Core Mathematical Function of SlimR v1.0.9

Zhaoqing Wang

zhaoqingwang@mail.sdu.edu.cn

Section 1 Parameter Calculate

Step 1: Coefficient of Variation (CV)

Coefficient of Variation (CV) quantifies the relative variability in gene expression, serving as a critical feature for parameter prediction. It normalizes expression dispersion by accounting for mean expression levels, enabling fair comparison across genes with different expression magnitudes.

$$CV = \frac{\mu}{\sigma + \epsilon}$$

Where σ is the standard deviation of gene expression, μ is the mean expression, and $\epsilon = 10^{-6}$ prevents division by zero.

Step 2: Expression Skewness

Expression Skewness captures the asymmetry of gene expression distributions across cells, reflecting the non-normality of biological expression patterns. It influences the minimum expression threshold by accounting for highly skewed distributions that may contain biological signals in the tail.

$$Skew = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)3}{\sigma^3}$$

Where x_i are individual expression values, μ is the mean expression, and μ is the standard deviation.

Step 3: Cluster Variability

Cluster Variability quantifies the separation between cell clusters by measuring the average distance between cluster centroids. It directly informs the specificity weight,

as greater cluster separation requires higher weighting to preserve distinct cell identities during annotation.

$$Variability = \frac{1}{\binom{n}{2}} \sum_{i<j} d(\mu_i, \mu_j)$$

Where $d(\mu_i, \mu_j)$ is the Euclidean distance between cluster centroids μ_i and μ_j, and n is the number of clusters.

Step 4: Optimal min_expression Calculation

To determine the minimum expression threshold that balances noise filtering with biological signal preservation. The following formula incorporates dataset sparsity and expression distribution shape to set the threshold for meaningful gene expression adaptively.

$$base_{min} = 0.05 + 0.15 \times zero\_frac$$

$$adj_{min} = base_{min} \times (1 + 0.2 \times Skew)$$

$$min_{expression} = clamp(adj_{min}, 0.01, 0.3)$$

Where $zero\_frac$ is the global fraction of zero expression values, and $clamp$ constrains values to [0.01, 0.3].

Step 5: Optimal specificity_weight Calculation

To set the weight for cluster specificity, balance the importance of cluster separation in annotation. The following formula accounts for both cluster separation strength and gene expression variability to optimize the confidence in cluster assignments.

$$base_{weight} = 1 + 2 \times \frac{Variability}{1 + sparsity}$$

$$adj_{\ weight} = base_{weight} \times (1 + 0.5 \times mean\_CV)$$

$$specificity_{weight} = clamp(adj_{weight}, 0.5, 8)$$

Where $sparsity$ is the expression sparsity (fraction of zero values), $mean\_CV$ is the mean coefficient of variation across genes, and $clamp$ constrains values to [0.5, 8].

Step 6: Dataset-Specific Post-Processing

To adjust and improve the prediction parameters based on the characteristics of extreme data sets. Functions as follows ensure parameters remain biologically meaningful by compensating for unusual data properties that the base formulas might not fully capture.

- For highly sparse datasets ($zero_frac > 0.8$):

$$min_{expression} = min_{expression} \times 1.2$$

- For poor cluster separation ($Variability < 1$):

$$specificity_{weight} = specificity_{weight} \times 1.3$$

Step 7: Model Performance Metric (R-squared)

Finally, the following statistical indicators were used to evaluate the accuracy of the machine learning model in predicting the optimal parameters. It quantifies how well the model captures the relationship between the dataset's characteristics and the target parameters, thereby providing a reliable basis for parameter recommendations.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Where y_i are actual target values, $\hat{y}_i$ are predicted values, and $\bar{y}$ is the mean of actual values.

Section 2 Celltype Calculate

Variables

- C: Number of cell clusters from input cluster_col.
- A: Number of cell clusters from the Markers_list.
- N: Number of genes in specific cell types from the Markers_list.
- $i \in \{1, C\}$: cluster_col's cluster index.
- $a \in \{1, A\}$: Markers_list's cluster index.
- $g \in \{1, N\}$: Markers_list's gene index corresponding to the current cell type.
- $x_{g,i}$: Expression value of gene g in cluster i .
- $\mu_{g,i}$: Average expression of gene g in cluster i .
- $\sigma_{g,i}$: Standard deviation of gene g in cluster i .
- $f_{g,i}$: Fraction of cells in cluster i where $x_{g,i} > m$ (minimum expression threshold m; default: 0.1).
- w: Specificity weight parameter (default: 3).
- $\sigma_{g,i}$: Average standard deviation of all genes in cluster i .
- ε: Small constant to avoid division by zero (default: 1×10^{-6}).

Step 1: Specificity Score Calculation

For each gene g in cluster i corresponding to the current cell type a in the Markers_list:

$$s_{g,i} = \mu_{g,i} \cdot f_{g,i} \cdot (1 + w \cdot \frac{\sigma_{g,i}}{\overline{\sigma}_i + \varepsilon})$$

Explanation:

- $\mu_{g,i}$: Mean expression level.
- $f_{g,i}$: Proportion of cells expressing g.
- $\sigma_{g,i}i$: Normalized variability of g compared to other genes in the cluster.
- w: Amplifies the impact of high variability.

Step 2: Normalization of Specificity Scores

Normalize $s_{g,i}$ across genes per cluster i :

$$s'_{g,i} = \frac{s_{g,i} - min(s_{g,i})}{max(s_{g,i}) - min(s_{g,i})} \ (if\ max(s_{g,i}) \neq min(s_{g,i}))$$

Purpose: Ensures scores are comparable across genes within the same cluster.

Step 3: Gene Weight Calculation

Compute weights for genes based on their variability-to-mean ratio:

$$g_w = \frac{sd(\sigma_{g,i})}{mean(\mu_{g,i})} \ (if\ mean(\mu_{g,i}) \neq 0)$$

Purpose: Prioritize genes with higher variability and lower mean expression.

Step 4: Cluster-Specific Gene Expression Score

Aggregate normalized scores $s'_{g,i}$ into a final cluster score $p_{i,a}$ corresponding to the current cell type a in the Markers_list:

$$p_{i,a} = \sum_{g=1}^{N} g_w \cdot s'_{g,i}$$

Interpretation:

- $p_{i,a}$ reflects the weighted sum of gene-specificity scores for cluster i.
- Higher $p_{i,a}$ indicates stronger evidence for the cluster corresponding to the current cell type a .

Step 5: Final Probability Matrix

For cluster_col's cluster $i \in \{1, C\}$ and Markers_list's clusters $a \in \{1, A\}$, the function outputs a probability matrix R where:

$$R_{i,a} = \frac{p_{i,a} - \min\limits_{b \in A}(p_{i,b})}{\max\limits_{b \in A}(p_{i,b}) - \min\limits_{b \in A}(p_{i,b})} \ (if\ \max_{b \in A}(p_{i,b}) \neq \min_{b \in A}(p_{i,b}))$$

Note: For each cluster_col's cluster i , the cell type with the highest normalized score p_i above threshold is selected as predicted cell type.

Step 6: AUC Validation

AUC Correction: ROC-AUC is computed using mean expression of signature genes to validate predictions:

$$AUC = \int_0^1 TPR(FPR)\, dFPR$$

Where：

$$TPR = f(mean(x_{g,i}))$$

Interpretation:

- True Positive Rate (TPR): $Sensitivity = \frac{TP}{TP + FN}$.
- False Positive Rate (FPR): $1 - Specificity = \frac{FP}{FP + TN}$.
- $x_{g,i}$: Expression value of gene g in cluster i .

Section 3 Celltype Verification

Variables

- $c \in \{1, A\}$: After annotation cluster index.
- $g \in \{1, \mathrm{N}\}$: Gene index.
- $x_{g,i}$: Expression value of gene g in cluster i .
- $f_{g,i}$: Fraction of cells in cluster i where $x_{g,i} > m$ (minimum expression threshold m; default: 0.1).
- k: Top gene count (default: 5).

Gene Scoring System

When the cell type c is in "Markers_list", verification markers uses the markers $g \in \{1, \mathrm{N}\}$ corresponding to the specific cell type in it.

Screening of verification markers for cell types c not located in "Markers_list", compute each gene g in after annotation cell types c :

$$G_c^k = \top_k\left(\sum_{c \neq j} \log_2\left(\frac{\mu_{g,c}}{\overline{\mu}_{g,j}}\right) \cdot f_{g,c}\right)$$

Note: $\overline{\mu}_{g,j}$ = mean($\mu_{g,j}$ for all $j \neq c$): Average mean expression in other clusters.

Feature Significance Score (FSS)

Feature Significance Score, FSS, product value of ‘log2FC’ and ‘Expression ratio’:

$$FSS = \Delta \log_2\left(\mu_{g,c}\right) \cdot f_{g,c}$$

Where:

$$\Delta\log_2\left(\mu_{g,c}\right) = \log_2\left(\frac{\mu_{g,c}}{\overline{\mu}_{g,j}}\right)$$

Note: The 'FSS' parameter is also used in the 'Read_seurat_markers()' function for Markers screening.
