Supplemental Material
Optimal marker genes for c-separated cell types with SepSolve

Bartol Borozan[1], Tomislav Prusina[1], Luka Borozan[1], Domagoj Ševerdija[1], Francisca Rojas Ringeling[2], Domagoj Matijević*[1], and Stefan Canzar†[2]

[1]School of Applied Mathematics and Informatics, University of Osijek, Osijek, Croatia
[2]Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany

NP-hardness proof via a reduction from Partition

We reduce the classic NP-complete Partition problem to the decision version of our c-separability problem (and to the zero-optimum test of the associated optimization problem). Given a multiset S of strictly positive integers, let $T := \sum_{k=1}^{d} s_k$. For any chosen cardinality $m \in \{1, \ldots, d\}$, we build three d-dimensional Gaussian components with means $\boldsymbol{\mu}^{(i)}$ and (diagonal) standard-deviation vectors $\boldsymbol{\sigma}^{(i)}$, together with a separation threshold c. We then consider the following optimization problem: we introduce nonnegative slack variables β_{ij} for each pair $\{i, j\}$ and minimize $\sum_{(i,j)} \beta_{ij}$ subject to the c-separability constraints. The reduction is engineered so that:

- choosing a binary selector $\boldsymbol{\alpha} \in \{0,1\}^d$ with $\sum_{k=1}^{d} \alpha_k = \boldsymbol{\alpha}^\top \mathbf{1} = m$ picks m coordinates;
- with our specific $(\boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{(i)}, c)$, the constraints become simple inequalities in $\boldsymbol{\alpha}$ and the β_{ij};
- *crucially*, the objective can be driven to zero if and only if the chosen coordinates sum to $T/2$;

thus deciding whether the optimum is zero (or, equivalently, whether there exists a feasible choice with all $\beta_{ij} = 0$) decides Partition. Hence the optimization problem is NP-hard, and the decision version of c-separability is NP-complete.

The optimization model. Let $L := \big\{\{1,2\}, \{1,3\}, \{2,3\}\big\}$. Our optimization problem is

$$\text{minimize} \quad \sum_{\{i,j\} \in L} \beta_{ij} \tag{1}$$

$$\text{subject to} \quad \sum_{k=1}^{d} \alpha_k^2 \big(\mu_k^{(i)} - \mu_k^{(j)}\big)^2 \;\geq\; \big(c^2 - \beta_{ij}\big) \cdot \max\Big\{ \sum_{k=1}^{d} \alpha_k \sigma_k^{(i)},\, \sum_{k=1}^{d} \alpha_k \sigma_k^{(j)} \Big\} \quad \forall \{i,j\} \in L,$$

$$\beta_{ij} \geq 0 \quad \forall \{i,j\} \in L, \qquad \alpha_k \in \{0,1\} \quad \forall k \in \{1, \ldots, d\}, \qquad \sum_{k=1}^{d} \alpha_k = m.$$

*domagoj@mathos.hr
†stefan.canzar@ur.de

(We keep α_k^2 to match the original statement; note $\alpha_k^2 = \alpha_k$ for binary α_k.)

Partition. Given a nonempty multiset $S = \{s_1, \dots, s_d\}$ of strictly positive integers and $T \coloneqq \sum_{k=1}^{d} s_k$, decide whether there exists $S_1 \subseteq S$ with $\sum_{x \in S_1} x = \sum_{x \in S \setminus S_1} x = T/2$. Equivalently, decide whether there is a subset whose sum is $T/2$.

The reduction. Fix any $m \in \{1, \dots, d\}$ (we will handle the unknown m at the end). Let $s \in \mathbb{R}^d$ denote the vector of the elements of S (in any order) and let $\sqrt{s}$ be its elementwise square root. Define

$$\begin{aligned} c &= \sqrt{T/2}, && (2)\\ \boldsymbol{\mu}^{(1)} &= \mathbf{0} \in \mathbb{R}^d, & \boldsymbol{\sigma}^{(1)} &= \tfrac{1}{2}\mathbf{1} \in \mathbb{R}^d,\\ \boldsymbol{\mu}^{(2)} &= \sqrt{m}\,\sqrt{s} \in \mathbb{R}^d, & \boldsymbol{\sigma}^{(2)} &= \mathbf{1} \in \mathbb{R}^d,\\ \boldsymbol{\mu}^{(3)} &= -\tfrac{T}{2}\,\mathbf{1} \in \mathbb{R}^d, & \boldsymbol{\sigma}^{(3)} &= m\,s \in \mathbb{R}^d. \end{aligned}$$

With this choice, problem (1) becomes

$$\begin{aligned} \text{minimize} \quad & \beta_{12} + \beta_{13} + \beta_{23} && (3)\\ \text{subject to} \quad & m\,\boldsymbol{\alpha}^\top s \;\geq\; \left(\tfrac{T}{2} - \beta_{12}\right) \cdot \max\{\tfrac{1}{2}\,\boldsymbol{\alpha}^\top \mathbf{1},\, \boldsymbol{\alpha}^\top \mathbf{1}\},\\ & \tfrac{T^2}{4}\,\boldsymbol{\alpha}^\top \mathbf{1} \;\geq\; \left(\tfrac{T}{2} - \beta_{13}\right) \cdot \max\{\tfrac{1}{2}\,\boldsymbol{\alpha}^\top \mathbf{1},\, m\,\boldsymbol{\alpha}^\top s\},\\ & \sum_{k=1}^{d} \alpha_k \left(\sqrt{m}\sqrt{s_k} + \tfrac{T}{2}\right)^2 \;\geq\; \left(\tfrac{T}{2} - \beta_{23}\right) \cdot \max\{\boldsymbol{\alpha}^\top \mathbf{1},\, m\,\boldsymbol{\alpha}^\top s\},\\ & \beta_{12}, \beta_{13}, \beta_{23} \geq 0, \qquad \boldsymbol{\alpha} \in \{0,1\}^d, \qquad \boldsymbol{\alpha}^\top \mathbf{1} = m. \end{aligned}$$

Because $m \geq 1$, $s_k \geq 1$ and $\boldsymbol{\alpha}^\top \mathbf{1} = m$, we have $\max\{\frac{1}{2}\,\boldsymbol{\alpha}^\top \mathbf{1},\, \boldsymbol{\alpha}^\top \mathbf{1}\} = m$ and $\max\{\frac{1}{2}\,\boldsymbol{\alpha}^\top \mathbf{1},\, m\,\boldsymbol{\alpha}^\top s\} = m\,\boldsymbol{\alpha}^\top s$ and $\max\{\boldsymbol{\alpha}^\top \mathbf{1},\, m\,\boldsymbol{\alpha}^\top s\} = m\,\boldsymbol{\alpha}^\top s$. Dividing by these strictly positive quantities yields the equivalent program

$$\begin{aligned} \text{minimize} \quad & \beta_{12} + \beta_{13} + \beta_{23} && (4)\\ \text{subject to} \quad & \tfrac{T}{2} - \boldsymbol{\alpha}^\top s \;\leq\; \beta_{12}, && (5)\\ & \tfrac{T}{2} - \tfrac{T^2}{4\,\boldsymbol{\alpha}^\top s} \;\leq\; \beta_{13}, && (6)\\ & \tfrac{T}{2} - \tfrac{\sum_{k=1}^{d} \alpha_k \left(\sqrt{m}\sqrt{s_k} + \frac{T}{2}\right)^2}{m\,\boldsymbol{\alpha}^\top s} \;\leq\; \beta_{23}, && (7)\\ & \beta_{12}, \beta_{13}, \beta_{23} \geq 0, \qquad \boldsymbol{\alpha} \in \{0,1\}^d, \qquad \boldsymbol{\alpha}^\top \mathbf{1} = m. \end{aligned}$$

Lemma 1. In any optimal solution of (4), $\beta_{13} = 0 \implies \beta_{23} = 0$ and $\beta_{23} > 0 \implies \beta_{13} > 0$.

Proof. Since $\sqrt{m}\sqrt{s_k}$ and $T/2$ are strictly positive, the following holds,

$$\sum_{k=1}^{d} \alpha_k \left(\sqrt{m}\sqrt{s_k} + \tfrac{T}{2}\right)^2 \;\geq\; \sum_{k=1}^{d} \alpha_k \left(\tfrac{T}{2}\right)^2 = \tfrac{T^2}{4}\,\boldsymbol{\alpha}^\top \mathbf{1} = \tfrac{T^2}{4}\,m.$$

Dividing by $m\,\boldsymbol{\alpha}^\top s > 0$ we obtain

$$\frac{\sum_{k=1}^{d} \alpha_k \left(\sqrt{m}\sqrt{s_k} + \frac{T}{2}\right)^2}{m\,\boldsymbol{\alpha}^\top s} \;\geq\; \frac{T^2}{4\,\boldsymbol{\alpha}^\top s}.$$

Hence the left-hand sides of (7) and (6) satisfy

$$\tfrac{T}{2} - \frac{\sum_{k=1}^{d} \alpha_k \big(\sqrt{m}\sqrt{s_k} + \tfrac{T}{2}\big)^2}{m\,\boldsymbol{\alpha}^\top s} \;\le\; \tfrac{T}{2} - \tfrac{T^2}{4\,\boldsymbol{\alpha}^\top s}.$$

If (6) holds with $\beta_{13} = 0$, then the left-hand side of (7) is ≤ 0 and, by nonnegativity, setting $\beta_{23} = 0$ is feasible and strictly better for the objective. The contrapositive yields the second implication. □

Lemma 2. In any optimal solution of (4), exactly one of the following three mutually exclusive regimes holds (with $\boldsymbol{\alpha}^\top \mathbf{1} = m$):

$$\begin{aligned}
\boldsymbol{\alpha}^\top s \;&<\; \tfrac{T}{2} \iff \beta_{12} > 0,\ \beta_{13} = 0,\ \beta_{23} = 0 \quad \text{and} \quad \beta_{12} + \beta_{13} + \beta_{23} > 0,\\
\boldsymbol{\alpha}^\top s \;&>\; \tfrac{T}{2} \iff \beta_{12} = 0,\ \beta_{13} > 0,\ \beta_{23} \ge 0 \quad \text{and} \quad \beta_{12} + \beta_{13} + \beta_{23} > 0,\\
\boldsymbol{\alpha}^\top s \;&=\; \tfrac{T}{2} \iff \beta_{12} = \beta_{13} = \beta_{23} = 0.
\end{aligned}$$

Proof. From (5), if $\boldsymbol{\alpha}^\top s < T/2$ then β_{12} must be positive; if $\boldsymbol{\alpha}^\top s \ge T/2$ we can set $\beta_{12} = 0$. From (6), if $\boldsymbol{\alpha}^\top s > T/2$ then $T/2 - T^2/(4\,\boldsymbol{\alpha}^\top s) > 0$ and $\beta_{13} > 0$ is forced; if $\boldsymbol{\alpha}^\top s \le T/2$ we can set $\beta_{13} = 0$. Lemma 1 aligns β_{23} with β_{13} in the optimal solution. The three cases cover all possibilities for $\boldsymbol{\alpha}^\top s$ and are mutually exclusive. □

Theorem 1. Let $T = \sum_{k=1}^{d} s_k$ and fix $m \in \{1, \ldots, d\}$. There exists a subset of exactly m elements of S summing to $T/2$ if and only if the optimal value of (4) equals 0.

Proof. (⇒) If there exists $\boldsymbol{\alpha} \in \{0,1\}^d$ with $\boldsymbol{\alpha}^\top \mathbf{1} = m$ and $\boldsymbol{\alpha}^\top s = T/2$, then by (5)–(7) we can set $\beta_{12} = \beta_{13} = \beta_{23} = 0$ and obtain objective value 0.

(⇐) Conversely, if the optimal value is 0, then all $\beta_{ij} = 0$. By Lemma 2 this forces $\boldsymbol{\alpha}^\top s = T/2$ with $\boldsymbol{\alpha}^\top \mathbf{1} = m$. Thus the chosen m elements sum to $T/2$. □

Corollary 1 (Decision version). The decision version of c-separability ("do there exist $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha}^\top \mathbf{1} = m$ such that all constraints hold with $\beta_{ij} = 0$?") is NP-complete.

Proof. Membership in NP: a certificate is $\boldsymbol{\alpha}$; verifying the three inequalities with $\beta_{ij} = 0$ is polynomial-time.

NP-hardness: Given an instance S of Partition with $d = |S|$, build the Gaussians as in (2). For each $m \in \{1, \ldots, d\}$, query the decision oracle once. By Theorem 1, the oracle answers YES for some m iff S is a YES-instance of Partition. This is a polynomial number of polynomial-time checks, so a polynomial-time algorithm for the decision version would yield a polynomial-time algorithm for Partition. Therefore the decision problem is NP-complete. □

Remark (optimization version). Since determining whether the optimum of (4) equals 0 is NP-complete, the optimization problem (1) is NP-hard.

Supplemental Figures

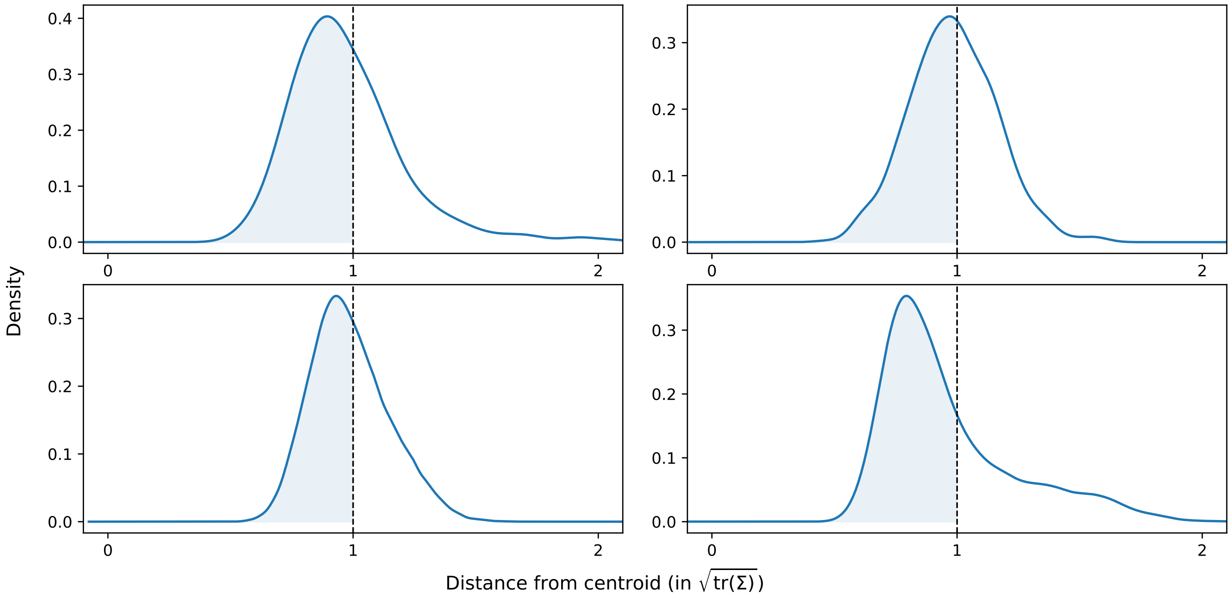

Figure 1: Probability mass of the distances to the centers (sample means) in units of $\sqrt{tr(\Sigma)}$ for the two largest cell types in the Zheng8eq (top) and IPF (bottom) datasets. The fraction of data lying within one standard deviation of the mean is reported as 63%, 55%, 57%, and 68%, from left to right and top to bottom.

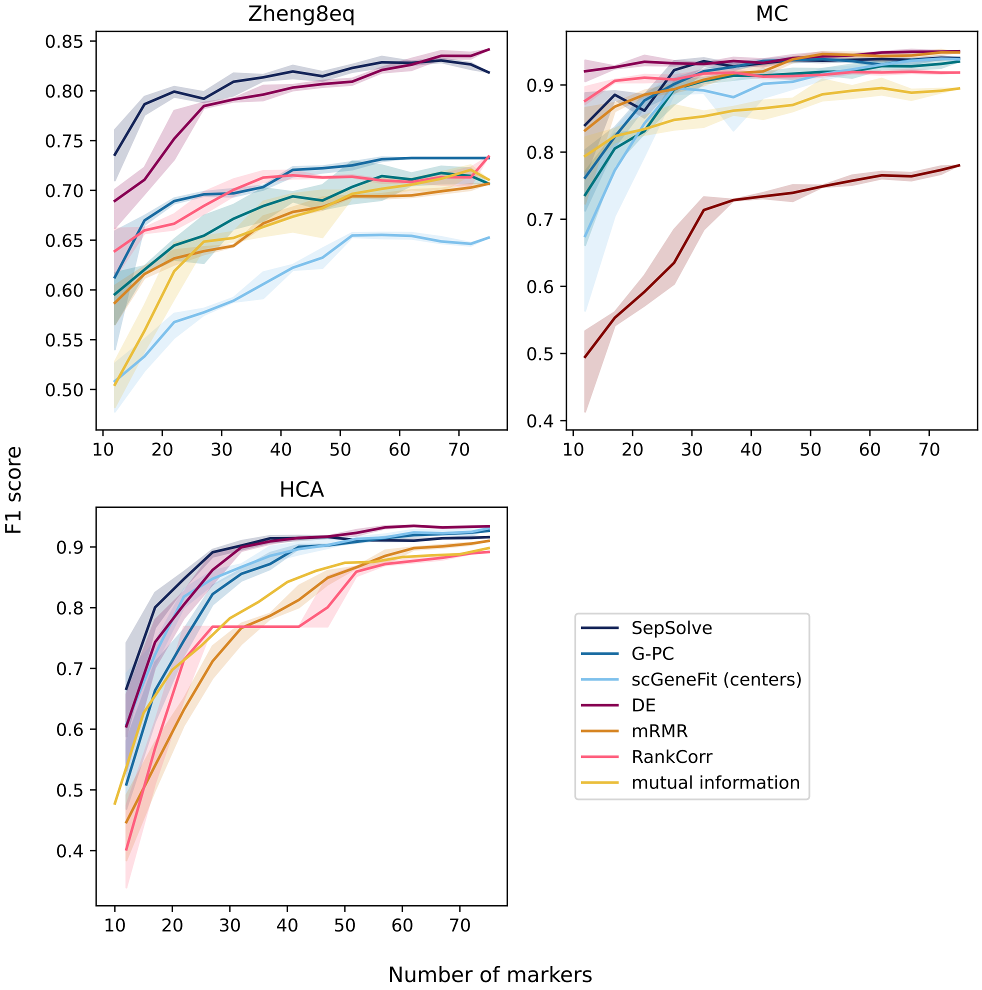

Figure 2: F1 scores of a logistic regression when provided varying numbers of marker genes computed by the different methods on the two smaller datasets (Zheng8eq, MC) and on HCA. On HCA, cell type labels were merged across tissues. scGeneFit was run in pairwise and centers mode on Zheng8eq and MC, and in centers mode only on HCA. Shaded regions depict standard deviation.

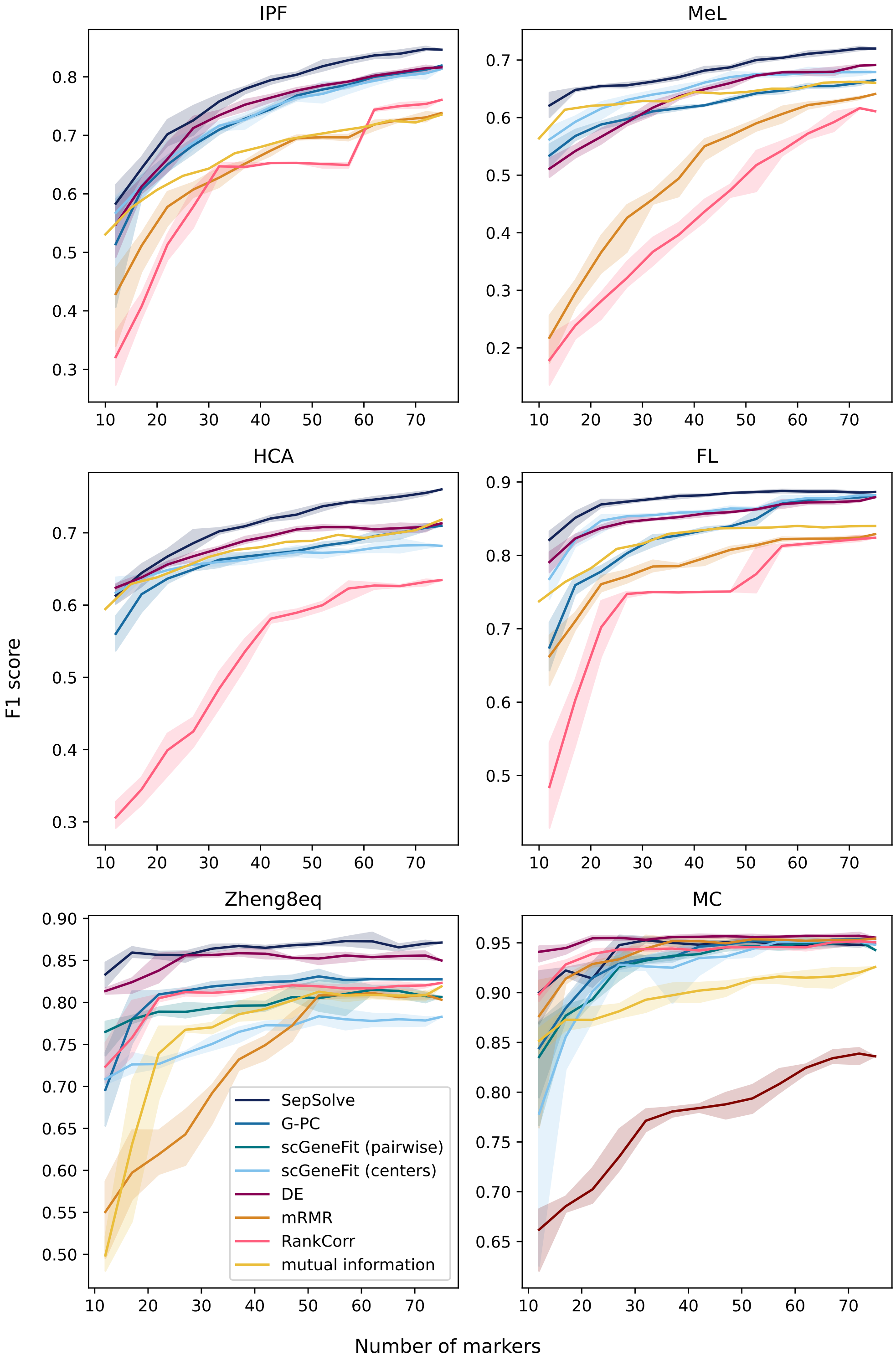

Figure 3: F1 scores of a k-NN classifier when provided varying numbers of marker genes computed by the different methods on all six datasets. On HCA, cell type labels distinguished the tissue of origin. scGeneFit was run in pairwise and centers mode on Zheng8eq and MC, and in centers mode only on the remaining datasets. Shaded regions depict standard deviation.

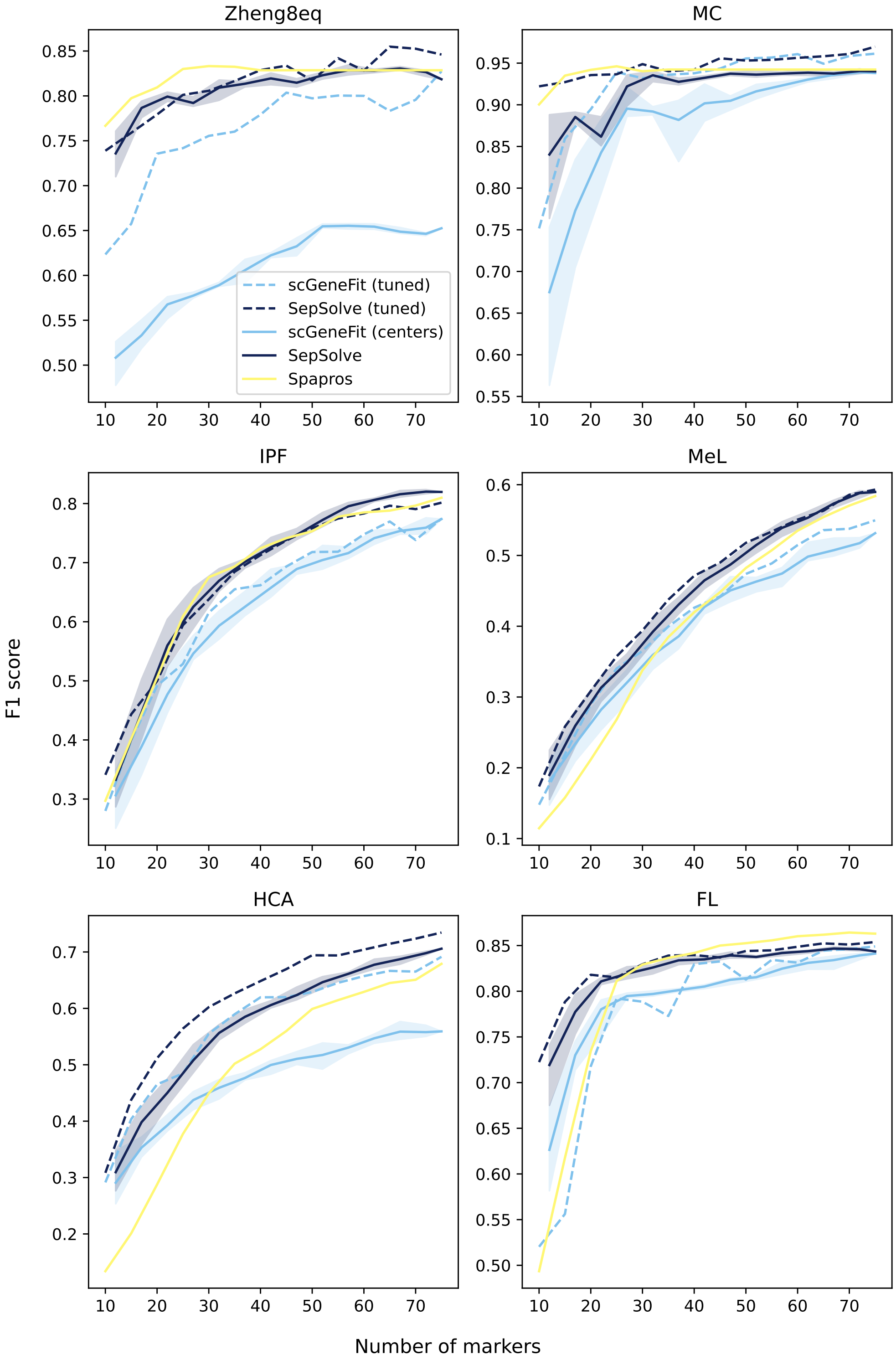

Figure 4: F1 scores of a logistic regression classifier when provided varying numbers of marker genes computed by SepSolve, Spapros, and scGeneFit. On HCA, cell type labels distinguished the tissue of origin. Dashed lines indicate results obtained when SepSolve and scGeneFit hyperparameters were tuned using grid search and dual annealing, respectively (Methods). Shaded regions depict standard deviation.

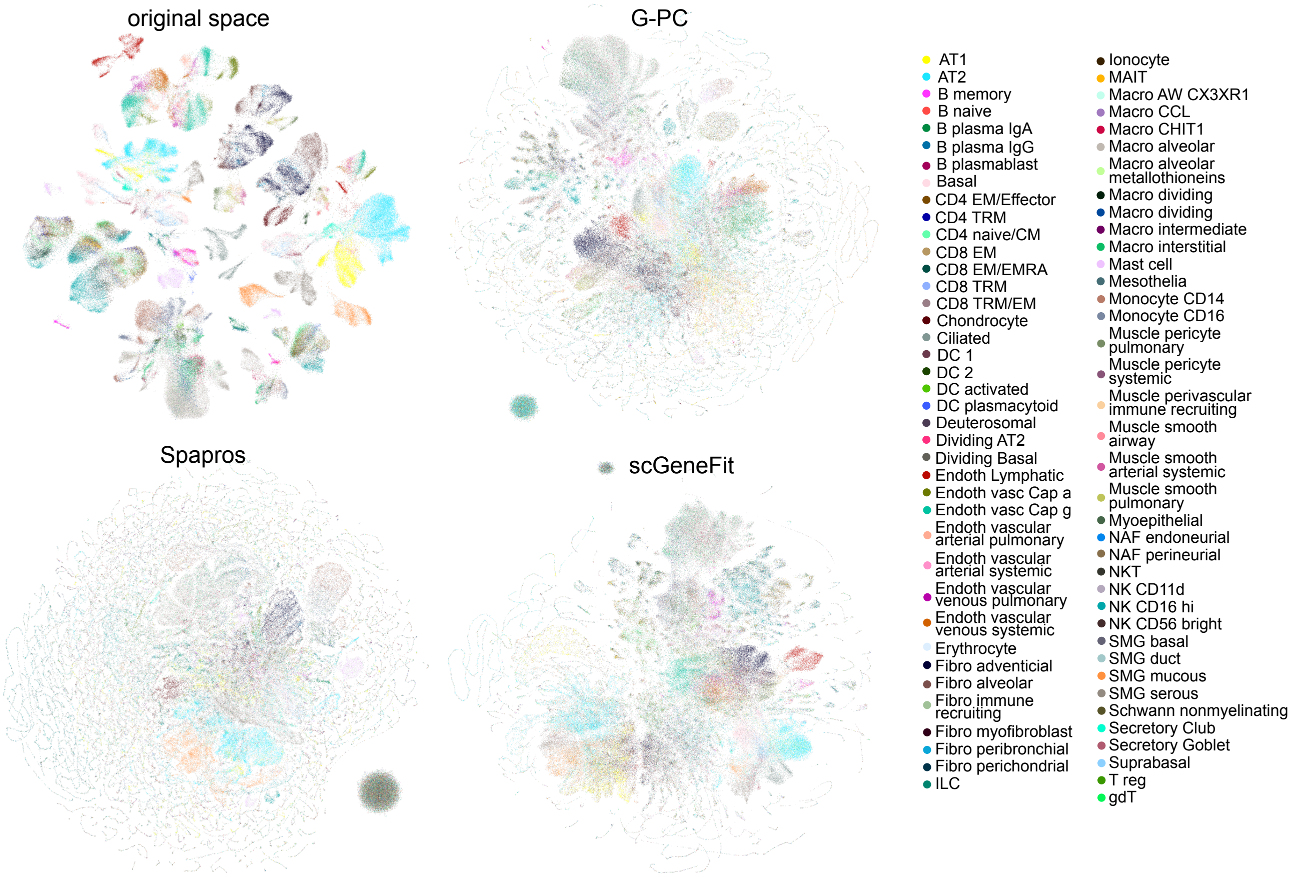

Figure 5: UMAP embeddings generated from 10,000 highly variable genes (original space) and from the 20 marker genes selected by G-PC, Spapros and scGeneFit on the human lung dataset MeL.

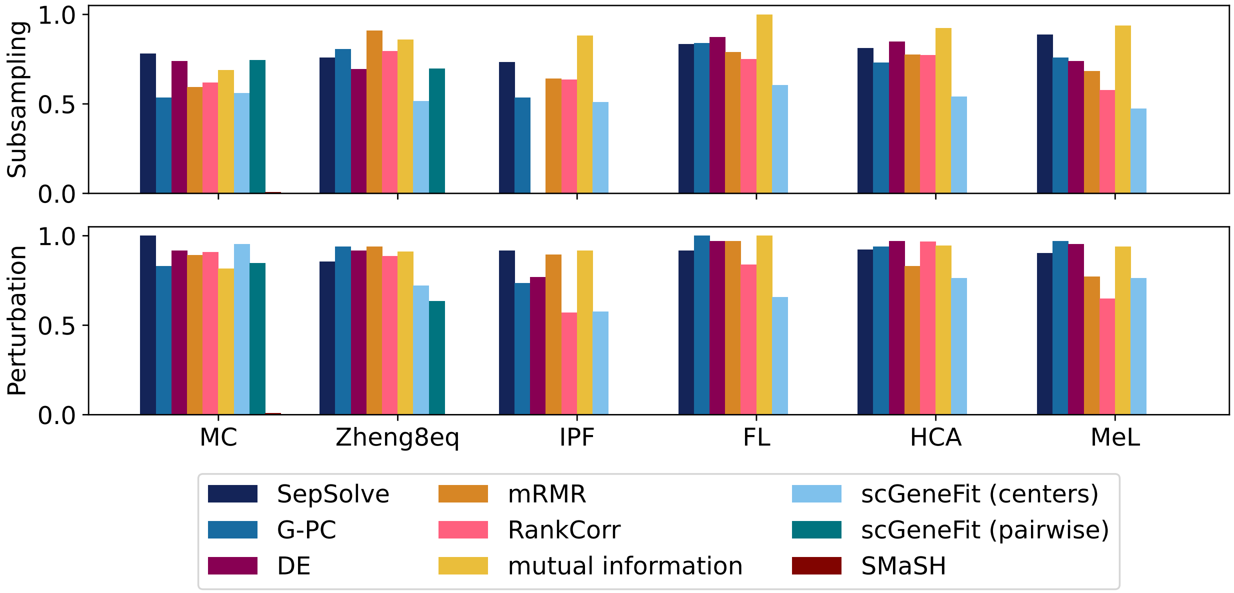

Figure 6: Stability of 25 marker genes computed by the different methods on random subsamples of cells (top) or perturbed counts (bottom). DE crashed on subsampled IPF data since an insufficient number of cells per cell type remained.

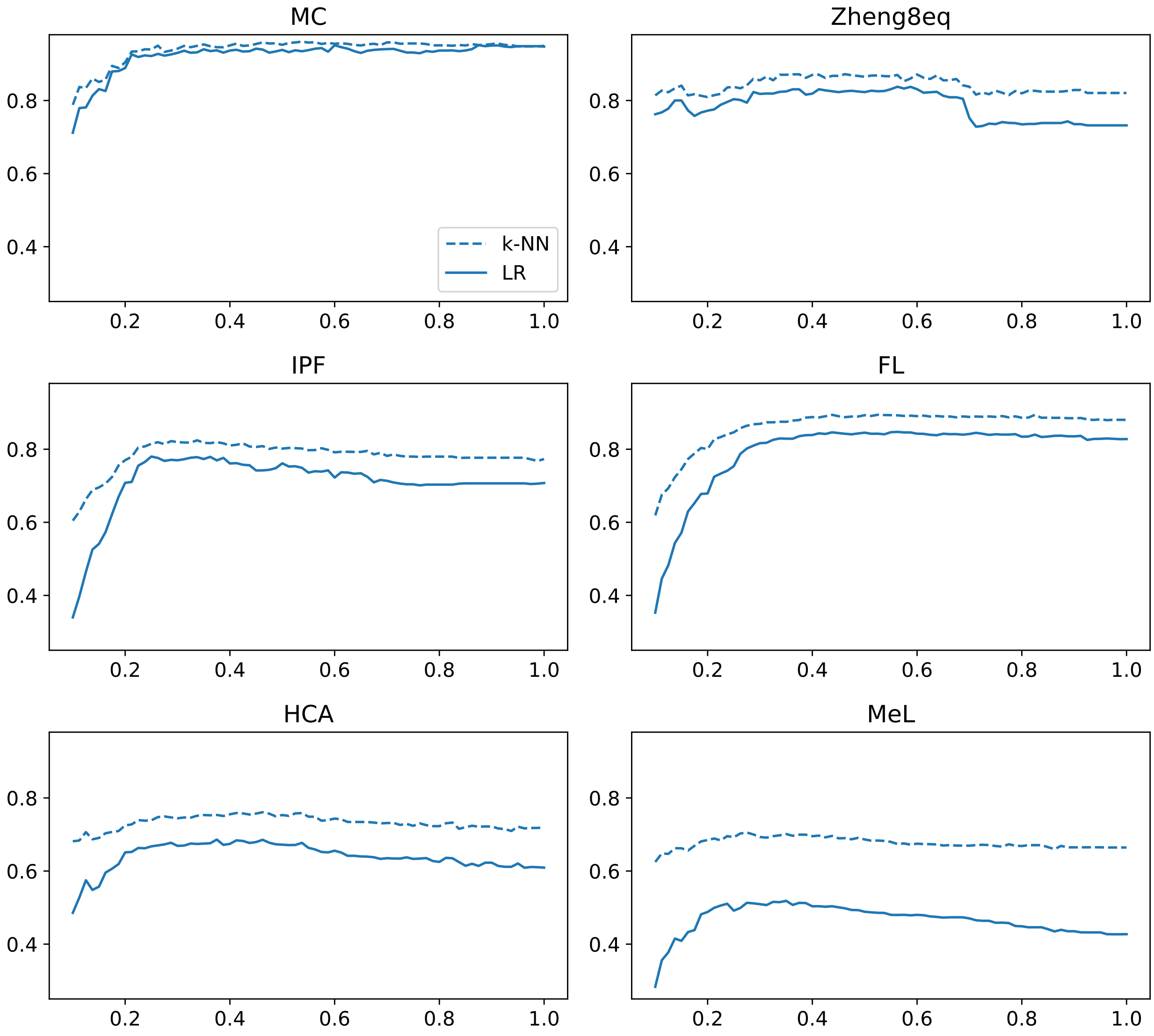

Figure 7: F1 scores of a logistic regression and a k-NN classifiers across all datasets when using 50 marker genes selected by SepSolve for varying separation constant c.

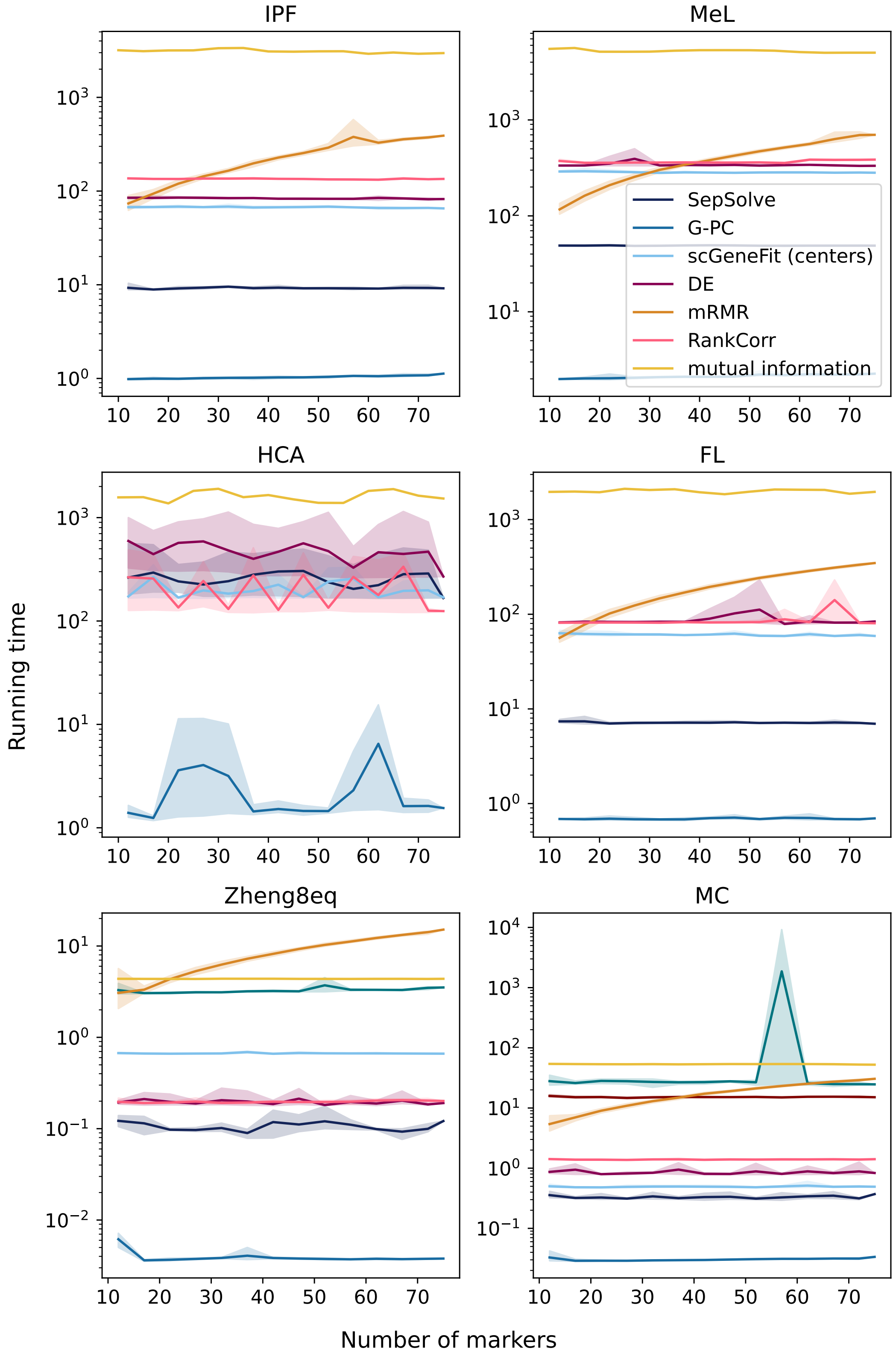

Figure 8: Running times (log scale, in seconds) with respect to the number of target genes across datasets.