Submitted to the Annals of Applied Statistics

SUPPLEMENTARY TO "CO-CLUSTERING OF SPATIALLY RESOLVED TRANSCRIPTOMIC DATA"

BY ANDREA SOTTOSANTI[1,*] AND DAVIDE RISSO[1,†]

[1] *University of Padova,* [*] *andrea.sottosanti@unipd.it,* [†] *davide.risso@unipd.it*

1. Derivation of the ICL for SpaRTaCo. Let m be the current model, and K and R be the number of row and column clusters. The *integrated classification likelihood* (Biernacki, Celeux and Govaert, 2000) is defined as

$$\begin{aligned} p(\mathbf{X}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{W}}; m, K, R) &= p(\mathbf{X}|\boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{W}}; m, K, R) p(\boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{W}}; m, K, R) \\ &= p(\mathbf{X}|\boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{W}}; m, K, R) p(\boldsymbol{\mathcal{Z}}; m, K) p(\boldsymbol{\mathcal{W}}; m, R). \end{aligned} \tag{1}$$

According to Biernacki, Celeux and Govaert (2000), the logarithm of the conditional distribution of $\mathbf{X}$ given the clustering labels can be approximated as

$$\log p(\mathbf{X}|\boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{W}}; m, K, R) \approx \max_{\boldsymbol{\Theta}} \log p(\mathbf{X}|\boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{W}}; \boldsymbol{\Theta}, m, K, R) + \frac{\lambda_{m,K,R}}{2} \log np,$$

where the first component is the classification log-likelihood evaluated in its maximum, and $\lambda_{m,K,R}$ is the number of free parameters in model m with K and R clusters. Thus, under the identifiability constraint in Section 3.1, $\lambda_{m,K,R} = 4KR + \dim(\boldsymbol{\phi})R$. The distribution of both $\boldsymbol{\mathcal{Z}}$ and $\boldsymbol{\mathcal{W}}$ is Multinomial with probabilities $1/K$ and $1/R$, respectively. It follows that

$$\log p(\boldsymbol{\mathcal{Z}}; m, K) = -n \log K, \qquad \log p(\boldsymbol{\mathcal{W}}; m, R) = -p \log R.$$

Finally, taking the logarithm of (1) and replacing $\boldsymbol{\mathcal{Z}}$ and $\boldsymbol{\mathcal{W}}$ with their estimates $\hat{\boldsymbol{\mathcal{Z}}}$ and $\hat{\boldsymbol{\mathcal{W}}}$, we obtain the ICL.

2. Spatial covariance functions. The following isotropic spatial covariance functions have been employed to generate the spatial experiments proposed in Section 4 of the manuscript:

$$k_1^{\text{true}}(d; \boldsymbol{\phi}_1^{\text{true}} = \{\theta_E\}) = \exp\left(-\frac{d}{\theta_E},\right), \qquad k_2^{\text{true}}(d; \boldsymbol{\phi}_2^{\text{true}} = \{\theta_R, \alpha_R\}) = \left(1 + \frac{d^2}{2\alpha_R \theta_R^2}\right)^{-\alpha_R},$$

$$k_3^{\text{true}}(d; \boldsymbol{\phi}_3^{\text{true}} = \{\theta_G\}) = \exp\left(-\frac{d^2}{2\theta_G^2},\right).$$

$k_1^{\text{true}}(\cdot; \theta_E)$ is the *Exponential* kernel with scale θ_E, $k_2^{\text{true}}(\cdot; \{\theta_R, \alpha_R\})$ the *Rational Quadratic* kernel with non-negative parameters (α_R, θ_R), and $k_3^{\text{true}}(\cdot; \theta_G)$ is the *Gaussian* kernel (known also as *Squared Exponential*) with *characteristic length-scale* θ_G.

3. Covariance matrices of the genes. We describe here the main characteristics of the covariance matrices simulated as in Formula (4.7) of the manuscript. The degrees of freedom of a Wishart distribution have to be at least equal to the matrix dimension, that is 200. Both the scales and the degrees of freedom are selected in such a way that the values in $\boldsymbol{\Sigma}_k^{\text{true}}$ have the same order of magnitude of c^{true}. For example, using the illustrated setup, the elements on the diagonals of $\boldsymbol{\Sigma}_1^{\text{true}}$ and $\boldsymbol{\Sigma}_2^{\text{true}}$ have expected values 6.3 and 11.5, respectively. The top line of Figure 2 displays the histogram of the diagonal values of a single realization of $\boldsymbol{\Sigma}_k^{\text{true}}$, for $k = 1, 2, 3$. The values are globally comparable across the three simulations. The bottom line of Figure 2 illustrates the elements out of the diagonal of $\boldsymbol{\Sigma}_k^{\text{true}}$. The difference between the first and the two other matrices is graphically visible: $\boldsymbol{\Sigma}_1^{\text{true}}$ is in fact the one with the smallest covariance values. The second and the third appear similar: in $\boldsymbol{\Sigma}_2^{\text{true}}$, the elements out of the diagonal are in the range $(-3.2, 3.1)$, while in $\boldsymbol{\Sigma}_3^{\text{true}}$ they are in the range $(-3.88, 3.81)$.

4. The PCA-k-means method for selecting the number of co-clusters. We describe a method for selecting the number of row and column clusters of a data matrix $\mathbf{X}$ separately by combing a dimension reduction method with K-MEANS. Let $\mathbf{A}$ be the matrix obtained by rotating $\mathbf{X}$ with respect to its principal components. The procedure fits K-MEANS on the first two variables of the rotated data, i.e., the first two columns of $\mathbf{A}$, using from 1 to $m_{\max}$ numbers of clusters. Let $\omega_m^{\mathbf{A}}$ be the total within sum of squares obtained fitting K-MEANS with m clusters: the integer m^* that solves the following minimization problem,

$$\min_{m^* \in \{1,\dots,m_{\max}\}} \min_{\beta_0,\beta_1,\beta_2} \sum_{m=1}^{m_{\max}} \left\{ \omega_m^{\mathbf{A}} - \beta_0 - \beta_1 (m - m^*) \mathbb{1}(m < m^*) - \beta_2 \mathbb{1}(m \geq m^*) \right\}^2,$$

is the selected number of row clusters. The number of column clusters can be determined by applying the same procedure on $\mathbf{X}^T$. The method can be applied also imposing $\beta_2 = 0$ to guarantee the continuity between the downward-sloping line $\beta_1(m - m^*)\mathbb{1}(m < m^*)$ and the flat line β_0.

We implemented this algorithm into the function `PCA.Kmeans.KR` of the `R` package `spartaco`.

5. Computational burden. In this section, we illustrate the computational time spent to perform 3,000 iterations of the CS-EM algorithm on the spatial experiment described in Section 5 of the manuscript. For every iteration, we run the SE Step for 150 times consecutively to favor the exploration of the clustering configurations and speed-up convergence. The time spent (in hours) is given in Figure 15 for the models with $(K = 2, R \in \{7, \dots, 12\})$.

The SE Step is the most computationally expensive phase because it requires the computation of the classification log-likelihood of every proposed clustering configuration $\boldsymbol{\mathcal{W}}^*$, and thus, to invert the covariance matrices of the clusters that differ from the former configuration $\boldsymbol{\mathcal{W}}^{(t-1)}$. The larger is R, the smaller is the size of the clusters and, consequently, of the matrices to invert. For this reason, models with large R are faster to be estimated.

6. Additional figures.

Figures from Section 2.

- Figure 1 gives a representation of the relations across co-clustering models described in Section 2.2 of the manuscript.

Figures from Section 4.

- Figure 3 shows the expression of three genes measured on the tissue sample 151507, whose spots have been used to build our simulations.
- Figure 4 gives the boxplots of the quantities $\varepsilon_k^{\text{rows}}$ and $\varepsilon_r^{\text{cols}}$, the row and column clustering uncertainties, measured over the 10 replicates of the first four simulation experiments proposed.
- Figure 5 shows the results of the model selection performed in Section 4.3 using the ICL criterion.
- Figure 6 gives an example of spatial experiments simulated under the frameworks discussed in Sections 4.4 and 4.5.
- Following the notation used in Section 4.6 of the manuscript, Figure 7 shows a single realization of $\mathbf{X}_s$, $\mathbf{X}_b$ and $\mathbf{X}$ using $\lambda_s = \lambda_b = \sqrt{0.5}$.
- Figure 8 shows the results of the model selection performed in Section 4.7. Panel (a) compares the classification log-likelihood and the ICL, for any model dimension proposed. Panel (b) gives the CER values obtained on the unique replicate of the simulation experiment proposed, using different co-clustering models.

Figures from Section 5.

- Figure 9 displays the genes ordered according to the deviance criterion proposed by Townes et al. (2019). The red line denotes the number of genes selected for our analysis ($n = 500$), the blue line is the "ideal" number of genes that should be used ($n = 200$), based on where the deviance curve has a significant change in the decay.

- Figure 10 displays the boxplots of the first 100 row vectors of the spatial experiment matrix $\mathbf{X}$, corresponding to the gene expressions measured on the cortical tissue sample analyzed in Section 5, transformed and sorted according to the procedure of Townes et al. (2019).
- Figure 11 illustrates some model fitting results. Panel (a) gives log-likelihood and the ICL values of the models with $K = 2$ and $R \in \{7, \ldots, 12\}$; Panel (b) gives the clustering uncertainty measures $\varepsilon_k^{\text{rows}}$ and $\varepsilon_r^{\text{cols}}$ of the model with $K = 2$ and $R = 9$.
- Figure 12 displays the conditional distributions of $\sigma^2_{\cdot r,i}$, for $i = 1, \ldots, n$, given the data and the parameter estimates. In addition, Table 1 lists the most variable genes in each spot cluster that appear also in Figure 12.
- Figures 13 and 14 display the expression of some genes that are highly variable in specific regions of the analyzed prefrontal tissue sample.

REFERENCES

BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence* **22** 719–725.

TOWNES, F. W., HICKS, S. C., ARYEE, M. J. and IRIZARRY, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome biology* **20** 1–16.

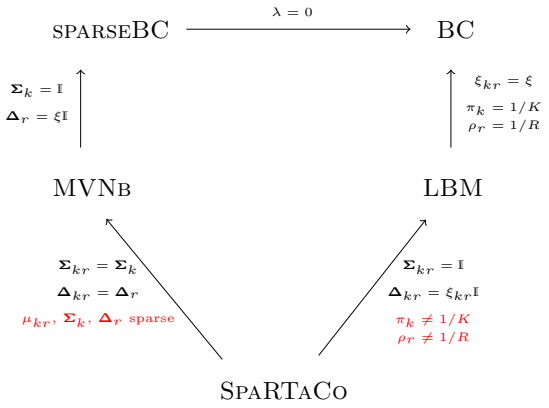

FIG 1. *Map of the co-clustering models described in Section 2.2 of the manuscript. An arrow from model A to model B means that B is a special case of A. Details of how to pass from model A to model B are written in black. A red label denotes a difference between two models A an B which does not make B a special case of A.*

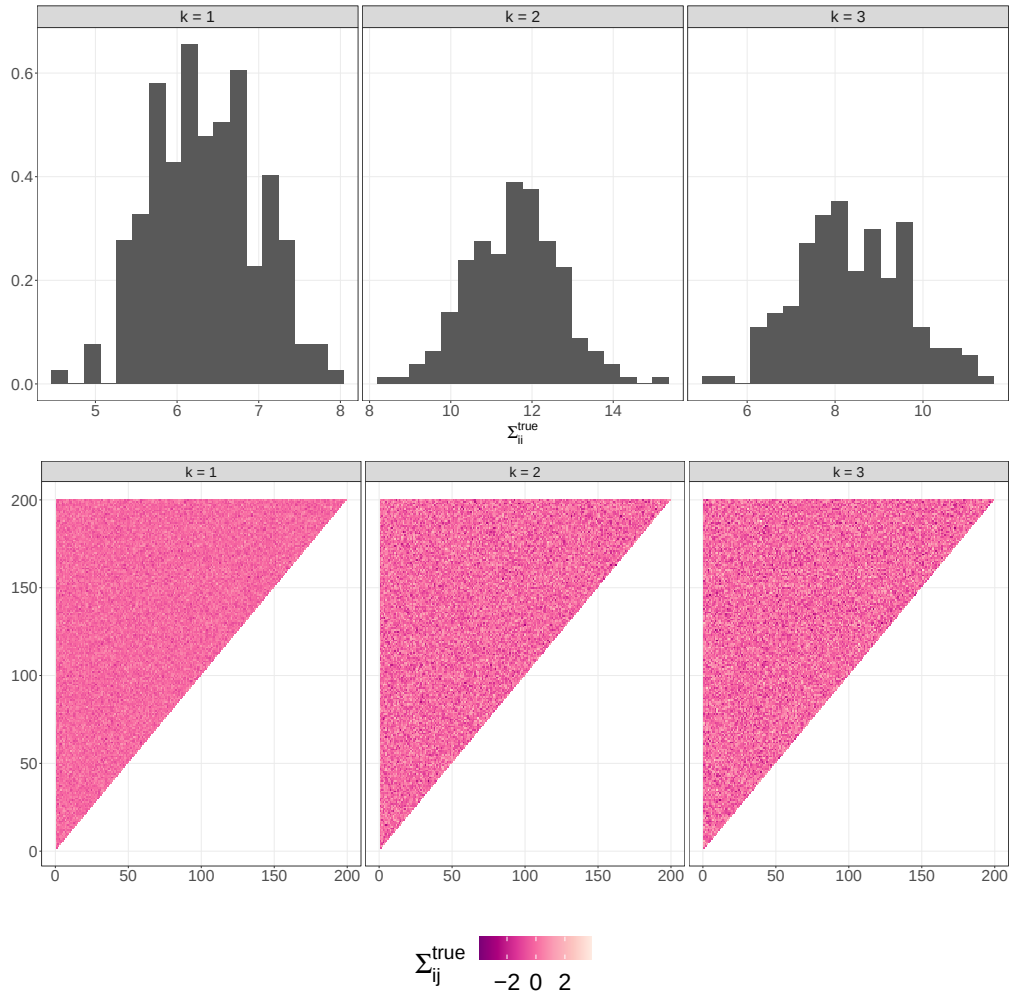

FIG 2. *Plot of the row covariance matrices used in Section 2.3 of the manuscript. The top line displays the histogram of the diagonal values of* $\mathbf{\Sigma}_k^{\text{true}}$*, the bottom line displays the upper triangular matrix of* $\mathbf{\Sigma}_k^{\text{true}}$*, for* $k = 1, 2, 3$.

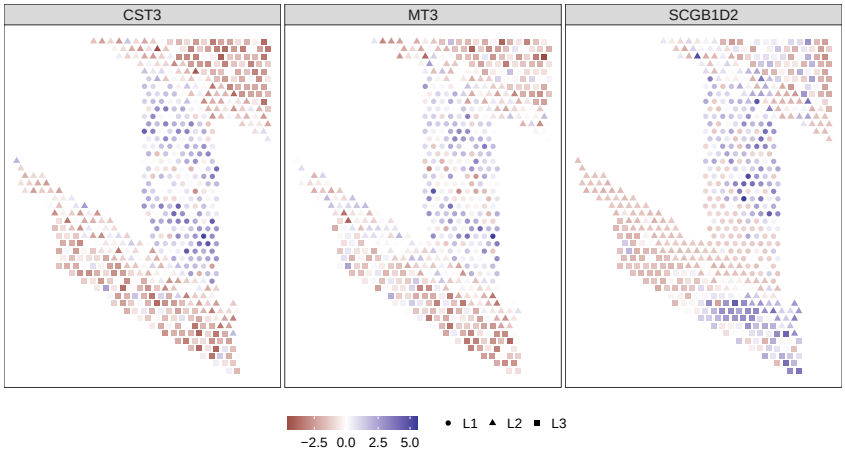

FIG 3. *Plot of the expression of three genes in the area used for the simulations, taken from the tissue with ID 151507. The symbols denote three different layers of the tissue. The gene expression was transformed from counts to a continuous measurement through the pre-processing procedure of Townes et al. (2019). More details of this transformation are given in Section 5 of the manuscript.*

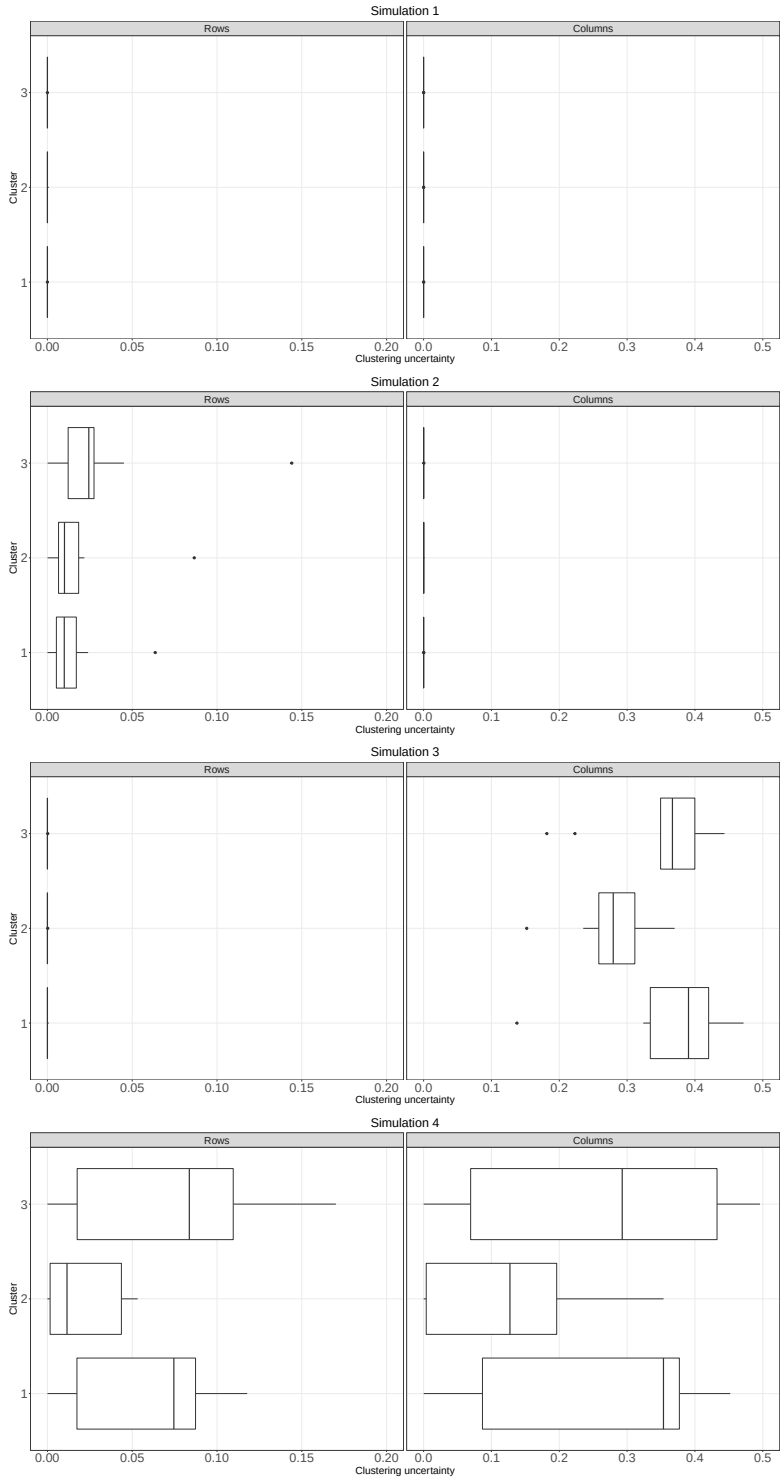

FIG 4. *Clustering uncertainty from Simulations 1-4. For each scenario, we fitted* SPARTACO *using five parallel runs and we estimated the quantities* $\varepsilon_k^{\text{rows}}$ *and* $\varepsilon_r^{\text{cols}}$ *on each of the 10 replicates. Every figure gives the box-plots of* $\varepsilon_k^{\text{rows}}$ *(left panel) and* $\varepsilon_r^{\text{cols}}$ *(right panel). Since the same cluster might take different labels across the replicates, we had to relabel the estimated clusters using the true clustering labels as reference.*

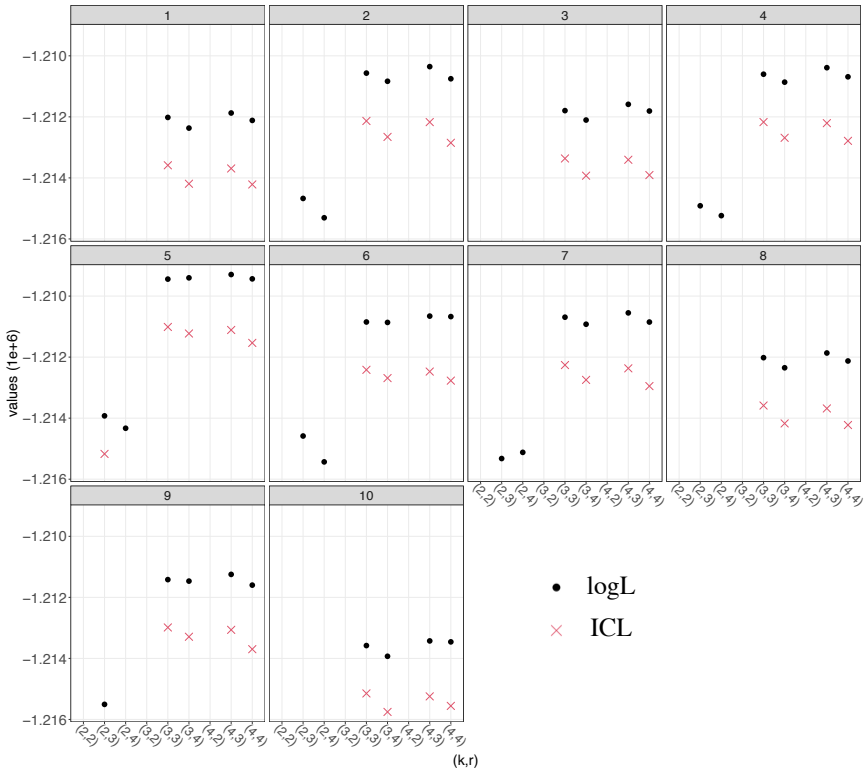

FIG 5. *Detail of Simulation 1. The graphs give the log-likelihood and the ICL values on each of the 10 replicates of the experiment, using different configurations of* SPARTACO. *We truncate on purpose the extremes of the y-axis to show only the largest log-likelihood and ICL values.*

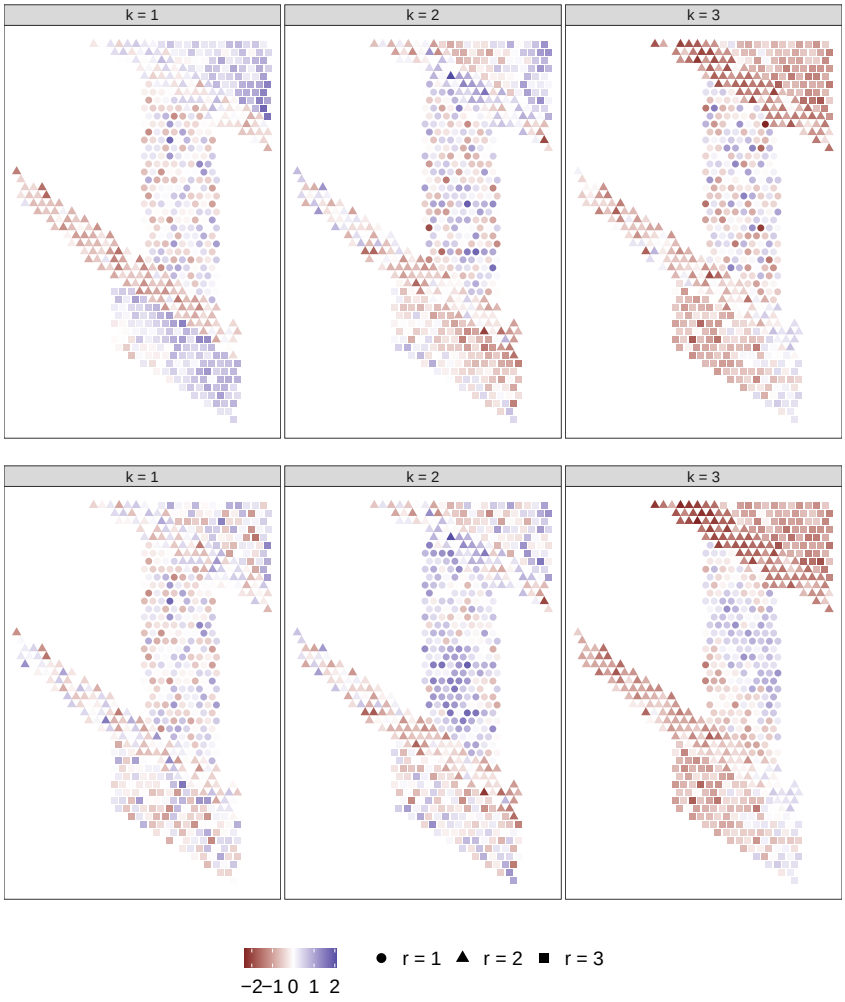

FIG 6. *Examples of a spatial experiment generated under Simulation 2 (top row) and 3 (bottom row). The spots are coloured according to* $n_k^{-1}(\mathbf{X}^{k\cdot})^T\mathbf{1}_{n_k}$*, the average expression of the* k*-th gene cluster. The three spot clusters are displayed with different symbols. The co-clusters with no spatial expression are the ones associated to* $r = 1$ *in Simulation 2, and the ones associated to* $k = 1$ *in Simulation 3. The co-clusters with the largest spatial signal-to-noise ratio are the ones associated to* $r = 3$ *in Simulation 2, and the ones associated to* $k = 3$ *in Simulation 3.*

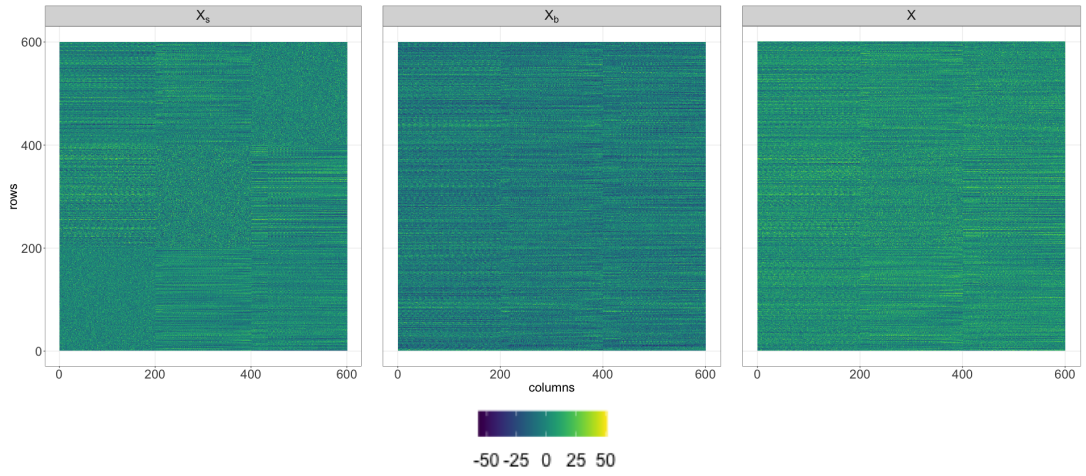

FIG 7. *Simulation 4. The matrices $\boldsymbol{X}_s$, $\boldsymbol{X}_b$ and $\boldsymbol{X}$ appear from the left to the right, using $\lambda_s = \lambda_b = \sqrt{0.5}$.*

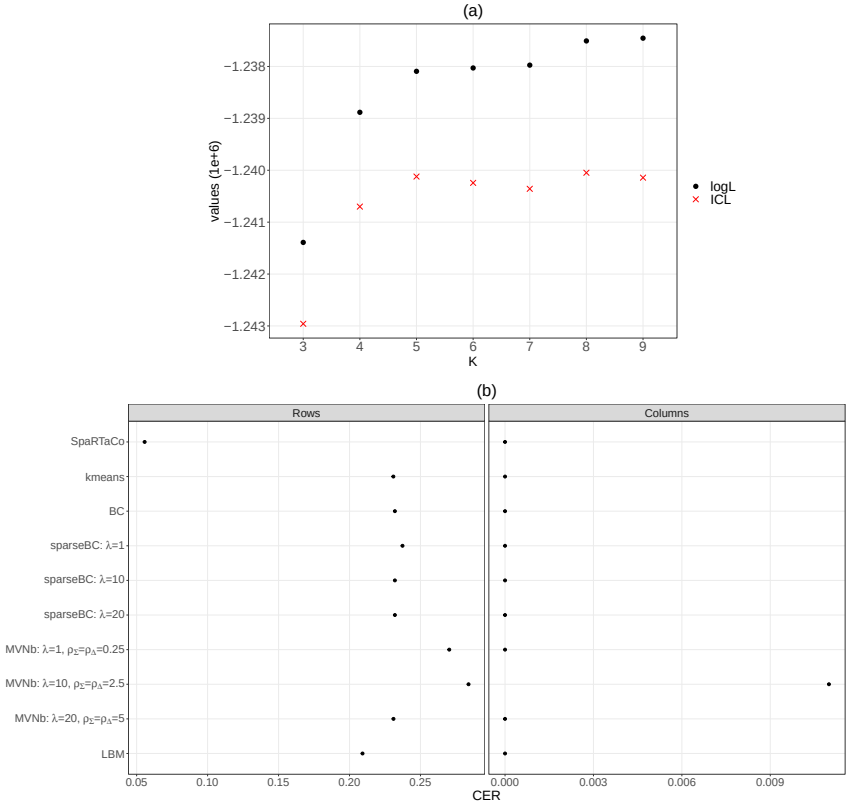

FIG 8. *Results from Simulation 5. Figure (a) compares the classification log-likelihood and the ICL of different* SPARTACO *models with K varying from 3 to 8 and with $R = 3$. The best model according to the ICL criterion is the one with $K = 8$ row clusters. Panel (b) gives the CER obtained on the rows and on the columns using* SPARTACO *with and the competing models, all with $K = 5$ and $R = 3$.*

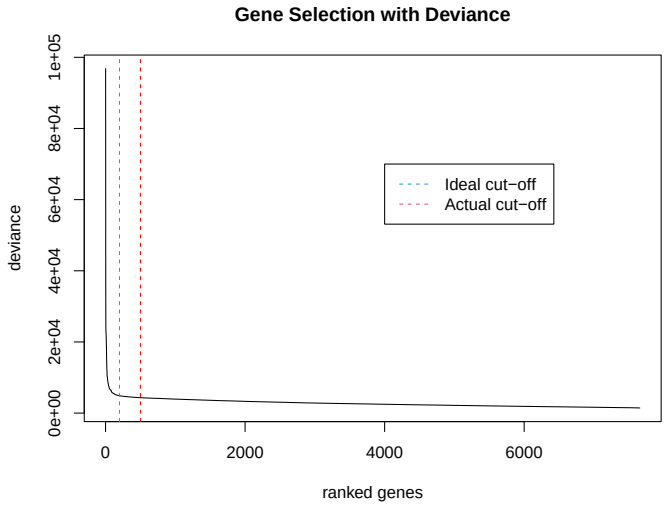

FIG 9. *Graph of the genes measured on the prefrontal cortex sample analyzed in Section 5, sorted in decreasing order according to the deviance value. High deviance values are associated to informative genes. Even if from a graphical evaluation the ideal number of genes is around 200, we included in the analysis the 500 genes with the largest deviance.*

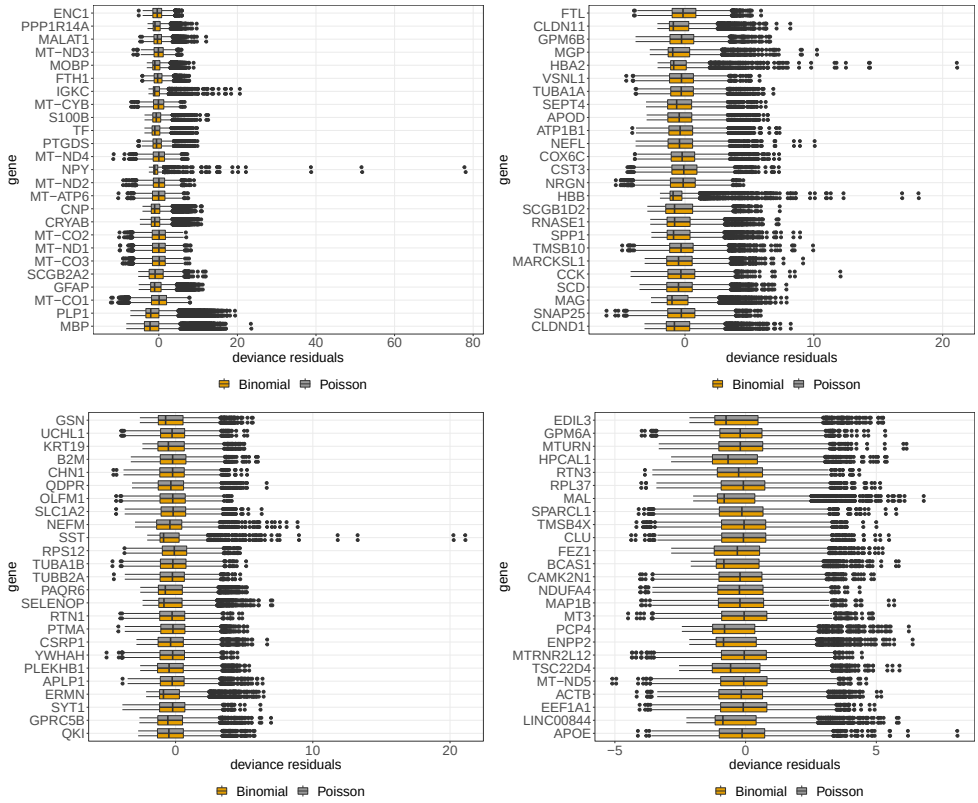

FIG 10. *Boxplots of the first 100 row vectors $\boldsymbol{x}_i$ from the prefrontal cortex tissue sample analyzed in Section 5. For every gene, we plot the deviance residuals using an approximation of the multinomial model based on both the binomial and the Poisson distributions, showing that the two methods are in practice equivalent on this dataset. It is worth remembering that, due to the column clustering performed by* SPARTACO*, each boxplot must be seen as the collection of $r = 1, ..., R$ different subvectors, each of length p_r. Therefore, it is not required to the distributions of the x_i to be symmetric.*

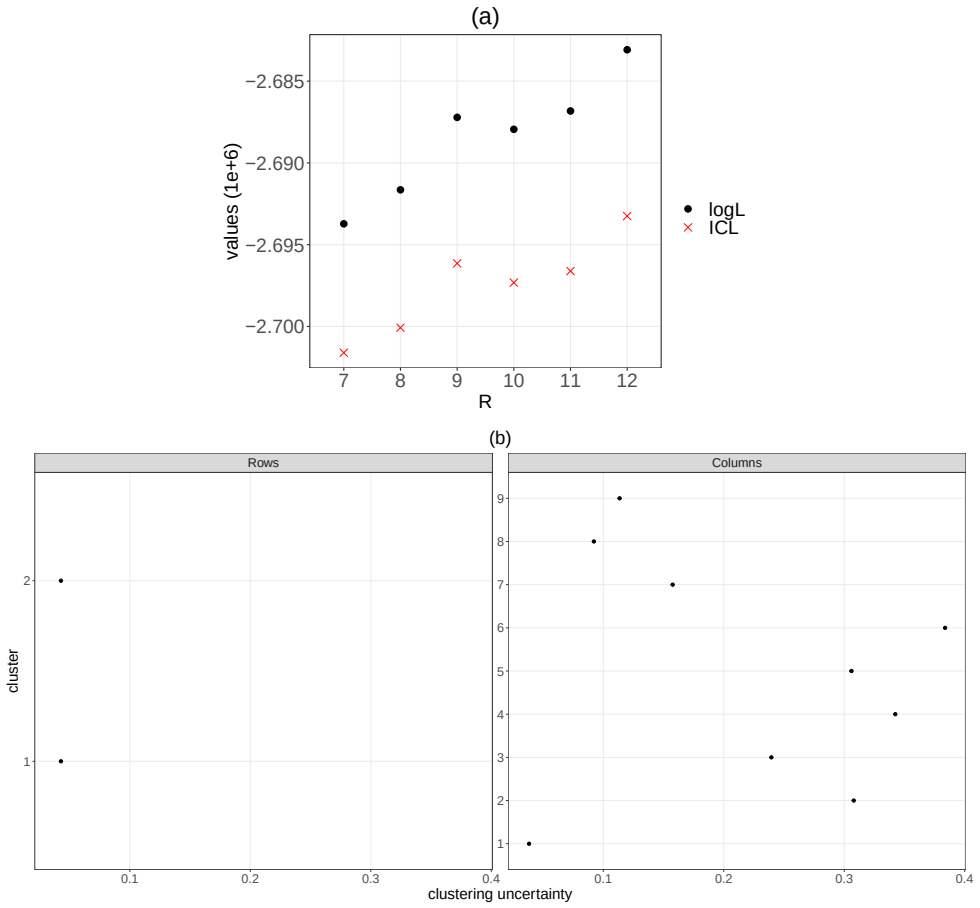

FIG 11. *Results from Section 5. Panel (a) compares the classification log-likelihood and the ICL of different* SPARTACO *models with R varying from 7 to 12 and with $K = 2$. Panel (b) displays the clustering uncertainty measures $\varepsilon_k^{\text{rows}}$ and $\varepsilon_r^{\text{cols}}$ for the selected model ($K = 2$, $R = 9$).*

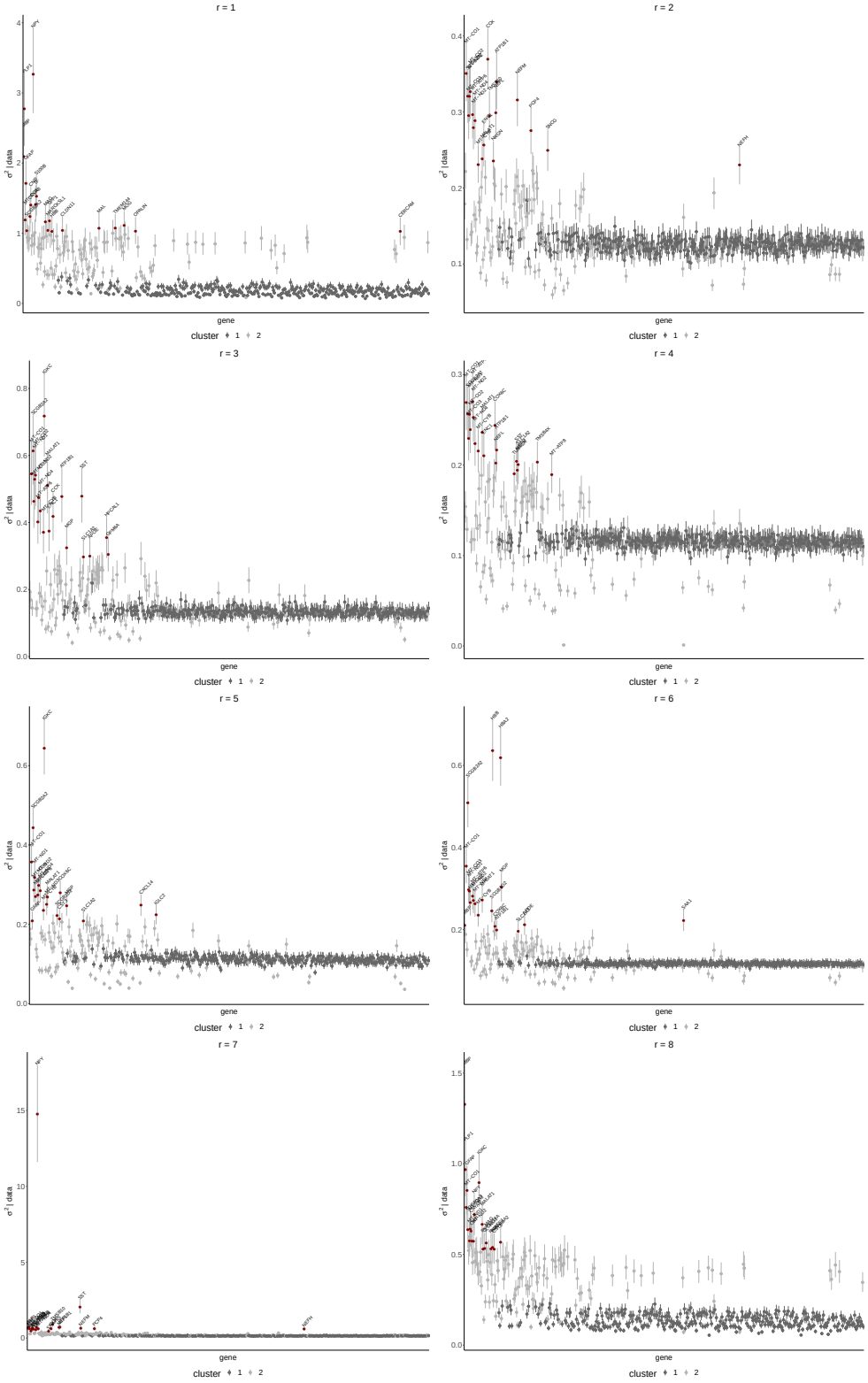

FIG 12. *Results from Section 5. Each panel gives the distribution of* $\sigma^2_{\cdot r,i}|\mathrm{data}$*, where* data *denotes both the input data and the estimated quantities. The dots denote the expected values and the error bars denote the 95% credible intervals. For each spot cluster, the twenty genes with the largest expectation are shown in red (see also Table 1).*

	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$	$r=8$	$r=9$
1	NPY	CCK	IGKC	MT-ATP6	IGKC	HBB	NPY	MBP	MBP
2	PLP1	MT-CO1	SCGB2A2	MT-CO1	SCGB2A2	HBA2	SST	PLP1	IGKC
3	MBP	ATP1B1	MT-CO1	SCGB2A2	MT-CO1	SCGB2A2	TMSB10	IGKC	MT-CO1
4	GFAP	MT-CO2	MT-CO2	MT-ND1	MT-ND1	MT-CO1	PLP1	GFAP	PLP1
5	S100B	SCGB2A2	MT-ND1	MT-ND2	MT-ND2	MGP	ATP1B1	MT-CO1	SCGB2A2
6	TF	MT-ND1	MALAT1	COX6C	MT-CO3	MT-CO3	NEFL	NPY	MT-CO3
7	CNP	NEFM	SST	MT-CO2	MT-ND4	MT-ND1	MT-ND2	MALAT1	MT-CO2
8	CRYAB	NEFL	ATP1B1	MALAT1	COX6C	MT-ATP6	MBP	MT-CO2	MALAT1
9	MT-CO1	MT-ATP6	MT-ND2	MT-CO3	MT-ATP6	MALAT1	NEFM	SCGB2A2	MT-ATP6
10	SPP1	MT-CO3	MT-CO3	MT-ND4	MT-CO2	MT-ND2	MT-ND1	CRYAB	MT-ND1
11	MAG	TMSB10	MT-ND4	ATP1B1	MALAT1	MT-CO2	MT-CO1	MT-ND1	MT-ND4
12	MOG	MT-ND4	CCK	MT-CYB	MT-ND3	MT-ND4	PCP4	CNP	GFAP
13	TMEM144	MT-ND2	MT-ATP6	ENC1	CXCL14	SCGB1D2	CCK	MT-ND2	MT-ND2
14	MAL	PCP4	ENC1	SST	MGP	MT-CYB	MT-ND4	HBA2	IGLC2
15	MARCKSL1	ENC1	MT-CYB	TMSB4X	MT-CYB	SAA1	NEFH	MAG	HBB
16	CLDN11	SNCG	HPCAL1	NEFL	IGLC2	APOE	MT-CO2	HBB	B2M
17	SCGB2A2	MALAT1	MGP	SLC1A2	SCGB1D2	MBP	MT-CO3	CLDND1	SPP1
18	OPALIN	NRGN	GPM6A	NEFM	CST3	COX6C	MT-ATP6	RNASE1	CNP
19	CERCAM	MT-CYB	APOE	TUBB2A	GFAP	ATP1B1	SCGB2A2	PPP1R14A	TF
20	HBB	NEFH	SLC1A2	MT-ATP8	SLC1A2	SLC1A2	SNAP25	CST3	SCD

TABLE 1. *List of the highly variable genes within each of the nine spot clusters discovered on the human dorsolateral prefrontal cortex tissue sample analyzed in Section 5. The genes listed here are the ones that appear in red in Figure 12.*

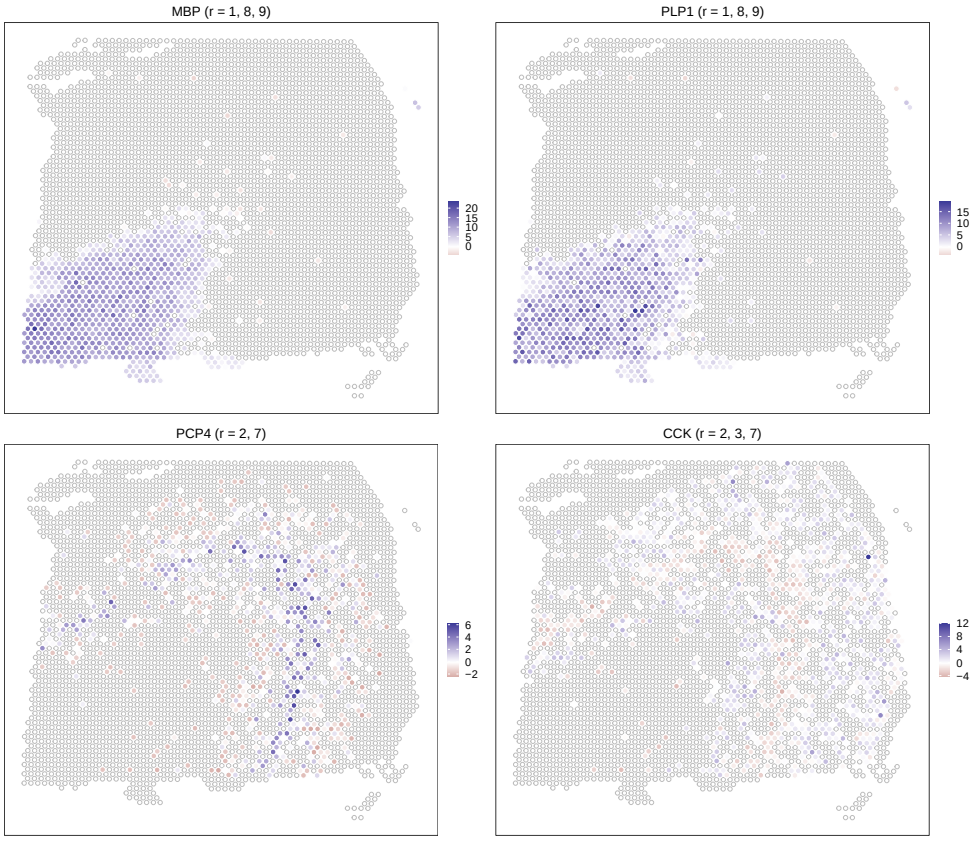

FIG 13. *Plot of the genes MBP, PLP1, PCP4 and CCK, discussed in Section 5 of the manuscript and selected among the most highly variable genes in specific areas of the tissue sample. The title of each figure gives both the displayed gene and the image clusters where the expression is shown.*

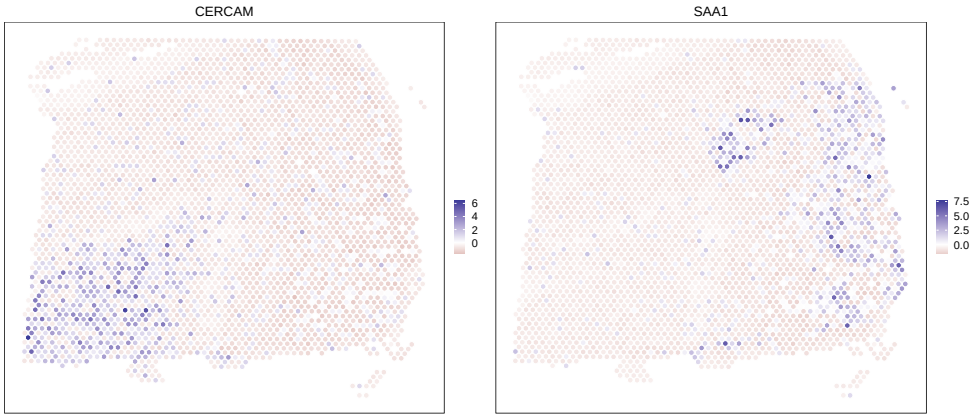

FIG 14. *Plot of the genes CERCAM and SAA1 over the whole tissue analyzed in Section 5.*

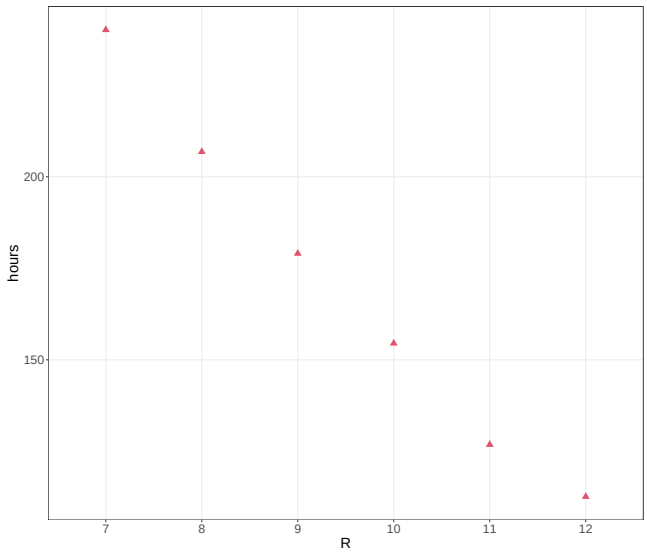

FIG 15. *Computation time (in hours) to perform 3,000 iterations of the CS-EM algorithm to estimate* SPARTACO *with* $K = 2$ *and* $R \in \{7, \ldots, 12\}$ *on the spatial experiment studied in Section 5 of the manuscript.*