

Results Summary 2023.10.11

Yu Yan

2023-10-11

Simulation Results Summary (Date: 2023.10.11)

In this report, we provide a comprehensive overview of simulation results conducted on October 11, 2023, focusing on the evaluation of the cyclical coordinate descent risk score algorithm's performance. The simulation aimed to assess the algorithm's efficacy in handling complex, real-world data scenarios by introducing enhanced variability in the relationship between covariates.

Simulation Methodology:

The simulated data were generated using a sophisticated simulation function tailored for statistical robustness. The function accepted various input parameters, including sample size (n), the number of variables with coefficients between 0 and 1 (p_1), variables with coefficients between 1 and 5 (p_2), and variables with coefficients equal to 0 (p_3). The feature set, structured as $n \times p \times (p_1 + p_2 + p_3)$, was meticulously sampled from a multivariate Gaussian distribution. Notably, the correlation between columns was introduced through an exponential correlation mechanism, characterized by the parameter ρ (defaulting to 0.5).

In the generation process, the outcome variable (y) was modeled as $y = xb + \epsilon$, where b represented a vector with 'supp_size' elements, randomly set to 1 and the remaining set to 0. To emulate real-world complexities, random noise (ϵ) was incorporated. The magnitude of this noise, controlled by the signal-to-noise ratio parameter (SNR), played a pivotal role in evaluating the algorithm's resilience in noisy environments.

Simulation Settings and Variability Parameters:

The simulation was conducted under three distinct settings to comprehensively analyze the algorithm's behavior:

1. **No Cross Validation:** The algorithm's performance was analyzed without employing cross-validation.
2. **Cross Validation with λ_{\min} :** Cross-validation was employed, and λ_{\min} was selected as λ_0 in the risk model.
3. **Cross Validation with λ_{1se} :** Cross-validation was applied, and λ_{1se} was chosen as λ_0 in the risk model.

The simulation spanned various combinations of parameters, including a sample size of $N = 1000$, varying feature dimensions ($P = 10, 50, 100$), proportions of coefficients ($P_1 = 0, p_2 = 10\%, 50\%, 90\%$ of P), and correlation strengths ($\rho = 0.1, 0.5, 0.9$). Additionally, the SNR parameter was set to 5 to represent different levels of noise in the data. For each unique combination, ten data frames were generated for meticulous analysis.

Background Significance:

Understanding the algorithm's behavior under diverse and challenging conditions is crucial for real-world applications. By introducing intricate relationships and noise structures, we aimed to simulate scenarios encountered in complex datasets. This comprehensive evaluation provides insights not only into the algorithm's accuracy but also its stability and adaptability in addressing real-world complexities.