

- 1 Motivation
- 2 Methodology
- 3 Simulation Study
 - Setup
 - Results
- 4 Conclusion

Motivating Problem

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

$$Y_i | \mathbf{x}_i \stackrel{\text{ind}}{\sim} f(\mathbf{x}_i^T \boldsymbol{\beta}, \epsilon)$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$, ϵ 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?

¹Xingqiu 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. DOI: 10.1080/01621459.2022.2156350.

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 $I - P_1$, where $P_1 = \frac{1}{n} \mathbf{1} \mathbf{1}^T$
- ③ For $d = 1, \dots, D$, create a $p \times k$ matrix with random entries from a $N(0, 1)$. Then we obtain our **random projection matrix** P_k , which is $p^{-1/2} \times$ the mean of each entry across the D matrices.
- ④ Obtain $U_k = (I - P_1)XP_k$, which is a $n \times k$ matrix.
- ⑤ Utilize U_k in our test statistics instead of X .

Proposed Test

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

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

where:

- \mathbf{P}_1 , \mathbf{P}_k , and \mathbf{U}_k are as described on the previous slide
- $\mathbf{H}_k = \mathbf{U}_k (\mathbf{U}_k^T \mathbf{U}_k)^{-1} \mathbf{U}_k^T$ 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 \mathbf{X} into \mathbf{U}_k , where $\mathbf{U}_k \in \mathbb{R}^{n \times k}$ and $k < n$. We then construct the test statistics in the usual way, but using \mathbf{U}_k instead of \mathbf{X} .

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

We will see later that T_W , T_S , T_{LR} do not actually converge to χ_k^2 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 $\Sigma = \mathbf{O}\mathbf{D}\mathbf{O}^T$, our covariance matrix, based off of the specified sparsity in the setting of the simulation.
- 2 Generate $\mathbf{X} = \mathbf{Z}\Sigma^{1/2}$ where \mathbf{Z} is generated from $N(0, 1)$
- 3 Generate $\beta \in \mathbb{R}^p = b\delta/\sqrt{\delta^T \Sigma \delta}$ in one of two ways based off of δ which controls the sparsity of β :
 - 1 δ_1 is a sparse vector with 10 non-zero values
 - 2 δ_2 is randomly selected from the span of the first 100 columns from the orthogonal \mathbf{O} matrix
- 4 Randomly generate Y from our SIM model using \mathbf{X} and β

Setup

For $L = 1000$ iterations on each setting:

- 1 Generate data (Y, \mathbf{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_k^2 as the null distribution)

Results

β Setting	b^2	T_{RP}	T_{LR}	T_W	T_S	
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

- 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	
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	
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.