Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 54 (4) (2025), 1518 – 1532
DOI : 10.15672/hujms.1679033

Research Article

Robust gene co-expression networks via partial robust M regression

Ayca Olmez[1], Aylin Alin[*2]

[1] *Department of Statistics, The Graduate School of Natural and Applied Sciences, Dokuz Eylul University, Izmir, 35390, Türkiye*
[2] *Department of Statistics, Faculty of Science, Dokuz Eylul University, Izmir, 35390, Türkiye*

Abstract

Gene expression data provide valuable information on the regulation and interactions of thousands of genes. However, constructing robust gene co-expression networks in the presence of outliers remains an open challenge. We propose a partial robust M regression based-method for building ene co-expression networks, which downweights extreme observations instead of discarding them. This preserves critical biological information while safeguarding the overall network structure from distortion. Through comprehensive simulations on the syntren300 dataset - including various outlier distributions (e.g. N(0, 5), N(1, 5), N(100, 10) and t(2)) and contamination levels up to 30%, the partial robust M regression-based approach outperforms widely used methods (weighted gene co-expression network analysis, bi-weighted midcorrelation and partial least squares regression-based connectivity) in terms of precision, F1 and Matthews correlation coefficient. Real-data analysis of mouse liver gene expression further validates the stability and biological relevance of partial robust M regression-based gene co-expression networks, as it accurately identifies functionally enriched genes even under data contamination. These findings underscore the potential of partial robust M regression-based network construction to enhance reliability and uncover novel insights in high-dimensional genomic studies, offering a robust alternative to traditional correlation-based or partial least squares regression-based methods.

Mathematics Subject Classification (2020). 62H99, 92-10, 92-08

Keywords. Bi-weight mid-correlation, gene co-expression network analysis, outliers, partial robust m regression, robust, weighted gene co-expression network analysis

1. Introduction

Biological networks offer a robust framework for analyzing complex interactions in biological systems. Gene Co-Expression Networks (GCNs), derived from high-throughput expression data, reveal patterns of gene associations between conditions. These networks help uncover novel gene relationships, functional modules, and disease-related genes, advancing our understanding of biology and informing therapeutic strategies. The construction of GCNs typically involves two primary steps:

*Corresponding Author.
Email addresses: olmezayca@gmail.com (A. Olmez), aylin.alin@deu.edu.tr (A. Alin)
Received: 18.04.2025; Accepted: 21.06.2025

1. Computing a similarity measure for each gene pair, represented as network edges.
2. Grouping genes with similar expression patterns using hierarchical clustering.

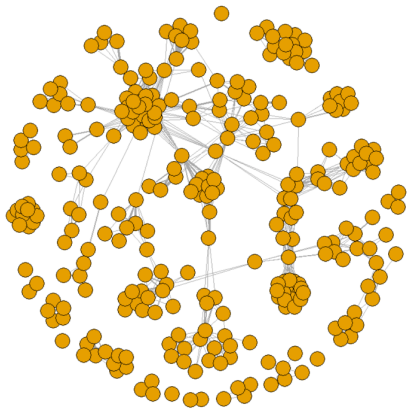

Figure 1. Visualization of Gene Network

Figure 1 presents the gene co-expression network of the simulated dataset used in the study. Nodes represent genes, while edges indicate pairwise connections.

Gene Set Enrichment Analysis (GSEA) is widely used to evaluate whether predefined gene sets exhibit statistically significant differences between experimental conditions. By highlighting overrepresented or underrepresented pathways or gene sets, GSEA uncovers underlying biological mechanisms and assesses the significance of identified genes [1]. Weighted Gene Co-Expression Network Analysis (WGCNA) [2], which employs Pearson's correlation as a similarity measure, is one of the most commonly used methods to construct GCNs. WGCNA has been applied successfully in various fields, including cancer research, cardiovascular studies, and mental health [4, 9, 10, 13, 18, 19, 24, 25]. However, Pearson's correlation is highly sensitive to outliers and leverage points, which can distort network structures and module assignments. These issues arise because Pearson correlation assumes normally distributed data and equal importance of all observations, assumptions that biological data often violate. To address these limitations, robust correlation measures, such as Biweight Midcorrelation (BICOR), have been proposed. BICOR reduces sensitivity to outliers and leverage points, making it a valuable alternative for constructing GCNs [2, 11, 15, 20, 28]. Another promising approach is the Partial Least Squares Regression (PLSR)-based connectivity scores [8]. PLSR is well-suited for high-dimensional biological data with missing values or noisy observations [27]. However, it also suffers from sensitivity to outliers and leverage points, which can compromise the reliability and predictions of the model. In this study, we propose a novel method for calculating robust connectivity scores in GCNs that directly addresses these challenges. Our approach is based on Partial Robust M-Regression (PRMR), which simultaneously downweights the influence of outliers and leverage points [23]. Unlike traditional similarity measures or PLSR-based connectivity scores, PRMR ensures reliable and accurate network structures without requiring the removal of anomalous data points.

To validate our method, a simulation study was conducted using the syntren300 data set from the R package grndata to ensure consistency of the results obtained from real data [24]. The dataset contains gene expression levels and the true underlying network used to generate them. In the true network structure, relationships are defined as present (1) or absent (0). Furthermore, we analyzed gene coexpression data from the liver of 135 female mice [21], a benchmark dataset that includes clinical traits and gene annotation information. This dataset enables a rigorous evaluation of the performance of

our method in identifying biologically significant modules and hub genes. We compare our method against WGCNA, BICOR, and PLSR-based approaches under both clean and outlier-contaminated conditions. Our findings demonstrate that the PRMR-based method consistently outperforms other approaches, particularly in the presence of outliers.

The rest of this paper is organized as follows. In the following Section, we detail the algorithms used in this study, including WGCNA, BICOR, PLSR-based scores, and our proposed PRMR-based approach. We describe their implementation details, parameter selection, and applications. Section 3 presents the applications of these methods in a simulation study and in real gene expression data analysis. The simulation study evaluates the robustness and performance of different methods under varying levels of outlier contamination, while the real data analysis explores gene co-expression networks associated with mouse body weight. We summarize key results, compare the performance of different methods, and discuss the advantages of the PRMR-based approach in gene network analysis in Section 4 where we also highlight future research directions and the potential applications of our method in biological studies. Supplementary materials, including the complete dataset and the implementation code, are provided for reproducibility and further exploration.

2. Methods

2.1. Topological Overlap Measure

The topological overlap measure (TOM), given with Eq. (2.1), is a similarity measure that captures higher-order interactions and identifies densely interconnected gene modules, which can provide information on gene regulation, biological functions, and disease mechanisms. It is calculated from the adjacency matrix considering the number of shared neighbors and their connectivity, and ranges from 0 to 1, with higher values indicating stronger topological overlap or interconnectedness between genes.

$$TOM_{ij} = \frac{\sum_u a_{iu}a_{uj} + a_{iu}}{min(k_i, k_j) + 1 - a_{ij}} \tag{2.1}$$

In Eq. (2.1), k is the row sum of the adjacency matrix with elements a_{ij} given in Eq. (2.3). Transformation to DistTOM in Eq. (2.2) allows the TOM values to be used as a dissimilarity measure in clustering analyses.

$$DistTOM_{ij} = 1 - TOM \tag{2.2}$$

Gene network analysis is performed using similarity measures calculated by any method, along with TOM and DistTOM matrices.

2.2. Weighted Gene Co-Expression Network

The Pearson correlation coefficient is calculated for each gene pair to construct weighted gene co-expression networks. Since directionality is not relevant, the absolute value of the correlation is used to construct the adjacency matrix in Eq. (2.3) where β is the soft threshold. The higher β the more emphasis is placed on high correlations [2].

$$a_{ij} = |Corr(i, j)|^{\beta} \tag{2.3}$$

2.3. Bi-weight Mid-correlation

Pearson correlation coefficient is the most common choice for similarity measures. However, it is sensitive to outliers. The bi-weight mid-correlation is considered a good alternative to Pearson's correlation, as it is more robust to outliers [16]. Calculation of bi-weight

mid-correlation for all possible gene pairs is given with Eq. (2.4) where $\boldsymbol{x} = (x_1, ..., x_n)$ and $\boldsymbol{y} = (y_1, ..., y_n)$ with $i = 1, ..., n$.

$$bicor = \frac{\sum_{i=1}^{n}(x_i - med(\boldsymbol{x}))\omega_i^{(x)}(y_i - med(\boldsymbol{y}))\omega_i^{(y)}}{\sqrt{\sum_{i=1}^{n}[(x_i - med(\boldsymbol{x}))\omega_i^{(x)}]^2}\sqrt{\sum_{i=1}^{n}[(y_i - med(\boldsymbol{y}))\omega_i^{(y)}]^2)}} \tag{2.4}$$

u_i and v_i (Eq. (2.5) and Eq.(2.6)) are calculated to obtain the robust weight $\omega_i^{(x)}$ and $\omega_i^{(y)}$ (Eq. (2.7)).

$$u_i = \frac{x_i - med(\boldsymbol{x})}{9MAD(\boldsymbol{x})} \tag{2.5}$$

$$v_i = \frac{y_i - med(\boldsymbol{y})}{9MAD(\boldsymbol{y})} \tag{2.6}$$

$$\omega_i^{(x)} = (1 - u_i^2)^2 I(1 - |u_i|) \qquad \omega_i^{(y)} = (1 - v_i^2)^2 I(1 - |v_i|) \tag{2.7}$$

$med(\boldsymbol{x})$ is the median of $\boldsymbol{x}$, $MAD(\boldsymbol{x}) = med|x_i - med(\boldsymbol{x})|$ is the median absolute deviation of $\boldsymbol{x}$, and $I(1 - |.|)$ is the indicator taking 1 if $1 - |.| > 0$, and 0 otherwise. Thus, the weight $\omega_i^{(x)}$ is close to 1 if x_i is close to $med(\boldsymbol{x})$, and is 0 if x_i differs from $med(\boldsymbol{x})$ by more than $9MAD(\boldsymbol{x})$. The weight $\omega_i^{(y)}$ is the counterpart of $w_i^{(x)}$ for y_i.

2.4. Proposed Algorithm

The PRMR, first proposed by Hubert and Verboven[14], combines the robustness of M-estimators with partial regression techniques. It was initially developed to handle outliers and leverage points in chemometric calibration models by integrating the strengths of partial least squares regression with robust weighting functions, and is particularly suited for high-dimensional data affected by multicollinearity and noise.

Pihur et al. [21] proposed the connectivity score in Eq. (2.8), a PLSR-based similarity measure, for the reconstruction of genetic association networks from microarray data. If there is an edge between two nodes (i^{th} and j^{th} genes), this edge is formed by statistically significant connectivity scores. A high connectivity score indicates a strong positive association or coexpression between the genes. Although this method is effective, it is sensitive to outliers and leverage points. To address this, we propose a robust connectivity score based on Partial Robust M-Regression (PRMR), designed to mitigate these limitations and provide reliable results. The connectivity score in Eq. (2.8) is obtained by calculating the association scores between the i^{th} and j^{th} genes, in the presence of the other genes

$$\hat{s}_{ij} = \frac{\sum_{a=1}^{A} c_{(i)}^a \nu_{(i)j}^a + \sum_{a=1}^{A} c_{(j)}^a \nu_{(j)i}^a}{2} \tag{2.8}$$

where in the first part of the numerator, the gene i is the response variable with $c_{(i)}^a$ representing the loading of the i^{th} gene on the a^{th} component. These loadings help us to understand the relationships between genes and components that capture the most significant patterns in the data. $\nu_{(i)j}^a$, on the other hand, is the contribution of the i^{th} gene. Once connectivity scores are calculated for each gene pair, the gene network can be constructed with the significant scores. To decide whether a connectivity score is significant or not, all scores are normalized from 1 to -1.

Unlike Pihur et al. [21], we calculate scores using robust weights and loadings obtained by the PRMR method [17]. Let the PRMR model be

$$y_i = \boldsymbol{t_i}\boldsymbol{\eta} + \delta_i \tag{2.9}$$

with $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_a)^T$ is the regression coefficients vector. The main advantage of the PRMR is its robustness to both leverage points and outliers. Robustness is achieved by weighting each observation using

$$w_i = \sqrt{w_i^r w_i^x} \tag{2.10}$$

where w_i^r addresses residual outliers and w_i^x accounts for leverage points. These weights are calculated as

$$w_i^r = f(h_i, c), \quad w_i^x = f(g_i, c) \tag{2.11}$$

where, f is the Fair function, $f(u,c) = \frac{1}{(1+\frac{u}{c})^2}$. The c is a tuning constant set to 4, h_i and g_i are given by Eq. (2.12) with $med_{L1}(\boldsymbol{X})$ is the robust center of design matrix $\boldsymbol{X}$, $\| \, . \, \|$ stands for the Euclidean norm. Using the Fair function provides a balance between efficiency and robustness. Unlike other robust loss functions, such as Huber or Tukey's bisquare, which can be too aggressive in complex biological datasets where extreme values may still maintain biological significance, the Fair function does not completely exclude the effect of outliers but step by step downweights them. In addition, the adjusting parameter c serves as a cut-off value in fair functions such as the Huber function and controls the level of robustness. In the literature, c is typically selected within the range of 1.5 to 4.5. In our study, we used the value c=4 because it provides sufficient robustness against outliers while maintaining statistical efficiency. These adjustments make it particularly suitable for biological data, where extreme values may represent a meaningful variation rather than simple noise. In other studies a variety of alternative weight functions and tuning constants have been explored [7, 17, 22, 23].

$$h_i = \frac{y_i - \boldsymbol{y}_{median}}{\underset{i}{\text{median}}|y_i - \boldsymbol{y}_{median}|}, \quad g_i = \frac{\| \, x_i - med_{L1}\boldsymbol{X} \, \|}{\underset{i}{\text{median}} \, \| \, x_i - med_{L1}\boldsymbol{X} \, \|} \tag{2.12}$$

Steps for calculating PRMR-based connectivity scores (gene i is treated as a response):

(1) Compute robust starting values w_i using Eq. (2.11) and Eq. (2.12).
(2) Perform NIPALS (Nonlinear Iterative Partial Least Squares) algorithm [18] on weighted observations $w_i x_i$ and $w_i y_i$.
(3) Recompute the weights w_i^r from the residuals and w_i^x from the scores $\boldsymbol{T} = (\boldsymbol{t_1}, \ldots, \boldsymbol{t_A})$ and update w_i.
(4) Return to Step (ii) and repeat until convergence.
(5) Calculate similarity scores, s_{ij}, from the final PLSR results using $c_i^a = y_i \boldsymbol{t}_a / \boldsymbol{t}_a' \boldsymbol{t}_a$ and $\nu_{(i)j}^a = \boldsymbol{X}' y_i / y_i' y_i$, where $a = 1, ..., A$ denotes the index of latent components.
(6) Construct the adjacency matrix, $a_{ij} = |s_{ij}|^\beta$.
(7) Derive TOM and DistTOM matrices from the adjacency matrix.

2.5. Performance Metrics

The majority of biological data consists of unbalanced data sets. This imbalance arises because, in a high-dimensional setting, the number of genes significantly related is relatively small. To compare the performance of the model, confusion matrices were utilized. In statistical modeling, a confusion matrix that provides a detailed comparison of the prediction of the model with the actual outcomes of a data set is commonly used to evaluate the performance of a classification model. It categorizes predictions into four main groups: correct classifications for both classes (true positives and true negatives) and incorrect classifications (false positives and false negatives). The matrix represents the number of instances that the model classified in the test dataset.

(1) True Positive (TP): The model correctly predicts a positive outcome when the actual outcome is positive.

(2) True Negative (TN): The model correctly predicts a negative outcome when the actual outcome is negative.
(3) False Positive (FP): The model incorrectly predicts a positive outcome when the actual outcome is negative, also known as a Type I error.
(4) False Negative (FN): The model incorrectly predicts a negative outcome when the actual outcome is positive, also known as a Type II error.

The key performance metrics Specificity, Precision, F1 Score, and Matthews Correlation Coefficient (MCC) can be calculated from the confusion matrix. These measures, with higher values indicating better model effectiveness, are crucial for evaluating model performance, particularly in large, unbalanced datasets with a high number of non-significant relationships [28]. The F1 score is the harmonic mean of Precision and Recall, balancing both metrics to provide a more comprehensive performance assessment. This metric is particularly useful in imbalanced datasets, where relying only on Precision or Recall may not provide an accurate representation of the effectiveness of a model. An F1 score with a high value indicates that the model performs well in terms of both identifying relevant instances (Precision) and capturing most of the actual positive cases (Recall). Since it penalizes extreme disparities between Precision and Recall, the F1 score serves as a reliable indicator of the classification performance of a model, especially when false positives and false negatives have significant implications. Furthermore, MCC is superior to the well-known Area Under Curve (AUC) metric for evaluating binary classifications because it considers sensitivity, specificity, and precision providing a balanced assessment. Unlike AUC, which can be overly optimistic by ignoring predictive values, MCC ensures that a high score reflects strong performance in all four metrics. Furthermore, while a high MCC always corresponds to a high AUC, the reverse is not always true, making MCC more reliable, especially with imbalanced datasets [29]. Table 1 displays the definitions and computations of the metrics used.

Table 1. Performance metrics: definitions and calculations.

Metric	Description	Equation
Specificity	The proportion of actual negative cases that are correctly identified as negative.	$\frac{TN}{TN+FP}$
Precision	The proportion of predicted positive instances that are true positives.	$\frac{TP}{TP+FP}$
Recall	The proportion of actual positive cases that are correctly identified by the model.	$\frac{TP}{TP+FN}$
F1 Score	A harmonic mean of precision and recall.	$2 \times \frac{Recall \times Precision}{Recall+Precision}$
MCC	Measures the overall performance of the model by considering all classification errors.	$\frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

3. Numerical Comparisons

3.1. Simulation With and Without Outliers

We conducted a simulation study to compare the methods using the syntren300 data set, which consists of 800 samples and 300 genes, from the package "grndata" in R. The dataset was generated using the publicly available SynTReN generator, which constructs synthetic gene expression data based on an E. coli source network [24]. This data set is selected for its suitability for benchmarking as it includes the ground-truth network structure. This characteristic allows for a direct and objective evaluation of algorithmic performance by comparing inferred network structures with the known underlying connections, thus

assessing the effectiveness of the proposed methods. The imbalance problem in the data arises from the presence of 728 relationships compared to 8,927 non-relationships. Here, we aim to correctly infer regulatory interactions (edges) between genes in a known synthetic E. coli gene regulatory network (GRN) generated by SynTReN. In this setup:

- Positives = gene pairs with a regulatory interaction (i.e., edges in the true E. coli network),
- Negatives = gene pairs with no interaction.

Initially, the dataset was analyzed in its original form without outliers. Subsequently, artificial outliers were introduced with contamination probabilities of 0.05, 0.10, 0.15, 0.20, 0.25, and 0.30. Outliers were generated from four different distributions: 1) Normal distribution with μ=0 and σ=5; 2) Normal distribution with μ=1 and σ=5; 3) Normal distribution with μ=100 and σ=10; and finally 4) A heavy-tailed Students t-distribution with 2 degrees of freedom. The third set of outliers is particularly extreme because of their significantly higher mean compared to the other distributions and because its low standard deviation ensures that the outliers are tightly clustered around a highly deviant value. These scenarios allow for comparison between mild and severe deviations in data analysis. Analyses were conducted on both original and contaminated datasets to evaluate the robustness of the methods. The study specifically focuses on robust network inference approaches, comparing the performance of proposed PRMR-based, PLSR-based, WGCNA, and BICOR networks under different levels of contamination.

Table 2 presents the number of TP, TN, FP, and FN for the proposed method alongside the WGCNA, BICOR, and PLSR-based methods under various distributions and contamination levels, while Figures 2-5 display the corresponding performance metrics listed in Table 1. The results are summarized as follows: Although the PLSR-based network has fewer true positives compared to the other methods, it achieves the highest values across all metrics in the no-contamination setting, regardless of the distribution. This is due to its lower number of false positives and false negatives, as well as a higher number of true negatives. The well-known WGCNA and BICOR methods yield a higher number of false positives and a lower number of true negatives across all scenarios. This suggests that they are prone to falsely predicting regulatory interactions where none exist, leading to reduced specificity and precision. Consequently, their overall performance in terms of precision, specificity, F1 score, and MCC is consistently inferior to that of the proposed method. Except in the no-contamination setting, the proposed PRMR-based method yields a higher number of true negatives and a lower number of false positives, demonstrating that it is highly effective at correctly ruling out non-interacting gene pairs. Although its true positive count may be slightly lower than WGCNA and BICOR, it achieves higher precision, indicating that its positive predictions are more likely to be correct. For example, under the N(100,10) distribution with 15% contamination, we obtain

- WGCNA precision = 689 / (689 + 15,627) = 0.042,
- PRMR precision = 477 / (477 + 4,679) = 0.093,

and under the t(2) distribution with 5% contamination, we get

- WGCNA precision = 688 / (688 + 15,373) = 0.043,
- PRMR precision = 473 / (473 + 4,841) = 0.089.

These results show that the PRMR method makes more accurate positive predictions, i.e., when it predicts an interaction, it is more likely to be correct. Although the PRMR-based method may occasionally yield more false positives, it achieves a higher F1 score, which reflects a good balance between precision and recall. This suggests that it can identify many of the true regulatory interactions while avoiding excessive inclusion of false ones. With a consistently higher number of true negatives, the proposed method correctly identifies non-interacting gene pairs, thus avoiding false discoveries. Despite occasional increases in FP, its higher specificity provides greater confidence that selected interactions

are not spurious or irrelevant, which is particularly important in imbalanced data where most potential interactions are, in fact, negatives. Finally, the higher MCC values achieved by our method demonstrate overall strong and reliable performance, even in the presence of class imbalance and noise. MCC accounts for all four components of the confusion matrix (TP, TN, FP, FN), making it a robust and informative measure in our context.

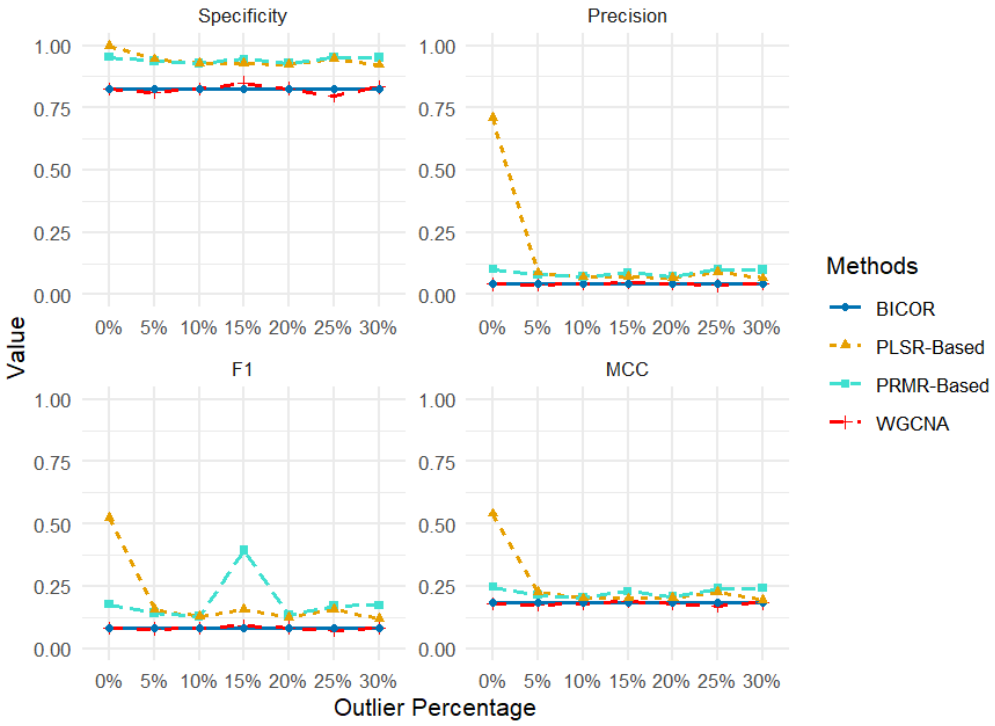

Figure 2. Performance evaluation of methods on the syntren300 dataset across different outlier levels with a distribution of N(μ=0, σ=5).

3.2. Application With and Without Outliers

We also evaluated the four methods on a real-world mouse liver gene expression dataset, a widely used benchmark that includes auxiliary and ontology data and is often used in gene co-expression network (GCN) analysis tutorials. The data set comprises 3,600 genes measured in 135 female mice and has been used to explore associations between gene expression and mouse body weight [16].

To focus on the primary variables of interest, the auxiliary data were removed and the expression matrix was transposed so that the rows corresponded to the genes and the columns to the samples. This preprocessing facilitates module detection and the investigation of genetrait relationships. During preprocessing, the dataset was examined for potential outliers and missing values. A hierarchical clustering dendrogram was generated to assess sample similarity, revealing sample F2-221 as a clear outlier. Missing values were addressed using the goodSamplesGenes function, and imputation was performed using gene-wise medians. All analyses were repeated with and without this outlier to assess robustness. The "mergeCloseModules" function was applied with a similarity threshold of 0.75 to identify and merge similar modules to ensure compact and meaningful module structures [16]. The comparative evaluation of the methods was based on their ability to detect biologically and statistically relevant gene modules.

Table 2. Detailed TP, FP, FN, and TN values for each method across varying outlier contamination rates (030%) and different data distributions (N(0,5), N(1,5), N(100,10), and t(2)).

	Distributions			**N(0,5)**				**N(1,5)**	
Outlier(%)	**Methods**	**TP**	**FN**	**FP**	**TN**	**TP**	**FN**	**FP**	**TN**
0	**WGCNA**	690	38	15630	73642	690	38	15630	73642
	BICOR	691	37	15345	73927	691	37	15345	73927
	PLSR-Based	302	426	124	89148	302	426	124	89148
	PRMR-Based	478	250	4200	85072	478	250	4200	85072
5	**WGCNA**	700	28	16892	72380	695	33	16645	72627
	BICOR	691	37	16335	73937	691	37	15335	73937
	PLSR-Based	487	241	6687	82585	473	255	4871	84401
	PRMR-Based	497	231	8115	81157	479	249	5541	83731
10	**WGCNA**	691	37	15519	73753	687	41	15033	74239
	BICOR	691	37	15347	73925	691	37	15349	73923
	PLSR-Based	485	243	6387	82885	478	250	5554	83718
	PRMR-Based	492	236	6376	82896	484	244	5156	84116
15	**WGCNA**	671	57	13341	75931	689	39	15627	73645
	BICOR	691	37	15343	73929	691	37	15345	73927
	PLSR-Based	486	242	6364	82908	480	248	6114	83158
	PRMR-Based	483	245	4937	84335	477	251	4679	84593
20	**WGCNA**	688	40	15330	73942	689	39	15471	73801
	BICOR	691	37	15339	73933	691	37	15341	73931
	PLSR-Based	496	232	6304	82968	497	231	7767	81505
	PRMR-Based	505	223	6743	82429	477	251	4337	84935
25	**WGCNA**	709	19	18035	71237	689	39	15455	73817
	BICOR	691	37	15345	73927	691	37	15345	73927
	PLSR-Based	475	253	4845	84427	489	239	6687	82585
	PRMR-Based	477	251	4343	84929	483	245	4597	84675
30	**WGCNA**	684	44	14896	74376	684	44	14912	74360
	BICOR	691	37	15343	73929	691	37	15345	73927
	PLSR-Based	489	239	7013	82259	481	247	6065	83207
	PRMR-Based	477	251	4307	84965	479	249	4641	84631
	Distributions			**N(100,10)**				**t(2)**	
Outlier(%)	**Methods**	**TP**	**FN**	**FP**	**TN**	**TP**	**FN**	**FP**	**TN**
0	**WGCNA**	690	38	15630	73642	690	38	15630	73642
	BICOR	691	37	15345	73927	691	37	15345	73927
	PLSR-Based	302	426	124	89148	302	426	124	89148
	PRMR-Based	478	250	4200	85072	478	250	4200	85072
5	**WGCNA**	564	164	16620	72652	688	40	15420	73852
	BICOR	691	37	15337	73935	691	37	15333	73939
	PLSR-Based	415	313	4731	84541	473	255	5403	83869
	PRMR-Based	396	332	2834	86438	473	255	4841	84431
10	**WGCNA**	648	80	14672	74600	689	39	15373	73899
	BICOR	691	37	15347	73925	691	37	15345	73927
	PLSR-Based	500	228	6275	82996	481	247	5649	83623
	PRMR-Based	491	237	4939	84333	473	255	4841	84431
15	**WGCNA**	689	39	15627	73645	689	39	15557	73715
	BICOR	691	37	15347	73927	691	37	15349	73923
	PLSR-Based	480	248	6114	83158	477	251	5419	83853
	PRMR-Based	477	251	4679	84593	482	246	5220	84052
20	**WGCNA**	637	91	12435	76837	689	39	15459	73813
	BICOR	691	37	15343	73929	691	37	15343	73929
	PLSR-Based	510	218	6164	83108	469	259	4705	84567
	PRMR-Based	513	215	7147	82125	478	250	4658	84614
25	**WGCNA**	688	40	15542	73730	690	38	15540	73732
	BICOR	693	35	15583	73689	691	37	15345	73927
	PLSR-Based	470	258	4618	84654	480	248	5836	83436
	PRMR-Based	478	250	4476	84796	478	250	4760	84512
30	**WGCNA**	671	57	13965	75307	790	19	18035	71237
	BICOR	691	37	15343	73929	691	37	15345	73927
	PLSR-Based	491	237	5223	84049	476	252	5038	84234
	PRMR-Based	472	256	3618	85654	478	250	4414	84858

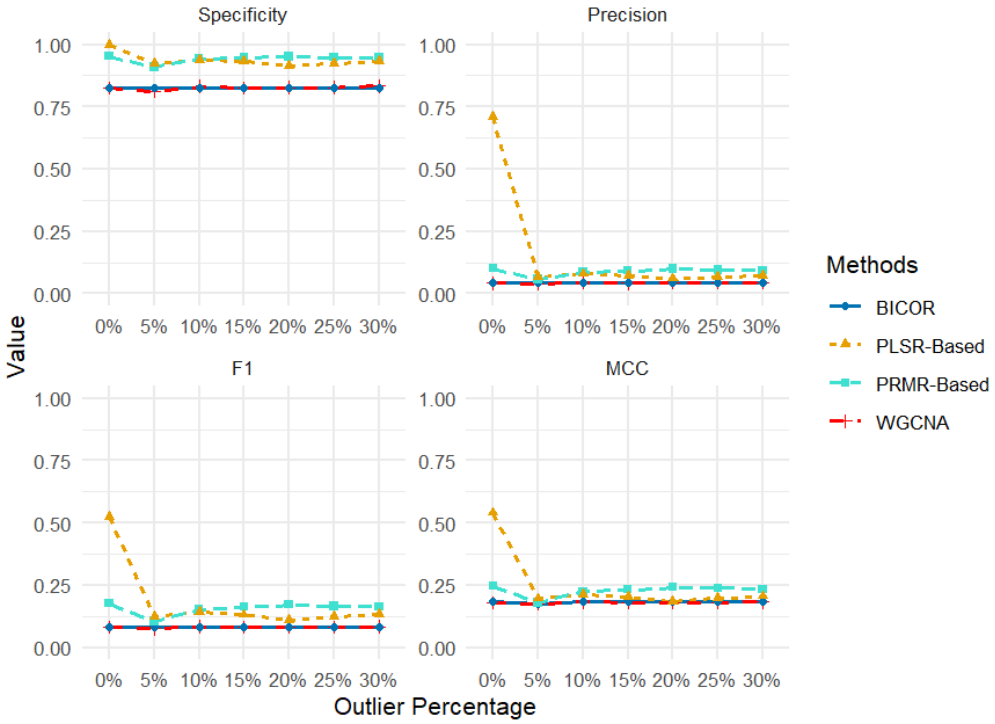

Figure 3. Performance evaluation of methods on the syntren300 dataset across different outlier levels with a distribution of N(μ=1, σ=5).

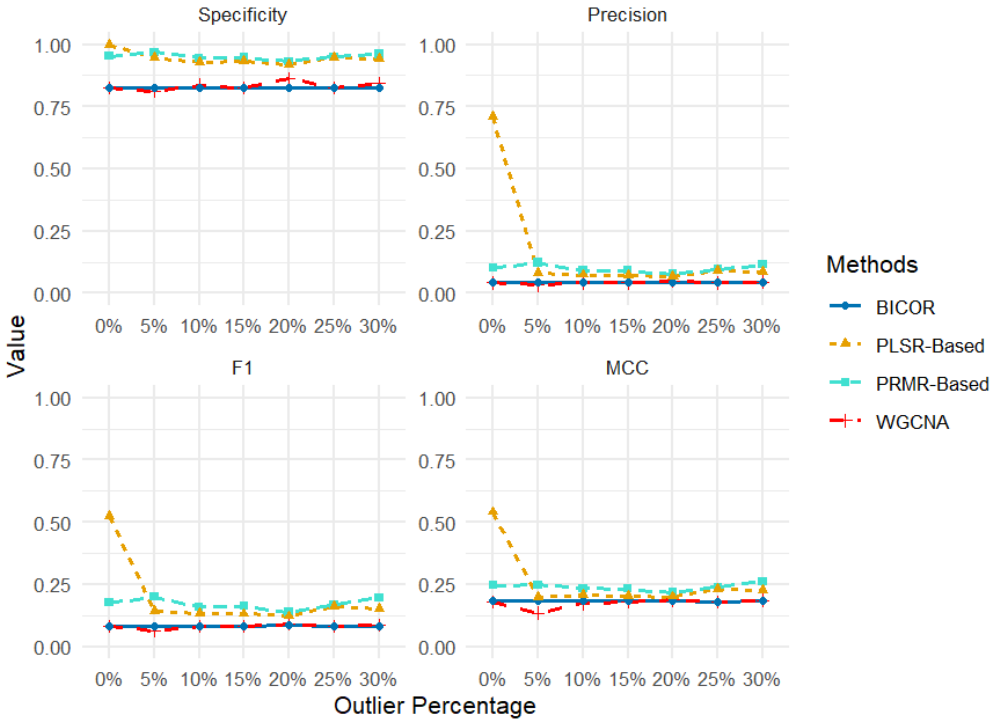

Figure 4. Performance evaluation of methods on the syntren300 dataset across different outlier levels with a distribution of N(μ=100, σ=10).

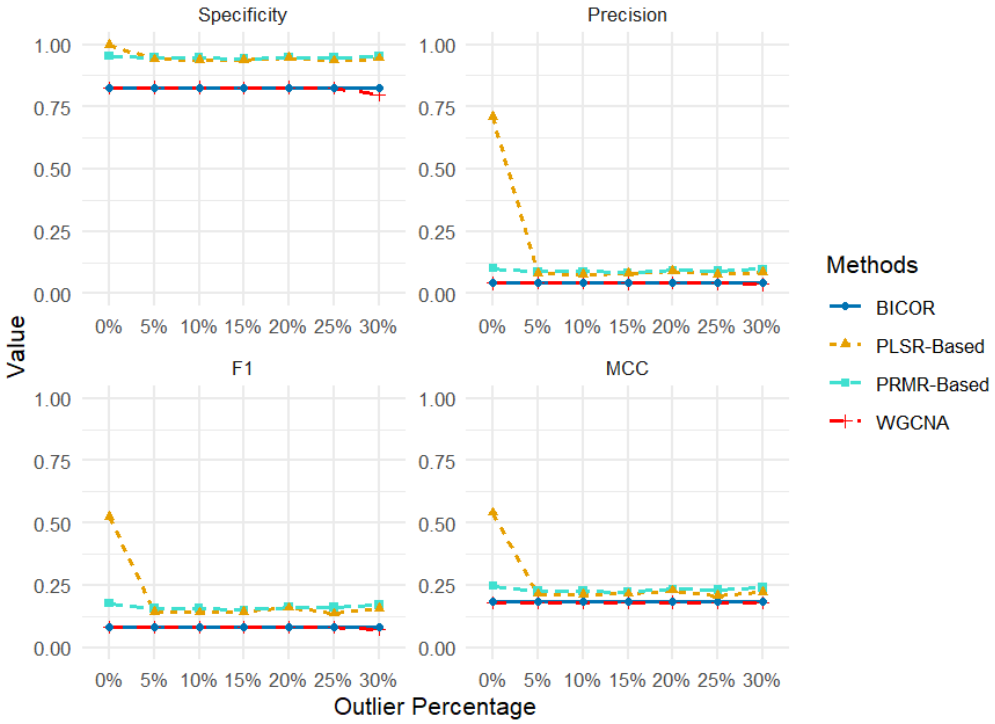

Figure 5. Performance evaluation of methods on the syntren300 dataset across different outlier levels with a distribution of t(2).

Following the construction of network structures (for both outlier-included and outlier-excluded datasets), we examined moduletrait relationships to identify modules significantly associated with mouse body weight. To assess biological relevance, Gene Ontology (GO) enrichment analysis was performed for each identified module. Only modules showing statistically significant associations with the trait ($p < 0.05$) were considered. Enrichment was considered successful if at least one GO term was significantly enriched at an adjusted p-value < 0.05 (Benjamini-Hochberg correction). The number of genes in such modules served as a proxy for the number of "correctly identified genes". This approach integrates both statistical association and biological validation.

It is important to note that, unlike in the simulation study where the ground truth of gene-gene or gene-trait associations is known, in real biological data the exact set of truly relevant genes is essentially unknown. As a result, classical performance metrics such as TP, TN, FP, and FN cannot be directly calculated for real datasets. Instead, we adopt a composite evaluation strategy that combines statistical significance (module-trait correlation) with biological relevance (significant GO enrichment) to define "correctly identified genes". This approach provides a practical and interpretable proxy for evaluating biological validity in the absence of ground truth.

Significant module-trait relationships were observed for the blue, magenta, salmon, cyan, brown, and purple modules ($p < 0.05$). Scatter plots for the blue and magenta modules (Figs. 6B and 6C) revealed a strong positive correlation between gene significance and module membership, indicating that these modules contained genes strongly associated with both the trait and their network module.

The number of genes correctly identified for each method is summarized in Table 3, based on the full set of 3,600 genes. For example, the blue module, identified by the PRMR-based method (using the dataset excluding the outlier), consisted of 428 genes, of which 423 were eligible for GO enrichment analysis. This module was significantly associated

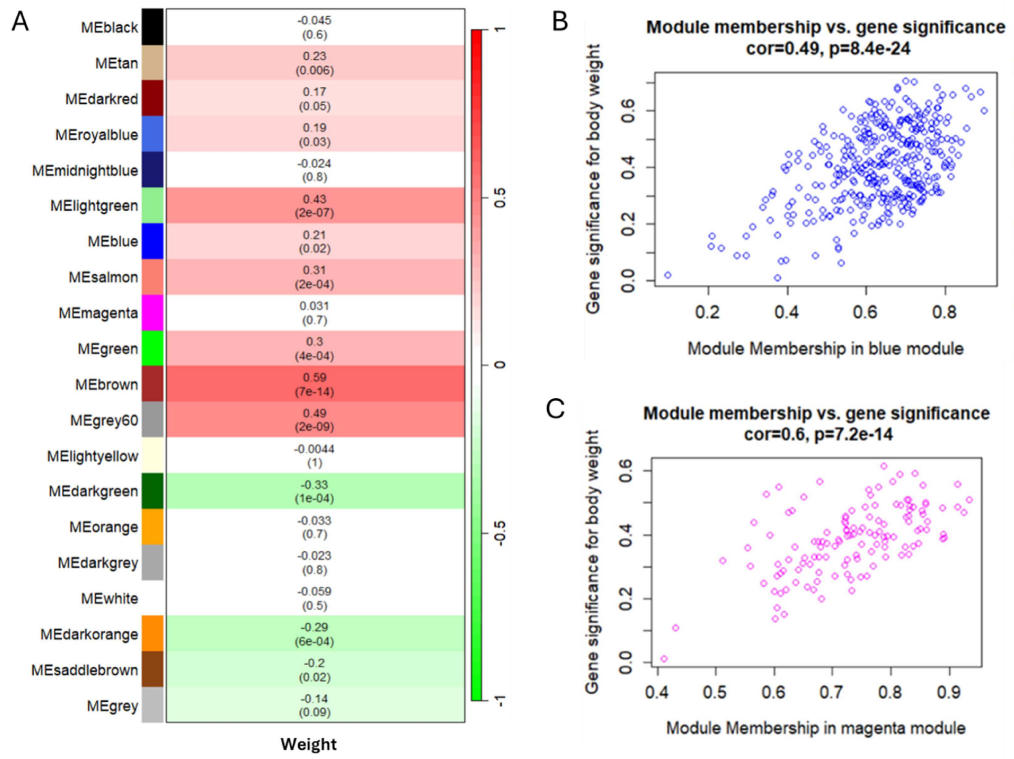

Figure 6. Module-trait relationships: correlation analysis between gene modules and body weight. The red boxes indicate a statistically significant positive relationship between a module and weight, while the green boxes suggest a significant negative relationship. (A) Module-phenotype correlation diagram. (B) and (C) Scatter plot showing the correlation between the blue and magenta module genes and weight. The horizontal axis represents the degree of module membership, and the vertical axis represents gene significance.

Table 3. Number of correctly identified genes out of the full set of 3600 genes by different methods, with and without outliers.

Outlier	WGCNA	BICOR	PLSR-based	PRMR-based
Without	1539	1053	1028	1053
With	944	1016	978	1080

with body weight and showed enrichment for several biological functions. The PRMR-based method again demonstrated strong robustness to outliers or leverage points. While WGCNA identified more genes when the outlier was excluded, its performance declined markedly when the outlier was included. In contrast, the PRMR-based method maintained high classification performance and biological interpretability even in the presence of the outlier. This robustness was not observed in the BICOR or PLSR-based methods, both of which also showed degraded performance when outliers were present. Thus, the PRMR-based approach outperformed the WGCNA, BICOR, and PLSR-based methods under noisy, real-world conditions in terms of gene identification consistency and biological validation through GO enrichment.

4. Conclusion

This study introduces a Partial Robust M-Regression (PRMR)-based approach for constructing gene co-expression networks (GCNs), specifically designed to handle outliers and leverage points common challenges in gene expression data. By iteratively down-weighting outlying observations, the method computes connectivity scores that remain robust and stable even under substantial data contamination, thereby preserving biologically meaningful structures without discarding potentially informative variation. Conventional methods such as weighted gene co-expression network analysis (WGCNA), bi-weighted mid-correlation (BICOR), and partial least squares regression (PLSR)-based connectivity scores generally perform well on clean datasets. However, their performance degrades markedly in the presence of moderate to severe outliers. WGCNA and BICOR show relatively stable, yet lower, classification performance, while PLSR-based approaches suffer sharp declines in metrics such as precision and F1 score with even minimal contamination.

In contrast, the PRMR-based method consistently demonstrates superior robustness, strong reliability under contamination, and balanced performance. By down-weighting outliers instead of excluding them, it protects the global network structure while preserving signals from extreme but potentially informative observations. Across different contamination scenarios and outlier-generating distributions (e.g., large mean shifts, increased variance), PRMR maintains high performance in simulation studies, particularly on F1 and Matthews Correlation Coefficient (MCC). In the simulation setup using the E. coli syntren300 network, the PRMR-based method achieves higher or comparable specificity, precision, F1 score, and MCC relative to the WGCNA, BICOR, and PLSR-based networks, especially when the data deviate from ideal conditions.

In the real-data application, using mouse liver gene expression data, the PRMR-based network effectively identified biologically significant genes through module-trait correlation and Gene Ontology (GO) enrichment analysis. In particular, it outperformed other methods when a known outlier was included, maintaining biological interpretability and stability, while WGCNA and others showed a decrease in performance.

Together, the results of both simulation and real data analysis provide compelling evidence that the PRMR-based network is statistically robust, biologically valid, resistant to contamination and imbalance, and suitable for high-dimensional genomic data analysis. Future directions can include extending the PRMR-based framework to more complex datasets and integrating domain-specific prior knowledge (e.g., known pathways or phenotypic traits) to further enhance interpretability and biological insight. In summary, the PRMR-based network method addresses a critical gap in GCN construction by offering a reliable, interpretable, and outlier-resilient alternative to standard approaches in modern gene expression studies.

Acknowledgements

We would like to thank the editor and the reviewers for their constructive comments, which significantly improved the quality of the original manuscript.

Author contributions. All the co-authors have contributed equally in all aspects of the preparation of this submission.

Conflict of interest statement. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding. The first author is supported by the Department of Scientific Research Projects (BAP Project No: 2021.KB.FEN.038)

Data availability. All codes and data are available on GitHub (https://github.com/olmezayca/Simulation-Study and https://github.com/olmezayca/GNA-with-Robust-Methods).

References

[1] M. Ackermann and K. Strimmer. A general modular framework for gene set enrichment analysis. BMC Bioinform. **10**, 120, 2009.

[2] I.J. Broce, L. Stein, E. Jones and T. Kwan. C9orf72 gene networks in the human brain correlate with cortical thickness in C9-FTD and implicate vulnerable cell types. Front. Neurosci. **18**, 1258996, 2024.

[3] J.B. Brown. Classifiers and their metrics quantified. Mol. Inform. **37**, 1700127, 2018.

[4] Q. Chen, R. Liu, X. Wang and H. Chen. Identification and analysis of spinal cord injury subtypes using weighted gene co-expression network analysis. Ann. Transl. Med. **9**, 466, 2021.

[5] D. Chicco and G. Jurman. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Min. **16**, 4, 2023.

[6] R Core Team et al., R: A language and environment for statistical computing, R Found. Stat. Comput., Vienna, 2013.

[7] D.J. Cummins and C.W. Andrews. Iteratively reweighted partial least squares: a performance analysis by Monte Carlo simulation. J. Chemom. **9**, 489507, 1995.

[8] S. Datta. E xploring relationships in gene expressions: a partial least squares approach. Gene Expr. **9**, 249, 2018.

[9] Y. Di, D. Chen, W. Yu and L. Yan. Bladder cancer stage-associated hub genes revealed by WGCNA co-expression network analysis. Hereditas **156**, 111, 2019.

[10] A.S. Feltrin, P.R. Castro, L.F. Silva and R. Costa. Assessment of complementarity of WGCNA and NERI results for identification of modules associated to schizophrenia spectrum disorders. PLoS One **14**, e0210431, 2019.

[11] E. Galán-Vásquez and E. Perez-Rueda. Identification of modules with similar gene regulation and metabolic functions based on co-expression data. Front. Mol. Biosci. **6**, 139, 2019.

[12] A. Ghazalpour, B. Bennett, S. Plaisier and J. Chen. Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet. **2**, e130, 2006.

[13] D.J. Guo, R. Lin and Y. Li. Identification of breast cancer prognostic modules via differential module selection based on weighted gene co-expression network analysis. BioSystems. **199**, 104317, 2021.

[14] M. Hubert and S. Verboven. Robust methods for partial least squares regression. J. Chemom. **17**, 537-549, 2003.

[15] S. Langfelder and S. Horvath. Fast R functions for robust correlations and hierarchical clustering. J. Stat. Softw. **46**, 117, 2012.

[16] S. Langfelder and S. Horvath. WGCNA: an R package for weighted correlation network analysis. BMC Bioinform. **9**, 113, 2008.

[17] B. Liebmann, P. Filzmoser and K. Varmuza. Robust and classical PLS regression compared. J. Chemom. **24**, 111120, 2010.

[18] W. Li, Y. Zhang, L. Zhang and Q. Wang. Weighted gene co-expression network analysis to identify key modules and hub genes associated with atrial fibrillation. Int. J. Mol. Med. **45**, 401416, 2020.

[19] S. Pan, Y. Lin, X. Liu and L. Wang. A comprehensive weighted gene co-expression network analysis uncovers potential targets in diabetic kidney disease. J. Transl. Intern. Med. **10**, 359368, 2023.

[20] Y. Peng, X. Chen, Q. Zhang and D. Liu. Identification of immune-related biomarkers in adrenocortical carcinoma: immune-related biomarkers for ACC. Int. Immunopharmacol. **88**, 106930, 2020.
[21] V. Pihur, S. Datta and S. Datta. Reconstruction of genetic association networks from microarray data: a partial least squares approach. Bioinformatics **24**, 561568, 2008.
[22] E. Polat, The effects of different weight functions on partial robust m-regression performance: a simulation study. Comms. in Stats. - Sim. and Comput. **49:4**, 1089-1104, 2020.
[23] S. Serneels, K. Suykens, B. De Moor and S. Van Huffel. Partial robust M-regression. Chemom. Intell. Lab. Syst. **79**, 5564, 2005.
[24] T. Van Den Bulcke, K. Van Leemput, B. Naudts and P. Van Remortel. SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC Bioinform. **7**, 112, 2006.
[25] F. Wang, H. Miao, S. Zhang, X. Hu, Y. Chu, W. Yang and J. Chen. Weighted gene co-expression network analysis reveals hub genes regulating response to salt stress in peanut. BMC Plant Biol. **24**, 425, 2024.
[26] B. Wang, X. Li, Y. Zheng and R. Zhang. Research on a weighted gene co-expression network analysis method for mining pathogenic genes in thyroid cancer. PLoS One **17**, e0272403, 2022.
[27] S. Wold, M. Sjöström and L. Eriksson. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. **58**, 109130, 2001.
[28] C.H. Zheng, Y. Li, L. Wang and W. Zhang. Gene differential coexpression analysis based on biweight correlation and maximum clique. BMC Bioinform. **15**, 17, 2014.
[29] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. **4**, 129, 2005.