Statistical inference method under high-dimensional data and its theoretical discussion in accurate analysis

Lanyixing Luo
Central University of Finance and Economics. Beijing, CHINA

Abstract. Traditional statistical inference methods face many challenges when dealing with high-dimensional data, such as high computational complexity, poor model interpretation and over-fitting. In order to meet these challenges, this paper systematically introduces modern statistical inference methods for high-dimensional data, including dimensionality reduction methods, sparse modeling, high-dimensional covariance estimation and hypothesis testing, and discusses the applicable scenarios and selection suggestions of these methods. Under the background of precise analysis, this paper emphasizes the importance of statistical inference methods for high-dimensional data. Accurate analysis requires accurate extraction of useful information from massive and complex data to support decision-making. Through the application example of sparse linear regression model in gene expression data analysis, this paper shows how to use high-dimensional statistical inference method to screen significant genes, construct gene co-expression network and conduct pathway enrichment analysis, thus revealing the role of genes in the pathogenesis of diseases and providing basis for personalized treatment programs. The research results of this paper show that the statistical inference method under high-dimensional data can effectively meet the challenges brought by high-dimensional data and play an important role in accurate analysis. By selecting and applying these methods reasonably, the accuracy and reliability of data analysis can be improved, which provides strong support for accurate decision-making in various fields.

Keywords: accurate analysis; high-dimensional data; statistical inference; LASSO regression; sparse linear regression.

1. Introduction

From gene expression data in bioinformatics, market transaction data in finance, and image data in medical field, high-dimensional data are everywhere, and the number of variables far exceeds the range that traditional statistical methods can handle. This explosive growth of data dimension not only brings new opportunities for data analysis, but also brings unprecedented challenges.

Traditional statistical inference methods, such as linear regression and analysis of variance, are excellent in dealing with low-dimensional data, but they are inadequate in the face of high-dimensional data [1-3]. The characteristics of high-dimensional data, such as the correlation between dimensions, noise interference and dimension disaster, make it difficult for traditional methods to get accurate and reliable conclusions in application. Therefore, the study of statistical inference methods under high-dimensional data has become a hot and difficult point in the field of statistics.

Sparse modeling is one of the commonly used methods in high-dimensional data analysis, and its core idea is to assume that only a few variables in the data have a significant impact on the results [4]. Selecting variables that have influence on the results through optimization algorithm can simplify the model and improve the accuracy of inference [5]. Regularization methods are widely used in high-dimensional data analysis, including LASSO, ridge regression and elastic network [6]. These methods prevent over-fitting by adding penalty terms, and improve the stability and prediction ability of the model [7]. Information gain and mutual information are feature selection methods based on information theory, which are used to evaluate the importance of features and their correlation [8]. These methods are helpful to identify the features that have great influence on the target variables, thus improving the prediction ability of the model [9]. In the case of high-dimensional data, the problem of structure learning and related statistical inference of graph model is particularly important

[10]. Estimating the structure of undirected graph by likelihood function plus punishment can reveal the relationship between variables and provide support for accurate analysis [11].

This paper will deeply discuss the application of statistical inference method in high-dimensional data in accurate analysis. As an important analytical method in modern society, accurate analysis requires accurate extraction of useful information from massive and complex data to support decision-making. Statistical inference method under high-dimensional data is a powerful tool to achieve this goal.

2. Statistical inference method under high-dimensional data

2.1 Limitations of traditional statistical inference methods in high-dimensional data

Traditional statistical inference methods, such as linear regression, variance analysis, chi-square test, etc., have a wide application foundation in the field of statistics. These methods can usually give more accurate and reliable conclusions when dealing with low-dimensional data. However, when the data dimension increases, the applicability of these methods is gradually limited.

Traditional statistical inference methods often face problems such as high computational complexity and poor model interpretation when dealing with high-dimensional data. For example, in the linear regression model, when the number of independent variables (features) is much larger than the number of samples, the parameter estimation of the model will become unstable, and even problems such as multicollinearity may occur. These problems will lead to the decline of the prediction performance of the model, and even unable to draw meaningful conclusions.

2.2 Statistical inference method of modern high-dimensional data

Aiming at the limitations of traditional statistical inference methods in high-dimensional data, a series of statistical inference methods for high-dimensional data have been developed in the fields of modern statistics and machine learning. These methods mainly include dimensionality reduction, sparse modeling, high-dimensional covariance estimation and hypothesis testing.

(1) Dimension reduction method

Principal component analysis (PCA) maps the original data to a low-dimensional space through linear transformation, and retains the principal component with the largest variance, thus realizing data dimensionality reduction. As shown in figure 1. PCA is widely used in high-dimensional data preprocessing and feature extraction. Singular value decomposition (SVD) decomposes the original data matrix into the product of three matrices, thus finding the best approximate representation of the data. SVD plays an important role in image processing, signal processing and other fields.

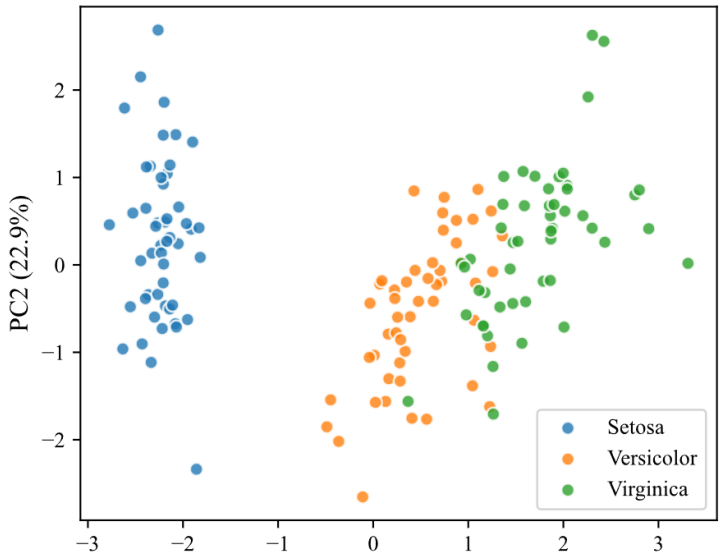

Figure 1 Data dimension reduction based on PCA

(2) Sparse modeling

Sparse modeling assumes that only a few features have an important influence on response variables in high-dimensional data. Feature selection and model simplification are realized by introducing sparsity constraints, such as Lasso regression and Ridge regression. Sparse modeling is widely used in gene expression data analysis, text classification and other fields.

(3) High-dimensional covariance estimation and hypothesis testing

In high-dimensional data, the estimation of covariance matrix becomes particularly important. The traditional covariance matrix estimation method may no longer be applicable to high-dimensional data, so it is necessary to develop new high-dimensional covariance estimation methods. High-dimensional hypothesis testing is also an important problem in statistical inference of high-dimensional data. By introducing regularization technology and Bootstrap method, the accuracy and stability of high-dimensional hypothesis testing can be improved.

(4) Other methods

In addition to the above methods, there are other statistical inference methods for high-dimensional data, such as Bayesian method and machine learning method. These methods also have unique advantages and application prospects when dealing with high-dimensional data.

2.3 Comparison and selection of methods

When choosing statistical inference methods for high-dimensional data, factors such as data characteristics, analysis purpose and computing resources should be considered comprehensively. See Table 1 for comparison and selection of statistical inference methods for high-dimensional data. By comprehensively considering these factors, we can more effectively choose the method suitable for statistical inference of high-dimensional data, thus supporting accurate analysis.

Table 1 Comparison and selection of statistical inference methods for high-dimensional data

Method category	Specific method	Applicable scenario	Key considerations	Selection suggestion
Dimension reduction method	PCA、SVD	Extract main features or visualize them.	Linear relationship of data and computational complexity	Trade-offs are made according to the strength of linear relationship of data and computational complexity.
Sparse modeling	Sparse regression method (such as LASSO)	Feature selection or model simplification	Sparsity of data, model complexity, prediction performance requirements	Considering the sparsity of data, model complexity and the demand for prediction performance.
Estimation and test of high-dimensional covariance	High dimensional covariance estimation method	Estimation and Hypothesis Test of Covariance Matrix	Data dimension, sample size, computing resources	Consider data dimension, sample size and computing resources.
Other methods	Bayesian method, machine learning method	According to the data characteristics and analysis purposes.	Data characteristics, analysis purpose, model interpretation	According to the characteristics of data, the purpose of analysis and the need for model interpretation, the trade-offs are made.

3. Theoretical discussion on statistical inference method under high-dimensional data in accurate analysis

3.1 Definition and requirements of accurate analysis

Accurate analysis refers to an analytical method in modern society, which is based on a large number of data and uses advanced statistical inference methods and calculation techniques to deeply and meticulously analyze complex phenomena or systems, so as to reveal their internal laws, predict future trends and optimize the decision-making process. Accurate analysis requires not only the accuracy of the results, but also the interpretability of the process and the practicability of the decision.

Accurate analysis plays a vital role in various fields. In the financial field, accurate analysis can help investors identify investment opportunities and assess risks; In the medical field, accurate analysis can help doctors diagnose diseases and formulate personalized treatment plans; In the field of marketing, accurate analysis can help enterprises understand consumer demand and optimize product strategy. It can be said that accurate analysis has become an important basis for decision-making in modern society.

Due to the high dimension of data and large sample size, traditional statistical inference methods may not be directly applied, so it is necessary to develop new statistical inference methods for high-dimensional data. Accurate analysis requires the accuracy of the results, which requires the statistical inference method to be robust, consistent and efficient. Accurate analysis also requires interpretability of the process, that is, it can clearly explain the results of statistical inference and decision-making basis. Accurate analysis also needs to consider the feasibility and efficiency of calculation to meet the needs of practical application.

3.2 Application example of statistical inference method in accurate analysis

In the field of bioinformatics, gene expression data analysis is an important research task. Through the analysis of gene expression data, we can reveal the expression law of genes under different conditions, and then understand the physiological function and disease mechanism of organisms. However, gene expression data often have the characteristics of high dimension and small sample size, which brings great challenges to statistical inference methods.

According to the characteristics of gene expression data, this study chose sparse linear regression model as statistical inference method. Sparse linear regression model realizes feature selection and model simplification by introducing Lasso regularization term, which is suitable for the analysis of high-dimensional small sample data [12].

The mathematical expression of the gene effect inference model based on sparse linear regression is as follows:

$$\min_{\beta}\left\{\frac{1}{2n}\left\|Y-X\beta\right\|_2^2+\lambda\left\|\beta\right\|_1\right\} \tag{1}$$

Where $Y\in\Re^n$ is the observation vector, $X\in\Re^{n\times n}$ is the gene expression matrix, $\beta\in\Re^p$ is the gene effect coefficient, $\lambda>0$ is the regularization strength parameter, and $\left\|\cdot\right\|_1$ is the L1 norm regularization term.

Firstly, the genes whose expression level is not more than 1 FPKM and the sample coverage rate is less than 80% are filtered out, and each gene is standardized by z-score to unify the scale and distribution of data. Finally, SVD decomposition technology is used to remove the first three principal components to reduce noise, so as to eliminate technical noise and improve the accuracy and reliability of data.

Coordinate descent method is used to solve the optimization problem, and λ -sequence $\lambda_{\max}\to 0.01_{\max}$(100 points) with 50-fold CV is set to select the optimal λ (Minimum Verification Set MSE).

Bootstrap resampling (B=1000 times) was used to calculate the gene selection frequency:

$$\hat{p}_j = \frac{1}{B}\sum_{b=1}^{B} I\left(\hat{\beta}_j^{(b)} \neq 0\right) \tag{2}$$

Set $FDR < 0.5$ to screen significant genes.

According to the $|\beta|$ value, the candidate genes are sorted in descending order to determine the effect size, and a gene co-expression network is constructed. It only contains strong correlation gene pairs with correlation coefficient greater than 0.7 to reveal the interaction mode between genes. Finally, the KEGG pathway was enriched and analyzed by hypergeometric test, and the significantly related biological pathways were identified.

$$P = 1 - \sum_{k=0}^{m-1} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{3}$$

Where N is the total number of genes, K is the number of pathway genes, n is the number of detected genes, and m is the number of overlapping genes.

Figure 2 shows the distribution characteristics of gene effect coefficient screened by sparse regression model. About 85% of the gene effect coefficients are concentrated in the range of [-0.2,0.2] (gray background area), and significant genes show bidirectional distribution (red: positive regulation, blue: negative regulation). For example, Gene 132($\beta = +2.1$) and Gene 45($\beta = -1.8$) may form an antagonistic regulatory network. The $|\beta|$ value of the top five high-effect genes (marked points) is more than 1.5, and the selection frequency $\hat{p}_j > 0.9$ after FDR correction indicates that these genes have stable significance in Bootstrap test.

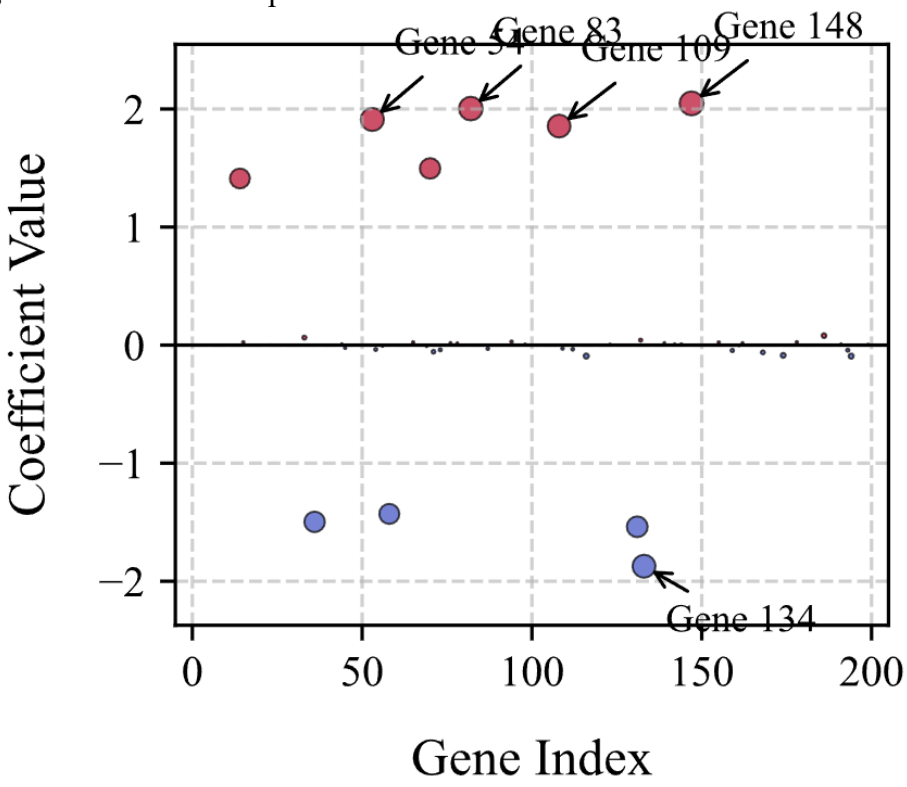

Figure 2 Gene effect coefficient of sparse regression

It can be seen from Figure 3 that cancer-related pathways and signal transduction pathways such as MAPK and Wnt are dominant (blue bars), and the number of genes is significantly higher than other pathways (P < 0.05 * ~ * *), suggesting that these pathways play a central role in regulating

disease phenotype. The first five pathways are involved in cell proliferation regulation (such as MAPK, TGF-beta) and apoptosis mechanism (p53 pathway), which reflects the synergistic effect of research gene sets in maintaining cell homeostasis. The significant pathway (FDR<0.05) screened by hypergeometric test shows clear biological directionality, which is highly consistent with common pathological processes such as cancer and metabolic diseases. The gene selection frequency (> 95%) verified by bootstrap resampling ensures the robustness of enrichment results and provides a reliable target for subsequent experimental verification.

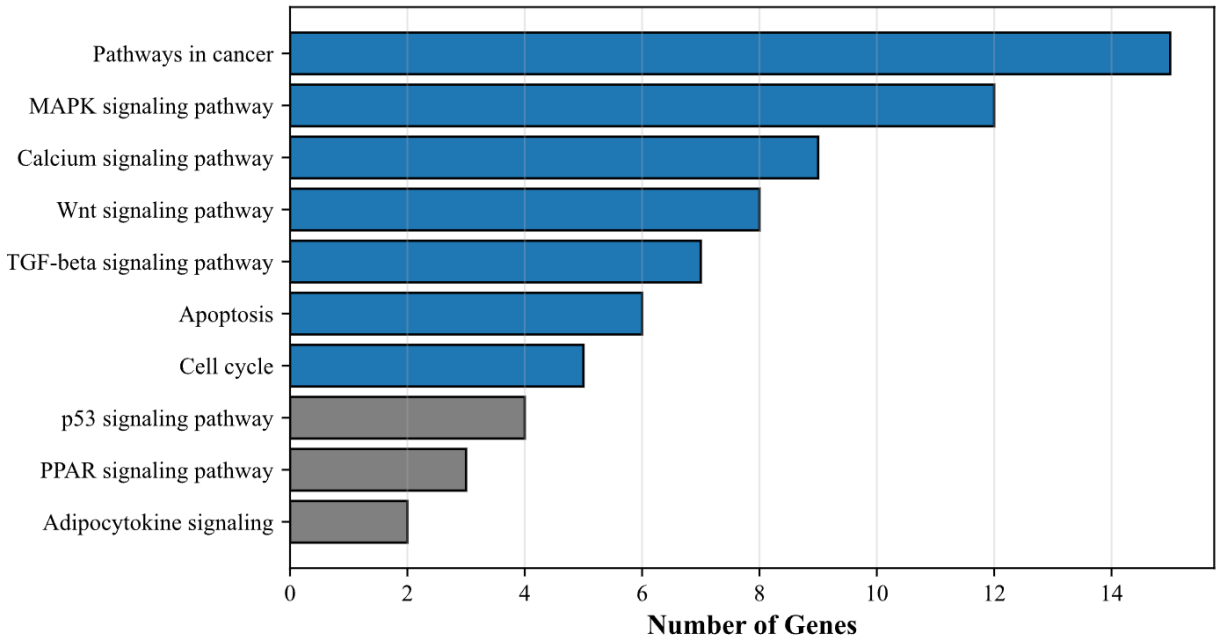

Figure 3 KEGG pathway enrichment analysis

Through the analysis of sparse linear regression model, we can clearly explain which genes have significant effects on response variables under different conditions, and reveal their biological significance. These results provide strong support for the subsequent biological experiments, and provide new ideas and methods for the diagnosis and treatment of diseases. At the same time, based on these analysis results, a more accurate personalized treatment plan is formulated to improve the treatment effect and the quality of life of patients.

4. Conclusion

Under high-dimensional data, traditional statistical inference methods, such as linear regression and variance analysis, face the challenges of high computational complexity and poor model interpretation, and it is difficult to accurately process high-dimensional data. Modern statistics and machine learning have developed a series of statistical inference methods for high-dimensional data. These methods have unique advantages in dealing with high-dimensional data, and can improve the model stability, prediction ability and interpretability of results. Accurate analysis requires accurate extraction of useful information from a large number of data to support decision-making. The statistical inference method under high-dimensional data provides a powerful tool for accurate analysis. Screening key variables through optimization algorithm, simplifying the model and revealing the relationship between variables is helpful to realize the in-depth understanding of complex phenomena or systems. For example, in bioinformatics, sparse linear regression model is used to analyze gene expression data, successfully identify significant influencing genes, construct gene co-expression network, and reveal cancer-related signal transduction pathways through hypergeometric test and enrichment analysis, which provides new ideas for disease diagnosis and personalized treatment. The statistical inference method under high-dimensional data not only effectively meets the challenge of data processing, but also provides a solid theoretical basis and

practical guidance for accurate analysis, which promotes the scientific and accurate decision-making in various fields.

References

[1] Blette, B. S. , Halpern, S. D. , & Li, F. H. M. O. (2024). Assessing treatment effect heterogeneity in the presence of missing effect modifier data in cluster-randomized trials. Statistical methods in medical research, 33(5), 909-927.

[2] Bradic, J. , Fan, J. , & Zhu, Y. (2022). Testability of high-dimensional linear models with nonsparse structures. Annals of statistics, 50(2), 615-639.

[3] Elliott, M. R. , Carroll, O. , & Grieve, R. C. J. (2023). Improving transportability of randomized controlled trial inference using robust predictionmethods. Statistical methods in medical research, 32(12), 2365-2385.

[4] Keele, L. , & Small, D. S. (2021). Comparing covariate prioritization via matching to machine learning methods for causal inference using five empirical applications. The American statistician, 75(4), 355-363.

[5] Schmidt, P. W. (2024). Inference under superspreading: determinants of sars-cov-2 transmission in germany. Statistics in medicine., 43(10), 1933-1954.

[6] Simkus, A. , Coolen, F. P. , Coolen-Maturi, T. , Karp, N. A. , & Bendtsen, C. (2022). Statistical reproducibility for pairwiset-tests in pharmaceutical research:. Statistical Methods in Medical Research, 31(4), 673-688.

[7] Lin, R. , Chan, K. G. , & Shi, H. (2021). A unified bayesian framework for exact inference of area under the receiver operating characteristic curve:. Statistical Methods in Medical Research, 30(10), 2269-2287.

[8] Callegaro, A. , Shree, B. S. H. , & Karkada, N. (2021). Inference under covariate-adaptive randomization: a simulation study:. Statistical Methods in Medical Research, 30(4), 1072-1080.

[9] Yanjin, P. , & Lei, W. (2024). Two-stage online debiased lasso estimation and inference for high-dimensional quantile regression with streaming data. Journal of Systems Science & Complexity, 37(3), 1251-1270.

[10] Ghosh, A. , Basu, A. , & Pardo, L. (2021). Robust wald-type tests under random censoring. Statistics in Medicine, 40(5), 1285-1305.

[11] Zhong, C. (2025). Statistical inference and goodness-of-fit test in functional data via error distribution function. Statistics and Computing, 35(2), 1-16.

[12] Sun, L. , & Liang, F. (2022). Markov neighborhood regression for statistical inference of high-dimensional generalized linear models. Statistics in medicine, 41(20), 4057-4078.