Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control

BY JELLE J. GOEMAN, HANS C. VAN HOUWELINGEN

Department of Medical Statistics and Bioinformatics, Leiden University Medical Centre, Postzone S5-P, Postbus 9600, 2300 RC Leiden, The Netherlands
j.j.goeman@lumc.nl  h.c.van_houwelingen@lumc.nl

AND LIVIO FINOS

Department of Statistical Sciences, University of Padua, Via Cesare Battisti 241, 35121 Padova, Italy
livio@stat.unipd.it

SUMMARY

Testing a low-dimensional null hypothesis against a high-dimensional alternative in a generalized linear model may lead to a test statistic that is a quadratic form in the residuals under the null model. Using asymptotic arguments, we show that the distribution of such a test statistic can be approximated by a ratio of quadratic forms in normal variables, for which algorithms are readily available. For generalized linear models, the asymptotic distribution shows good control of type I error for moderate to small samples, even when the number of covariates in the model far exceeds the sample size.

Some key words: Global test; High-dimensional data; Logistic regression; Score test.

1. Introduction

For the problem of testing a low-dimensional null hypothesis against a high-dimensional alternative in generalized linear models, Goeman et al. (2004, 2006) proposed a score test statistic that is a quadratic form in the residuals of the null model. This statistic was shown to be locally most powerful on average in a neighbourhood of the null hypothesis. Unlike classical tests, this test does not degenerate when the number of model parameters under the alternative exceeds the sample size, making the test useful for applications in genetics and genomics (Goeman et al., 2004; Liu et al., 2007; Pan, 2009). However, the theory of Goeman et al. (2006) only suggests a test statistic to be used, but does not give a general null distribution. Applications of this test (Goeman et al., 2004) resorted to permutation testing or moment matching to calculate p-values.

This paper derives an exact finite sample distribution for the test in the linear model with normal errors and an asymptotic distribution in generalized linear models with canonical link functions. Accounting for estimation of nuisance parameters leads to an approximate distribution, which is that of a ratio of quadratic forms in normal variables.

A quadratic form in the residuals of a generalized linear model has been proposed for several other applications, e.g., in the variance components literature (Verbeke & Molenberghs, 2003), in goodness-of-fit testing (Azzalini & Bowman, 1993; Le Cessie & Van Houwelingen,
Locally Most Powerful Tests in Generalized Linear Models

Suppose we have a sample of $n$ independent subjects, for which we have an $n$-vector response $y$ and an $n \times (p + q)$ design matrix. We partition the design matrix into an $n \times p$ design matrix $X$ of the null hypothesis and an $n \times q$ matrix $Z$ with the additional columns of the design matrix under the alternative. Assuming a generalized linear model, we write

$$E(y) = g^{-1}(X\alpha + Z\beta)$$

for a $p$-vector $\alpha$ and a $q$-vector $\beta$, where $g$ is the canonical link function for the distribution of $y$, taken to apply componentwise on a vector. We are interested in investigating the association of the covariates in $Z$ with the response $y$ by testing

$$H_0 : \beta = 0, \quad H_A : \beta \neq 0.$$  \hspace{1cm} (2)

As we are interested in possibly high-dimensional models, we allow a large number of covariates in $Z$, possibly more than the number of subjects; although we assume that $p < n$, we make no such assumption on $q$.

Goeman et al. (2006) derive a general test statistic for the problem (2), i.e.,

$$L = s^T s - \text{trace}(I),$$  \hspace{1cm} (3)

where $s = \partial \ell / \partial \beta$ and $I = -\partial^2 \ell / \partial \beta \partial \beta^T$ are the score and the observed information of $\beta$, respectively, and $\ell$ is the loglikelihood function. Among all possible tests, this test has optimal power, on average, to detect alternatives uniformly distributed on a ball $\|\beta\| = \varepsilon$, for $\varepsilon \downarrow 0$. The test is not invariant to the parametrization of the covariates in $Z$, so some standardization may be necessary, depending on the application.

In the framework of generalized linear models with a canonical link function, the test statistic $L$ becomes, up to a constant,

$$L = (y - \mu)^T ZZ^T(y - \mu) - \text{trace}(Z^T W Z)$$

where $\mu$ and $W$ are, respectively, the expectation and the diagonal covariance matrix of $y$ under the null hypothesis. As a consequence of the choice of link function, as the second term on the
right-hand side of (3) does not depend on the response \( y \), \( L \) is equivalent to

\[
Q = n^{-1} (y - \mu)^T ZZ^T (y - \mu).
\]

We refer to Goeman et al. (2004) for interpretations and decompositions of the test statistic for the case of the logistic model. The division by \( n \) is useful to avoid asymptotic degeneracy; we return to this in § 3.

3. Asymptotics

In regression models, asymptotic theory involves making some additional assumptions, because model (1), being conditional on the covariates, does not specify how the design matrices \( X \) and \( Z \) grow as the sample size grows. Theorem 1 assumes that \( q \), the column dimension of \( Z \), is fixed and does not depend on the sample size \( n \). This assumption is appropriate in regression models used for prediction, where the set of covariates is generally given, and unrelated to the sample size. The assumption of fixed \( q \) contrasts with the assumptions of Commenges & Jacqmin-Gadda (1997) who let \( q \) grow to infinity with \( n \); we return to this distinction below.

Theorem 1 gives an asymptotically equivalent distribution of the test statistic \( Q \) when the values of all nuisance parameters are known. It is essentially a corollary to the central limit theorem for nonidentically distributed variables of Lindeberg–Feller (Van der Vaart, 1998, Prop. 2.27). The proof is given in the Appendix.

**Theorem 1.** Let \( u = (U_1, \ldots, U_n)^T \), where \( U_1, \ldots, U_n \) are independent random variables with \( E(U_i) = 0 \) and \( \text{var}(U_i) = 1 \) \( (i = 1, \ldots, n) \). Let \( A = \{a_{ij}\} \) be an \( n \times q \) matrix, whose components, but not its column dimension \( q \), may depend on \( n \), for which \( \lim_{n \to \infty} n^{-1} A^T A \) exists and is nonzero, and for which

\[
\max_{i,j} \frac{a_{ij}^2}{n} \to 0.
\]

Then for the sequences

\[
T_n = n^{-1} u^T A A^T u, \quad G_n = \sum_{i=1}^{r} \lambda_n, i C_i^2,
\]

where \( r = \text{rank}(A) \), \( \lambda_{n,1} \geq \cdots \geq \lambda_{n,r} \) are the positive eigenvalues of \( n^{-1} A^T A \) and \( C_1^2, \ldots, C_r^2 \) are independent \( \chi^2_1 \) distributed random variables, we have

\[
\text{pr}(T_n \leq x) - \text{pr}(G_n \leq x) \to 0, \quad n \to \infty.
\]

The actual asymptotic distribution of \( T_n \) is that of \( G_\infty \), which is not computable because \( \lim_{n \to \infty} n^{-1} A^T A \) is usually not known. The distribution of \( G_n \), however, is asymptotically equivalent to that of \( T_n \) and serves as a practical and computable approximation to it. In case of the linear model, the distributions of \( T_n \) and \( G_n \) are identical.

Nondegeneracy of the distributions in Theorem 1 is guaranteed by the condition that \( U = \lim_{n \to \infty} n^{-1} A^T A \) exists and is nonzero, as this ensures that the mean of the variables \( G_n \) converges to \( \text{trace}(U) < \infty \) and its variance to \( 0 < 2 \text{trace}(U^2) < \infty \). The assumption on \( \max_{i,j} a_{ij}^2 / n \) guarantees that no covariate is tied to a single subject. Both assumptions are common in regression models with fixed dimension (Van der Vaart, 1998, p. 21).

Applying Theorem 1 to the test statistic \( Q \), taking \( A = W^{1/2} Z \), we see that the distribution of \( Q \) is asymptotically equivalent to that of a weighted sum of \( \chi^2_1 \) distributed variables with
weights given by the positive eigenvalues of \( n^{-1} Z^T W Z \), i.e., a test statistic that approximates the distribution of the response \( y \) by a \( \mathcal{N}(\mu, W) \) distribution. A heuristic argument for this writes
\[
n^{-1/2} W^{1/2} Z Z^T W^{1/2} = \sum_{i=1}^{\min(n,q)} \lambda_i v_i v_i^T,
\]
where \( \lambda_1, \ldots, \lambda_n \) and \( v_1, \ldots, v_n \) are the eigenvalues and eigenvectors of \( W^{1/2} Z Z^T W^{1/2} \), respectively, giving
\[
Q = \sum_{i=1}^{\min(n,q)} \lambda_i \| v_i^T W^{-1/2} (y - \mu) \|^2.
\]
If \( n \) is large enough, each \( v_i^T W^{-1/2} (y - \mu) \) is a weighted mean of many independent variables, motivating their approximation by normally distributed random variables, and thus \( Q \) by a weighted sum of \( \chi^2_1 \) variables.

Commenges & Jacqmin-Gadda (1997) derived an asymptotic normal distribution of \( Q \) for a special case under different assumptions, motivated by applications in homogeneity testing and random effects modelling. Their assumptions differ in important respects from ours. Our assumption that the dimension of \( Z \) remains fixed as \( n \to \infty \), which is appropriate in prediction models, is inappropriate in random effects models in which a covariate in \( Z \) is typically the random effect of a single observation or group of observations. In the random effect model context, it is natural to assume that the number of random effects grows with the sample size, whereas the number of subjects that is influenced by each random effect remains bounded. This leads to a different asymptotic eigenvalue structure from that assumed in Theorem 1. There, the dimension of the matrix \( n^{-1} A^T A \) is bounded, and the matrix converges to a limit with at least one nonzero eigenvalue. In the asymptotic theory of Commenges & Jacqmin-Gadda (1997), the same matrix grows to infinite dimension as \( n \to \infty \), while each individual eigenvalue converges to zero. Commenges & Jacqmin-Gadda (1997) also require the test statistics to be scaled differently: under their assumptions, the variance of \( Q \) converges to zero as \( n \to \infty \), rather than to a constant as under the assumptions of this paper, and \( Q \) should be scaled by \( n^{-1/2} \), after subtracting the mean, rather than by \( n^{-1} \).

Despite this, as finite sample approximations the distribution suggested in Theorem 1 may be compared to that proposed by Commenges & Jacqmin-Gadda (1997). Two remarks can be made. First, unlike the normal distribution of Commenges & Jacqmin-Gadda (1997), the approximate distribution of Theorem 1 is the exact finite sample distribution in the case of the linear model. Secondly, if the assumptions on the asymptotic eigenvalue structure of Commenges & Jacqmin-Gadda (1997) hold, it can be shown using their Theorem 1 that the distribution of \( G_n \), properly normalized, converges to the asymptotic normal distribution of Commenges & Jacqmin-Gadda (1997) as \( n \to \infty \). This may suggest that the approximate distribution of Theorem 1 is in some sense closer to the data and that the conditions of Theorem 1 might be relaxed to encompass the case of homogeneity testing, possibly along the lines of Götzte & Tikhomirov (2002).

4. Nuisance parameters and approximate pivoting

4.1. The linear model

Both the test statistic \( Q \) and its approximate distribution function may depend on nuisance parameters, in this case \( \alpha \) and, depending on the distribution of \( y \), possibly a dispersion parameter.
We therefore rescale $Q$ to provide an exact or approximate pivot and insert maximum likelihood estimates under the null hypotheses, although algorithms for the distribution of the pivot must be available.

We first discuss the linear model. Let $y \sim \mathcal{N}(\mu, \sigma^2 I)$, where $\mu = X\alpha$. The distribution of test statistic (5) depends on the eigenvalues of $Z^T W Z = \sigma^2 Z^T Z$, and therefore on the nuisance parameter $\sigma^2$. We can avoid this dependence by transforming $Q$ to the pivot

$$S_{\text{lin}} = \frac{(y - \mu)^T Z Z^T (y - \mu)}{n \sigma^2} ,$$

whose distribution, a linear combination of $\chi_1^2$ variables with weights given by the positive eigenvalues of $n^{-1} Z Z^T$, does not depend on $\mu$ or $\sigma^2$. Next, we plug in the maximum likelihood estimates $\hat{\mu} = H y$ and $\hat{\sigma}^2 = n^{-1} y^T (I - H) y$ under the null hypothesis, where $H = X (X^T X)^{-1} X^T$ is the hat matrix of the null model. Noting that $H \mu = \mu$, we can write the test statistic as

$$\hat{S}_{\text{lin}} = \frac{(y - \hat{\mu})^T Z Z^T (y - \hat{\mu})}{n \hat{\sigma}^2} = \frac{(y - \mu)^T (I - H) Z Z^T (I - H) (y - \mu)}{(y - \mu)^T (I - H) (y - \mu)} .$$

The right-hand expression shows that the distribution of $\hat{S}_{\text{lin}}$ does not depend on the unknown $\mu$ or $\sigma^2$, but only on the known matrices $H$ and $Z$. The distribution of $\hat{S}$ is that of a ratio of two dependent quadratic forms in normal variables, for which algorithms are available; see § 4·3.

### 4.2. The general case

In generalized linear models, we propose to use the approximate pivot

$$S = \frac{Q}{E(Q)} = \frac{(y - \mu)^T Z Z^T (y - \mu)}{\text{trace}(Z^T W Z)} .$$

This is analogous to the solution for the linear model, as (8) reduces to (6), up to a constant, if $W = \sigma^2 I$. It gives us an approximate pivot. The mean of $S$ is equal to 1, independent of the values of the nuisance parameters. Its asymptotic variance

$$\text{var}(S) \rightarrow \frac{2 \text{trace}(Z^T W Z Z^T W Z)}{\text{trace}^2(Z^T W Z)} , \quad n \rightarrow \infty ,$$

is not fully invariant to the values of the nuisance parameters, but is invariant to multiplication of $W$ by a constant, so its distribution is invariant to the values of the dispersion parameter of the distribution of $y$ and of the regression coefficient of the intercept term.

Next, we insert estimates of the unknown parameters in expression (8). First, consider the problem of estimating $W$. Write the denominator of (8) as

$$M = n^{-1} \text{trace}(Z^T W Z) = \sum_{i=1}^n d_i w_i ,$$

where each $d_i = \sum_{j=1}^q z_{ij}^2$ is the $i$th diagonal element of $Z Z^T$ and $w_i = E(y_i - \mu_i)^2$ is the $i$th diagonal element of $W$. It follows that a moment estimator of $M$ is

$$\hat{M} = n^{-1} (y - \mu)^T D (y - \mu) ,$$

where $D$ is a diagonal $n \times n$ matrix that is equal to $ZZ^T$ on the diagonal and zero otherwise. The estimate $\hat{M}$ is an unusual but consistent estimator of $M$. It is a mean of $n$ independent random
variables, which, by the assumptions of Theorem 1, have bounded and nondegenerate variation. By the law of large numbers, as \( n \to \infty \), \( M \) converges to \( M' \) with a variance that converges to 0.

The moment estimator \( \hat{M} \), being a quadratic form like the test statistic itself, is computationally very convenient to work with. Plugging in this estimator yields the test statistic

\[
\frac{Q}{\hat{M}} = \frac{(y - \mu)^T ZZ^T(y - \mu)}{(y - \mu)^T D(y - \mu)},
\]

which is asymptotically equivalent to \( Q \) by Slutsky’s lemma (Van der Vaart, 1998, Lemma 2.8, Prop. (iii)).

It is essential for Slutsky’s lemma to be applied that the variance of \( \hat{M} \) is negligible relative to that of \( Q \). This is always asymptotically true by the assumptions of Theorem 1, but may not hold for finite samples, especially if \( ZZ^T \) is nearly diagonal. This suggests that the quality of the approximate distribution may be poor in situations with near-diagonal \( ZZ^T \), which happens if \( q \) is large and a strong eigenvalue structure is absent.

Dividing the test statistic by \( \hat{M} \) may also entail some loss of information in finite samples. Commenges & Jacqmin-Gadda (1997) decomposed their test statistic into an overdispersion test statistic \( \hat{M} \) and a weighted pairwise correlations test statistic \( M' - M \). Even though the overdispersion statistic \( \hat{M} \) has some power to reject the null hypothesis (Jacqmin-Gadda & Commenges, 1995), they recommend using \( M' - M \) rather than \( \hat{M} \) itself, because it is more robust to nuisance overdispersion. Dividing by \( \hat{M} \) is asymptotically equivalent to subtracting it.

The other nuisance parameter that has to be estimated is \( \mu \). Plugging in the maximum likelihood estimate \( \hat{\mu} \) of \( \mu \) under the null hypothesis, we obtain

\[
\hat{S} = \frac{(y - \hat{\mu})^T ZZ^T(y - \hat{\mu})}{(y - \hat{\mu})^T D(y - \hat{\mu})}.
\]

Assuming \( \text{rank}(X) = p < n \), we can account for the estimation of \( \hat{\mu} \) in the distribution of \( \hat{S} \) by remarking that

\[
\hat{\alpha} - \alpha = (X^T WX)^{-1} X^T (y - \mu)\{1 + o_P(1)\},
\]

where the \( o_P(1) \) term is asymptotically negligible relative to the leading term. By the delta method,

\[
\hat{\mu} - \mu = \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \alpha} (\hat{\alpha} - \alpha)\{1 + o_P(1)\} = H(y - \mu)\{1 + o_P(1)\},
\]

where \( H = WX(X^T WX)^{-1} X^T \) is the asymmetric hat matrix of the generalized linear model. Noting that \( y - \hat{\mu} = (y - \mu) - (\hat{\mu} - \mu) \), we finally obtain

\[
y - \hat{\mu} = (I - H)(y - \mu)\{1 + o_P(1)\},
\]

so that, analogous to the linear model case,

\[
\hat{S} = \frac{(y - \mu)^T (I - H)^T ZZ^T (I - H)(y - \mu)}{(y - \mu)^T (I - H)^T D(I - H)(y - \mu)}\{1 + o_P(1)\}.
\]

We find the distribution of \( \hat{S} \) by remarking that, by Theorem 1 and by the asymptotic normality of the denominator, the distribution of both the numerator and the denominator are asymptotically equivalent to that of a corresponding statistic that takes \( y \) to have a normal \( \mathcal{N}(\mu, W) \) distribution. The distribution of \( \hat{S} \) is therefore asymptotically equivalent to that of a ratio of two dependent quadratic forms in normal variables, just as in the linear model. The resulting test statistic is
partially pivoted: its distribution is invariant to the dispersion parameter and to the regression coefficient of the intercept, but is not fully invariant of the nuisance parameter $\alpha$.

4.3. Algorithms

To calculate the distribution function of a ratio of two dependent quadratic forms in normal variables, it can be helpful to use that

$$\Pr\left(\frac{u^T A u}{u^T B u} \geq k\right) = \Pr\left\{ u^T (A - kB) u \geq 0 \right\} = \Pr\left(\frac{u^T C^+ u}{u^T C^- u} \geq 1\right),$$

(9)

for $A$, $B$ positive semidefinite matrices, $u$ a random vector and $C^+$ and $C^-$ positive semidefinite matrices, obtained by diagonalization, such that $C^+ - C^- = A - kB$ and $C^+ C^- = C^- C^+ = 0$. Using (9), we may use algorithms by Imhof (1961) and by Robbins & Pitman (1949). We found that Imhof’s algorithm is generally faster, but Robbins’ algorithm can be numerically more accurate for the very small $p$-values that are of interest in genomics research. Quick and accurate approximate solutions are available from Kuonen (1999).

5. Scaling and interpretation

The scaling of the test statistic is arbitrary: multiplying the test statistic by any constant results in a completely equivalent test statistic. It can be worthwhile, however, to find a useful general scaling so that not only the $p$-value of the test statistic is interpretable, but also the test statistic itself.

In the linear case, the Schwartz inequality yields

$$\hat{S}_{\text{lin}} = \frac{(y - \mu)^T (I - H) ZZ^T (I - H) (y - \mu)}{(y - \mu)^T (I - H) (y - \mu)} \leq \text{tr}ZZ^T (I - H).$$

(10)

Dividing $\hat{S}_{\text{lin}}$ by the right-hand term in (10), we obtain a test statistic that always takes values between 0 and 1. This is asymptotically degenerate under the null hypothesis, under which both its mean and standard deviation converge to zero as $n \to \infty$, but not under the alternative. It can be interpreted as a weighted, average squared partial correlation coefficient between each of the covariates and the response, corrected for the null covariates. Written in this way, the test statistic itself gains an immediate and useful interpretation as an estimate of a weighted partial $R^2$.

A direct analogy to (10) for the test statistic $\hat{S}$ in generalized linear models is possible but complicated. We prefer to scale $\hat{S}$ simply by $n - p$, which is asymptotically equivalent, and interpret $\hat{S}/(n - p)$ as a weighted $R^2$ even though it is not guaranteed to be smaller than 1 in finite samples.

6. Simulation study

The test based on $\hat{S}$ can be used when the dimensionality of the alternative hypothesis is very large and can then be quite powerful (Goeman et al., 2006). We performed a simulation to investigate how well the approximate distribution of the global test keeps the nominal level in this situation.

In order to obtain simulated high-dimensional models with a realistic collinearity structure, we based the simulation on three real data sets: two well-known microarray data sets, of Van de Vijver et al. (2002), with 4919 covariates for 295 subjects, and of Golub et al. (1999), with 7129 covariates for 72 subjects; and a proteomics data set (Mertens, 2008; Van der Werff et al., 2008), with 11 205 covariates for 153 subjects. To be able to vary the sample size, $n = 5$,
10, 20, 50, 100, we resampled covariate vectors from these data sets. We only used the gene expression measurements and protein measurements, but not the response or phenotype variables. As a fourth data set we took independent random normal covariates.

The first three data sets have moderately strong principal component structures. The proteomics data have the strongest, with only eight components accounting for 80% of the variation. Accounting for 80% requires 20 principal components in the data of Golub et al. (1999) and 85 in the data of Van de Vijver et al. (2002). The random normal data have no strong principal components.

For each run of the simulation, we randomly sampled \(n\) subjects, by subsampling for the first three data sets and by random generation for the fourth. For these subjects, we randomly generated a response based on the model in question. The means of the response variable were chosen as a function of the index of each individual and were chosen in such a way as to get reasonable values of \(E(y)\); for the linear model \(E(y_i)\) ranging between 1 and \(N\); for the Poisson model between 1 and \(e^2 = 7.4\), for the logistic between \(\logit(-1.5) = 0.18\) and \(\logit(1.5) = 0.82\). \(P\)-values were subsequently calculated for the global test based on the null model with the subject index as the only covariate, but against an alternative model containing all measured covariates. Average rejection rates for 5000 random data sets are given in Table 1.

The performance of the approximate distribution in this simulation illustrates that the dimensionality of the alternative hypothesis is of little importance to the asymptotics, which only depends on the assumption that the number of covariates under the null hypothesis is small relative to the sample size. The linear model results are accurate for all sample sizes. The Poisson model clearly performs better than the logistic, which seems to oscillate around the nominal level. The most striking feature of the simulation is the relatively slow convergence to the asymptotic distribution in the random normal model, likely due to the absence of a strong principal component structure in the random normal data, as explained in §4·2. In high-dimensional data, the approximate distribution requires a strong eigenvalue structure for good performance.

### Acknowledgement

The work was supported by a Netherlands Organisation for Scientific Research Veni grant and a Netherlands Bioinformatics Centre Biorange grant.

---

**Table 1. High-dimensional alternatives. Achieved type I error, nominal level \(\alpha = 0.05\). Number of rejections per 1000. Standard error for \(\alpha = 0.05\) is 0.3**

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Linear ((q = 4919))</th>
<th>Poisson ((q = 4919))</th>
<th>Logistic ((q = 4919))</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4.8</td>
<td>5.2</td>
<td>3.8</td>
</tr>
<tr>
<td>10</td>
<td>5.2</td>
<td>5.0</td>
<td>7.2</td>
</tr>
<tr>
<td>20</td>
<td>4.9</td>
<td>4.9</td>
<td>5.6</td>
</tr>
<tr>
<td>50</td>
<td>5.1</td>
<td>5.0</td>
<td>5.5</td>
</tr>
<tr>
<td>100</td>
<td>4.6</td>
<td>5.1</td>
<td>5.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Linear ((q = 7129))</th>
<th>Poisson ((q = 7129))</th>
<th>Logistic ((q = 7129))</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5.0</td>
<td>5.2</td>
<td>3.9</td>
</tr>
<tr>
<td>10</td>
<td>5.2</td>
<td>5.0</td>
<td>5.8</td>
</tr>
<tr>
<td>20</td>
<td>4.8</td>
<td>5.2</td>
<td>5.4</td>
</tr>
<tr>
<td>50</td>
<td>5.1</td>
<td>5.0</td>
<td>5.7</td>
</tr>
<tr>
<td>100</td>
<td>4.9</td>
<td>5.1</td>
<td>5.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Linear ((q = 11205))</th>
<th>Poisson ((q = 11205))</th>
<th>Logistic ((q = 11205))</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5.0</td>
<td>5.0</td>
<td>3.8</td>
</tr>
<tr>
<td>10</td>
<td>5.0</td>
<td>5.0</td>
<td>4.4</td>
</tr>
<tr>
<td>20</td>
<td>4.9</td>
<td>5.0</td>
<td>5.2</td>
</tr>
<tr>
<td>50</td>
<td>5.1</td>
<td>5.0</td>
<td>5.3</td>
</tr>
<tr>
<td>100</td>
<td>4.9</td>
<td>5.1</td>
<td>5.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Linear ((q = 1000))</th>
<th>Poisson ((q = 1000))</th>
<th>Logistic ((q = 1000))</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5.0</td>
<td>5.0</td>
<td>2.8</td>
</tr>
<tr>
<td>10</td>
<td>5.0</td>
<td>5.2</td>
<td>15.7</td>
</tr>
<tr>
<td>20</td>
<td>4.9</td>
<td>4.5</td>
<td>7.1</td>
</tr>
<tr>
<td>50</td>
<td>5.1</td>
<td>4.9</td>
<td>5.1</td>
</tr>
<tr>
<td>100</td>
<td>4.7</td>
<td>4.7</td>
<td>4.7</td>
</tr>
</tbody>
</table>
Appendix

Proof of Theorem 1. The proof makes use of the Lindeberg–Feller central limit theorem. Without loss of generality, we may assume that
\[ \lim_{n \to \infty} \lambda_{n,r} > 0. \] (A1)

If necessary, reduce the dimension \( q \) of \( A \) until (A1) holds.

Let \( V = (v_{ij}) = A(A^tA)^{-1/2} \), where superscript plus denotes the Moore–Penrose generalized inverse, and define \( t = (t_1, \ldots, t_N)^t = V'u \). Then \( E(t) = 0 \) and \( E(tt^t) = A^tA \), a \( q \times q \) rank \( r \) projection matrix that converges to the \( q \times q \) identity matrix as \( n \to \infty \), by assumption (A1).

To be able to apply the theorem of Lindeberg–Feller to the vector \( t \), we must check that for every \( \varepsilon > 0 \),
\[ \sum_{i=1}^{n} \sum_{j=1}^{q} v_{ij}^2 E\left(U_i^2 \mathbb{1}_{\{U_i^2 > \varepsilon\}}\right) \to 0, \quad n \to \infty, \] (A2)
where \( \mathbb{1}_{\{\cdot\}} \) is an indicator function. As \( \sum_{i=1}^{n} \sum_{j=1}^{q} v_{ij}^2 = \text{trace} \left(A(A^tA)^{-1}A\right) = r \to q \), condition (A2) is fulfilled when \( \max_i E\left(U_i^2 \mathbb{1}_{\{U_i^2 > \varepsilon\}}\right) \to 0 \), i.e., when \( \max_i \sum_{j=1}^{q} v_{ij}^2 \to 0 \) (cf. Van der Vaart, 1998, p. 21). This is guaranteed because
\[ \sum_{j=1}^{q} v_{ij}^2 = a_i^t(A^tA)^{-1}a_i \leq \frac{n^{-1}a_i^tAa_i}{\lambda_{n,r}}, \] (A3)
where \( a_i \) is the \( i \)th row of \( A \). By assumptions (4) and (A1), the right-hand side of (A3) converges to 0 as \( n \to \infty \).

Consequently, by the Lindeberg–Feller theorem, the difference between the distribution function of \( t \) and a normal \( \mathcal{N}(0, I) \) distribution converges to 0 as \( n \to \infty \). The theorem now follows from the continuous mapping theorem, remarking that \( T_n = n^{-1}t^tAAt \), and using standard theory on quadratic forms in normal variables. \( \square \)

References


*[Received November 2009. Revised February 2011]*