A Comparison of Random Projection-Based Test Statistics in High Dimensions

Simulation Study & Further Results

Connor McNeill* Miles Woollacott*

*Department of Statistics, North Carolina State University

December 3, 2024

Outline

Motivating Problem

- Suppose that we observe data $(Y_i, \boldsymbol{x_i}), i = 1, ..., n$.
- With these data, we are interested in studying a single-index model (SIM) such as a logistic or Poisson model in the form

$$Y_i | \boldsymbol{x_i} \overset{\text{ind}}{\sim} f(\boldsymbol{x_i}^T \boldsymbol{\beta}, \boldsymbol{\epsilon})$$

  where $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\epsilon}$ is our error term, and $p \gg n$.
- We are interested in testing $H_0 : \boldsymbol{\beta} = 0$ vs. $H_1 : \boldsymbol{\beta} \neq 0$.
- Our classical tests such as the Wald, Score, and Likelihood Ratio cannot be used directly in this case as $p > n$.

Key Ideas

- Random projection-based approaches perform dimension reduction on the covariates from p to k so that $n > k$.
- Liu et al.[1] proposed the random projection (RP) test statistic and compared it with the Wald, Score, and Likelihood Ratio tests for a logistic SIM.

Main Idea

How do all four test statistics compare in Type I error and power when constructed via the random projection approach for a Poisson SIM?

[1] Xingqiu Zhao Changyu Liu and Jian Huang. "A Random Projection Approach to Hypothesis Tests in High-Dimensional Single-Index Models". In: *Journal of the American Statistical Association* 119.546 (2024), pp. 1008–1018.

Random Projection Approach

The random projection approach is as follows

1. For a set projection ratio ρ, obtain k, the projection dimension. $k = \lceil \rho n \rceil$.
2. Obtain $\boldsymbol{I} - \boldsymbol{P}_1$, where $\boldsymbol{P}_1 = \frac{1}{n}\mathbf{1}\mathbf{1}^T$
3. For $d = 1, \ldots, D$, create a $p \times k$ matrix with random entries from a $N(0, 1)$. Then we obtain our **random projection matrix** $\boldsymbol{P}_k$, which is $p^{-1/2}\times$ the mean of each entry across the D matrices.
4. Obtain $\boldsymbol{U}_k = (\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{X}\boldsymbol{P}_k$, which is a $n \times k$ matrix.
5. Utilize $\boldsymbol{U}_k$ in our test statistics instead of $\boldsymbol{X}$.

Proposed Test

The proposed random projection test statistic, T_{RP}, is

$$T_{RP} = \frac{\boldsymbol{y}^T \boldsymbol{H}_k \boldsymbol{y}/k}{\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_1 - \boldsymbol{H}_k)\boldsymbol{y}/(n-k-1)},$$

where:

- $\boldsymbol{P}_1, \boldsymbol{P}_k,$ and $\boldsymbol{U}_k$ are as described on the previous slide
- $\boldsymbol{H}_k = \boldsymbol{U}_k(\boldsymbol{U}_k^T \boldsymbol{U}_k)^{-1} \boldsymbol{U}_k^T$ is our Hat matrix

Under H_0 and other assumptions, we reject H_0 when

$$\frac{T_{RP} - 1}{\sqrt{2/n\rho(1-\rho)}} > z_\alpha, \text{ where } \rho = \frac{k}{n} \in (0,1)$$

and z_α is the upper α-quantile of the standard normal.

Classical Tests in High Dimensions

As stated earlier, the Wald (T_W), Score (T_S), and Likelihood Ratio (T_{LR}) tests cannot be formed when $p > n$.

Using the random projection approach, we transform $\boldsymbol{X}$ into $\boldsymbol{U}_k$, where $\boldsymbol{U}_k \in \mathbb{R}^{n \times k}$ and $k < n$. We then construct the test statistics in the usual way, but using $\boldsymbol{U}_k$ instead of $\boldsymbol{X}$.

Liu et al. proved that while $\boldsymbol{X}^T\boldsymbol{X}$ is not invertible, $\boldsymbol{U}_k^T\boldsymbol{U}_k$ is invertible with probability 1.

We will see later that T_W, T_S, T_{LR} do not actually converge to χ^2_k as would typically be seen.

Our Approach

We will use a simulation-based approach to compare Type I Errors and power for T_{RP}, T_W, T_S, and T_{LR}. For the simulation, we will use $n = 400$ and $p = 1000$.

Data Generation

Overview:

1. Generate $\boldsymbol{\Sigma} = \boldsymbol{ODO}^T$, our covariance matrix, based off of the specified sparsity in the setting of the simulation.
2. Generate $\boldsymbol{X} = \boldsymbol{Z}\Sigma^{1/2}$ where $\boldsymbol{Z}$ is generated from $N(0, 1)$
3. Generate $\boldsymbol{\beta} \in \mathbb{R}^p = b\boldsymbol{\delta}/\sqrt{\boldsymbol{\delta}^T\boldsymbol{\Sigma}\boldsymbol{\delta}}$ in one of two ways based off of $\boldsymbol{\delta}$ which controls the sparsity of $\boldsymbol{\beta}$:
   1. $\boldsymbol{\delta}_1$ is a sparse vector with 10 non-zero values
   2. $\boldsymbol{\delta}_2$ is randomly selected from the span of the first 100 columns from the orthogonal $\boldsymbol{O}$ matrix
4. Randomly generate Y from our SIM model using $\boldsymbol{X}$ and $\boldsymbol{\beta}$

Setup

For $L = 1000$ iterations on each setting:

1. Generate data $(Y, \boldsymbol{X})$ as described previously with a sparse covariance matrix
2. Compute the four test statistics for each iteration

Then compute the rejection rate for each setting. (NOTE: for the classical tests, use the typical χ^2_k as the null distribution)

Results

β Setting	b^2	T_{RP}	T_{LR}	T_W	T_S	
$\mathbf{0}$	0	0.062	0.011	0.205	0.068	Type I Error
$\boldsymbol{\delta}_1$	0.1	0.534	0.533	0.928	0.802	Power
	0.2	0.940	0.974	0.999	0.995	Power
$\boldsymbol{\delta}_2$	0.1	0.525	0.516	0.910	0.806	Power
	0.2	0.929	0.972	1.000	0.997	Power

- T_W is not a valid test in this setting.
- The Type I Error of T_{RP} was closest to 0.05.
- T_S is more powerful test than T_{LR} and T_{RP} in this setting
- T_{LR} performs similarly to T_{RP} when $b^2 = 0.1$ but is more powerful when $b^2 = 0.2$.

Comparison with Logistic Model

Poisson

β Setting	b^2	T_{RP}	T_{LR}	T_W	T_S	
$\mathbf{0}$	0	0.062	0.011	0.205	0.068	Type I Error
δ_1	0.1	0.534	0.533	0.928	0.802	Power
	0.2	0.940	0.974	0.999	0.995	Power
δ_2	0.1	0.525	0.516	0.910	0.806	Power
	0.2	0.929	0.972	1.000	0.997	Power

Logistic

β Setting	b^2	T_{RP}	T_{LR}	T_W	T_S	
$\mathbf{0}$	0	0.059	0.830	0.000	0.021	Type I Error
δ_1	0.4	0.469	0.993	0.000	0.227	Power
	0.8	0.831	0.973	0.026	0.581	Power
δ_2	0.4	0.480	0.987	0.003	0.214	Power
	0.8	0.830	0.981	0.019	0.613	Power

Conclusion

- Certain test statistics are still effective in high-dimensional settings with random projections.
- Choice of model impacts performance of test statistics.
- The novel test statistic proposed by Liu et al. (T_{RP}) was effective in controlling Type I Error, but at the cost of a sub-optimal power.