span}\{\mathbf{D}(\boldsymbol{\beta})\mathbf{X}_S\}$:

$$\boldsymbol{\Pi}_S(\boldsymbol{\beta}) := \mathbf{D}(\boldsymbol{\beta})\mathbf{X}_S(\mathbf{X}_S^\top \mathbf{D}(\boldsymbol{\beta})^2 \mathbf{X}_S)^{-1} \mathbf{X}_S^\top \mathbf{D}(\boldsymbol{\beta}),$$

and the orthogonal complement $\boldsymbol{\Pi}_S^\perp(\boldsymbol{\beta}) := \mathbf{I}_n - \boldsymbol{\Pi}_S(\boldsymbol{\beta})$, which are both well-defined under the technical assumptions in the Appendix.

To keep the notation minimal, we will write $\hat{\mathbf{Z}}, \hat{\boldsymbol{\mu}}, \hat{\mathbf{D}}, \hat{\boldsymbol{\Pi}}_S^\perp$ as a shorthand for $\mathbf{Z}(\hat{\boldsymbol{\beta}}_S), \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_S), \mathbf{D}(\hat{\boldsymbol{\beta}}_S), \boldsymbol{\Pi}_S^\perp(\hat{\boldsymbol{\beta}}_S)$ at the MLE $\hat{\boldsymbol{\beta}}_S$. Similarly, we write $\mathbf{Z}, \boldsymbol{\mu}, \mathbf{D}, \boldsymbol{\Pi}_S^\perp$ as a shorthand notation for the values of $\mathbf{Z}(\boldsymbol{\beta}^*), \boldsymbol{\mu}(\boldsymbol{\beta}^*), \mathbf{D}(\boldsymbol{\beta}^*), \boldsymbol{\Pi}_S^\perp(\boldsymbol{\beta}^*)$ at the true coefficient vector $\boldsymbol{\beta}^*$. Note that under the null hypothesis it does not matter whether we evaluate all of the above at $\boldsymbol{\beta}^*$ or $\boldsymbol{\beta}_S^*$.

With this notation, we define the collinearity metric:

$$R_i := \frac{\langle \hat{\Pi}_{\mathcal{S}}^\perp \hat{\mathbf{D}} \mathbf{v}_i, \hat{\Pi}_{\mathcal{S}}^\perp \hat{\mathbf{Z}} \rangle}{\|\hat{\Pi}_{\mathcal{S}}^\perp \hat{\mathbf{D}} \mathbf{v}_i\| \|\hat{\Pi}_{\mathcal{S}}^\perp \hat{\mathbf{Z}}\|}, \quad \text{for all } i \in \mathcal{S}^c, \quad (3)$$

so that $R_i \in [-1, 1]$, whenever the norms in the denominator are not zero.

Generally, if $\mathbf{v}_i$ is an irrelevant model vector with no explanatory capacity, then we expect its collinearity with the residual $\hat{\Pi}_{\mathcal{S}}^\perp \hat{\mathbf{Z}}$ to be close to zero. Conversely, if $\mathbf{v}_i$ has explanatory capacity, its collinearity with the residual is expected to be significantly different from zero. This reasoning then leads us to the simple collinearity test statistic:

$$T_n := \max_{i \in \mathcal{S}^c} nR_i^2.$$

In order to use the collinearity test statistic, we need some understanding of its limiting distribution under the null hypothesis. To this end, we have the following result, whose proof is delegated to the Appendix.

Theorem 1 (Asymptotic distribution of nR_i^2 under the Null) *Assuming the null hypothesis is true, and under standard technical assumptions given in the Appendix that ensure that, for example, the MLE converges in probability ($\hat{\beta}_{\mathcal{S}} = \beta_{\mathcal{S}}^* + o_{\mathbb{P}}(1)$), the large-sample approximate distribution of each R_i^2 is $\text{Beta}(1/2, (n - |\mathcal{S}| - 1)/2)$. In other words, nR_i^2 converges in distribution to the χ_1^2 law:*

$$nR_i^2 \xrightarrow{d} \chi_1^2.$$

Note that, were the adjusted response (2) exactly Gaussian distributed (as is the case for the Gaussian linear model), then the $\text{Beta}(1/2, (n - |\mathcal{S}| - 1)/2)$ distribution in Theorem 1 is no longer the approximate large-sample distribution, but the exact finite- n distribution of R_i^2 .

In order to study the tail distribution of the collinearity statistic T_n , we use the following bounds (here $s = |\mathcal{S}|$):

$$\max_{i \in \mathcal{S}^c} \mathbb{P}[nR_i^2 > t] \leq \mathbb{P}[T_n > t] \leq (p - s) \max_{i \in \mathcal{S}^c} \mathbb{P}[nR_i^2 > t].$$

Note that, while the nR_i^2 have the same asymptotic χ_1^2 marginal distribution, they are not necessarily independent. Nevertheless, from Theorem 1 we can conclude that the $1 - \alpha/(p - s)$ quantile $q_{1-\alpha/(p-s)}$ of the χ_1^2 distribution guarantees that $\mathbb{P}[T_n > q_{1-\alpha/(p-s)}]$ will be less than $\alpha + \varepsilon$ for sufficiently large n (here $\varepsilon > 0$ is arbitrarily small).

The above discussion and Bonferroni quantile approximation is valid when $n \uparrow \infty$ and p is fixed, but not necessarily when $p = p_n > n$. This is because the asymptotic result in Theorem 1 does not necessarily hold uniformly over all $i \in \mathcal{S}^c$ and for a growing $p = p_n$.

When $p = p_n > n$ as $n \uparrow \infty$, the quantile $q_{1-\alpha/p_n}$ of the χ_1^2 distribution has the asymptotic approximation:

$$q(n) = 2 \ln \frac{p_n}{\alpha \sqrt{\pi}} - \ln \left[\ln \frac{2p_n}{\alpha} \right].$$

Under a few additional assumptions that ensure the uniformity of the χ_1^2 approximation over all $i \in \mathcal{S}^c$, the following result shows that the quantile approximation $q(n)$ of the χ_1^2 distribution also works for high-dimensional hypothesis testing.

Theorem 2 (Validity of χ_1^2 approximation in high dimensions.) *Assume that $p = p_n > n$ such that $\frac{(\ln p_n)^4}{n} \rightarrow 0$ as $n \uparrow \infty$. Then for a given small $\varepsilon > 0$ and under the additional “3. Stronger Lipschitz condition”, “4. Lindeberg projection decay rate”, and “5. Uniform quadratic approximation to cumulant” assumptions in the appendix, we have that*

$$\mathbb{P}[T_n \geq (1 + \varepsilon)q(n)] = o(1), \quad n \uparrow \infty.$$

In other words, irrespective of whether p is fixed or $p = p_n > n$, the interval $[(1 + \varepsilon)q_{1-\alpha/(p-s)}, \infty)$ is a large-sample conservative approximation of the critical or rejection region for our hypothesis test at the α level of significance. Our numerical experience is that these large-sample approximations are excellent when the p-values are small.

In case more precise estimates are needed of the $1 - \alpha$ quantile of T_n for small to moderate p , we recommend using Monte Carlo simulation. Namely, we simulate m iid vectors $\Theta_1, \dots, \Theta_m$ from $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and replace the adjusted response $\hat{\mathbf{Z}}$ in (3) with Θ_k for each k to obtain the Monte Carlo sample:

$$T_n^{[k]} = n \max_{i \in \mathcal{S}^c} \frac{\langle \hat{\Pi}_S^\perp \hat{\mathbf{D}} \mathbf{v}_i, \hat{\Pi}_S^\perp \Theta_k \rangle^2}{\|\hat{\Pi}_S^\perp \hat{\mathbf{D}} \mathbf{v}_i\|^2 \|\hat{\Pi}_S^\perp \Theta_k\|^2}, \quad k = 1, \dots, m.$$

We then estimate the large-sample $1 - \alpha$ quantile of T_n from the empirical distribution of $T_n^{[1]}, \dots, T_n^{[m]}$. The cost of the Monte Carlo approximation is essentially equivalent to m evaluations of (3), which cost $O(nmp)$ when using the QR decomposition of $\hat{\mathbf{D}} \mathbf{X}_S$. This is the approach we use in the next numerical section to estimate p-values larger than 0.01, for which the analytical bounds may be too loose.

Finally, we point out a classical method for computing the MLE $\hat{\beta}_S$ as the limit of a sequence $\beta_S^{(0)}, \beta_S^{(1)}, \dots$, generated by the iterative reweighted least squares (IRLS):

$$\beta_S^{(t+1)} = \underset{\beta_S}{\operatorname{argmin}} \|\mathbf{Z}(\beta_S^{(t)}) - \mathbf{D}(\beta_S^{(t)}) \mathbf{X}_S \beta_S\|^2, \quad t = 0, 1, \dots$$

Convergence is guaranteed provided the initial $\beta_S^{(0)}$ lies in a sufficiently small neighborhood of the MLE. Since both the iteration above and the statistic T_n use the function $\mathbf{Z}(\cdot)$, working with the collinearity test statistic is computationally as straightforward as applying the classical IRLS method to compute the MLE.

3 Numerical Comparison with LRT and Wald Test

In line with existing works on high-dimensional hypothesis testing for GLMs, we provide numerical examples for the logistic regression model using simulated datasets, where the true model coefficients β^* are known a priori [5–7]. This permits us to easily estimate the statistical power of any given test.

Recall that for the Logistic regression, the link function is $\ln(\mu \div (1 - \mu)) = \mathbf{X}\beta^*$, which is equivalent to stating that $\phi = 1, c(y_i, 1) = 0, b(x) = \ln(1 + \exp(x))$ in the exponential-family model (1). Hence, the conditional variance of the i -th response is $b''(\eta_i(\beta^*)) = \mu_i(1 - \mu_i)$, providing us with all the necessary formulas for the estimation of quantities in (2).

Simulation Setup for Logistic Regression.

Next, consider the following simple simulation setup. We let $n = 100$, $\mathcal{S} = \{1, 2, 3\}$, so that $s = 3$. We then vary p and the size of $|\mathcal{A}|$ under the alternative hypothesis. Varying $|\mathcal{A}|$ is essentially the same as changing the value of $\mathbf{X}\beta^*$. Here

$$\mathbf{X}\beta^* = \underbrace{-3\mathbf{v}_1 + \mathbf{v}_2 + 2\mathbf{v}_3}_{\text{null hypothesis}} + 2 \sum_{j=4}^{3+|\mathcal{A}|} \mathbf{v}_j,$$

where $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are constructed from iid draws from the uniform distribution on $[0, 1]$ and for the remaining model vectors under the alternative $\mathbf{v}_j \mid \mathbf{v}_1 \sim_{\text{iid}} \mathcal{N}(\mathbf{v}_1, \mathbf{I}_n)$ for $j \geq 4$. In this setting the term $2 \sum_{j=4}^{3+|\mathcal{A}|} \mathbf{v}_j$ represents the pollution or violation of the null hypothesis with model vectors from the set $\mathcal{A}$ under the alternative hypothesis.

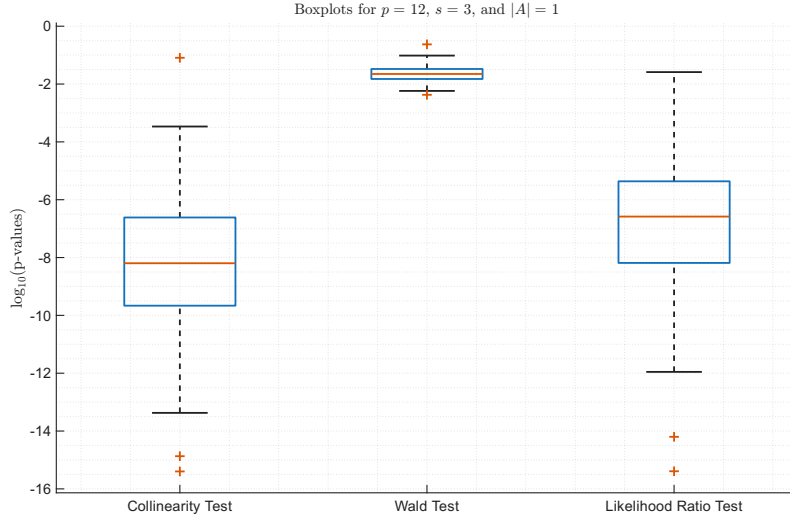


Fig. 1 Boxplots of the distribution of the p-values of the three tests on a logarithmic scale for a sparse alternative: $p = 12, s = 3, |\mathcal{A}| = 1$.

Classical Regime of relatively small p with $p = o(n)$.

We begin by considering the classical regime with a few predictors. Figure 1 shows the boxplots of the distribution of the p-values across 100 replications. Here the pollution of the null hypothesis is sparse, because $|\mathcal{A}| = 1$, that is, the null hypothesis has incorrectly omitted one predictor. All tests correctly reject the null hypothesis on average, but the collinearity and LRT tests produce the smallest p-values.

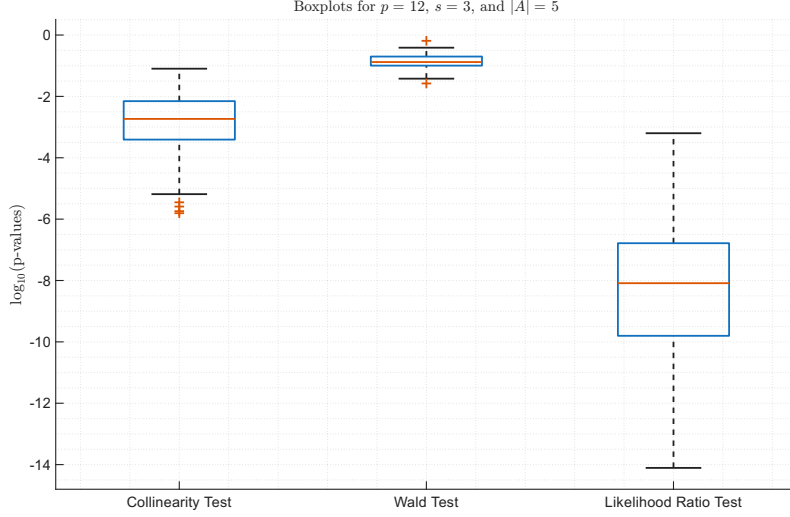


Fig. 2 Boxplots of the distribution of the p-values of the three tests on a logarithmic scale for a non-sparse alternative: $p = 12$, $s = 3$, $|\mathcal{A}| = 5$.

In contrast, Figure 2 shows the same situation, but with $|\mathcal{A}| = 5$. The power of the collinearity and Wald tests has diminished, but the power of the LRT test has improved. This is the setting for which the LRT asymptotic properties hold and, unsurprisingly, the LRT performs best.

Moderate Dimensions: $p/n \rightarrow \kappa \in [0, 1/2)$.

In this setting, we consider the case where the number of predictors is not negligible relative to n . In this case, the likelihood can become highly unstable and all estimators highly biased. For this reason here the LRT and Wald tests are computed from the widely-used stabilized/penalized log-likelihood that uses the Jeffreys' invariant prior penalty [10, 11]. This approach reduces small-sample bias for better overall performance. Despite this standard improvement, Figure 3 shows that the Wald test is not competitive and the LRT has substantially smaller statistical power compared to the collinearity test. The relative performance of the tests is unchanged when $|\mathcal{A}|$ is increased from $|\mathcal{A}| = 1$ to $|\mathcal{A}| = 6$; see Figure 4. However, note that a larger non-sparse alternative increases the (numerically estimated) p-values of all tests.

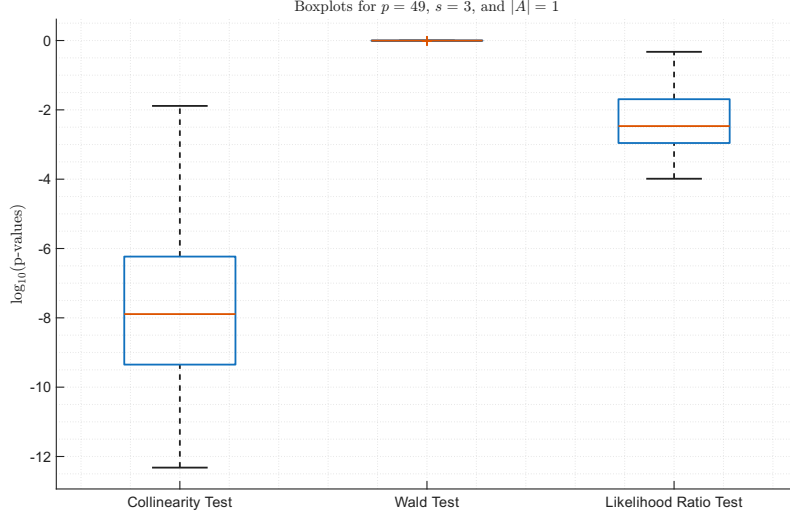


Fig. 3 Boxplots of the distribution of the p-values of the three tests on a logarithmic scale for a sparse alternative in moderate dimensions: $p = 49$, $s = 3$, $|A| = 1$.

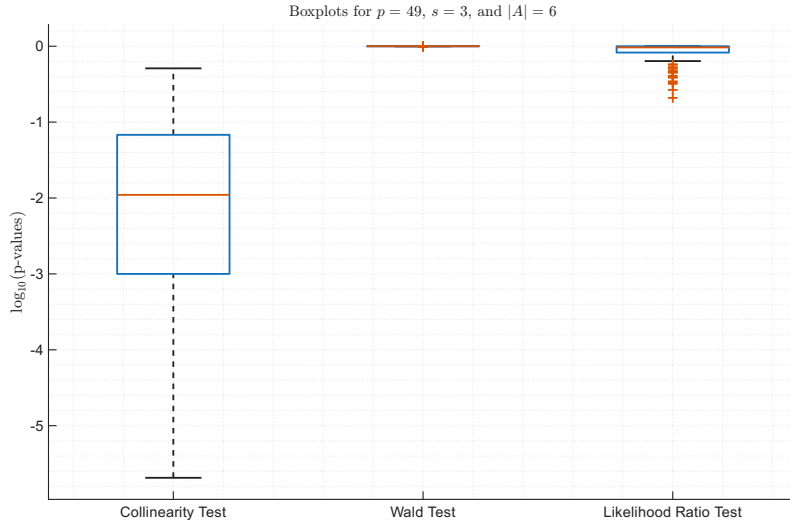


Fig. 4 Boxplots of the distribution of the p-values of the three tests on a logarithmic scale for a nonsparse alternative in moderate dimensions: $p = 49$, $s = 3$, $|A| = 6$.

High Dimensions: $p > n$.

In the high-dimensional case, the classical LRT and Wald methods are inapplicable, so we only report results for the collinearity test. Figure 5 illustrates the point that

the collinearity test works in the $p > n$ regime without any modification. Here we have $p = 1000 > 100 = n$ and the size of the hypothesized model is $s = 3$. We compute the p-values when the alternative is sparse ($|\mathcal{A}| = 1$) and not too sparse ($|\mathcal{A}| = 10$). The results in Figure 5 illustrate the observation that the power of the collinearity test appears to be higher when the alternative is sparse. In particular, the median p-value for $|\mathcal{A}| = 1$ is smaller than 10^{-6} , while for $|\mathcal{A}| = 10$ it is merely smaller than 10^{-1} . The p-values in the non-sparse case can be reduced by increasing n from 100 to, for example, $n = 500$ (not shown here).

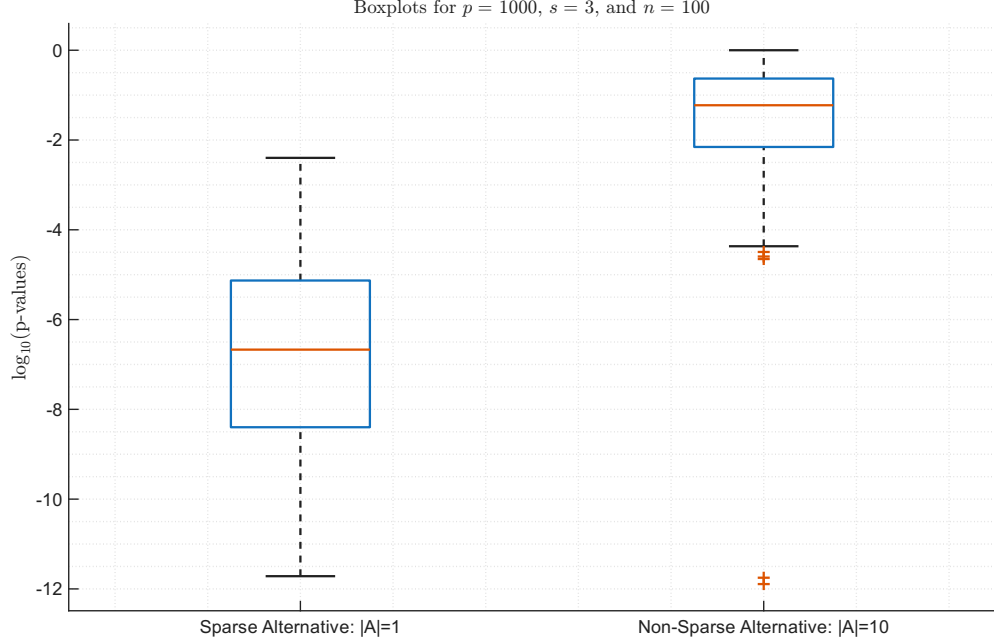


Fig. 5 Boxplots of the distribution of the p-values of the collinearity test for $p = 1000$, $s = 3$, $n = 100$ and with either a sparse ($|\mathcal{A}| = 1$) or non-sparse ($|\mathcal{A}| = 10$) alternative hypothesis.

Overall the Wald test is the worst performing, and although the collinearity test is not the most powerful in all the numerical examples we consider, in many of them it comes close to the power of the LRT.

How do we explain the observation that when the alternative is sparse ($|\mathcal{A}|$ is small), then the performance of the collinearity test is superior?

To explain this phenomenon, we consider the theoretical rate of decay of the Type 2 Error, under the simplest of all GLMs – the Gaussian linear model.

Theorem 3 (Type 2 Error Rate of Decay for Gaussian model) *For the Gaussian GLM with constant conditional response variance $\sigma_i^2 = \sigma^2$, let $q := q_{1-\frac{\alpha}{p-s}}$ and*

$$\boldsymbol{\nu} := \frac{1}{\sigma} \sum_{j \in \mathcal{A}} \beta_j^* \boldsymbol{\Pi}_S^\perp \mathbf{v}_j, \quad \mathcal{A} \subset \mathcal{S}^c$$

represent the deviation from the null hypothesis ($\|\boldsymbol{\nu}\| = 0$ when the null is true). Then,

$$\ln \mathbb{P}[T_n \leq q] \leq -\frac{\|\boldsymbol{\nu}\|_2^2}{2} + \Delta_n \sqrt{q} \|\boldsymbol{\nu}\|_2 + \frac{\Delta_n^2 q \|\boldsymbol{\nu}\|_2^2}{2n},$$

where $\Delta_n := \frac{1}{\sigma \|\boldsymbol{\nu}\|_2} \sum_{j \in \mathcal{A}} \|\beta_j^* \boldsymbol{\Pi}_S^\perp \mathbf{v}_j\|_2$ measures the “sparsity” of the alternative hypothesis.

To answer how Theorem 3 explains the greater power of the collinearity test when $|\mathcal{A}|$ is small, first note that $\Delta_n = \mathcal{O}_{\mathbb{P}}(\sqrt{|\mathcal{A}|})$ when $\{\beta_j^* \boldsymbol{\Pi}_S^\perp \mathbf{v}_j, j \in \mathcal{A}\}$ are either a set of orthogonal vectors, or $\{\mathbf{v}_j, j \in \mathcal{A}\}$ are iid $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ distributed. Since Δ_n can be thus linked to the size of $\mathcal{A}$, it can be interpreted as a measure of the “sparsity” of the alternative hypothesis. In fact, by Minkowski’s inequality $\Delta_n \geq 1$, so that the sparsest possible alternative $|\mathcal{A}| = 1$ corresponds to $\Delta_n = 1$. In general, Δ_n takes into account not only the sparsity of $\mathcal{A}$, but also the alignment and geometry of the vectors $\{\beta_j^* \boldsymbol{\Pi}_S^\perp \mathbf{v}_j, j \in \mathcal{A}\}$.

Second, if we make the natural assumption that $\|\boldsymbol{\nu}\|_2 = \Theta(\sqrt{n})$ and $\Delta_n^2 q = o(n)$, then Theorem 3 implies that there are finite positive constants c_1 and c_2 such that for large enough n :

$$\ln \mathbb{P}[T_n \leq q] \leq -c_1 n + c_2 \Delta_n \sqrt{q} \sqrt{n}.$$

In other words, while the leading order rate of decay is $\mathcal{O}(n)$, the second order decay is $\mathcal{O}(\sqrt{n})$ with Δ_n appearing as a multiplicative factor in the second order term. As a result, large values of Δ_n reduce the overall rate of decay of the Type 2 Error.

The above result not only explains why the collinearity test is powerful when Δ_n is small, but also why the test still works in the high-dimensional case of $p_n > n$ with $(\ln p_n)/n \rightarrow 0$. If in Theorem 3 we replace $q_{1-\frac{\alpha}{p-s}}$ with $q(n) = \mathcal{O}(\ln p_n)$, then for a bounded alternative ($\Delta_n = \mathcal{O}(1)$), we can write

$$\begin{aligned} \ln \mathbb{P}[T_n \leq q(n)] &\leq -c_1 n + c_2 \Delta_n \sqrt{q(n)} \sqrt{n} \\ &= -n \left(c_1 - \mathcal{O} \left(\sqrt{\frac{\ln p_n}{n}} \right) \right) \\ &= -n(c_1 - o(1)), \end{aligned}$$

indicating that even in the $p_n > n$ case, the leading rate of decay of the Type 2 Error can remain $\mathcal{O}(n)$.

The above theoretical results on the rate of decay of the Type 2 Error are established under the simplest GLM with identity link function and Gaussian error. Whether similar results hold asymptotically for the statistical power of the collinearity test for the logistic or Poisson regression is an open problem, which we intend to investigate in a follow-up article.

4 Conclusion

In this article we introduced a novel collinearity statistic for testing whether important predictors are missing from a hypothesized GLM. An advantage of our proposal is that the same test applies across different dimensional regimes, regardless of whether the number of predictors exceeds the number of observations ($p > n$) or vice versa ($p < n$). The collinearity test does not require optimization of nonconvex penalized log-likelihoods and is generally as straightforward to implement as the iterative reweighted least squares algorithm used to compute the MLE under the null hypothesis.

The numerical results indicate that our proposal compares favorably with the classical LRT and Wald tests. We observe that the collinearity test has its greatest power when the alternative hypothesis is sparse, in the sense that only a few important predictors are missing. This observation is consistent with our theoretical result on the power of the collinearity test in the simplest setting — the Gaussian linear model.

In future work, we intend to compare the collinearity test, both numerically and theoretically, with more specialized penalized log-likelihood methods [7, 8]. Although these are more difficult to implement and more computationally intensive, they remain viable alternatives. In addition, it remains to be investigated whether the power result in Theorem 3 can be extended to non-Gaussian GLMs, such as logistic and Poisson regression.

Appendix A Technical Arguments

In the following arguments, for a given coefficient vector $\beta \in \mathbb{R}^p$ and an arbitrary choice of $i \in \mathcal{S}^c$, we define the functions:

$$\begin{aligned} d_i(\beta) &:= \frac{\Pi_{\mathcal{S}}^\perp(\beta) \mathbf{D}(\beta) \mathbf{v}_i}{\|\Pi_{\mathcal{S}}^\perp(\beta) \mathbf{D}(\beta) \mathbf{v}_i\|_2} \\ F_{n,i}(\beta) &:= \frac{\langle d_i(\beta), \sqrt{n} \mathbf{V}(\beta) \rangle}{\|\mathbf{V}(\beta)\|}. \end{aligned}$$

Assumptions

We use the following technical assumptions.

1. **Fixed Model Size.** The set $\mathcal{S}$ has fixed size $|\mathcal{S}|$, independent of n . Under the null hypothesis $\beta_{\mathcal{S}^c}^* = \mathbf{0}$ and $\beta_i^* \neq 0$ for $i \in \mathcal{S}$.
2. **Bounded eigenvalues.** Assume a regularity condition from Fahrmeir and Kaufmann [12] :

$$0 < \underline{\lambda} \leq \frac{\|\mathbf{D}(\beta_{\mathcal{S}}) \mathbf{X}_{\mathcal{S}} \mathbf{v}\|^2}{n \|\mathbf{v}\|^2} \leq \bar{\lambda} < \infty,$$

for all sufficiently large n , all $\mathbf{v} \neq \mathbf{0}$, for some constants $\underline{\lambda}$ and $\bar{\lambda}$, and for all $\beta_{\mathcal{S}}$ in a fixed neighborhood of $\beta_{\mathcal{S}}^*$. This ensures, for example, that $\hat{\Pi}_{\mathcal{S}}$ and its orthogonal complement are well-defined.

3. Uniform Lipschitz Continuity. There exists a finite constant c such that:

$$\sup_i |F_{n,i}(\beta_S) - F_{n,i}(\tilde{\beta}_S)| \leq c \|\beta_S - \tilde{\beta}_S\|_2.$$

This condition can be deduced from more unwieldy and ponderous assumptions that include, for example, bounded fourth derivatives for $b(\cdot)$.

4. Lindeberg Projection Condition. As $n \uparrow \infty$, the irrelevant model vectors satisfy:

$$\max_{i \in \mathcal{S}^c} \frac{\|\Pi_S^\perp \mathbf{D} \mathbf{v}_i\|_\infty}{\|\Pi_S^\perp \mathbf{D} \mathbf{v}_i\|_2} = \max_{i \in \mathcal{S}^c} \|\mathbf{d}_i(\beta_S^*)\|_\infty \rightarrow 0.$$

In other words, the Lindeberg projection condition requires that the part of each $\mathbf{v}_i$ that lies outside of $\text{Span}\{\mathbf{D}\mathbf{X}_S\}$ must not have a coordinate that dominates all others.

5. Uniform Square Integrability of Response. The sequence of standardized regression responses $(\sqrt{n} V_i(\beta_S^*), i \geq 1)$ is (conditional on $\mathbf{X}$) uniformly square integrable, that is,

$$\limsup_{b \uparrow \infty} \mathbb{E}[n V_i^2(\beta_S^*) \mathbb{I}\{\sqrt{n} |V_i(\beta_S^*)| > b\}] = 0.$$

This condition naturally already holds for the Gaussian, Bernoulli and Poisson GLMs with uniformly bounded linear predictors, but may be needed if we consider a more exotic GLM.

Proof of Theorem 1

From the score equation $\mathbf{X}_S^\top (\mathbf{Y} - \hat{\boldsymbol{\mu}}) = \mathbf{0}$, we have

$$\begin{aligned} \hat{\Pi}_S^\perp \hat{\mathbf{Z}} &= \underbrace{\hat{\Pi}_S^\perp \hat{\mathbf{D}} \mathbf{X}_S \hat{\beta}_S}_{=0} + \hat{\Pi}_S^\perp \hat{\mathbf{D}}^{-1} (\mathbf{Y} - \hat{\boldsymbol{\mu}}) / \phi \\ &= \hat{\mathbf{D}}^{-1} (\mathbf{Y} - \hat{\boldsymbol{\mu}}) / \phi - \hat{\mathbf{D}} \mathbf{X}_S (\mathbf{X}_S^\top \hat{\mathbf{D}}^2 \mathbf{X}_S)^+ \underbrace{\mathbf{X}_S^\top (\mathbf{Y} - \hat{\boldsymbol{\mu}})}_{=0} / \phi \\ &= \sqrt{n} \mathbf{V}(\hat{\beta}_S). \end{aligned}$$

Therefore,

$$\sqrt{n} R_i = \frac{\langle \mathbf{d}_i(\hat{\beta}_S), \sqrt{n} \mathbf{V}(\hat{\beta}_S) \rangle}{\|\mathbf{V}(\hat{\beta}_S)\|_2} = F_{n,i}(\hat{\beta}_S).$$

Thus, the desired result $\sqrt{n} R_i \xrightarrow{d} \mathcal{N}(0, 1)$ is established using Slutsky's theorem [13, Theorem D.79, Page 651] if we show that under the null hypothesis: $F_{n,i}(\hat{\beta}_S) - F_{n,i}(\beta_S^*) = o_{\mathbb{P}}(1)$ and $F_{n,i}(\beta_S^*) \xrightarrow{d} \mathcal{N}(0, 1)$, as $n \uparrow \infty$.

Showing that $F_{n,i}(\widehat{\beta}_{\mathcal{S}}) - F_{n,i}(\beta_{\mathcal{S}}^*) = o_{\mathbb{P}}(1)$.

Assumptions 1 and 2 imply [12, Theorem 1] that $\widehat{\beta}_{\mathcal{S}} = \beta_{\mathcal{S}}^* + o_{\mathbb{P}}(\mathbf{1})$. Therefore, using Assumptions 1 and 3, we obtain:

$$\sup_i |F_{n,i}(\widehat{\beta}_{\mathcal{S}}) - F_{n,i}(\beta_{\mathcal{S}}^*)| \leq c \|\widehat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}^*\|_2 = o_{\mathbb{P}}(1).$$

Showing that $F_{n,i}(\beta_{\mathcal{S}}^*) \xrightarrow{d} \mathcal{N}(0, 1)$.

For simplicity of notation, for a given $i \in \mathcal{S}^c$ write $\mathbf{u} := \mathbf{d}_i(\beta_{\mathcal{S}}^*)$ and

$$W_i := u_i \sqrt{n} V_i(\beta_{\mathcal{S}}^*) = \frac{u_i}{\sqrt{\sum_{j=1}^n u_j^2}} \sqrt{n} V_i(\beta_{\mathcal{S}}^*), \quad i = 1, \dots, n.$$

Then, $W_1, \dots, W_n$ are independent zero-mean random variables and Assumption 4 implies the Lindeberg's condition [14] for $\mathbf{u}$:

$$\frac{\max_{i \leq n} |u_i|}{\sqrt{\sum_{j=1}^n u_j^2}} \rightarrow 0, \quad n \uparrow \infty.$$

Since by Assumption 5 the sequence $(nV_i^2(\beta_{\mathcal{S}}^*), i \geq 1)$ is uniformly integrable, we have:

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}[W_i^2 \mathbb{I}\{|W_i| > \epsilon\}] &= \sum_{i=1}^n \frac{u_i^2}{\sum_{j=1}^n u_j^2} \mathbb{E}\left[nV_i^2(\beta_{\mathcal{S}}^*) \mathbb{I}\left\{|\sqrt{n}V_i(\beta_{\mathcal{S}}^*)| > \epsilon \frac{\sqrt{\sum_{j=1}^n u_j^2}}{|u_i|}\right\}\right] \\ &\leq \sup_{i \geq 1} \mathbb{E}\left[nV_i^2(\beta_{\mathcal{S}}^*) \mathbb{I}\left\{|\sqrt{n}V_i(\beta_{\mathcal{S}}^*)| > \epsilon \frac{\sqrt{\sum_{j=1}^n u_j^2}}{\max_{i \leq n} |u_i|}\right\}\right] \rightarrow 0, \quad n \uparrow \infty. \end{aligned}$$

Hence, by the CLT result of Lindeberg [14], we conclude that

$$\sum_i W_i = \sqrt{n} \langle \mathbf{u}, \mathbf{V}(\beta_{\mathcal{S}}^*) \rangle \xrightarrow{d} \mathcal{N}(0, 1).$$

In addition, by the Law of Large Numbers, $\|\mathbf{V}(\beta_{\mathcal{S}}^*)\|_2^2 = 1 + o_{\mathbb{P}}(1)$. Finally, by Slutsky's theorem [13, Theorem D.79, Page 651], we can write:

$$F_{n,i}(\beta_{\mathcal{S}}^*) = \frac{\sqrt{n} \langle \mathbf{u}, \mathbf{V}(\beta_{\mathcal{S}}^*) \rangle}{\|\mathbf{V}(\beta_{\mathcal{S}}^*)\|} = \sqrt{n} \langle \mathbf{u}, \mathbf{V}(\beta_{\mathcal{S}}^*) \rangle + o_{\mathbb{P}}(1) \xrightarrow{d} \mathcal{N}(0, 1).$$

Proof of Theorem 2

We introduce the following simplifying notation: Define

$$\boldsymbol{\epsilon} := \sqrt{n} \mathbf{V}(\beta^*),$$

so that $\epsilon_1, \dots, \epsilon_n$ are independent zero mean with unit variance. Since the $Y_1, Y_2, \dots$ have density (1), we conclude that the cumulant generating function of ϵ_i is thus:

$$\psi_i(t) = \ln \mathbb{E} \exp(t\epsilon_i) = -\frac{t\mu_i}{\sigma_i} + \frac{b(\eta_i(\beta^*) + \phi t/\sigma_i) - b(\eta_i(\beta^*))}{\phi}.$$

We still use assumptions 1 to 5 (plus $p_n > n$ with $(\ln p_n)^4/n \rightarrow 0$ as $n \uparrow \infty$), but assumptions 3 to 5 are made more stringent, as follows:

3. Stronger Lipschitz Condition. The MLE and the true β^* satisfy $\|\hat{\beta}_S - \beta_S^*\|_2 = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$, so that the uniform Lipschitz continuity becomes:

$$\sup_i |F_{n,i}(\hat{\beta}_S) - F_{n,i}(\beta_S^*)| \leq c \|\hat{\beta}_S - \beta_S^*\|_2 = \mathcal{O}_{\mathbb{P}}(n^{-1/2}).$$

4. Lindeberg Projection Decay Rate. We have $\max_{i \in S^c} \|\mathbf{d}_i(\beta_S^*)\|_{\infty} = \mathcal{O}(\sqrt{\ln(p_n)/n})$. This condition is quite natural, because $\max_{i \leq p} \frac{\|\Theta_i\|_{\infty}}{\|\Theta_i\|_2} = \mathcal{O}_{\mathbb{P}}(\sqrt{\ln(np)/n})$ for standard Gaussian iid vectors $\Theta_1, \Theta_2, \dots$ in $\mathbb{R}^n$.

5. Uniform Quadratic approximation to Cumulant. Instead of the uniform (square) integrability of the scaled response ϵ , we require

$$\sup_i \left| \psi_i(t) - \frac{t^2}{2} \right| \leq c|t^3|, \quad t \in [-\delta, \delta]$$

for some positive constants c and δ . This condition holds in, for example, logistic and Poisson regression when the linear predictors are uniformly bounded: $\sup_i |\eta_i(\beta^*)| < \text{const.} < \infty$.

Proof: We first show that using $\|\epsilon\|/\sqrt{n} = 1 + \mathcal{O}_{\mathbb{P}}(n^{-1/2})$ and assumption 3, we have

$$\mathbb{P}[n \max_{i \in S^c} R_i^2 \geq (1 + \varepsilon)q(n)] \leq \mathbb{P}[\max_{i \in S^c} |\langle \mathbf{d}_i(\beta_S^*), \epsilon \rangle| > a_n] + o(1),$$

where

$$a_n := (\sqrt{(1 + \varepsilon)q(n)} - n^{-1/4})(1 - n^{-1/4}) = \sqrt{(1 + \varepsilon)q(n)} + \underbrace{\mathcal{O}\left(\frac{\ln p_n}{n^{1/4}}\right)}_{o(1)}.$$

To see this define the event

$$A_n := \left\{ \left| \|\epsilon\|/\sqrt{n} - 1 \right| \leq n^{-1/4}, \sup_i |F_{n,i}(\hat{\beta}_S) - F_{n,i}(\beta_S^*)| \leq n^{-1/4} \right\}.$$

From assumption 3, we have $\mathbb{P}[A_n^c] \rightarrow 0$ as $n \rightarrow \infty$. Therefore,

$$\begin{aligned} \mathbb{P}[\sqrt{n} \max_{i \in S^c} |R_i| \geq \sqrt{(1 + \varepsilon)q(n)}] &= \mathbb{P}[\max_{i \in S^c} |F_{n,i}(\hat{\beta}_S)| \geq \sqrt{(1 + \varepsilon)q(n)}] \\ &\leq \mathbb{P}[\max_{i \in S^c} |\langle \mathbf{d}_i(\beta_S^*), \epsilon \rangle| \geq a_n] + \mathbb{P}[A_n^c], \end{aligned}$$

establishing the claim.

Second, we show that under the assumptions 4 and 5:

$$\max_{i \in \mathcal{S}^c} \mathbb{P} \left[|\langle \mathbf{d}_i(\boldsymbol{\beta}_{\mathcal{S}}^*), \boldsymbol{\epsilon} \rangle| > a_n \right] \leq p_n^{-(1+\varepsilon+o(1))} = o(p_n^{-1}).$$

This follows from a standard Chernoff bound using the properties of the cumulant function of $\sum_j \epsilon_j d_{i,j}(\boldsymbol{\beta}_{\mathcal{S}}^*)$:

$$\begin{aligned} \ln \mathbb{P}[\langle \mathbf{d}_i(\boldsymbol{\beta}_{\mathcal{S}}^*), \boldsymbol{\epsilon} \rangle > a_n] &\leq -ta_n + \sum_j \psi_j(t d_{i,j}) \\ &= -ta_n + \frac{t^2}{2} \|\mathbf{d}_i(\boldsymbol{\beta}_{\mathcal{S}}^*)\|^2 + \mathcal{O}\left(t^3 \sum_j |d_{i,j}|^3\right) \\ &= -ta_n + \frac{t^2}{2} + \mathcal{O}\left(t^3 \|\mathbf{d}_i(\boldsymbol{\beta}_{\mathcal{S}}^*)\|_{\infty}\right). \end{aligned}$$

Setting $t = a_n$ and using $t^3 \max_{i \in \mathcal{S}^c} \|\mathbf{d}_i(\boldsymbol{\beta}_{\mathcal{S}}^*)\|_{\infty} = \mathcal{O}(\sqrt{(\ln p_n)^4/n}) = o(1)$, we thus obtain the desired result from:

$$\max_{i \in \mathcal{S}^c} \ln \mathbb{P}[\langle \mathbf{d}_i(\boldsymbol{\beta}_{\mathcal{S}}^*), \boldsymbol{\epsilon} \rangle > a_n] \leq -a_n^2/2 + o(1).$$

It then follows that

$$\begin{aligned} \mathbb{P}[T_n \geq (1+\varepsilon)q(n)] &\leq \sum_{i \in \mathcal{S}^c} \mathbb{P}[|F_{n,i}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}})| \geq \sqrt{(1+\varepsilon)q(n)}] \\ &\leq p_n \max_{i \in \mathcal{S}^c} \mathbb{P}\left[|F_{n,i}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}})| \geq \sqrt{(1+\varepsilon)q(n)}\right] \\ &\leq o(1) + p_n \max_{i \in \mathcal{S}^c} \mathbb{P}\left[|\langle \mathbf{d}_i(\boldsymbol{\beta}_{\mathcal{S}}^*), \boldsymbol{\epsilon} \rangle| \geq a_n\right] \\ &\leq o(1) + \mathcal{O}(p_n^{-\varepsilon}) = o(1), \quad n \uparrow \infty. \end{aligned}$$

Proof of Theorem 3

Recall that $\boldsymbol{\Pi}_{\mathcal{S}}^{\perp}$ and $\mathbf{D}$ stand for $\boldsymbol{\Pi}_{\mathcal{S}}^{\perp}(\boldsymbol{\beta}^*)$ and $\mathbf{D}(\boldsymbol{\beta}^*)$. In the linear Gaussian case with constant conditional variance $\sigma_i^2 = \sigma^2$ for all i , we have that

$$\begin{aligned} \boldsymbol{\mu}(\boldsymbol{\beta}) &= \boldsymbol{\eta}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta} \\ \phi &= \sigma^2 \\ \mathbf{D}(\boldsymbol{\beta}) &= \mathbf{I}_n/\sigma \\ \widehat{\boldsymbol{\Pi}}_{\mathcal{S}}^{\perp} \widehat{\mathbf{Z}} &= \sqrt{n} \mathbf{V}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}) = \boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon} + \boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \mathbf{X}_{\mathcal{A}} \boldsymbol{\beta}_{\mathcal{A}}^*/\sigma, \end{aligned}$$

where $\boldsymbol{\epsilon} = \sqrt{n} \mathbf{V}(\boldsymbol{\beta}^*) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. For simplicity of notation, let

$$\mathbf{w}_j := \boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \mathbf{v}_j/\sigma, \quad j \in \mathcal{S}^c$$

so that

$$\boldsymbol{\nu} = \sum_{j \in \mathcal{A}} \beta_j^* \mathbf{w}_j, \quad \mathcal{A} \subset \mathcal{S}^c, \quad \widehat{\boldsymbol{\Pi}}_{\mathcal{S}}^\perp \widehat{\mathbf{Z}} = \boldsymbol{\Pi}_{\mathcal{S}}^\perp \boldsymbol{\epsilon} + \boldsymbol{\nu},$$

and

$$\|\boldsymbol{\nu}\|^2 = \|\boldsymbol{\Pi}_{\mathcal{S}}^\perp \mathbf{X} \boldsymbol{\beta}^*\|^2 / \sigma^2.$$

Since $\boldsymbol{\epsilon}$ is standard Gaussian distributed, the cumulant function of $\boldsymbol{\epsilon}$ satisfies:

$$\sum_{i=1}^n \psi_i(t_i) = \frac{\|\mathbf{t}\|^2}{2}, \quad \forall \mathbf{t}.$$

The proof will use the method of exponential tilting, for which the above property of the cumulant function is crucial. An exponentially-tilted version of the density $f(\boldsymbol{\epsilon}_i)$ is:

$$f_{t_i}(\boldsymbol{\epsilon}_i) = \exp(t_i \boldsymbol{\epsilon}_i - \psi_i(t_i)) f(\boldsymbol{\epsilon}_i),$$

where t_i is a valid tilting parameter. Note that under tilted distribution $\mathbb{E}_{t_i} \boldsymbol{\epsilon}_i = \psi'_i(t_i)$. In the following arguments we will choose the tilting parameter value

$$\mathbf{t} = -\boldsymbol{\nu},$$

which is equivalent to shifting the mean of the Gaussian by $\mathbf{t} = -\boldsymbol{\nu} = -\boldsymbol{\Pi}_{\mathcal{S}}^\perp \boldsymbol{\nu}$. With this tilting, we have for any quantile q

$$\begin{aligned} \mathbb{P}[T_n \leq q] &= \mathbb{E}_{\mathbf{0}} \prod_{i \in \mathcal{S}^c} \mathbb{I}\left\{ \langle \mathbf{w}_i, \boldsymbol{\Pi}_{\mathcal{S}}^\perp \boldsymbol{\epsilon} + \boldsymbol{\nu} \rangle^2 \leq \frac{q}{n} \|\mathbf{w}_i\|^2 \|\boldsymbol{\Pi}_{\mathcal{S}}^\perp \boldsymbol{\epsilon} + \boldsymbol{\nu}\|^2 \right\} \\ &= \mathbb{E}_{\mathbf{t}} \exp\left(-\langle \mathbf{t}, \boldsymbol{\epsilon} \rangle + \sum_i \psi_i(t_i)\right) \prod_{i \in \mathcal{S}^c} \mathbb{I}\left\{ \langle \mathbf{w}_i, \boldsymbol{\Pi}_{\mathcal{S}}^\perp \boldsymbol{\epsilon} + \boldsymbol{\nu} \rangle^2 \leq \frac{q}{n} \|\mathbf{w}_i\|^2 \|\boldsymbol{\Pi}_{\mathcal{S}}^\perp \boldsymbol{\epsilon} + \boldsymbol{\nu}\|^2 \right\} \\ &= \exp(-\|\boldsymbol{\nu}\|^2/2) \mathbb{E}_{\mathbf{t}} \exp(-\langle \mathbf{t}, \boldsymbol{\Pi}_{\mathcal{S}}^\perp \boldsymbol{\epsilon} + \boldsymbol{\nu} \rangle) \mathbb{I}\{(\boldsymbol{\Pi}_{\mathcal{S}}^\perp \boldsymbol{\epsilon} + \boldsymbol{\nu}) \in \mathcal{C}_n\} \\ &= \exp(-\|\boldsymbol{\nu}\|^2/2) \mathbb{E}_{\mathbf{0}} \exp(-\langle \mathbf{t}, \boldsymbol{\Pi}_{\mathcal{S}}^\perp \boldsymbol{\epsilon} \rangle) \mathbb{I}\{\boldsymbol{\Pi}_{\mathcal{S}}^\perp \boldsymbol{\epsilon} \in \mathcal{C}_n\} \end{aligned}$$

where $\mathcal{C}_n := \{\mathbf{z} : |\langle \mathbf{w}_i, \mathbf{z} \rangle| \leq \sqrt{q/n} \|\mathbf{w}_i\| \|\mathbf{z}\|, \forall i \in \mathcal{S}^c\}$. Now, on the set $\mathcal{C}_n$ we have the bound:

$$|\langle \mathbf{t}, \boldsymbol{\Pi}_{\mathcal{S}}^\perp \boldsymbol{\epsilon} \rangle| = |\langle \boldsymbol{\nu}, \boldsymbol{\Pi}_{\mathcal{S}}^\perp \boldsymbol{\epsilon} \rangle| \leq \sum_{j \in \mathcal{A}} |\beta_j^* \langle \mathbf{w}_j, \boldsymbol{\Pi}_{\mathcal{S}}^\perp \boldsymbol{\epsilon} \rangle| \leq \frac{\|\boldsymbol{\epsilon}\|}{\sqrt{n}} \underbrace{\sqrt{q} \sum_{j \in \mathcal{A}} |\beta_j^*| \|\mathbf{w}_j\|}_{=: u_n}.$$

Thus, using the result for $X \sim \chi_n^2$ (see auxiliary Lemma below for proof):

$$\mathbb{E} \exp(u \sqrt{X/n}) \leq \exp(u + u^2/(2n)), \quad u \geq 0,$$

we have:

$$\mathbb{E}_{\mathbf{0}} \exp(-\langle \mathbf{t}, \boldsymbol{\Pi}_{\mathcal{S}}^\perp \boldsymbol{\epsilon} \rangle) \mathbb{I}\{\boldsymbol{\Pi}_{\mathcal{S}}^\perp \boldsymbol{\epsilon} \in \mathcal{C}_n\} \leq \mathbb{E}_{\mathbf{0}} \exp(u_n \|\boldsymbol{\epsilon}\|/\sqrt{n}) \leq \exp(u_n + u_n^2/(2n)).$$

Therefore,

$$\ln \mathbb{P}[T_n \leq q] \leq -\frac{\|\boldsymbol{\nu}\|^2}{2} + u_n + \frac{u_n^2}{2n},$$

completing the proof.

Lemma: $\mathbb{E} \exp(u\sqrt{X/n}) \leq \exp(u + u^2/(2n))$ for $X \sim \chi_n^2$.

This result is not original, but we provide its proof for completeness. Let $\boldsymbol{\Theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. Recall that if $f : \mathbb{R}^n \mapsto \mathbb{R}$ is L -Lipschitz continuous, that is, $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2$, then $f(\boldsymbol{\Theta})$ has a sub-Gaussian distribution [15] with parameter L :

$$\mathbb{E} \exp(u[f(\boldsymbol{\Theta}) - \mathbb{E}f(\boldsymbol{\Theta})]) \leq \exp((Lu)^2/2).$$

Since $\|\mathbf{x}\|_2 - \|\mathbf{y}\|_2 \leq \|\mathbf{x} - \mathbf{y}\|_2$, the Euclidean norm $\|\cdot\|_2$ is 1-Lipschitz continuous, so that

$$\mathbb{E} \exp(u\|\boldsymbol{\Theta}\|_2) \leq \exp(u^2/2 + u\mathbb{E}\|\boldsymbol{\Theta}\|_2).$$

Since X has the same distribution as $\|\boldsymbol{\Theta}\|_2^2$ and $\mathbb{E}\|\boldsymbol{\Theta}\|_2 \leq \sqrt{\mathbb{E}\|\boldsymbol{\Theta}\|_2^2} = \sqrt{n}$ we thus obtain:

$$\mathbb{E} \exp(u\sqrt{X}) = \mathbb{E} \exp(u\|\boldsymbol{\Theta}\|_2) \leq \exp(u^2/2 + u\sqrt{n}).$$

References

- [1] Fisher, R.A.: Statistical methods for research workers. (1934)
- [2] Neyman, J., Pearson, E.S.: On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **231**, 289–337 (1933)
- [3] Wilks, S.S.: The large-sample distribution of the likelihood ratio for testing composite hypotheses. The annals of mathematical statistics **9**(1), 60–62 (1938)
- [4] McCullagh, P., Nelder, J.A.: Generalized Linear Models. Chapman and Hall/CRC, Boca Raton, FL (2019)
- [5] Sur, P., Candès, E.J.: A modern maximum-likelihood theory for high-dimensional logistic regression. Proceedings of the National Academy of Sciences **116**(29), 14516–14525 (2019)
- [6] Zhao, Q., Sur, P., Candès, E.J.: The asymptotic distribution of the mle in high-dimensional logistic models: Arbitrary covariance. Bernoulli **28**(3), 1835–1861 (2022)
- [7] Ma, R., Tony Cai, T., Li, H.: Global and simultaneous hypothesis testing for high-dimensional logistic regression models. Journal of the American Statistical Association **116**(534), 984–998 (2021)
- [8] Shi, C., Song, R., Chen, Z., Li, R.: Linear hypothesis testing for high dimensional generalized linear models. Annals of statistics **47**(5), 2671 (2019)

- [9] Hill, J.B.: Testing many zero restrictions in a high dimensional linear regression setting. *Journal of Business & Economic Statistics* **43**(1), 55–67 (2025)
- [10] Kosmidis, I., Firth, D.: Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika* **108**(1), 71–82 (2021)
- [11] Firth, D.: Bias reduction of maximum likelihood estimates. *Biometrika*, 27–38 (1993)
- [12] Fahrmeir, L., Kaufmann, H.: Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics* **13**(1), 342–368 (1985)
- [13] Botev, Z.I., Kroese, D.P., Taimre, T.: *Data Science and Machine Learning: Mathematical and Statistical Methods*, 2nd edn. CRC Press, Boca Raton, FL (2025)
- [14] Lindeberg, J.W.: Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* **15**, 211–225 (1922)
- [15] Boucheron, S., Lugosi, G., Massart, P.: *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, ??? (2013). See Sections 5.3 and 10.5