University of Zurich[UZH]

Master Program in Biostatistics
www.biostat.uzh.ch

Simulation of Spatial Transcriptomic Data

STA 495 Master thesis

Supervision by Prof. Robinson

Zürich, August 27, 2023

Frederik Philipona, 16-810-657 (*frederik.philipona@uzh.ch*)

1 Introduction

In the process of transcription a segment of the DNA is copied into RNA. These RNAs can be categorized into protein-coding RNAs and nonprotein-coding RNAs. The protein-coding RNAs, also called messenger RNAs (mRNAs), might for example produce hormones, which influence distant parts of multicellular organisms, or enzymes, which accelerate chemical reactions, such as metabolic processes. The gene corresponding to the segment on the DNA transcribed is said to be expressed. Through gene expression the information encoded in the genome gives rise to the phenotype of the cell or organism. Studying gene expressions can give insight into their regulations, which ultimately influence the cell's function and/or its structure, and studying changes in gene expressions can for example give insights into the progression of diseases.

A way of studying gene expressions is through transcirptomic technologies which aim to collect the transcriptome, the momentary sum of all RNA transcripts in a cell or collection of cells. Contemporary methods to gather the transcriptome include microarrays and methods based on next-generation sequencing (NGS). A microarray consists of specific DNA sequences (called probes) attached to a solid surface which hybridize their complementary fluorescently labeled transcripts generating a signal to be picked up by a camera. NGS has four steps: first a sequencing library has to be constructed. This is done by fragmenting the DNA sample into smaller fragments. Then the library is attached to a surface and amplified to increase the signal that can ultimately be picked up. In the third step the library is sequenced by reading the individual bases (often by optical detection) as they are polymerized one after the other. In the final step the short DNA fragments now identified have to be aligned to a reference genome.

Spatial locations of gene expressions can give further insight into complex biological systems, such as the human brain. Spatial transcriptomics (ST) techniques allow for measuring gene expressions together with their spatial locations in a tissue.

Many different techniques to extract ST data exist and they can broadly be put into four different categories ((Asp et al., 2020)). Technologies based on microdissection first isolate different regions of the tissue sample and then profile the gene expression in the regions separately for example using microarrays (examples: LCM (Meier-Ruge et al., 1976) and tomo-seq (Junker et al., 2014)). In situ hybridization (ISH) technologies use gene-specific labeled probes to hybridize to the RNA of interest directly on the tissue sample (examples: MERFISH (Chen et al., 2015) and seqFISH+ (Eng et al., 2019)). In situ sequencing (ISS) technologies use non gene-specific probes to profile low numbers of bases at a time directly on the tissue sample (examples: ISS (Ke et al., 2013) and STARmap (Wang et al., 2018)). In situ capturing technologies capture the RNA molecules directly from the tissue sample but sequence them seperately from it. The capturing can for example be done by placing the tissue sample on a prepared surface, to which spatially barcoded primers are printed. The RNA molecules from the tissue sample hybridize to the primers. This combination of RNA molecules and primers can then be sequenced using for example NGS (examples: VISIUM (Ståhl et al., 2016) and GeoMx DSP (Merritt et al., 2020)).

The different techniques have different strengths and weaknesses. The size of the tissue processed, the number of genes counted, the fraction of gene expressions captured and the spatial resolution can vary between the different ST techniques. The most prevalent technique is VISIUM, a commercialized technique based on in situ capturing, which allows for transcriptome wide profiling, but does not attain single cell resolution (Moses and Pachter, 2022).

Many computational methods have emerged to process data obtained by ST techniques. Among others there are methods to find spatially variable genes (spatialDE (Svensson et al., 2018), SPARK (Sun et al., 2020), GPcounts (BinTayyash et al., 2021)), cell type deconvolution methods, that find,

in the case of data with less than single cell resolution, the proportion of each cell type in a spot (cell2location (Kleshchevnikov et al., 2022), SPOTlight (Elosua-Bayes et al., 2021)), spatial domain recognition methods, to find regions with similar gene expression patterns (BayesSpace (Zhao et al., 2021), HMRF (Zhu et al., 2018), STAGATE (Dong and Zhang, 2022)) and methods to find the presence of cell-cell interactions ((Pham et al., 2020)). With multiple methods existing for each data processing task, they have to be compared in order for a researcher to know which one to use.

In principle, there are two ways to judge a methods accuracy to predict the desired ground truth. The investigated method can be used on a real world data set with known ground truth (through a combination of other methods) and its prediction compared to that ground truth, or data can be simulated from any ground truth and the method's prediction is investigated on that simulated data set.

Typically, the articles first introducing the above mentioned methods don't include a thorough comparison of accuracy to other existing methods, based on both real world data with known ground truth and realistic simulated data.

For example, articles introducing methods to find spatially variable genes in many cases (spatialDE, SPARK, GPcounts) applied the new method on real world data without knowing the ground truth (here, which genes have spatially variable expressions) and compared the prediction to the prediction of other already existing methods. In addition, usually a simulation study is included, where the focus on the simulation is mostly on simplicity and not on realistic data (SPARK, spatialDE). An example of such a simulation procedure can be found in spatialDE: The assumption was, that the expression of a gene is either spatially variable or not. The spots were divided into two groups, the first group was set to have lower gene expressions for the spatially variable genes, the second higher. A gene's expression at a spot was then assumed to be Poisson distributed with a rate, which for the spatially variable genes depended on the group of spots this specific spot belonged to and a Gaussian noise term, and for the spatially non-variable genes was a Gaussian noise term plus an intercept independent of the spot's location. In BayesSpace, HMRF and STAGATE, which propose methods to find spatial domains, accuracy is assessed on data sets with known ground truths, simulation studies are however not included.

The lack of thorough simulation studies in these cases is likely due to the limited amount of models to simulate realistic ST data. One model that promises to produce realistic data is scDesign3 (Song et al., 2023). It uses an approach of estimating gene-wise spot conditional marginal expression distributions to then combine them using a vine or Gaussian copula. The spot conditional marginal gene expression distributions are negative binomials with the mean and dispersion parameters both modeled with a Gaussian process smoother to incorporate the spatial location of the spot, additionally to having parameters for batch and condition effects and gene specific intercepts. Concretely, the marginal distributions are of the form

$$y_{g,i} \sim NBin(\mu_{g,i}, \sigma_{g,i}),$$

with

$$log(\mu_{g,i}) = \alpha_{g0} + \alpha_{gb_i} + \alpha_{gc_i} + f^{GP}(\boldsymbol{s}_i)$$

and

$$log(\sigma_{g,i}) = \beta_{g0} + \beta_{gb_i} + \beta_{gc_i} + f^{GP}(\boldsymbol{s}_i).$$

The expression of gene g in spot i depends on the level of the (log) mean and (log) dispersion on a gene specific intercept α_{g0} (or β_{g0} respectively), batch and condition effects of batch b_i and condition c_i of spot i on the expression of g, and a Gaussian process smoother term, which incorporates the position $\boldsymbol{s}_i$ of spot i. Gaussian process smoothing (also known as kriging) is used to predict the value

of a function at a specific point as a weighted average of the values of that function at the other points. The weights are given by a covariance function (usually such that the further the distance between the points, the lower their covariance). For certain covariance functions Gaussian process smoothing predictions are equal to predictions obtained from spline regression (Handcock et al., 1994). Here, the estimation of the marginal gene expression distributions is done with the function gamlss from the R package gamlss (Rigby and Stasinopoulos, 2005), and a low rank version of the Gaussian process smoother based on thin plate splines is used (approach outlined in (Wood, 2017)).

Two problems with scDesign3's approach exist. The first problem is the limited known ground truth when simulating data this way. We know the true means of the genes and the batch and condition effects, meaning that only methods that try to find spatially variable genes or condition effects, or try to remove batch effects can be evaluated. Among others, cell type deconvolution methods, methods to find spatial domains and methods to find cell-cell interactions cannot be compared using scDesign3, as the desired ground truth is not part of the model. Secondly, a gene's expression at a specific spot depends on many different biological and technical sources of variation, the Gaussian process smoother term in scDesign3 is therefore hard to interpret, in addition to being difficult to change. In this thesis, I assume that the biological sources of variation are: the number of cells in that spot, the cell type composition of that spot and the spot's location in the tissue (Moses and Pachter, 2022). The technical sources of variation considered are: a spot-specific capture sensitivity (a term previously incorporated in cell2location), and the batch and condition effects of the experiment for that gene, as in scDesign3. Under these assumptions, the Gaussian process smoother term in scDesign3 would be the combination not only of an effect of the locations on gene expression, but also of the number of cells at the spots, the influence of cell types at spots and technical capture sensitivity of the spots.

In this thesis, I build on scDesign3 and present a model to simulate ST data incorporating all the different sources of technical and biological variation mentioned above in a semi-parametric fashion. The sources of technical and biological variation are easily interpretable and can be changed, allowing for the evaluation of methods in different scenarios. The different ground truths in this model can be used to test, for example, methods to detect spatially variable genes, spatial domains, cell types at spots, and batch and condition effects. The data generated with the model is evaluated for how realistic they are, and a small simulation study is conducted comparing four methods of spatial domain recognition.

2 Methods

2.1 Simulation Model

The model I propose in this thesis is, like scDesign3, based on marginal spot conditional gene expression distributions combined with a Gaussian copula. The marginal distributions are of the form:

$$y_{g,s} \sim NBin(\mu_{g,s}, \sigma_{g,s}),$$

with

$$log(\mu_{g,s}) = log(nc_s \times \boldsymbol{w_s^T} \times \boldsymbol{m_g} \times sd_s) + gd_g + se_{s,g}$$

and

$$log(\sigma_{g,s}) = \alpha_g + \beta_g \times log(nc_s \times \boldsymbol{w_s^T} \times \boldsymbol{m_g} \times sd_s).$$

The expression y of gene g in spot s, $y_{g,s}$, depends on the level of the mean $\mu_{g,s}$ on nc_s, the number of cells in spot s, $\boldsymbol{w_s^T}$, the cell type proportions of spot s, $\boldsymbol{m_g}$, the mean contributions of the cell types to the expression of gene g, sd_s, a spot detection efficiency for the experiment, gd_g, a (log) gene detection efficiency for the experiment, and $se_{s,g}$ the (log) spatial effect of spot s on the expression of gene g. The term $nc_s \times \boldsymbol{w_s^T} \times \boldsymbol{m_g}$ can be thought of as a theoretical mean, without taking into account technical errors in the experiment and a spatial effect on gene expressions beyond that which is due to spatial distribution of the cell types. The technical errors eluded to above are encompassed in sd_s and gd_g. The parameters in the dispersion of the marginal distributions are not as easily interpretable. Here, flexibility (when estimating the distributions) is more important.

The marginal spot conditional gene expression distributions are combined at a spot level with a Gaussian copula, such that a dependence structure between the expressions of the genes in a spot can be produced. A copula is a joint cumulative distribution function with marginal distributions being all uniforms on the interval [0,1]:

$$C(u_1, u_2, ..., u_m) = \mathbb{P}(U_1 \leq u_1, U_2 \leq u_2, ..., U_m \leq u_m),$$

where the U_is are uniformly distributed on the interval [0,1]. A Gaussian copula is defined as:

$$C(u_1, u_2, ..., u_m) = \Phi_m(\phi^{-1}(u_1), \phi^{-1}(u_2), ..., \phi^{-1}(u_m); \mathbf{R_m}),$$

where ϕ^{-1} is the quantile function of the standard normal distribution, Φ_m is the multivariate normal distribution with covariance matrix equal to the correlation matrix $\mathbf{R_m}$, and an m-dimensional vector of zeroes as the mean vector. In order to simulate from an m-dimensional joint gene expression distribution for a spot (m the number of genes) one would first simulate a sample from the multivariate normal distribution constituting the Gaussian copula $(x_1, x_2, ..., x_m)$, then, by applying the probability integral transform, transform the marginals to realizations from uniform distributions on the interval [0,1] $(\phi^{-1}(x_1), \phi^{-1}(x_2), ..., \phi^{-1}(x_m))$ and then plugging the resulting realizations from the uniform distributions into the cumulative distribution functions of the spot conditional marginal gene expression distributions.

A technical replicate of the same experiment would in this model only differ in the technical parameters, i.e. the gene and spot detection efficiencies. A biological replicate would additionally slightly differ in number of cells and cell types at spots, assuming the same condition for the biological replicates (e.g healthy), the mean contribution of the cell types to the expression of the genes, the spatial effect and the copula defining the dependence structure in spots would stay the same.

2.2 Estimation of model parameters

In principle every model parameter could be provided by the user before simulating. Because of the many parameters, a better approach is to first estimate the model parameters and then to change some of them if needed. Here, I estimated the parameters in a few steps. Matching scRNA-seq and ST datasets (e.g., derived from the same tissue) are needed to carry out this estimation.

2.2.1 Number of cells per spot, nc_s

I assumed that neighboring spots have a similar number of cells, and that the number of cells is directly proportional to the true librarysize of a spot. Because of my earlier assumption of one source of technical error being differing spot detection efficiencies, I first constructed a true librarysize per spot, by smoothing the measured librarysizes with Gaussian process regression, and taking the fitted values per spot. Then, by declaring a mean number of cells per spot in the experiment and setting this equal to the mean spot librarysize, the number of cells for each spot was estimated.

2.2.2 Cell type proportions per spot, $\boldsymbol{w_s}$

Using the scRNA-seq experiment the cell type deconvolution method SPOTlight was applied, to get estimated cell type proportions per spot, note that any other cell type deconvolution method could have been used.

2.2.3 Mean contributions of the cell types to the expression of a gene, $\boldsymbol{m_g}$

I estimated these contributions by taking the mean expression of a gene for a particular cell type in the scRNA-seq experiment.

2.2.4 Spot detection efficiency, sd_s

After having estimated the cell type proportions and the number of cells per spot and using the observed librarysizes, it is now possible to estimate the spot detection efficiency, by predicting a theoretically true librarysize for each spot and dividing the observed librarysize by this true one. I used out of sample random forest prediction with the cell types per spot scaled by the number of cells per spot and the entropy of the cell type proportion vectors as predictors, to get the true librarysize.

2.2.5 Other parameters for the marginal distributions

The (log) gene detection efficiency gd_g and the (log) spatial effect $se_{s,g}$ on the level of the mean, and α_g and β_g on the level of the dispersion were estimated using the gamlss R package (Rigby and Stasinopoulos, 2005). Using the already estimated other parameters the term $log(nc_s \times \boldsymbol{w_s^T} \times \boldsymbol{m_g} \times sd_s)$ entered the model as an offset term on the level of the mean, and as the only covariate on the level of the dispersion. The spatial effect was either a Gaussian process smoother term based on the coordinates of the spot (implemented in the R package mgcv (Wood, 2011)), or a spatial domain-specific intercept effect, if spatial domains on the tissue were defined. Using the Gaussian process smoother results in the spatial effect being a smooth surface that is a multiplicative effect on the level of the mean. This Gaussian process smoother term captures more accurately the spatial effect in the data than a simple regional intercept, however, it is difficult to change from one simulation to another.

2.2.6 Gaussian Copula

As stated above, a Gaussian copula is a multivariate normal distribution with mean vector of zeroes and with a correlation matrix as the covariance matrix. This correlation matrix has to be estimated. This is possible after the m spot conditional marginal gene expression distributions $F_g(\cdot|s)$ have been estimated. Following the approach justified in (Sun et al., 2021), the distributional transform (Rüschendorf, 2013) has to be applied to get marginal uniformly distributed $u_{g,s}$s. Then the desired correlation matrix can be estimated as the $m \times m$ sample correlation matrix of $(\phi^{-1}(u_{1,j}), \phi^{-1}(u_{2,j}), ..., \phi^{-1}(u_{m,j}))^T$, with $j = 1, ..., n$, indicating the n spots.

2.3 Simulation study

In this thesis, a simulation study is conducted to compare domain recognition methods. Data is simulated from two ST data sets in 8 different scenarios, and the ground truth is compared with the predictions from the domain recognition methods using the adjusted Rand index (ARI) (Hubert and Arabie, 1985).

2.3.1 Domain recognition methods

Four different domain recognition methods are compared in this simulation study: HMRF, BayesSpace, STAGATE and hierarchical correlation clustering using the Ward method (Ward Jr, 1963) to merge clusters.

HMRF is an approach to find spatial domains implemented in the R package Giotto (Dries et al., 2021) based on a hidden Markov random field model. The gene expressions at a spot are assumed to be uniquely influenced by the outcome of the underlying unobserved Markov random field at that spot. A Markov random field is a graph based model, where neighboring spots are connected to each other, and the outcome of a spot is independent on the outcome of all other spots given its neighbors. In this application, the hidden Markov random field models the hidden states of spatial domains underlying the observed gene expressions. The spatial domain of a spot is inferred from combining information from the gene expressions at the spot and from the neighborhood configuration. In the simulation study, HMRF was applied following the approach presented in the vignette (Giotto).

BayesSpace is a fully bayesian model where the principal components of the gene expressions in a spot are assumed to follow a normal distribution with mean vector depending on the spatial domain the spot belongs to. The different parameters are found using a Markov chain monte carlo method. For the parameter indicating what domain a spot belongs to, spatial prior information is implemented again using a Markov random field. In the simulation study, BayesSpace was used with settings recommended in its documentation on 15 principal components calculated from log-normalized counts of highly variable genes.

STAGATE uses a graph attention auto-encoder (Salehi and Davulcu, 2019). This neural network takes the gene expression matrix and the neighborhood structure of the spots as input and constructs through the encoder side a low dimensional embedding of the spots. The decoder side of the network takes these low dimensional embeddings and tries to reconstruct the gene expression matrix. Each spot takes, through trainable attention parameters, its neighborhood into account. Once the graph attention auto-encoder has been trained, the low dimensional embeddings can be clustered to get a prediction of the spatial domains. In this simulation study, STAGATE was applied following the approach presented in the tutorial (STAGATE, 2021).

Hierarchical correlation clustering with the Ward method to merge clusters was used in the simulation study to compare the methods specifically introduced to find spatial domains with a method that

takes no spatial information into account. The dissimilarity between two spots was set as 1 minus the correlation of the raw counts in the two spots.

2.3.2 Scenarios

To measure the performance of the methods to be evaluated in different situations, data was generated according to 8 different scenarios. The spot and gene detection efficiencies could either be as estimated or drawn from a distribution with higher variance, to examine how the different methods dealt with higher amounts of technical noise. Additionally, the cell type distributions per spot could either be as estimated or with a weaker spatial pattern. The combination of these factors allowed to vary resulted in 8 scenarios.

The distribution with higher variance used to draw new gene and spot detection efficiencies was a mixture distribution. With a probability of 50%, the new gene or spot detection efficiency was drawn from a uniform distribution and with probability 50% it was drawn from the empirical distribution of the gene or spot detection efficiencies estimated from the data. The uniform distribution's support was equal to the support of the corresponding empirical distribution.

The more homogeneous cell types per domain were generated by first randomly shuffling around the spot cell type proportion vectors within each domain. These shuffled cell type proportion vectors were further changed by adding a small amount to the proportion of each cell type and then normalizing. Then, for each spot in the domain a few samples were drawn from a multinomial distribution with probability vector equal to the changed cell type proportion vectors. The drawn vectors were then normalized to get the final cell type proportion vectors used for the scenarios.

	Gene detection efficiency		Spot detection efficiency		Cell types within domains	
Scenario	as estimated	higher variance	as estimated	higher variance	as estimated	weaker spatial pattern
scenario 1	X		X		X	
scenario 2		X	X		X	
scenario 3	X			X	X	
scenario 4	X		X			X
scenario 5		X		X	X	
scenario 6		X	X			X
scenario 7	X			X		X
scenario 8		X		X		X

Table 1: the setup of the scenarios

2.3.3 Datasets

Two ST and two matching scRNA-seq data sets were used for both the evaluation of the simulated data and for the comparison of the domain recognition methods in the simulation study.

The first ST data set was sample 151507 from the collection of (Maynard et al., 2021), consisting of gene expressions in the human dorsolateral prefrontal cortex (DLPFC) gathered using Visium. The raw data consisted of expressions of 33538 genes in 4221 spots, and the spots were manually annotated to 7 spatial domains (white matter and the six cortical layers). A matching snRNA-seq data set, also provided in the same collection, downloadable in the R package SLIBD (Pardo et al., 2022) and gathered using Chromium, was also used. This data set consisted of expressions of 36601 genes and 77604 cells of 25 cell types. To lower computational cost, in the snRNA-seq data, the cells were randomly downsampled to at most 500 cells per cell type. Additionally, only genes were kept whose variance of expression were in both the snRNA-seq and ST data among the highest 3000 gene-wise variances. This resulted in a final ST data set of expressions of 878 genes in 4221 spots. The relatively

low number of 878 retained genes is due to the fact that in the original ST data only 883 genes had a total expression over 2000 over the 4221 spots.

The second ST data set was downloaded using the R package TENxVisiumData (Crowell, 2023) and contains spot level gene expressions of a kidney slide of a mouse gathered by Visium. The matching scRNA-seq data used was from the Tabula Muris Senis project (tab, 2020) and was downloaded using the corresponding R package. The mouse kidney ST data had expressions of 32285 genes in 1436 spots, and the scRNA-seq data consisted of the expressions of 20138 genes and 2806 cells of 17 cell types. Here, cells were downsampled to at most 50 cells per cell type, and genes whose variances were not among the 3000 highest gene-wise variances in the scRNA-seq data set were discarded from the both data sets. For the simulation study spatial domains had to be found. This was done manually, by looking at the expression of genes with a clear spatial structure.

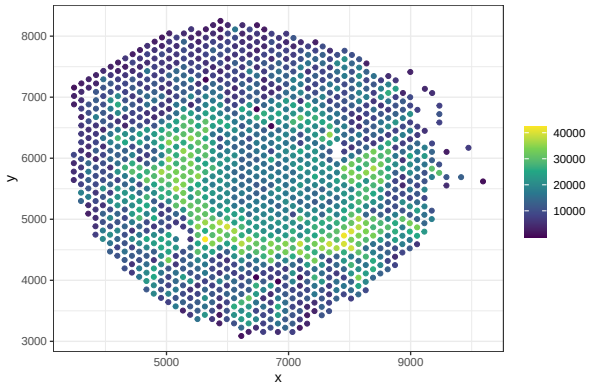

Figure 1: Librarysize per spot for the mouse kidney ST data

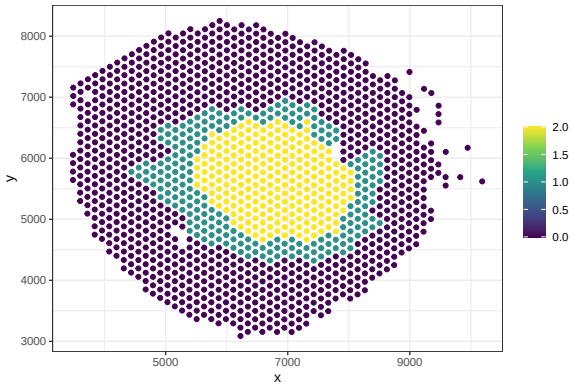

Figure 2: Spatial domains for the mouse kidney ST data

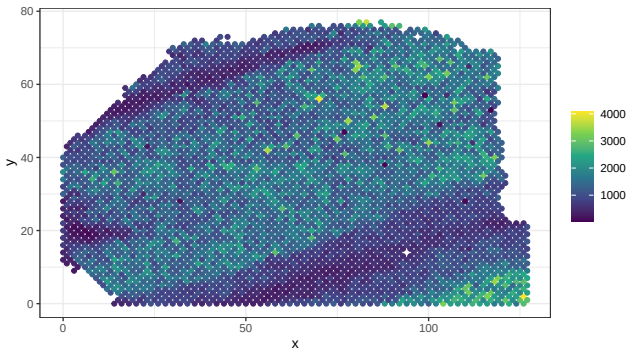

Figure 3: Librarysize per spot for the DLPFC ST data

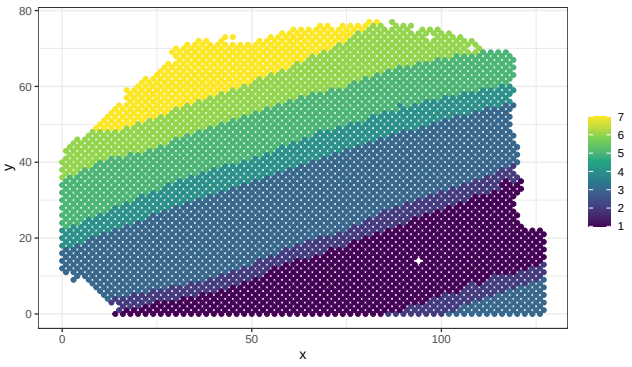

Figure 4: Spatial domains for the DLPFC ST data

3 Results

3.1 Estimated parameters for the DLPFC ST data

To get more intuition about the model introduced above, I present here some of the parameter estimates, before presenting statistics on the simulated data. The parameter estimates shown are for a model with domain-specific intercepts for each gene as the spatial effects, instead of Gaussian process smoother spatial effects. This choice was made, because the same estimated parameters will also be used in the simulation study. For reasons of conciseness, only parameter estimations for the DLPFC data are shown.

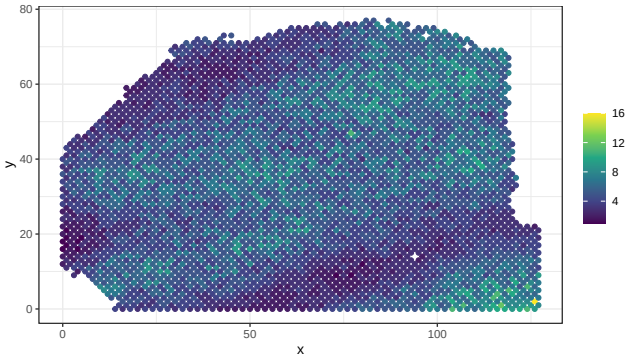

Figure 5: Estimated number of cells per spot for the DLPFC ST data

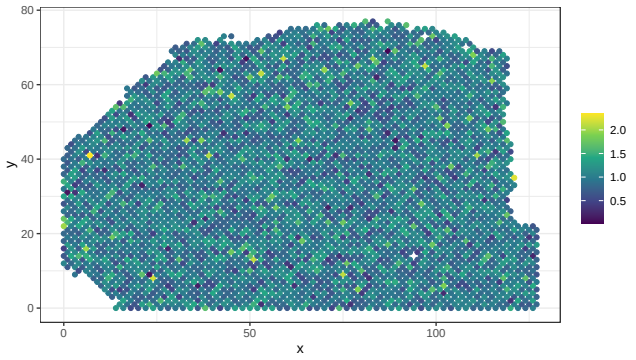

Figure 6: Estimated spot detection efficiencies for the DLPFC ST data

Figure 5 shows the estimated number of cells in each spot for the DLPFC data. The picture looks similar to figure 3 which shows the true librarysizes per spot for the same data. The assumption was, that neighboring spots have similar number of cells and that the number of cells and the librarysize per spot are proportional to each other. Recall that a mean number of cells has to be provided in order to estimate the number of cells per spot. Here, I chose 5 for the mean number of cells, and the resulting number of cells range from 1 to 16. The spot detection efficiencies, being technical noise, were assumed to not follow any spatial pattern. Figure 6 shows the estimated spot detection efficiencies for the DLPFC data, and no spatial pattern can be found. Because of the way the spot detection efficiencies are estimated, they are about centered around one.

The model assumes spatial domain effects to be multiplicative intercepts on the level of the mean for each gene. These domain effects are shown in figures 7 and 8 for two genes. The domain effects are centered around one to help with the identifiability problem when estimating the parameters via gamlss. The domain effects for the two genes differ strongly. While for the gene STMN1 the estimated domain effects are approximately proportional to the librarysizes per domain, this is not true for the domain effects of gene CAMK2N1.

The dispersion of the marginal gene expression distributions was fitted as $log(\sigma_{g,s}) = \alpha_g + \beta_g \times log(nc_s \times \boldsymbol{w_s^T} \times \boldsymbol{m_g} \times sd_s)$. This setup allows for flexibility as can be seen looking at figures 11 and 12. For gene CAMK2N1, we can see a modeled overdispersion as we would expect from a negative binomial model. However, for gene SLC1A3 we see that the modeled variance depends linearly on the modeled mean, which we would expect from a quasi-Poisson model. The different lines in the two plots show the mean variance relationships for the different spatial domains. The figures 9 and 10 show the means and standard deviations per spot for gene CAMK2N1.

Figure 13 shows a histogram of the estimated gene detection efficiencies for the DLPFC ST data. A

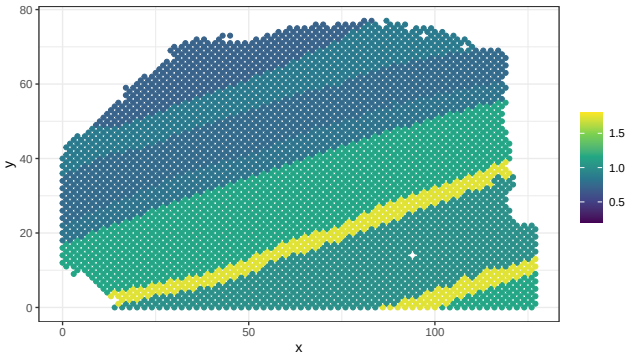

Figure 7: Domain effects of gene CAMK2N1 for the DLPFC ST data

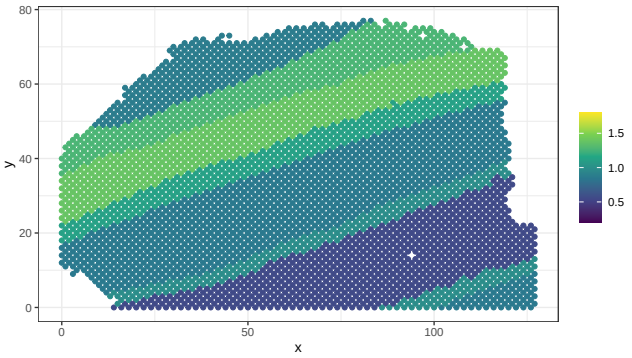

Figure 8: Domain effects of gene STMN1 for the DLPFC ST data

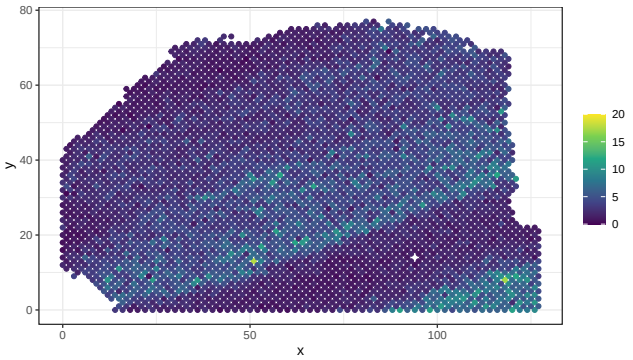

Figure 9: Fitted mean of gene CAMK2N1 for the DLPFC ST data

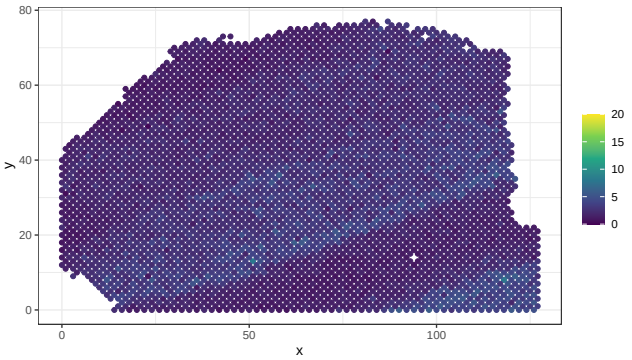

Figure 10: Fitted standard deviation of gene CAMK2N1 for the DLPFC ST data

large portion of them are near zero, suggesting that the overall detection efficiency of gene transcripts in the experiment compared with the corresponding snRNA-seq experiment was low.

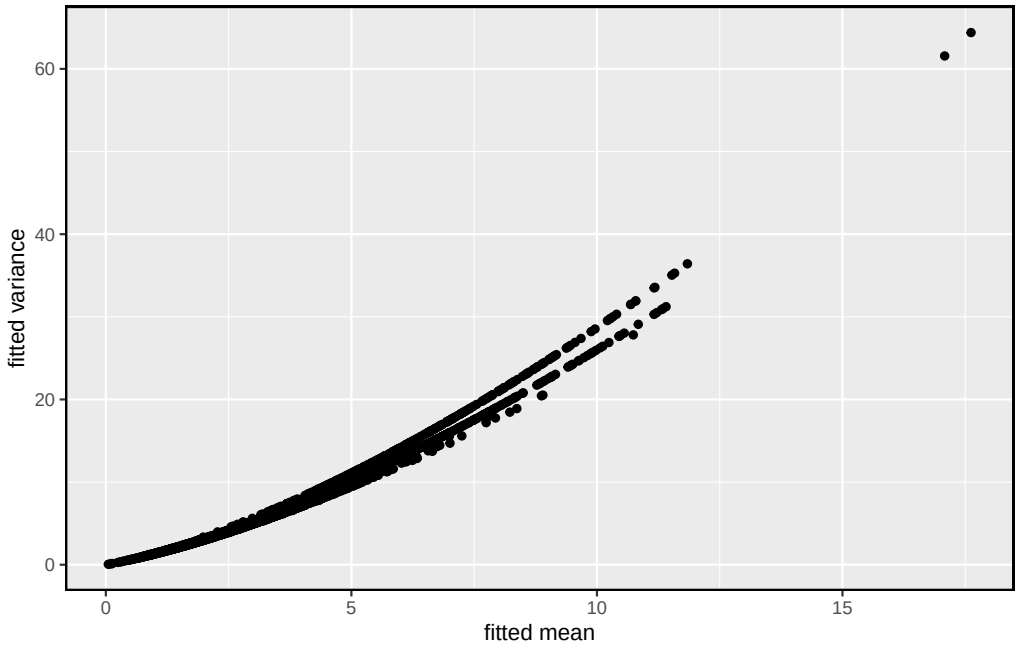

Figure 11: Fitted variance vs fitted mean of gene CAMK2N1 for the DLPFC ST data. The relationship of fitted variance and fitted mean shows overdispersion like it would be expected for a random variable being negative binomial distributed.

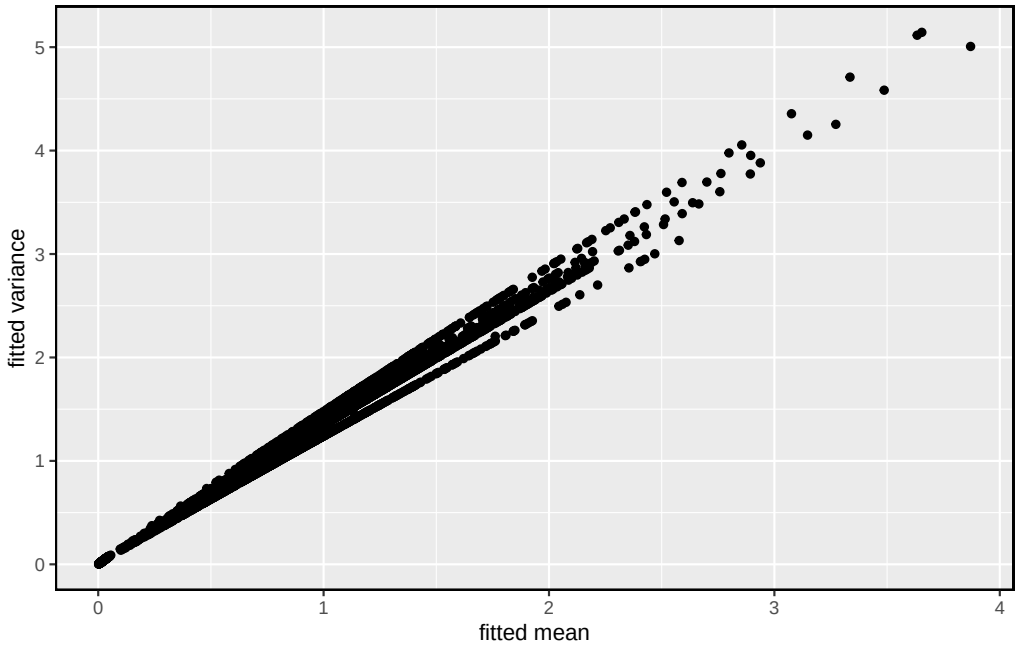

Figure 12: Fitted variance vs fitted mean of gene SLC1A3 for the DLPFC ST data. The relationship of fitted variance and fitted mean is like one we would expect from a random variable being quasi-Poisson distributed.

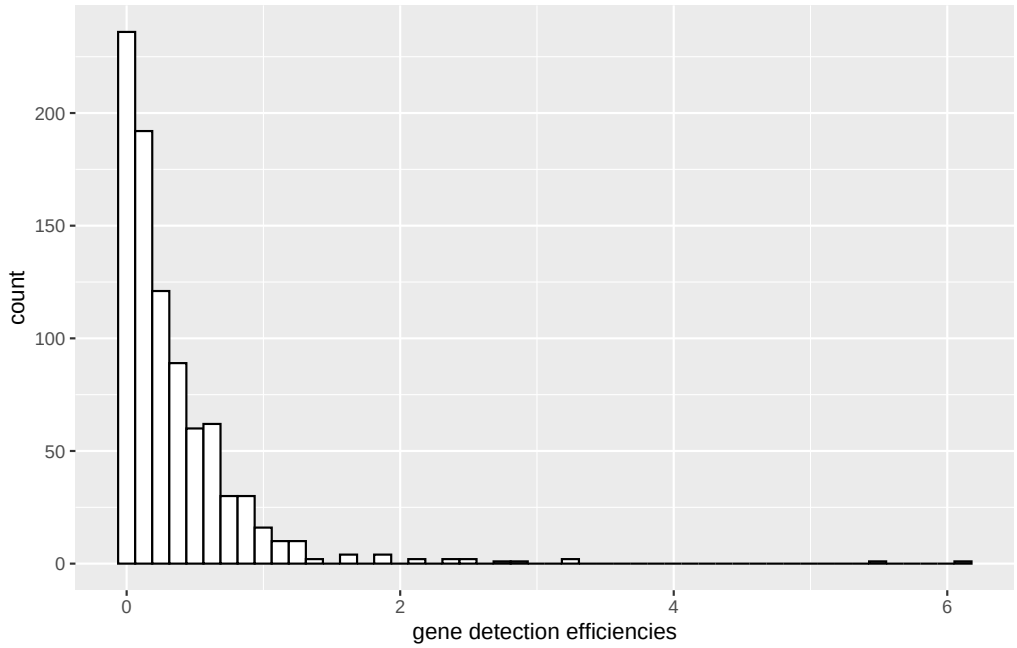

Figure 13: Histogram of the gene detection efficiecnies for the DLPFC ST data. Most of the genes have a low estimated detection efficiency.

3.2 Simulated DLPFC ST data

This section shows some statistics of the simulated DLPFC ST data. The data shown here was generated by simulating from the above specified model as it was estimated beforehand. Later on, in the simulation study, some parameters will be changed, for some of the scenarios. The simulated data shown here should therefore be the closest to the real data.

Figure 14 shows the gene-wise means for the simulated and the observed data. In the simulated data, the gene-wise means are a little bit higher than in the observed data, but the difference is small. The gene-wise standard deviations in the simulated data are also higher than in the observed data as can be seen in figure 15. For most genes, the difference is again small, however, here there are some outliers.

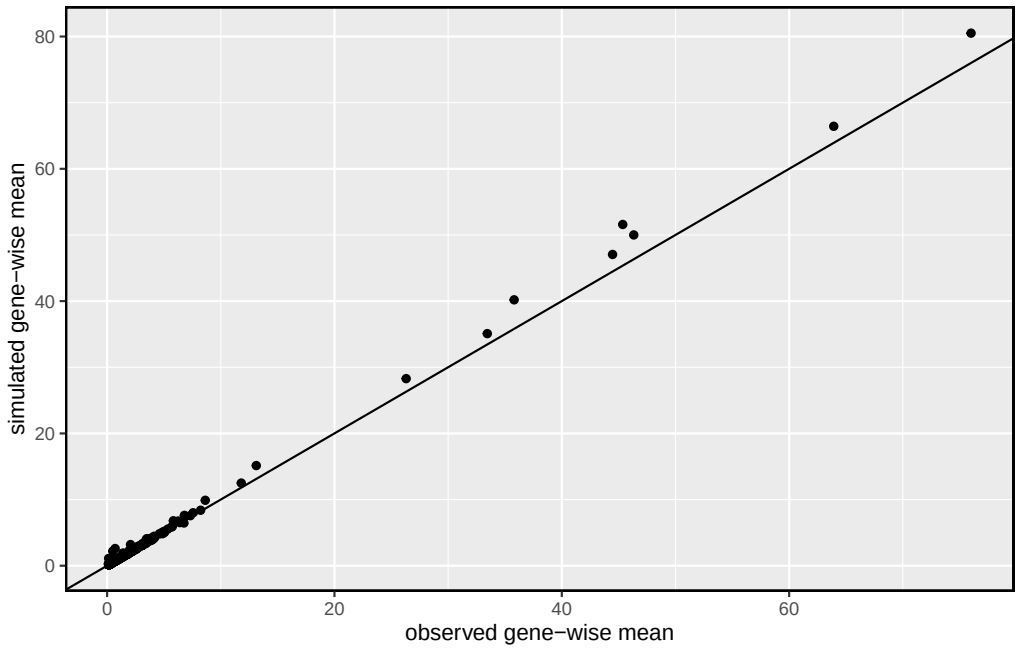

Figure 14: Gene-wise means of observed vs simulated data. The means are preserved in the simulated data.

For gene CALB2 the observed standard deviation in the original data is 0.42, while in the simulated data the standard deviation is 17.54. The maximal expression of CALB2 in the observed data is 5, with most expressions in most spots being either 0 or 1. Despite that fact, the fitted standard deviations, which can be inspected in figure 16 range from 0 to 115. In general, because of the multiple steps of estimations it would be difficult to say what went wrong. Here however, the contribution of cell types at spots for this gene was strongly overestimated (offset term for the mean of the marginal model, and covariate for its dispersion, inspect the difference in figures 18 and 19). Then the marginal model needed to correct that overestimation by assigning high amounts of variation to all spots, resulting in a badly calibrated marginal model, with unrealistic simulated expressions. The simulated expressions can be inspected in figure 17. More than 90% of the simulated expressions are zero, but their maximum is 717.

The distributions of observed and simulated librarysizes per spot, shown in figure 20, are quite similar. Again, as with the gene-wise means the spot-wise librarysizes are a little inflated in the simulated data. The same is true for the spot-wise standard deviations, shown in figure 21.

To further examine whether the simulated data preserves the spatial dependencies of the observed

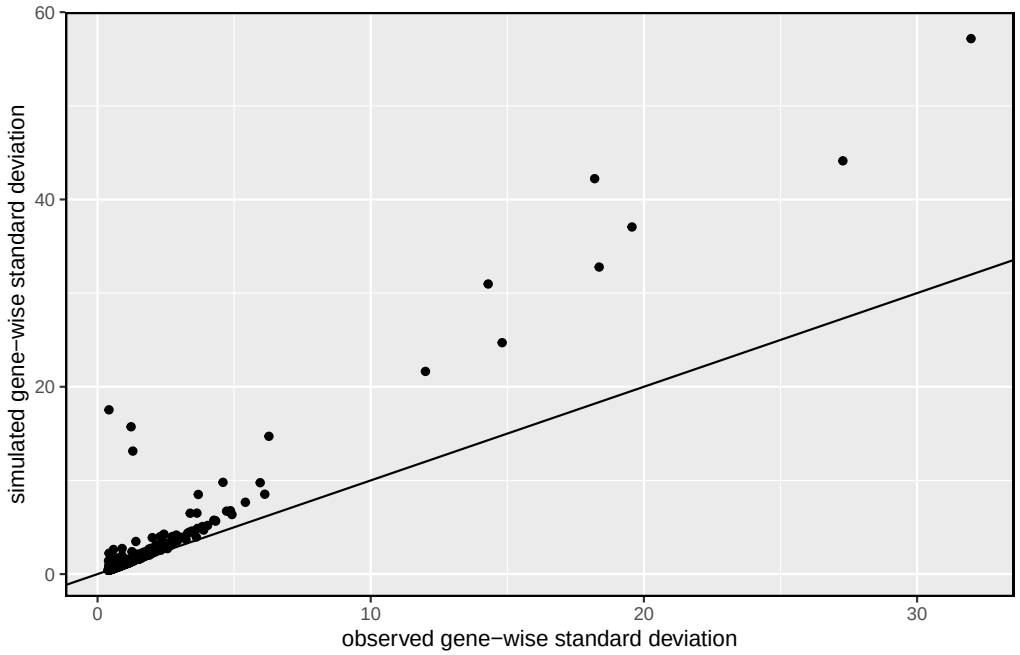

Figure 15: Gene-wise standard deviations of observed vs simulated data. For most genes, the standard devations are similar, but there are outliers for which the standard deviation in the observed data is very low, whereas it is high in the simulated data.

data, gene-wise Moran's Is are plotted against each other for the simulated and the observed data in figure 22. The Moran's Is were computed using a neighborhood matrix for spots obtained in a Visium experiment calculated with the R package DR.SC (Liu et al., 2023). The genes with higher spatial dependency measured by Moran's I in the observed data also tend to have a higher spatial dependency in the simulated data. However, the spatial dependency in the simulated data is lower than in the observed data for almost all genes. The model as estimated and used for simulation here can only include spatial dependency for gene expressions via the spatial distribution of cell types and the gene-wise spatial effects in form of domain-specific intercepts on the level of the log of the mean. Finer spatial dependencies like effects between neighboring spots cannot be included in the simulation.

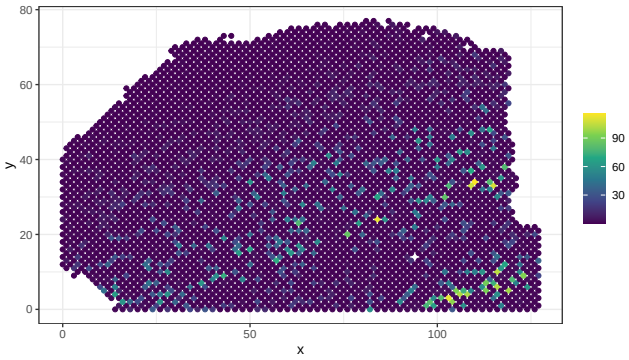

Figure 16: Fitted standard deviation for gene CALB2, a gene with large standard deviation in the simulated data and low standard deviation in the observed data. The fitted standard deviation can be very high, even though the observed counts are all low.

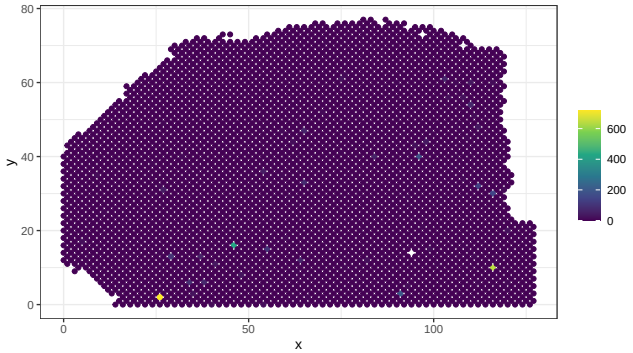

Figure 17: Simulated expressions for gene CALB2, a gene with large standard deviation in the simulated data and low standard deviation in the observed data. Very few spots have an extremely high simulated expression.

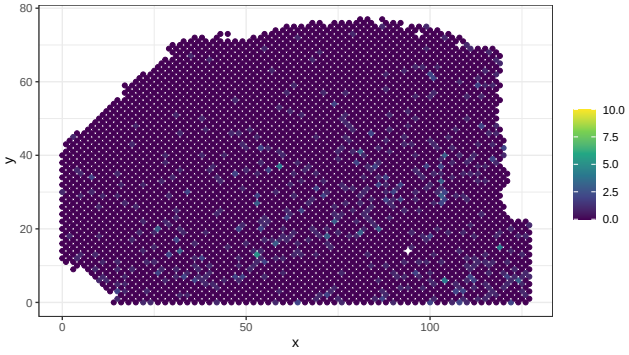

Figure 18: Observed expressions for gene CALB2, a gene with large standard deviation in the simulated data and low standard deviation in the observed data.

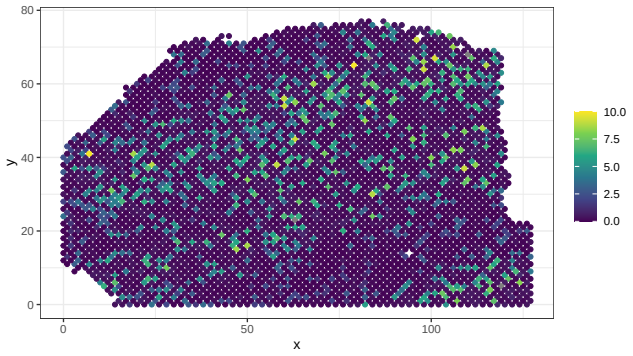

Figure 19: Estimated contribution of cell types per spots for gene CALB2, a gene with large standard deviation in the simulated data and low standard deviation in the observed data. This estimated contribution can be high in places where the observed expression is very low, likely leading the model to overestimate standard deviations.

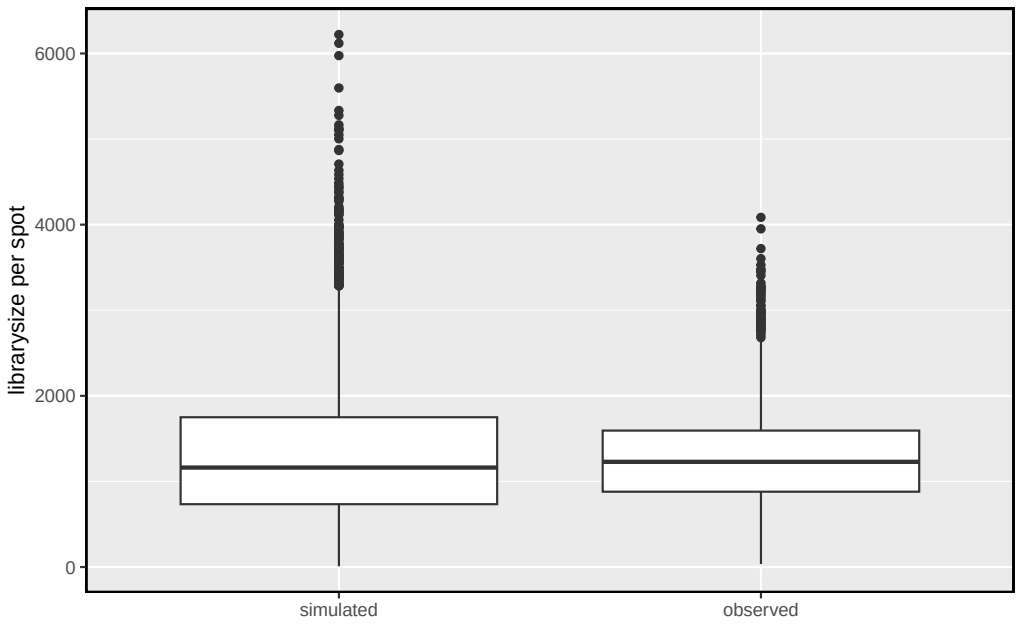

Figure 20: Observed vs simulated librarysizes per spot. The distribution of librarysizes per spots stays roughly the same in the simulated data, but it is a little more skewed to higher librarysizes.

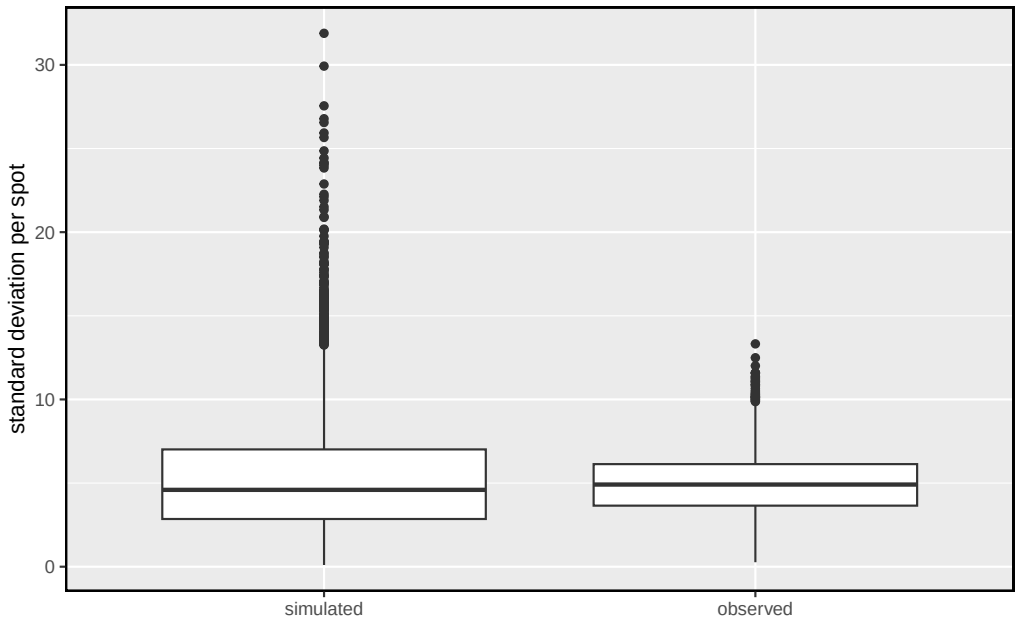

Figure 21: Observed vs simulated standard devations of expressions per spot. The distribution of standard deviations per spots is skewed to higher standard deviations in the simulated data.

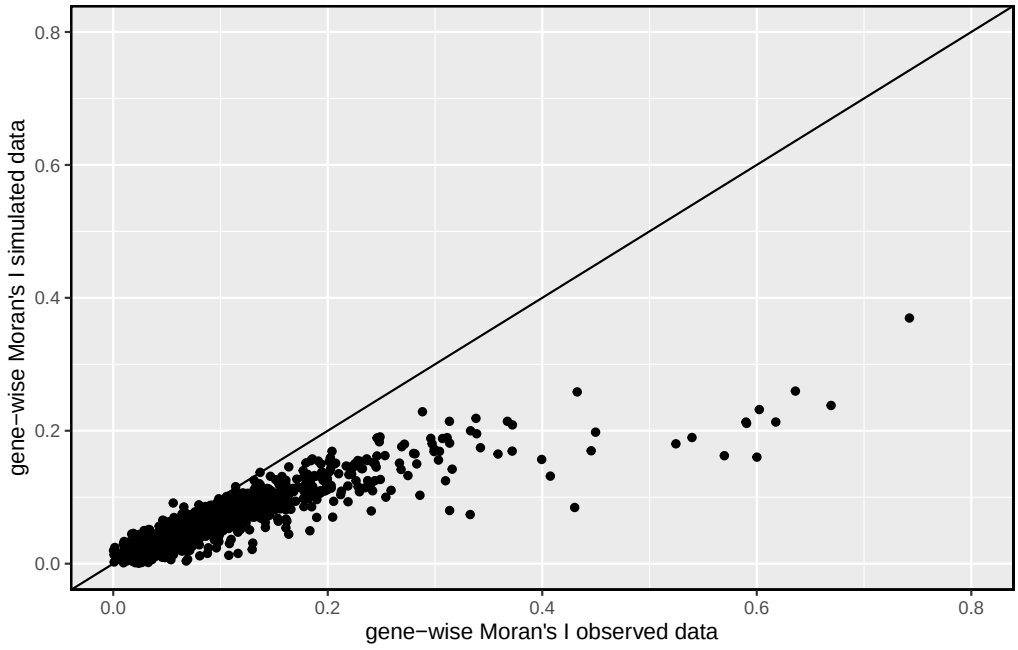

Figure 22: Gene-wise Moran's Is for the simulated and observed data. Genes with higher observed spatial dependency tend to have higher spatial dependency in the simulated data, too. However, spatial dependency in the simulated data, as measured by Moran's I, is lower than in the observed data.

3.3 Simulation Study

3.3.1 New parameters

Recall that for the simulation study comparing different methods to find spatial domains, the data is generated in different scenarios. The different scenarios comprise different amounts of technical noise and different amounts of spatial structure of the cell types. Table 1 shows the composition of each scenario. Figures 23 and 24 show the estimated and a examples of the newly generated spot and gene detection efficiencies. The newly generated efficiencies have a larger variance than the estimated ones, inducing higher technical noise for the respective scenarios. Note that for each generated sample in each scenario the spot and gene detection efficiencies were independently regenerated if needed for that scenario.

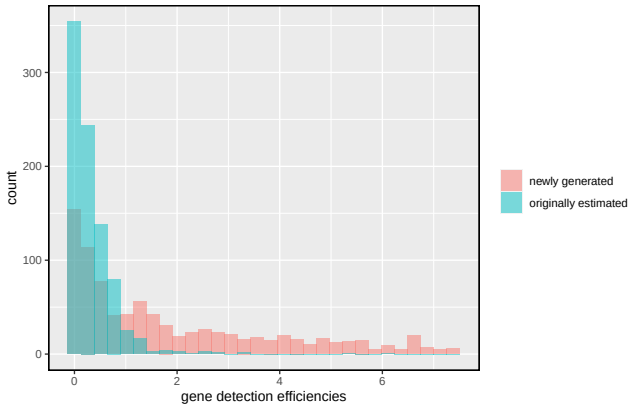

Figure 23: Histograms of the originally estimated and newly generated gene detection efficiencies

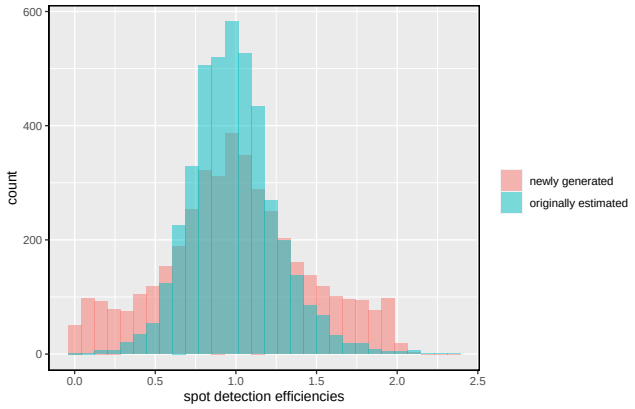

Figure 24: Histograms of the originally estimated and newly generated spot detection efficiencies

The estimated and newly generated cell types per spots for the DLPFC data are shown in scatterpies (figures 25 and 26) constructed with the R package SPOTlight (Elosua-Bayes and Crowell, 2023). These plots show the proportions of cell types at spots in a piechart for each spot. A very weak spatial pattern can be observed in the estimated cell types per spots, at most 3 spatial domains can be identified. For the newly generated ones no spatial pattern can be seen.

For the mouse kidney data, the estimated cell types per spots show a spatial pattern roughly correspondent to the 3 spatial domains (figure 27). The newly generated cell types per spots don't show any spatial pattern (figure 28).

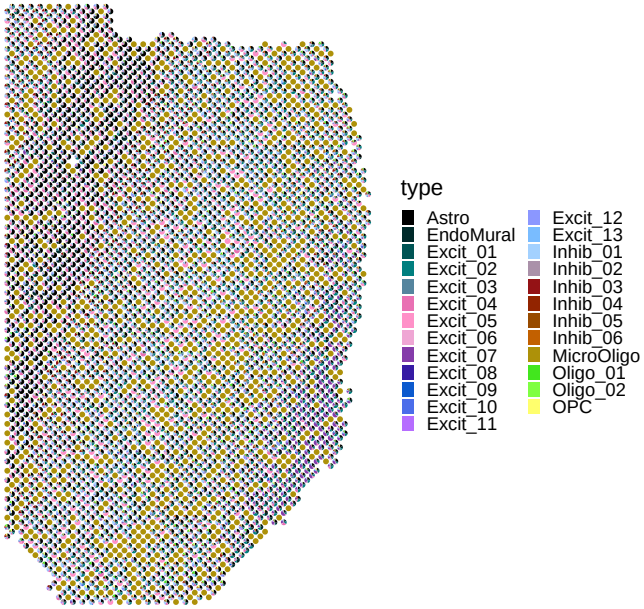

Figure 25: Scatterpie showing estimated celltypes per spot for the DLPFC data. A weak spatial pattern of cell types can be observed.

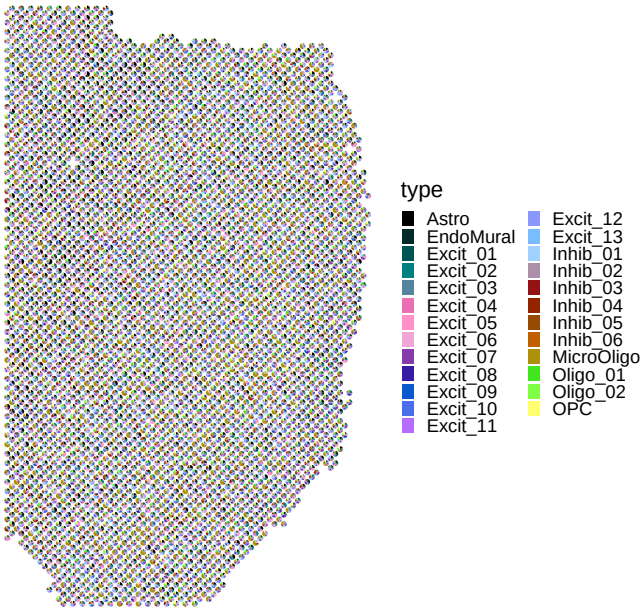

Figure 26: Scatterpie showing simulated celltypes per spot for the DLPFC data. No spatial pattern of cell types can be observed.

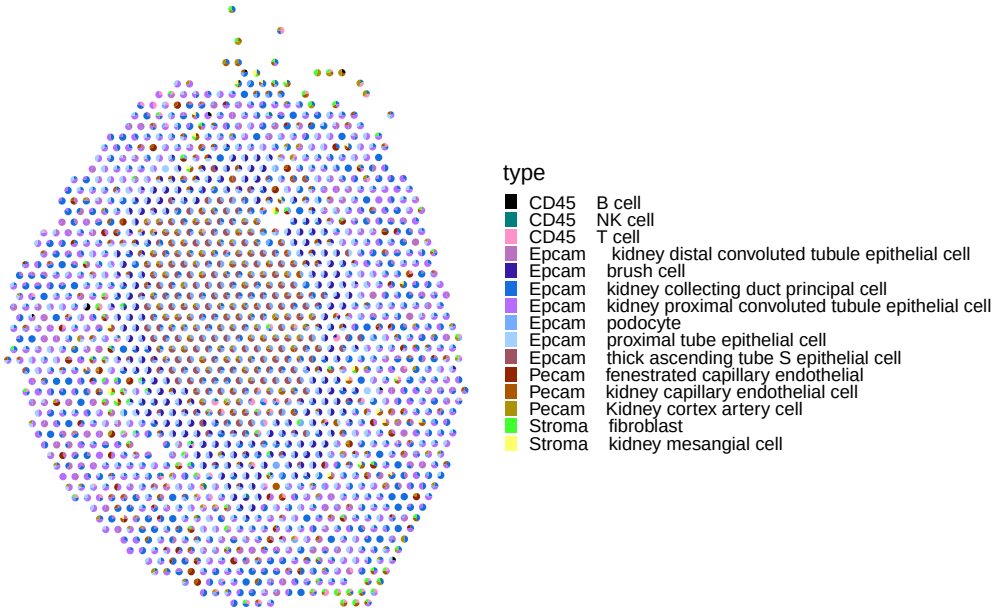

Figure 27: Scatterpie showing estimated celltypes per spot for the mouse kidney data. A clear spatial pattern of cell types corresponding roughly to the three spatial domains can be observed.

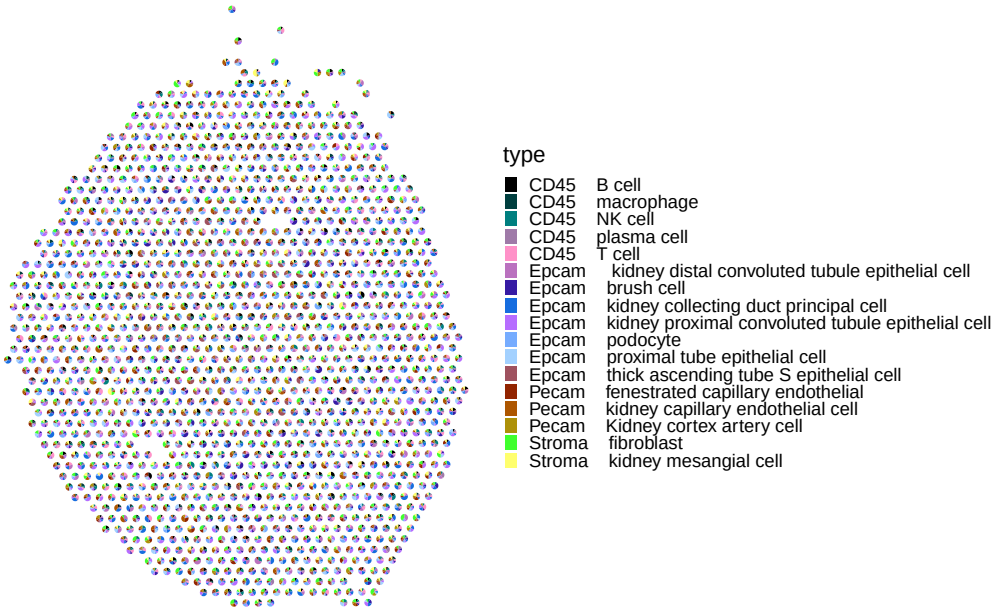

Figure 28: Scatterpie showing simulated celltypes per spot for the mouse kidney data. A spatial pattern of cell types can no longer be observed.

3.3.2 Performance of the domain recognition methods on the mouse kidney data

The performance of all the domain recognition methods tested here in scenario 1 for the mouse kidney data is very good. The ARIs for scenario 1 are shown in figure 29. The variance of the ARIs for the BayesSpace method is highest. When the gene detetcion efficiencies have higher variances as in scenario 2 (performances shown in figure 30), the performances of all methods goes up. This is because lots of genes had estimated gene detection efficiencies of nearly zero, and so there expressions were only very sparsely occurring. The new gene detection efficiencies allow for overall more gene expressions and thus more information for the domain recognition methods to use. With higher variance in spot detection efficiencies the performance of all methods decreases (figure 31). BayesSpace is most affected by this with the mean ARI dropping to about 0.5. STAGATE is least affected by the new spot detection efficiciencies. The weaker spatial pattern of cell types per domain also decreases the performance of all methods (figure 32). BayesSpace, again, has a large variance in its performance. STAGATE, again, is the least affected, while non-spatial clustering performs the worst in scenario 4. The performance plots of scnearios 5 to 8 can be viewed in the appendix. In general, the performances on the mouse kidney data is high, because there are only three different spatial domains to find.

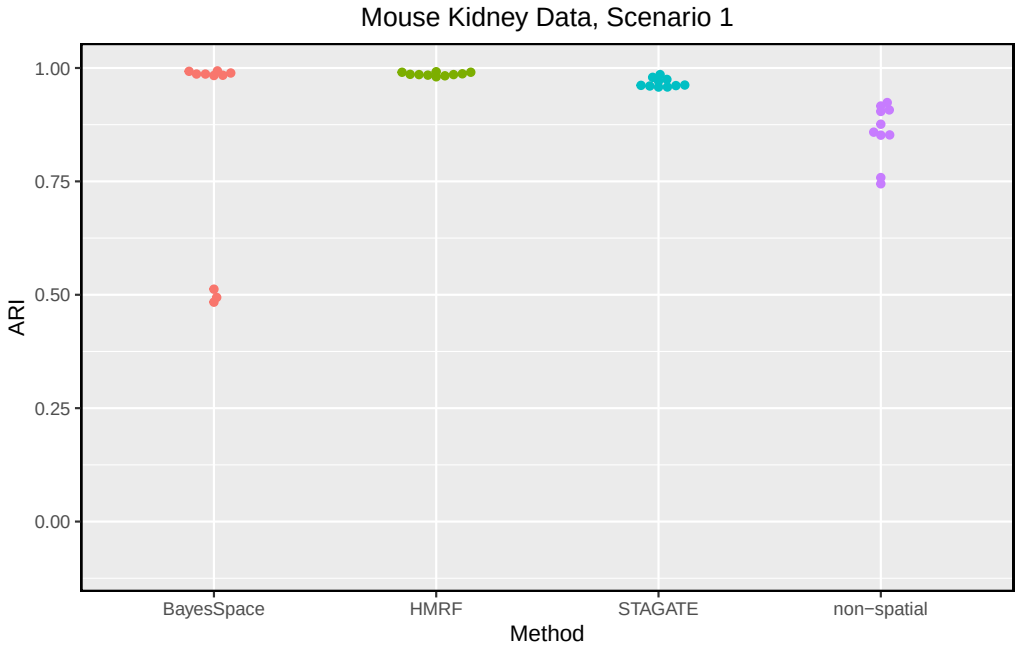

Figure 29: ARIs of the domain recognition methods on the mouse kidney data in scenario 1: model as estimated. All methods perform well, BayesSpace has highest variance in performance.

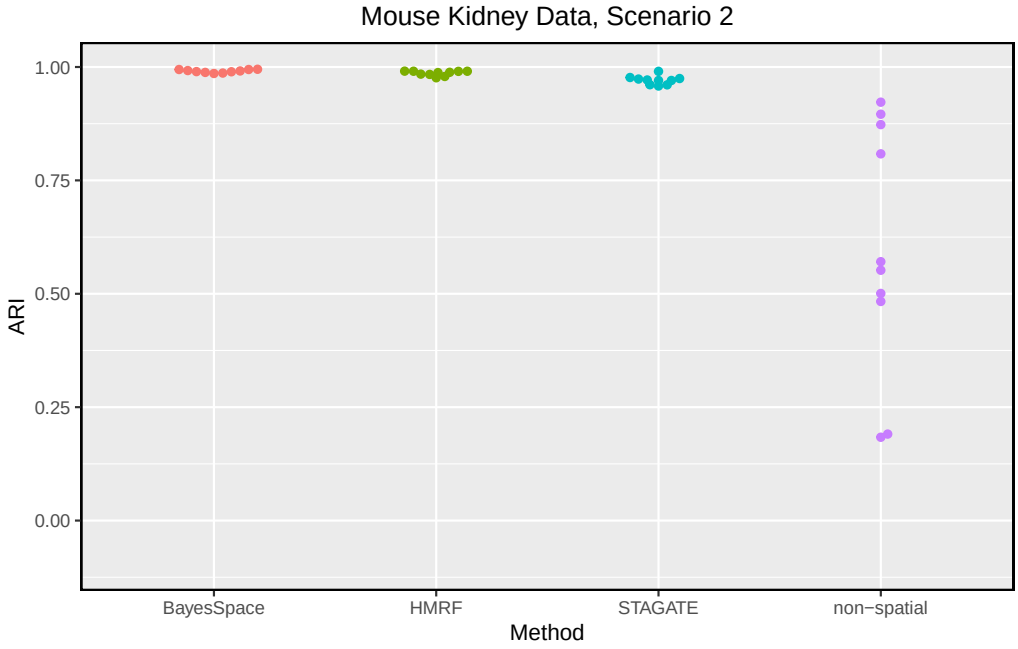

Figure 30: ARIs of the domain recognition methods on the mouse kidney data in scenario 2: increased variation in gene detection efficencies. The methods perform better than in scenario 1, because there are overall more gene expressions to work with.

Figure 31: ARIs of the domain recognition methods on the mouse kidney data in scenario 3: increased variation in spot detection efficiencies. All methods perform worse than in scenario 1, the difference is largest for BayesSpace.

Figure 32: ARIs of the domain recognition methods on the mouse kidney data in scenario 4: weaker spatial pattern of cell types per domains. All methods perform worse than in scenario 1. BayesSpace, again, has the highest variance in performance

3.3.3 Performance of the domain recognition methods on the DLPFC data

The performance of the domain recognition methods on the DLPFC data is, in general, lower than on the mouse kidney data, because the DLPFC data has 7 spatial domains to find. Overall, STAGATE performs by far the best on the DLPFC data, in second place is BayesSpace, which again has the highest variance in resulting ARIs. Scenario 2 (figure 34) with higher variance in gene detection efficiencies, again, makes the task simpler, because the expression of many genes is detected more often. Scenario 3 (higher variance in spot detection efficiencies, figure 35) decreases the performance of all methods compared to scenario 1 (as estimated, figure 33). Scenario 4 (figure 36), which has less spatial pattern in cell types per domains, has no impact on the performance of the methods compared to scenario 1. This is likely due to the already weak estimated spatial pattern in cell types per domains.

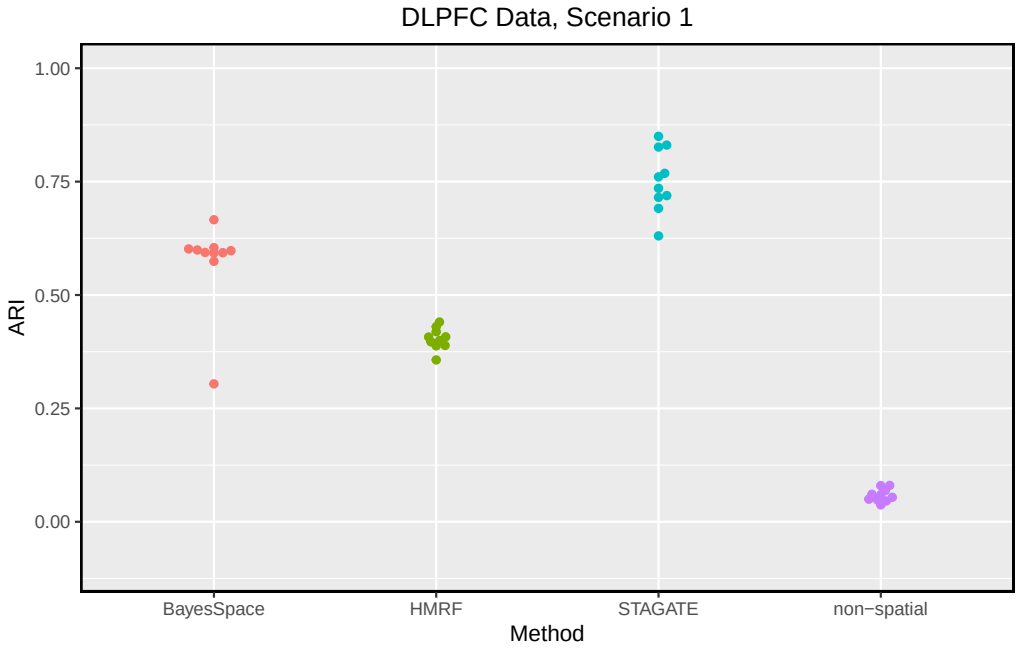

Figure 33: ARIs of the domain recognition methods on the DLPFC data in scenario 1: model as estimated. STAGATE's performance is highest. In general, the performance of all methods is worse than on the mouse kidney data, because here 7 spatial domains have to be identified instead of only 3.

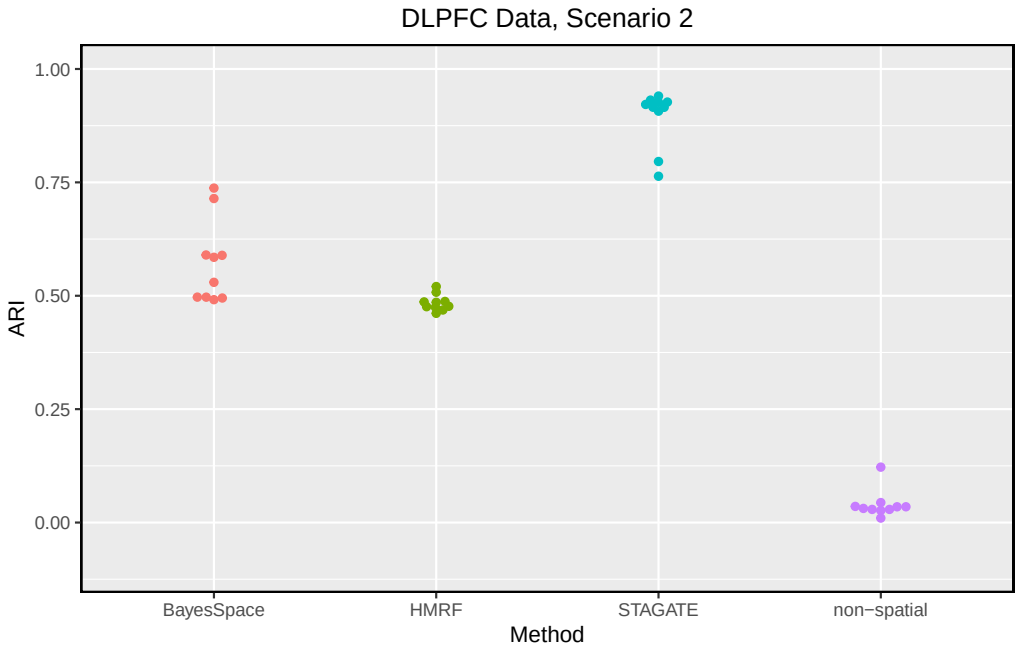

Figure 34: ARIs of the domain recognition methods on the DLPFC data in scenario 2: increased variation in gene detection efficencies. Compared to scenario 1 the performances are better.

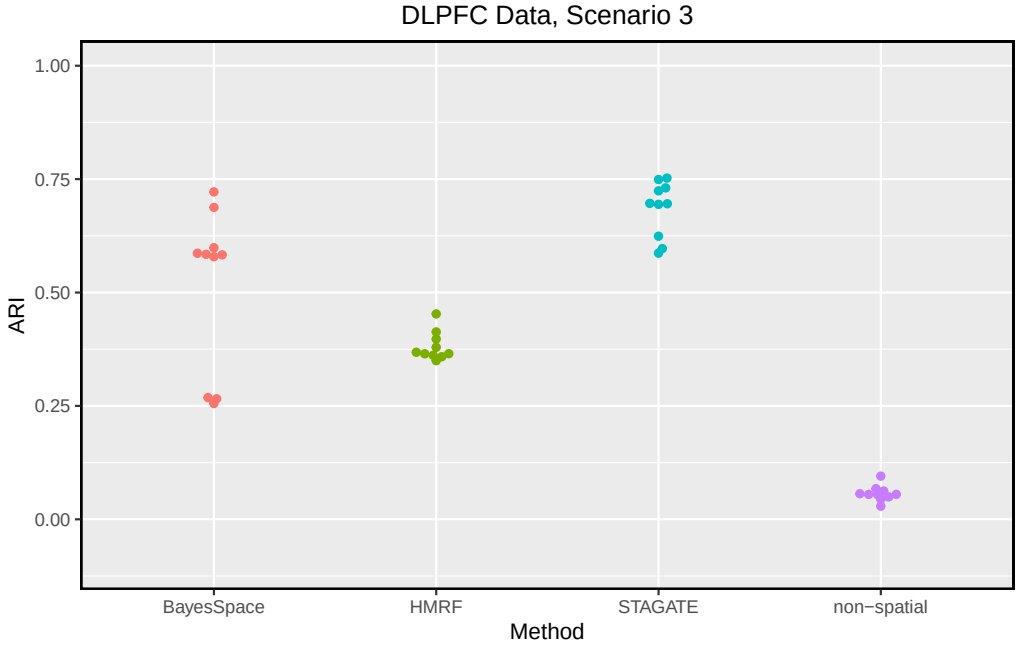

Figure 35: ARIs of the domain recognition methods on the DLPFC data in scenario 3: increased variation in spot detection efficiencies. Compared to scenario 1 the performances are worse.

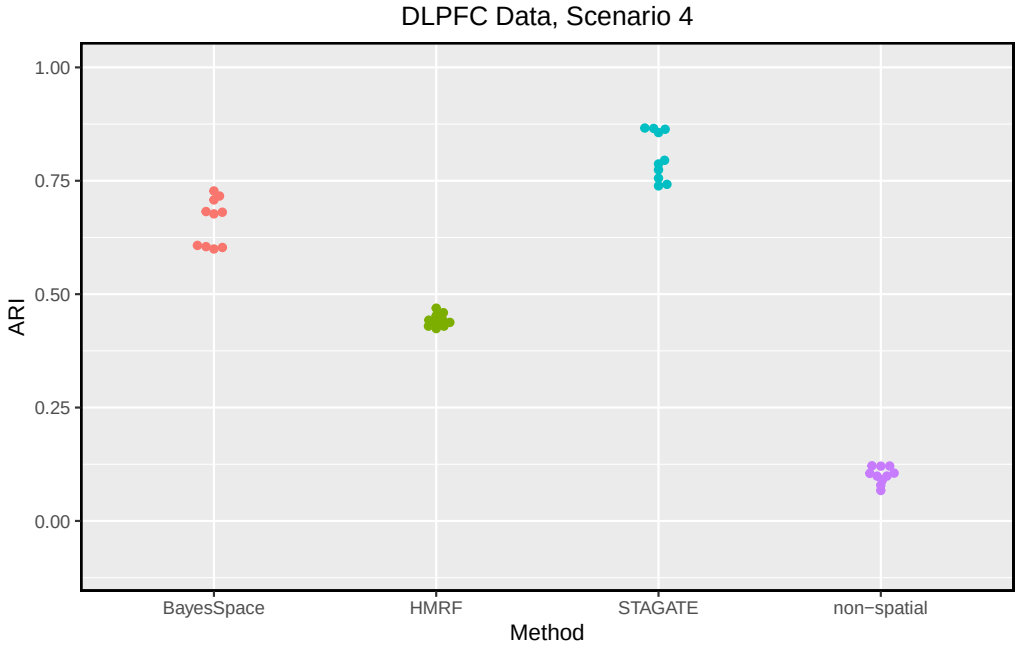

Figure 36: ARIs of the domain recognition methods on the DLPFC data in scenario 4: weaker spatial pattern of cell types per domains. Compared to scenario 1 the performances are about equal.

4 Conclusion

A parametric model with interpretable parameters to simulate ST data was presented. The gene-wise means were preserved and the gene-wise standard deviations were mostly preserved in the simulated data. For very few genes the estimated marginal distributions did not fit well to the data, resulting in unrealistic simulated gene expressions. Due to the convoluted nature of the estimation (different steps, of which some try to estimate similar effects) the cause for this bad fit isn't always clear, and the estimated parameters have to be treated with some caution with respect to their interpretability (because of identifiability issues when estimating parameters in this convoluted manner). The spatial dependency of gene expressions, as measured by Moran's I, was largely preserved in the simulated data, however, in general it was lower.

Of the investigated domain recognition methods STAGATE performed the best in the simulation study. BayesSpace had high amounts of variance in performance but had better mean performance than HMRF. BayesSpace and HMRF have both a relatively strong drop in performance with higher varying spot detection efficiencies in the data. The performance of all methods depends strongly on the number of spatial domains to be found.

References

(2020). A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583** 590–595.

Asp, M., Bergenstråhle, J. and Lundeberg, J. (2020). Spatially resolved transcriptomes—next generation tools for tissue exploration. *BioEssays* **42** 1900221.

BinTayyash, N., Georgaka, S., John, S., Ahmed, S., Boukouvalas, A., Hensman, J. and Rattray, M. (2021). Non-parametric modelling of temporal and spatial counts data from rna-seq experiments. *Bioinformatics* **37** 3788–3795.

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. and Zhuang, X. (2015). Spatially resolved, highly multiplexed rna profiling in single cells. *Science* **348** aaa6090.

Crowell, H. L. (2023). *TENxVisiumData: Visium spatial gene expression data by 10X Genomics*. R package version 1.8.0.
URL `https://bioconductor.org/packages/TENxVisiumData`

Dong, K. and Zhang, S. (2022). Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications* **13** 1739.

Dries, R., Zhu, Q., Dong, R., Eng, C.-H. L., Li, H., Liu, K., Fu, Y., Zhao, T., Sarkar, A., Bao, F. et al. (2021). Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology* **22** 1–31.

Elosua-Bayes, M. and Crowell, H. L. (2023). *SPOTlight: 'SPOTlight': Spatial Transcriptomics Deconvolution*. R package version 1.4.1.
URL `https://bioconductor.org/packages/SPOTlight`

Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. and Heyn, H. (2021). Spotlight: seeded nmf regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic acids research* **49** e50–e50.

Eng, C.-H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.-C. et al. (2019). Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature* **568** 235–239.

Giotto (????). Giotto vignette mouse seqfish cortex. `https://rubd.github.io/Giotto_site/articles/mouse_seqFISH_cortex_200914.html#part-11-hmrf-spatial-domains`. Accessed: 2023-06-10.

Handcock, M. S., Meier, K. and Nychka, D. (1994). Kriging and splines: An empirical comparison of their predictive performance in some applications: Comment. *Journal of the American Statistical Association* **89** 401–403.

Hubert, L. and Arabie, P. (1985). Comparing partitions journal of classification 2 193–218. *Google Scholar* 193–128.

Junker, J. P., Noel, E. S., Guryev, V., Peterson, K. A., Shah, G., Huisken, J., McMahon, A. P., Berezikov, E., Bakkers, J. and van Oudenaarden, A. (2014). Genome-wide rna tomography in the zebrafish embryo. *Cell* **159** 662–675.

Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C. and Nilsson, M. (2013). In situ sequencing for rna analysis in preserved tissue and cells. *Nature methods* **10** 857–860.

Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H. W., Li, T., Elmentaite, R., Lomakin, A., Kedlian, V., Gayoso, A. et al. (2022). Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature biotechnology* **40** 661–671.

Liu, W., Yang, Y. and Liu, J. (2023). *DR.SC: Joint Dimension Reduction and Spatial Clustering*. R package version 3.2.
URL `https://CRAN.R-project.org/package=DR.SC`

Maynard, K. R., Collado-Torres, L., Weber, L. M., Uytingco, C., Barry, B. K., Williams, S. R., Catallini, J. L., Tran, M. N., Besich, Z., Tippani, M. et al. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience* **24** 425–436.

Meier-Ruge, W., Bielser, W., Remy, E., Hillenkamp, F., Nitsche, R. and Unsöld, R. (1976). The laser in the lowry technique for microdissection of freeze-dried tissue slices. *The Histochemical Journal* **8** 387–401.

Merritt, C. R., Ong, G. T., Church, S. E., Barker, K., Danaher, P., Geiss, G., Hoang, M., Jung, J., Liang, Y., McKay-Fleisch, J. et al. (2020). Multiplex digital spatial profiling of proteins and rna in fixed tissue. *Nature biotechnology* **38** 586–599.

Moses, L. and Pachter, L. (2022). Museum of spatial transcriptomics. *Nature Methods* **19** 534–546.

Pardo, B., Spangler, A., Weber, L. M., Hicks, S. C., Jaffe, A. E., Martinowich, K., Maynard, K. R. and Collado-Torres, L. (2022). spatiallibd: an r/bioconductor package to visualize spatially-resolved transcriptomics data. *BMC Genomics* .
URL `https://doi.org/10.1186/s12864-022-08601-w`

Pham, D., Tan, X., Xu, J., Grice, L. F., Lam, P. Y., Raghubar, A., Vukovic, J., Ruitenberg, M. J. and Nguyen, Q. (2020). stlearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *BioRxiv* 2020–05.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics* **54** 507–554.

Rüschendorf, L. (2013). Mathematical risk analysis. *Springer Ser. Oper. Res. Financ. Eng. Springer, Heidelberg* .

Salehi, A. and Davulcu, H. (2019). Graph attention auto-encoders. *arXiv preprint arXiv:1905.10715* .

Song, D., Wang, Q., Yan, G., Liu, T., Sun, T. and Li, J. J. (2023). scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nature Biotechnology* 1–6.

STAGATE (2021). Stagate tutorial 1: 10x visium (dlpfc dataset). `https://stagate.readthedocs.io/en/latest/T1_DLPFC.html`. Accessed: 2023-06-10.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M. et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353** 78–82.

Sun, S., Zhu, J. and Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature methods* **17** 193–200.

Sun, T., Song, D., Li, W. V. and Li, J. J. (2021). scdesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome biology* **22** 163.

Svensson, V., Teichmann, S. A. and Stegle, O. (2018). Spatialde: identification of spatially variable genes. *Nature methods* **15** 343–346.

Wang, X., Allen, W., Wright, M., Sylwestrak, E., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J. et al. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. science 361, eaat5691.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **58** 236–244.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* **73** 3–36.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Zhao, E., Stone, M. R., Ren, X., Guenthoer, J., Smythe, K. S., Pulliam, T., Williams, S. R., Uytingco, C. R., Taylor, S. E., Nghiem, P. et al. (2021). Spatial transcriptomics at subspot resolution with bayesspace. *Nature biotechnology* **39** 1375–1384.

Zhu, Q., Shah, S., Dries, R., Cai, L. and Yuan, G.-C. (2018). Identification of spatially associated subpopulations by combining scrnaseq and sequential fluorescence in situ hybridization data. *Nature biotechnology* **36** 1183–1190.

A Appendix

A.1 Performance of domain recognition methods for scenarios 5-8 on the mouse kidney data

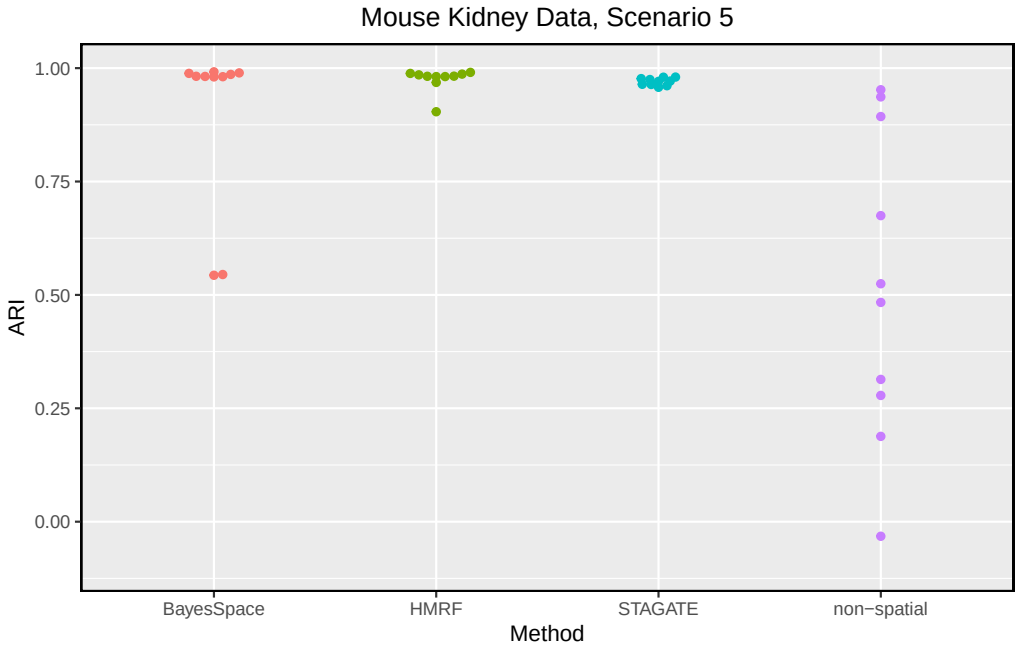

Figure 37: ARIs of the domain recognition methods on the mouse kidney data in scenario 5.

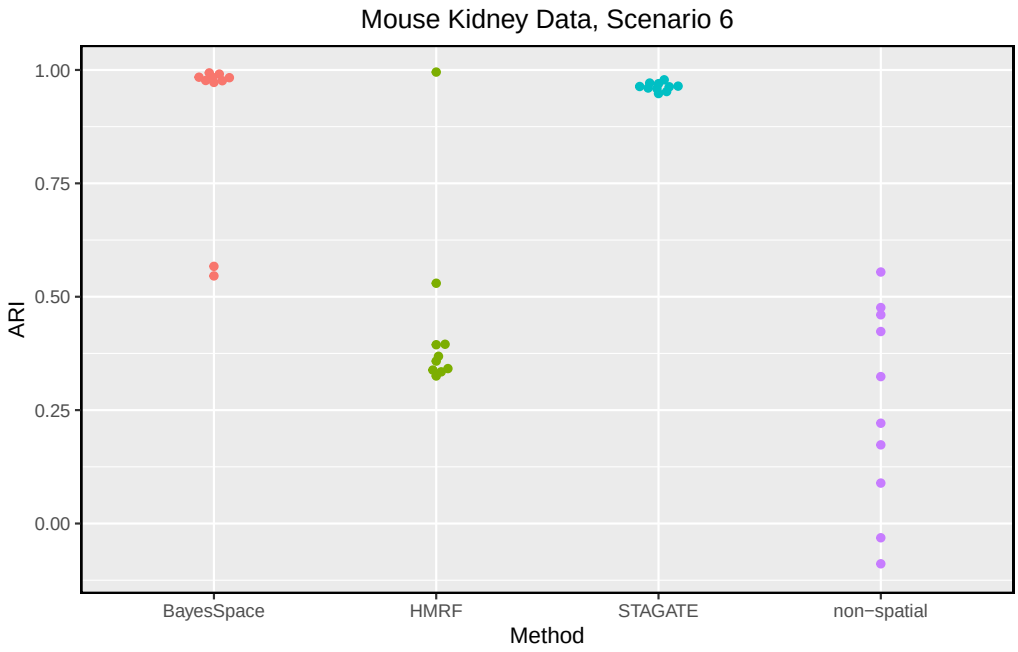

Figure 38: ARIs of the domain recognition methods on the mouse kidney data in scenario 6.

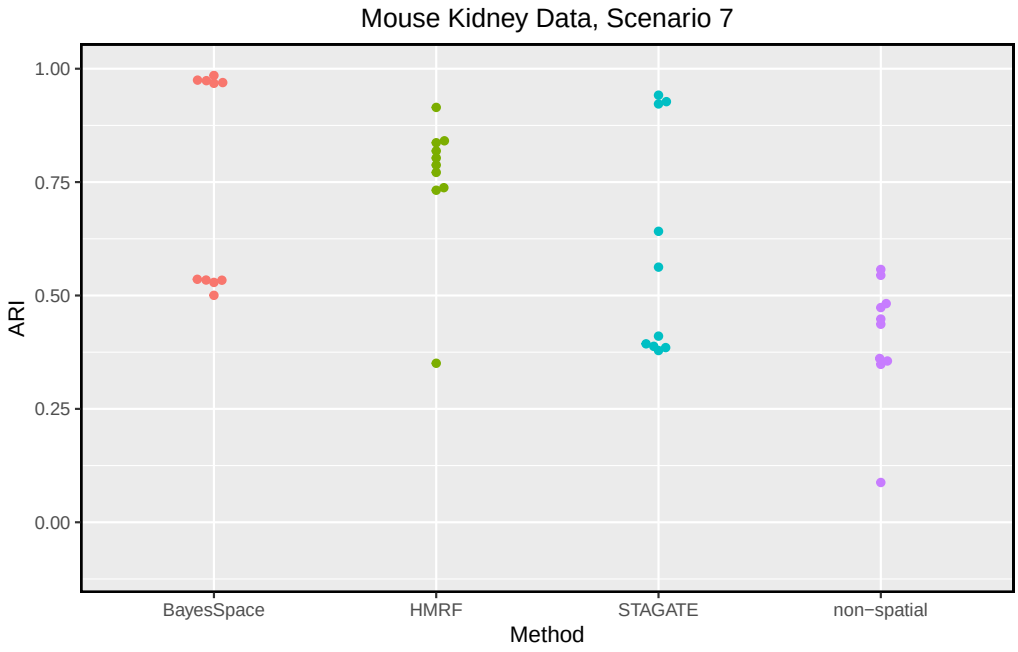

Figure 39: ARIs of the domain recognition methods on the mouse kidney data in scenario 7.

Figure 40: ARIs of the domain recognition methods on the mouse kidney data in scenario 8.

A.2 Performance of domain recognition methods for scenarios 5-8 on the DLPFC data

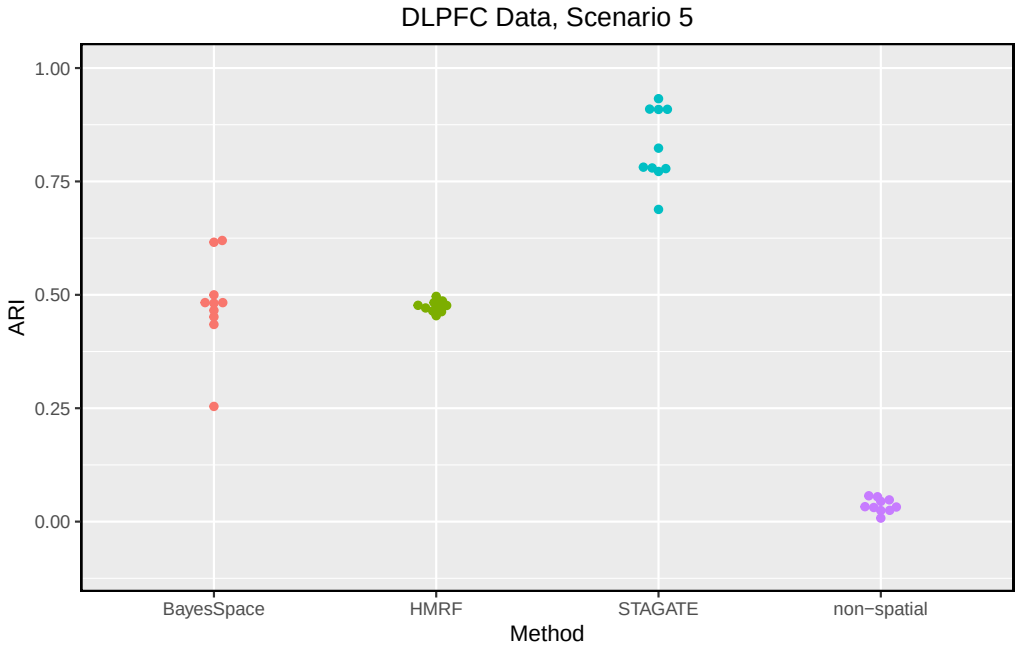

Figure 41: ARIs of the domain recognition methods on the DLPFC data in scenario 5.

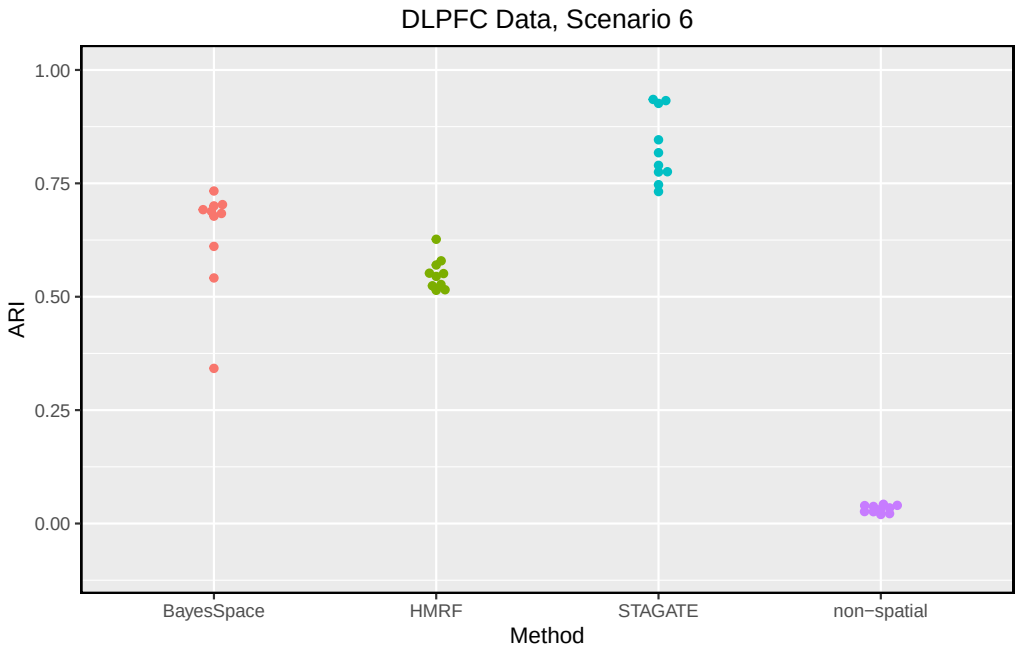

Figure 42: ARIs of the domain recognition methods on the DLPFC data in scenario 6.

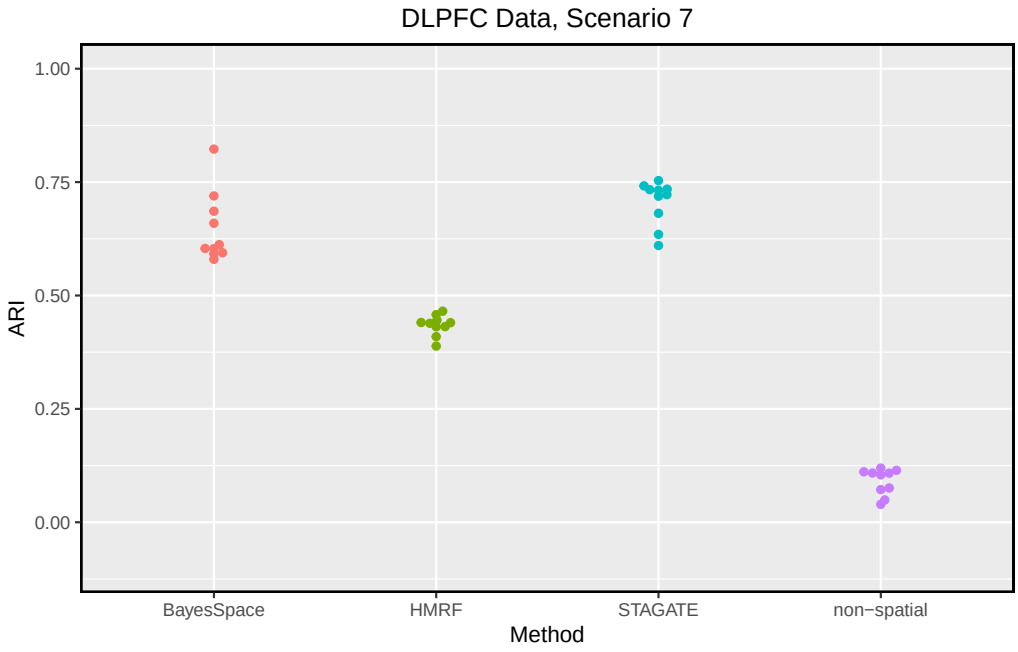

Figure 43: ARIs of the domain recognition methods on the DLPFC data in scenario 7.

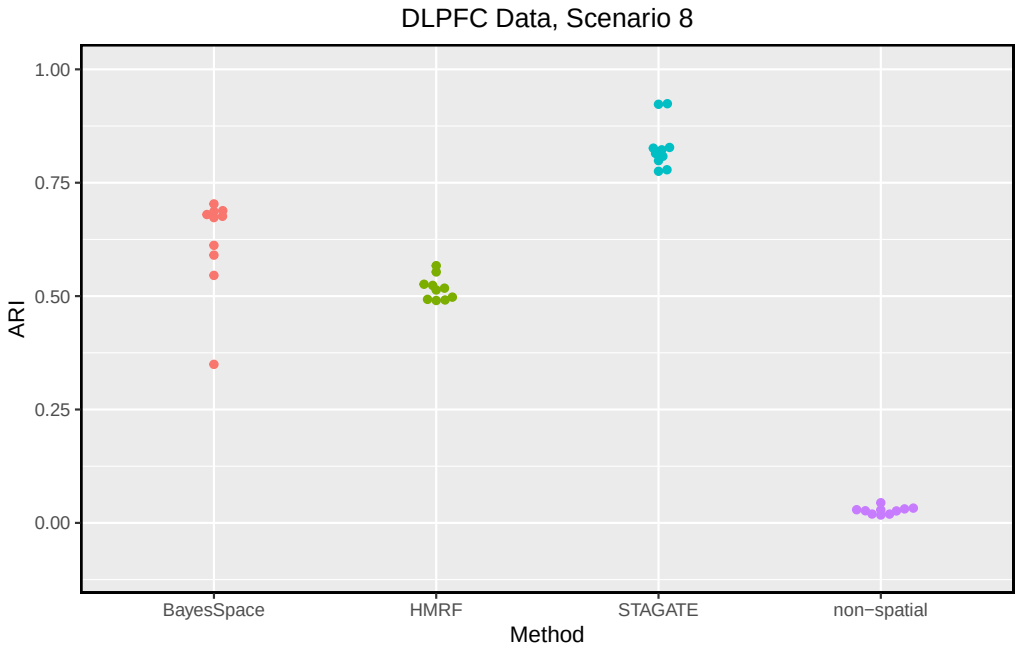

Figure 44: ARIs of the domain recognition methods on the DLPFC data in scenario 8.