The correlation between binomial deviance and gene Laplacian scores provides independent evidence that deviance is an appropriate filtering criterion.

Recall the analytic information flow described in this posting, specifically the path to obtaining gene Laplacian scores:

1. Calculate binomial deviance for genes from UMI counts.
2. Calculate binomial deviance for randomized counts to obtain a filtering threshold.
3. Select genes with binomial deviance exceeding a selected threshold.
4. Calculate standardized **null residuals**.
5. Apply random forest classification to obtain gene proximities.
6. Cluster genes with distance derived from these proximities.
7. Select an appropriate number of gene clusters.
8. Reduce the feature space dimension by using gene cluster means.
9. Apply random forest classification to obtain **cell** **proximities**.

Gene Laplacian scores are computed from the **null residuals** and **cell** **proximities**.

Choosing a filtering threshold

Step 3 suggests selecting a filtering threshold based on randomized counts.

For example, one randomization of the Zhengmix4eq counts gave 2827 as the maximum binomial deviance. Examples below suggest that the maximum is an appropriate threshold, but its variability is large: standard deviation 10 times that of the 99.9th percentile; 20 times that of the 99.5th percentile. Therefore, the results presented here do not rely on a single estimate of the maximum deviance. Rather, they are based on

* calculating binomial deviance for 11 sets randomized counts and
* choosing the median of the 11 maxima as a filtering threshold.

This table displays the relation between median binomial deviance and the number of genes selected for specific threshold values for the Zhengmix4eq data:

|  |  |  |
| --- | --- | --- |
| **percentile** | **binomial deviance** | **genes** |
| maximum | 2767 | 164 |
| 99.9 | 2247 | 268 |
| 99.5 | 2059 | 322 |
| 99 | 1981 | 345 |
| 95 | 1773 | 429 |
| 90 | 1672 | 477 |
| 75 | 1510 | 564 |
| 50 | 1346 | 689 |
| 25 | 1198 | 856 |

Read as

* Top row:
  + the median of the maxima from the 11 randomized sets of counts is 2767
  + 164 genes are selected with this threshold
* Bottom row:
  + the median of the 25th percentile of binomial deviance from the 11 randomized sets of counts is 1198
  + 856 genes are selected with this threshold

For the Zhengmix8eq data

* 195 genes are selected with the “maximum” threshold
* 925 genes are selected with the “25th percentile” threshold

Correlation between binomial deviance and Laplacian score

He et al. [1] show that the Laplacian and Fisher scores are related by

L = 1 / ( 1 + F )

That is

* The Laplacian score is bounded between 0 and 1.
* Features with smaller Laplacian scores are preferred.

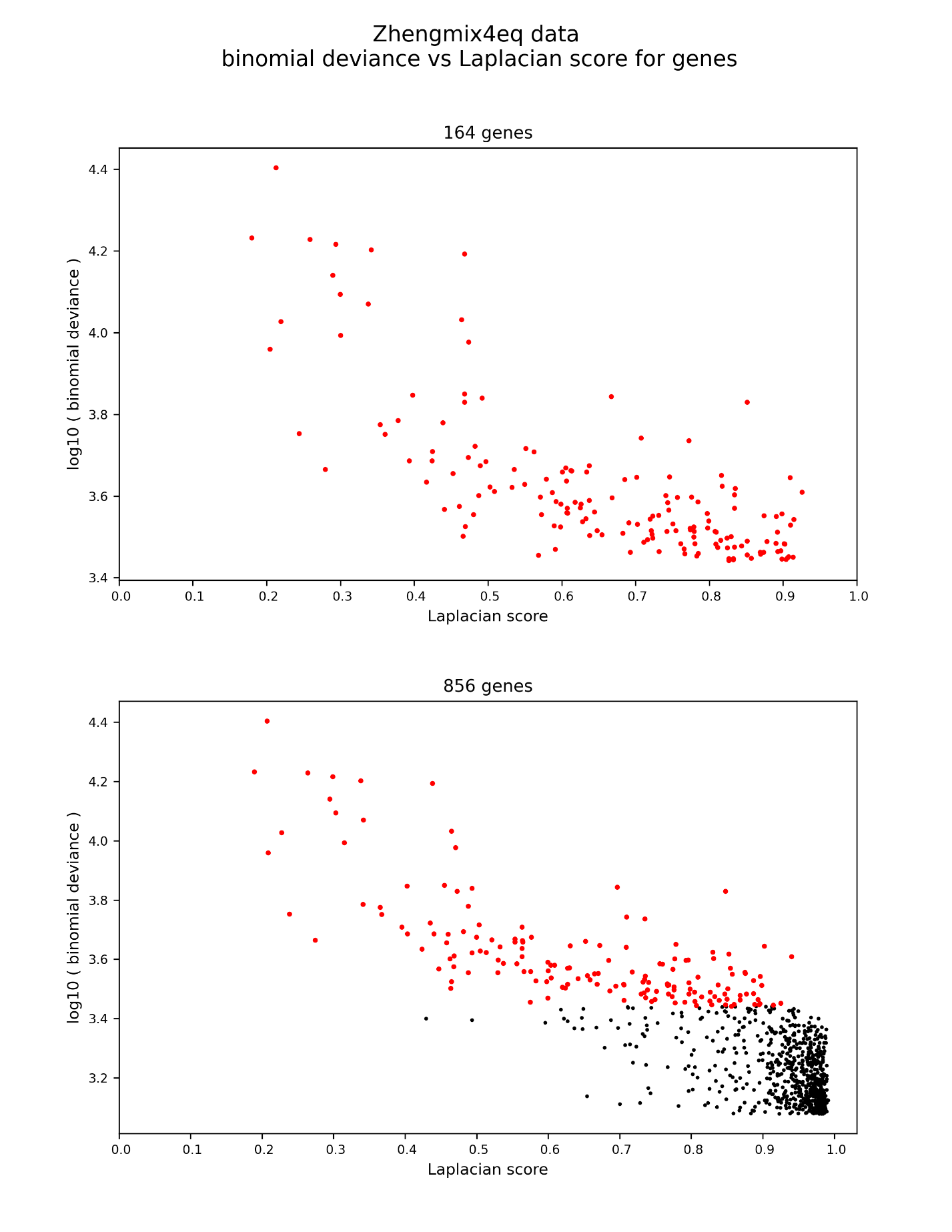
The relation between binomial deviance and Laplacian score is illustrated in scatter plots below. For each of the two Zhengmix data sets there are two plots:

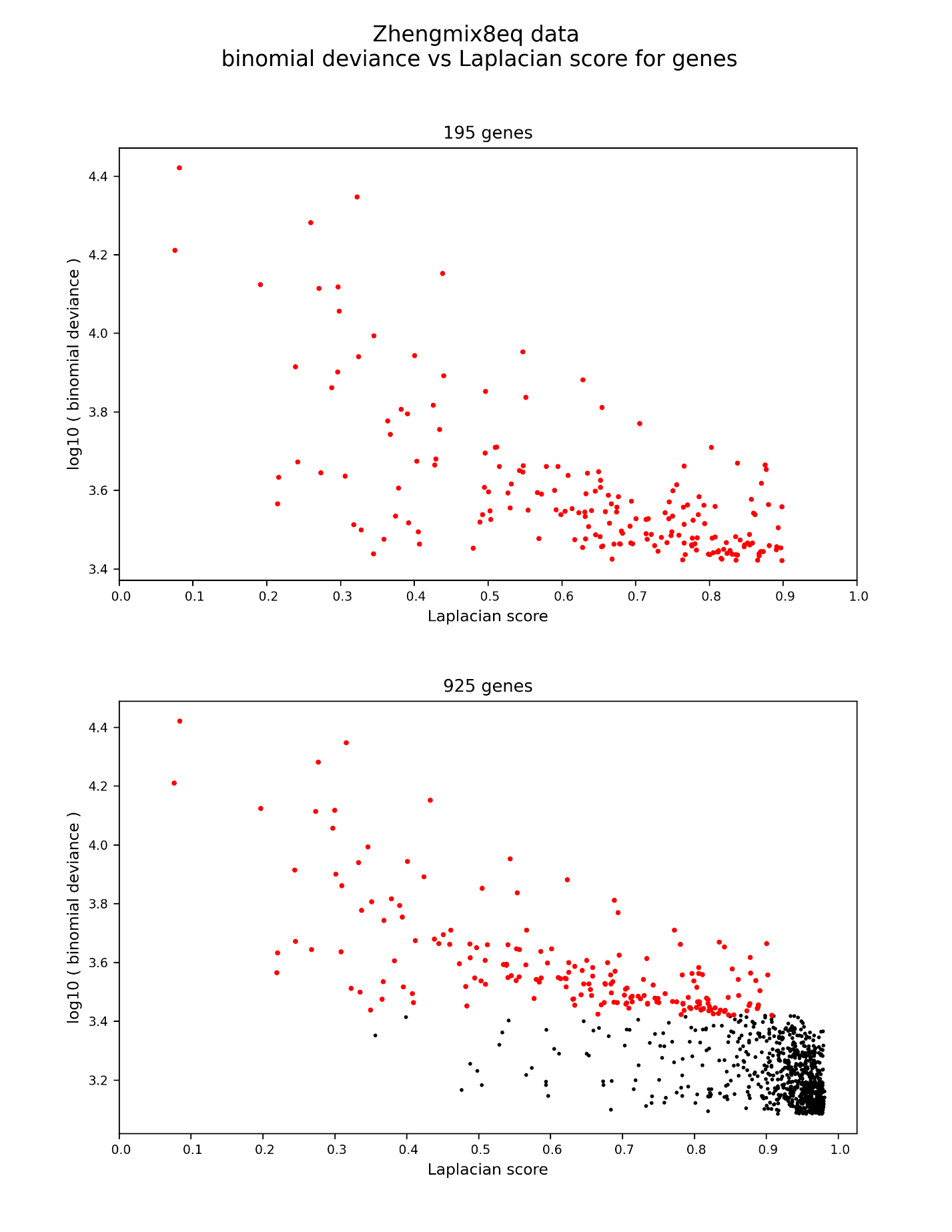
* Genes satisfying the “maximum” threshold
* Genes satisfying the “25th percentile” threshold

The “25th percentile” plots distinguish genes satisfying the “maximum” threshold. Most of the genes with deviance below the “maximum” threshold have Laplacian scores near 1. According to the criteria of **both** Townes et al. **and** He et al. they provide poor discrimination.

For completeness, a technical detail:

* For Zhengmix4eq data, Laplacian scores for genes satisfying the “maximum” threshold are calculated with null residuals for 164 genes
* Laplacian scores for the “25th percentile” genes require residuals for 856 genes.
* Hence, for the 164 genes satisfying the “maximum” threshold, two sets of Laplacian scores are calculated. The differences are negligible.





**Reference**

1. He X, Cai D, Niyogi P: Laplacian score for feature selection. In *Advances in neural information processing systems*. 2006: 507-514.