

Supplementary Material: “Spatially Regularized Gaussian Mixtures for Clustering Spatial Transcriptomic Data”

Andrea Sottosanti^{1*}, Davide Risso¹ and Francesco Denti¹

*Corresponding author(s). E-mail(s): andrea.sottosanti@unipd.it;
Contributing authors: davide.risso@unipd.it; francesco.denti@unipd.it;

This Supplementary Material includes:

- visualizations of the computational cost of SR-GMM estimation, measured on simulated data (Figure 1) and on the prostate cancer tissue sample analyzed in Section 5 of the manuscript (Figure 3);
- a summary of gene expression levels across the entire prostate cancer tissue sample, with values transformed using the deviance-based method proposed by Townes et al. (2019) (Figure 2).

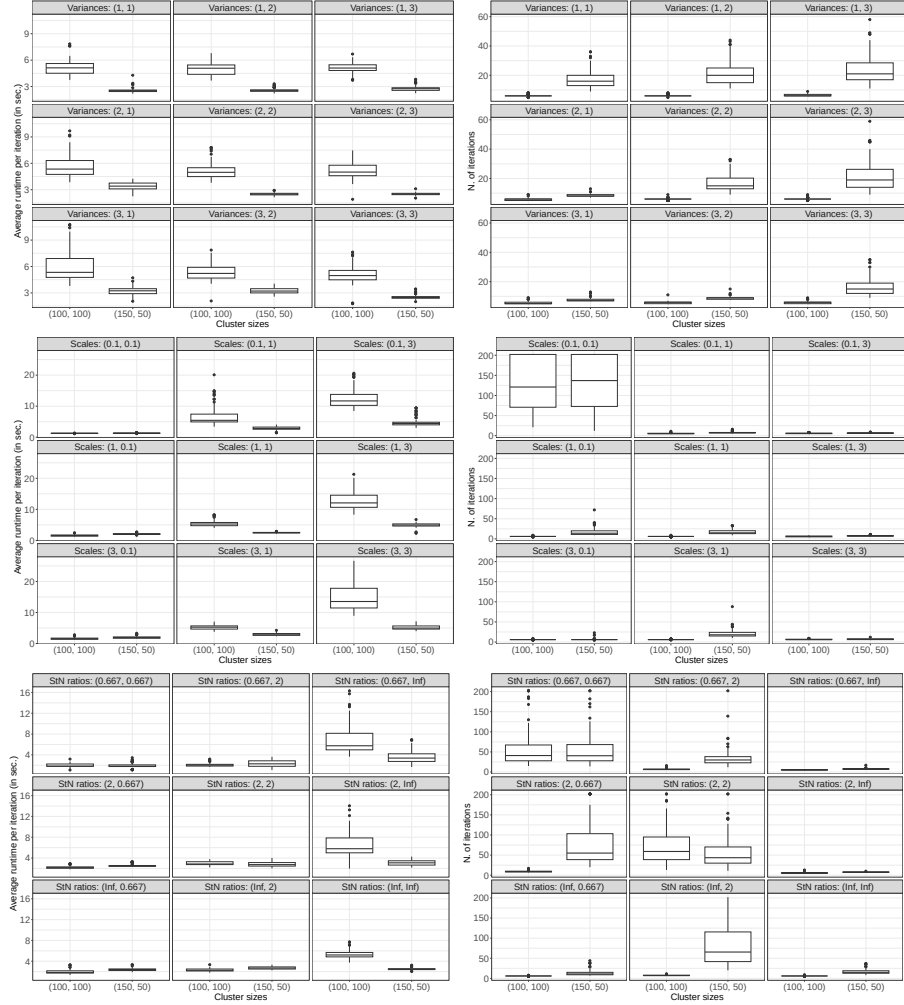


Fig. 1 Computational cost of SR-GMM estimation on simulated data. Each row corresponds to one of the three simulation scenarios described in Sections 4.1–4.3 of the manuscript. For each scenario, we generated 10 independent simulated datasets, and estimated SR-GMM using 10 different random initializations per dataset. Consequently, each boxplot is based on 100 runs. The left column shows the distribution of average runtime per iteration, while the right column depicts the distribution of the number of iterations required for convergence.

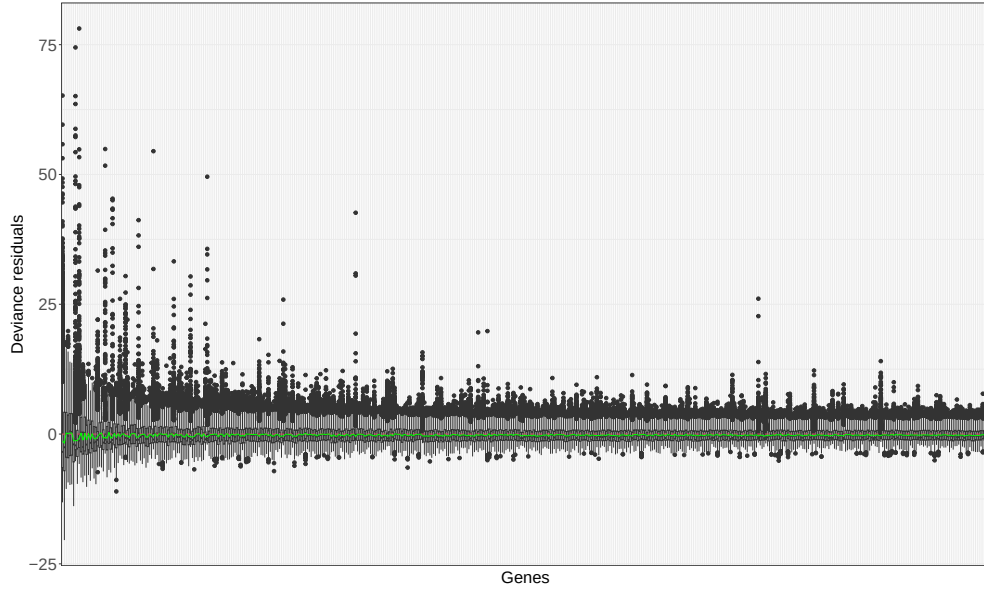


Fig. 2 Representation of gene expression levels measured across the entire prostate cancer tissue sample. The ℓ -th boxplot displays $\mathbf{y}_{\ell,\cdot}$, the expression vector of the ℓ -th gene across all spatial locations, transformed according to the deviance-based criterion proposed by Townes et al. (2019). The green line represents the empirical means of the genes.

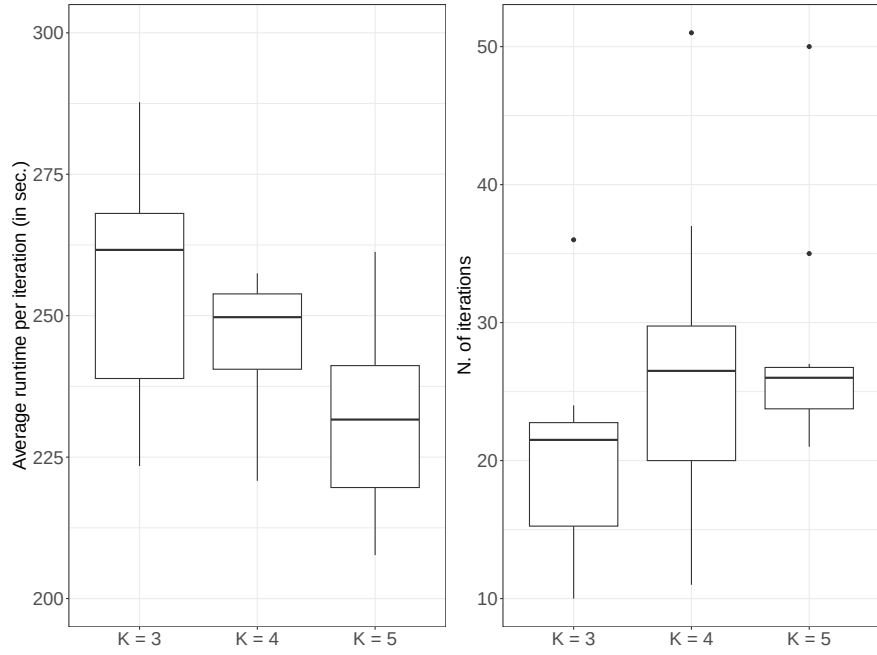


Fig. 3 Distribution of the average runtime per iteration (left) and the number of iterations required for SR-GMM to reach convergence (right) on the prostate cancer dataset described in Section 5 of the manuscript. SR-GMM was estimated using 3, 4, and 5 clusters, with each configuration run from 10 different initialization points. One of the ten runs with three clusters reported an average runtime per iteration greater than 500 seconds, likely due to a server issue. This point is not shown in the graph.

References

- Townes, F.W., S.C. Hicks, M.J. Aryee, and R.A. Irizarry. 2019. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome biology* 20(1): 1–16. <https://doi.org/10.1186/s13059-019-1861-6> .