FISHFactor: A Probabilistic Factor Model for Spatial Transcriptomics Data with Subcellular Resolution

Florin Walter[1], Oliver Stegle[1,2,3], and Britta Velten[1,3,⋆]

[1] Division of Computational Genomics and System Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany
[2] European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany
[3] Cellular Genetics Programme, Wellcome Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

Abstract. Factor analysis is a widely-used method for dimensionality reduction of high-throughput datasets in molecular biology and has recently been adapted to spatial transcriptomics data. However, existing methods assume (count) matrices as input and are therefore not directly applicable to single-molecule resolved data, which increasingly arise for example from multiplexed fluorescence in-situ hybridization or in-situ sequencing experiments. To address this, we here propose FISHFactor, a probabilistic model that combines the benefits of spatial, non-negative factor analysis with a Poisson point process likelihood to explicitly model and account for the nature of single-molecule resolved data. FISHFactor furthermore leverages principles of multi-modal factor analysis to enable dissecting the transcriptional heterogeneity between multiple groups of samples, such as different cells. Using simulated and real data, we show that our approach leads to improved estimates of the true spatial transcriptome landscape compared to existing methods that rely on aggregating information by spatial binning. Applied to a set of NIH/3T3 cells, FISHFactor identifies major subcellular expression patterns and accurately recovers known spatial gene clusters.

Keywords: spatial transcriptomics, single-molecule resolution, factor analysis, point process, Gaussian process

1 Introduction

Transcriptomic profiling of individual cells using single-cell RNA sequencing is now a widely accessible tool to study the cellular heterogeneity of tissues and has helped to uncover novel cell types. However, single-cell RNA sequencing protocols are based on a disassociation step, and therefore can only provide limited insights into the spatial organization of tissue and no information on the localization of RNA molecules within a cell. To address this, a growing number of spatially resolved transcriptomic technologies are being developed, which enable measurements of the gene expression while retaining spatial information [27]. For example, next-generation sequencing coupled with spatial barcodes provides whole transcriptome measurements of tissue regions [28, 29], but the resolution of current methods is at most at the level of individual cells and cannot resolve subcellular localizations. On the other hand, imaging-based techniques such as in-situ sequencing [9, 16, 17, 34] or fluorescence in-situ hybridization (FISH) achieve a subcellular resolution by measuring spatial positions of individual molecules. While FISH technologies were originally limited to the detection of a single or at most a handful of genes [14, 22, 26], advances in imaging technologies, sequential hybridization and barcoding strategies nowadays enable probing tens to thousands of genes in a single experiment [8, 10, 12, 13, 20, 21], thus rendering such techniques increasingly applicable for the identification of subcellular gene expression patterns at scale.

Despite the availability of technologies that offer single-molecule resolution, established analysis strategies to process these data do not fully exploit the given resolution. Instead, RNA quantifications are limited to cellular resolution, for example using image-based cell segmentation and aggregation of molecule counts per cell [8, 10, 12]. Such approaches fail to capture subcellular expression patterns of mRNAs, which can provide important insights into cellular states, heterogeneity within cell types [6, 36] and can modulate the function of genes [12]. While segmentation-free or hybrid approaches have been proposed [19, 23–25], these existing methods focus on the task of cell type inference or clustering of spots and hence do not explicitly

⋆ Correspondence: b.velten@dkfz.de

model subcellular localization patterns of individual RNA molecules within the cell. With the increasing throughput of single-molecule techniques, it will become ever more important to identify major subcellular gene expression patterns and to use them as additional dimension when dissecting cell-to-cell heterogeneity.

Factor models are already widely used for the unsupervised discovery of the principal sources of variation in high-dimensional molecular data sets [1,2,5,30,35] and recent extensions to spatial data have successfully identified spatial gene expression patterns at the *cellular* level [3,31,33]. However, these existing methods cannot leverage the *subcellular* resolution of the data, as they require a (normalized) count matrix as input. Consequently, they are not directly applicable to single molecule data and instead require crude aggregation steps, such as spatial binning or using cell segmentations, thereby resulting in a loss of the exact spatial information.

To address these shortcomings, we here propose FISHFactor, a principled factor analysis framework that opens up the application of factor models for spatially resolved single-molecule data and enables the unbiased identification and discovery of subcellular expression patterns (Fig. 1). In contrast to existing spatial factor models, FISHFactor employs spatial Poisson point processes as observation model to explicitly model the subcellular coordinates of each molecule and thereby can fully leverage the single-molecule resolution of the data. We combine this with a spatially aware inference of factors using Gaussian processes tailored to spatial transcriptomics data and impose interpretable factors and weights using non-negativity constraints. FISHFactor is designed as multi-modal factor model that jointly models the information from multiple cells, thereby enabling the integration and comparison of subcellular localization patterns across a population of cells. We assess our model using simulated and real data, where we demonstrate advantages of FISHFactor over conventional approaches based on binning. We illustrate the utility of FISHFactor to uncover major subcellular localization patterns of genes, compare these across cells and analyze the co-localization of genes within a cell.

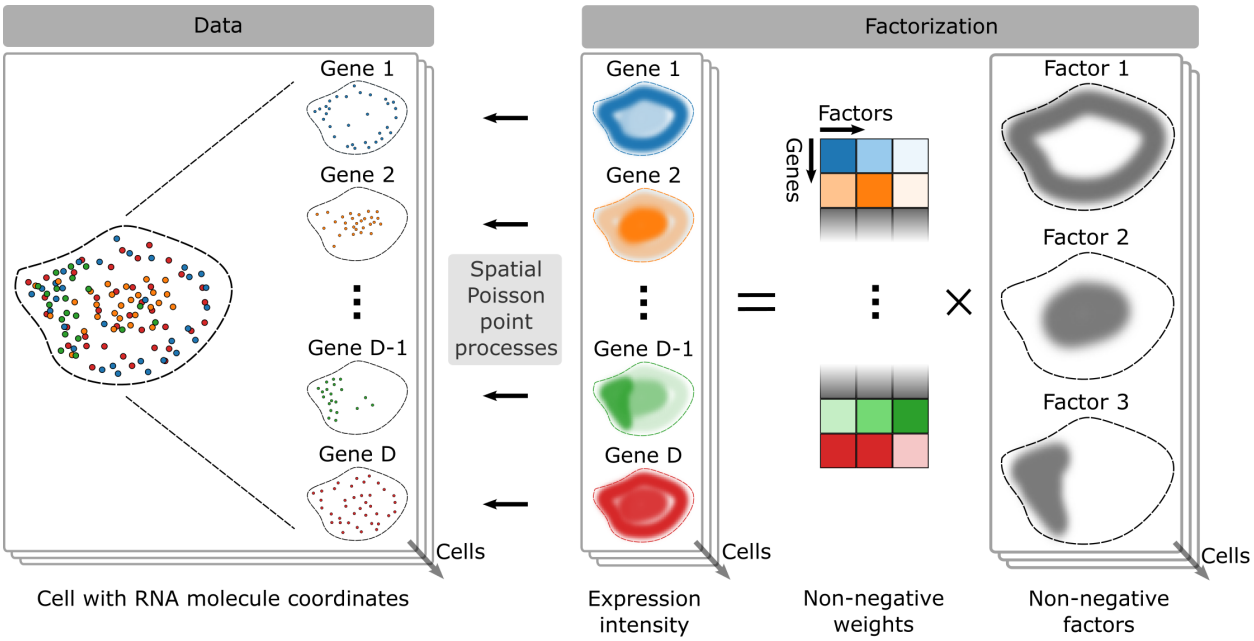

Fig. 1. Illustration of FISHFactor for the analysis of spatial transcriptomics data with subcellular resolution. The input dataset (left) consists of molecule coordinates, e. g. from multiplexed FISH measurements, for multiple genes in multiple groups or cells (one cell shown for illustration). FISHFactor models the observed RNA molecule coordinates as realizations of a spatial Poisson point process with a gene-wise intensity function in each cell. The intensity functions of all genes and cells are then governed by a low-dimensional decomposition into non-negative spatially-aware factors and non-negative weights (right, illustrated for 3 latent factors). Weights are shared across cells, whereas factor values are specific to individual cells.

2 Methods

2.1 Factor analysis for count data

Factor analysis is a dimensionality reduction technique that is frequently applied for the unsupervised analysis of high-dimensional omics data sets [30]. Based on correlation structures in a high-dimensional feature space, factor analysis aims to find a low-dimensional embedding in terms of a small number of latent factors representing the major axes of variation in the data. Starting from a high-dimensional dataset $\mathbf{Y} \in \mathbb{R}^{N \times D}$ with N observations of D features, factor analysis finds a factorization $\mathbf{Y} = \mathbf{Z}\mathbf{W}^T + \mathbf{E}$ with K latent factors $\mathbf{Z} \in \mathbb{R}^{N \times K}$ (typically $K \ll D$), associated factor weights $\mathbf{W} \in \mathbb{R}^{D \times K}$ and residual noise $\mathbf{E} \in \mathbb{R}^{N \times D}$. In contrast to non-linear dimensionality reduction methods, factor analysis can find a highly interpretable latent embedding, because the weights directly link each latent factor to molecular features. Formulated in a probabilistic framework, factor analysis furthermore enables to incorporate prior knowledge and different sparsity assumptions by the use of suitable prior distributions, provides uncertainty estimates for the inferred variables and can account for various data types by the use of appropriate likelihood models. As a baseline model for sequencing data we here consider a Poisson likelihood, which can account for the count nature of the data and has been successfully applied to transcriptomics data [32]. The decomposition in a Poisson factor model is given by

$$p(y_{nd} \mid \mathbf{W}, \mathbf{Z}) = \text{Poisson}(\lambda_{nd}) \tag{1}$$

$$\lambda_{nd} = g(\sum_{k=1}^{K} w_{dk} z_{nk}), \tag{2}$$

where g denotes a positive inverse-link function such as the exponential or softplus.

2.2 Non-negative factor models

To improve the interpretability and identifiability of factor analysis models, different sparsity assumptions on the factors and weights have been employed, including sparsity on the level of features or sets of features [1,2,35] as well as non-negativity constraints [5]. For spatial transcriptomics data, non-negativity constraints have been found to be particularly useful [31], as they allow to find additive non-negative spatial patterns and molecular signatures. In practice this is achieved by constraining the weights and factors to non-negative values, for example using non-negative parametrization or non-negative priors. One such model is *probabilistic non-negative matrix factorization (PNMF)* [31], where

$$\lambda_{nd} = \sum_{k=1}^{K} w_{dk} z_{nk} \tag{3}$$

$$z_{nk} = \exp(f_{nk}) \tag{4}$$

$$f_{nk} \sim \mathcal{N}(\mu_k, \sigma_k^2) \tag{5}$$

with non-negative weights and additional options to account for size difference if applied at the cellular level.

2.3 Gaussian process factor analysis

A limitation of classical factor analysis in applications to spatial data is its assumption of independent observations $n = 1, \ldots, N$. While this assumption may be appropriate for some data types, it generally does not hold for spatial data, where each observation comes with a spatial coordinate and spatial structures are present between the samples. For example, gene expression profiles at nearby points are expected to be more similar than at points that are far apart. This spatial covariance can be incorporated into factor analysis by replacing the univariate Gaussian priors on the factors in Eq. (5) by multivariate priors that can model covariation across samples. A flexible prior for this purpose are Gaussian process (GP) priors, which provide a non-parametric framework to model continuous dependencies between samples. This has given rise to the framework of *Gaussian process factor analysis (GPFA)* [7], where independent GP priors

are placed on the factors to model smooth temporal patterns. The same concept has recently been applied for the identification of patterns in spatial transcriptomics data, in combination with different likelihood models and sparsity constraints [31, 33]. In particular, [31] defines *non-negative spatial factorization (NSF)* as an extension to PNMF, where the factor prior in Eq. (5) is replaced with

$$f_{nk} = f_k(\mathbf{c}_n) \tag{6}$$

$$f_k \sim \mathrm{GP}(\mu_k, \kappa_k). \tag{7}$$

Here, $\mathbf{c}_n \in \mathbb{R}^2$ is the spatial coordinate of sample n and μ_k a mean function, which in NSF is given by a linear function in the spatial coordinates. The choice of the kernel function κ_k determines the covariance structure. For example, a squared exponential kernel produces very smooth patterns, whereas a Matérn kernel leads to a more angular appearance as often observed for spatial expression patterns [31].

2.4 Poisson point process likelihood

In contrast to (spatial) transcriptomics data sets at the cellular level, single-molecule resolved data consists of a list of N coordinate vectors $\{\mathbf{c}_n\}_{n=1,\dots,N}$ for each gene with $\mathbf{c}_n \in \mathbb{R}^2$. For such data, the Poisson likelihood used in the previous models can only be employed after a pre-processing step that aggregates the number of molecules in a certain spatial region or cell and ignores the exact spatial information. A more suitable likelihood model for single-molecule resolved data are Poisson point processes, which directly model the coordinates of each molecule. Poisson point processes have already been successfully applied in GPFA with temporal data in neuroscience [11] and for cell typing in spatial data [25] but so far have not been considered in factor models for spatial transcriptomics data. Formally, an inhomogeneous spatial Poisson point process is characterized by a non-negative intensity function $\lambda : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$. A set of point coordinates then has the probability density

$$p(\{\mathbf{c}_n\}_{n=1,\dots,N}) = \exp\left(-\int \lambda(\mathbf{c})\,\mathrm{d}\mathbf{c}\right) \prod_{n=1}^{N} \lambda(\mathbf{c}_n). \tag{8}$$

Intuitively, this means that more points are expected in regions where λ is high, and vice versa.

3 The FISHFactor model

FISHFactor is a probabilistic factor model that combines (i) spatially aware inference of factors using Gaussian processes tailored to spatial transcriptomics data, (ii) interpretable factors and weights using non-negativity constraints, (iii) a likelihood model accounting for the nature of single-molecule data using inhomogeneous Poisson point processes and (iv) the ability to integrate and compare inferred patterns across data groups (e.g. cells or regions of space). Given a list of spatial molecule coordinates $\{\mathbf{c}_n^{dm}\}_{n=1,\dots,N_{dm}}$ with $\mathbf{c}_n^{dm} \in \mathbb{R}^2$ for a set of genes $d = 1, \dots, D$ and data groups $m = 1, \dots, M$, FISHFactor models the coordinates as realization of a spatial Poisson point process, where the gene- and group-wise intensity function λ_{dm} is given by a decomposition into group-specific factors and a weight matrix that is shared between groups. The probabilistic model of FISHFactor is defined as

$$p(\{\mathbf{c}_n^{dm}\}_{n=1,\dots,N_{dm}} \mid \lambda_{dm}) = \exp\left(-\int \mu_{dm}\lambda_{dm}(\mathbf{c})\,\mathrm{d}\mathbf{c}\right) \prod_{n=1}^{N_{dm}} \mu_{dm}\lambda_{dm}(\mathbf{c}_n^{dm}) \tag{9}$$

$$\lambda_{dm}(\mathbf{c}) = \sum_{k=1}^{K} w_{dk} z_{mk}(\mathbf{c}) \tag{10}$$

$$w_{dk} = \mathrm{softplus}(q_{dk}) \tag{11}$$

$$z_{mk}(\mathbf{c}) = \mathrm{softplus}(f_{mk}(\mathbf{c})) \tag{12}$$

$$q_{dk} \sim \mathcal{N}(0, 1) \tag{13}$$

$$f_{mk} \sim \mathrm{GP}(0, s_{mk}\kappa_{mk}). \tag{14}$$

To determine the relevant region of space where the integral in Eq. (9) is evaluated, the method determines kernel density estimates on all data points of a group and uses a threshold to create group-specific masks. The average intensity μ_{dm} per gene and group is computed as the number of molecules N_{dm} divided by the mask area of the corresponding group. This value is then multiplied with the intensity function λ_{md} in Eq. (9) to account for differences in the overall expression intensities of genes. The GP prior in Eq. (14) uses a Matérn kernel κ_{mk} with fixed smoothness parameter $\nu = 1.5$ and learnable lengthscales ℓ_{mk}. Factor- and group-specific scale parameters s_{mk} are included to enable factors to have different activity levels in each group. For example, using cells as groups in the model, this allows to dissect cell-to-cell heterogeneity.

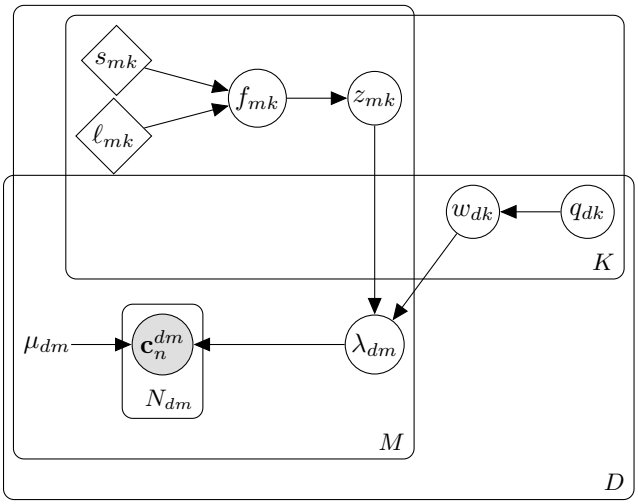

Fig. 2. Graphical model of FISHFactor with K factors, D genes, M sample groups (e. g. cells) and N_{dm} molecules per gene and sample group. Gray nodes indicate observed variables, white nodes latent variables and rhombuses learnable parameters.

Implementation FISHFactor is implemented in the probabilistic programming language Pyro [4] and the low-level Pyro interface of GPyTorch [15] for inference with sparse variational Gaussian processes. Stochastic variational inference is done with a clipped adam optimizer, optionally with learning rate decrease. Optimization of the evidence lower bound is stopped as soon as it does not increase by a specified value within a specified number of iterations. If multiple groups are used, the stopping criterion is determined for every group separately and needs to be fulfilled for all groups. It is possible to run the model on either cpu or gpu.

4 Results

4.1 Validation and benchmarking on simulated data

First, we validated FISHFactor on simulated datasets comprising a single cell ($M = 1$) and compared its performance to the most closely related factor models PNMF and NSF.

We simulated data from the generative model of FISHFactor, employing additional sparsity in the gene weights and fixed latent factors that mimicked spatial expression patterns observed in real data. In particular, we used 10 sets of 3 images with 100 pixels side length as latent factors, resembling a nucleus, a cytoplasm and one arbitrary shape (Fig. 3c). For each set of factors, we simulated a random non-negative weight matrix for 60 genes from a log-normal distribution with mean 0 and scale 1 and multiplied each entry with a Bernoulli random variable with parameters 0.6, 0.8 and 0.7 to induce factor-specific sparsity. The intensity function λ_d of the spatial Poisson point process of gene d was then obtained from the model (Eq. (10)). To generate datasets of different overall molecule abundances, we used 5 different intensity scales $\mu = (20, 50, 80, 100, 150)$, which were identical for all genes, resulting in a total of 5×10 scenarios. Based on λ_d, molecule coordinates were simulated from the point process by thinning [18]. As the intensity function was only available on a discrete grid, we used the closest grid point for a continuous evaluation.

While FISHFactor can be applied directly to molecule coordinates, PNMF and NSF require data aggregation in spatial bins. To evaluate the sensitivity to the selected bin size, we used different resolutions of 5,

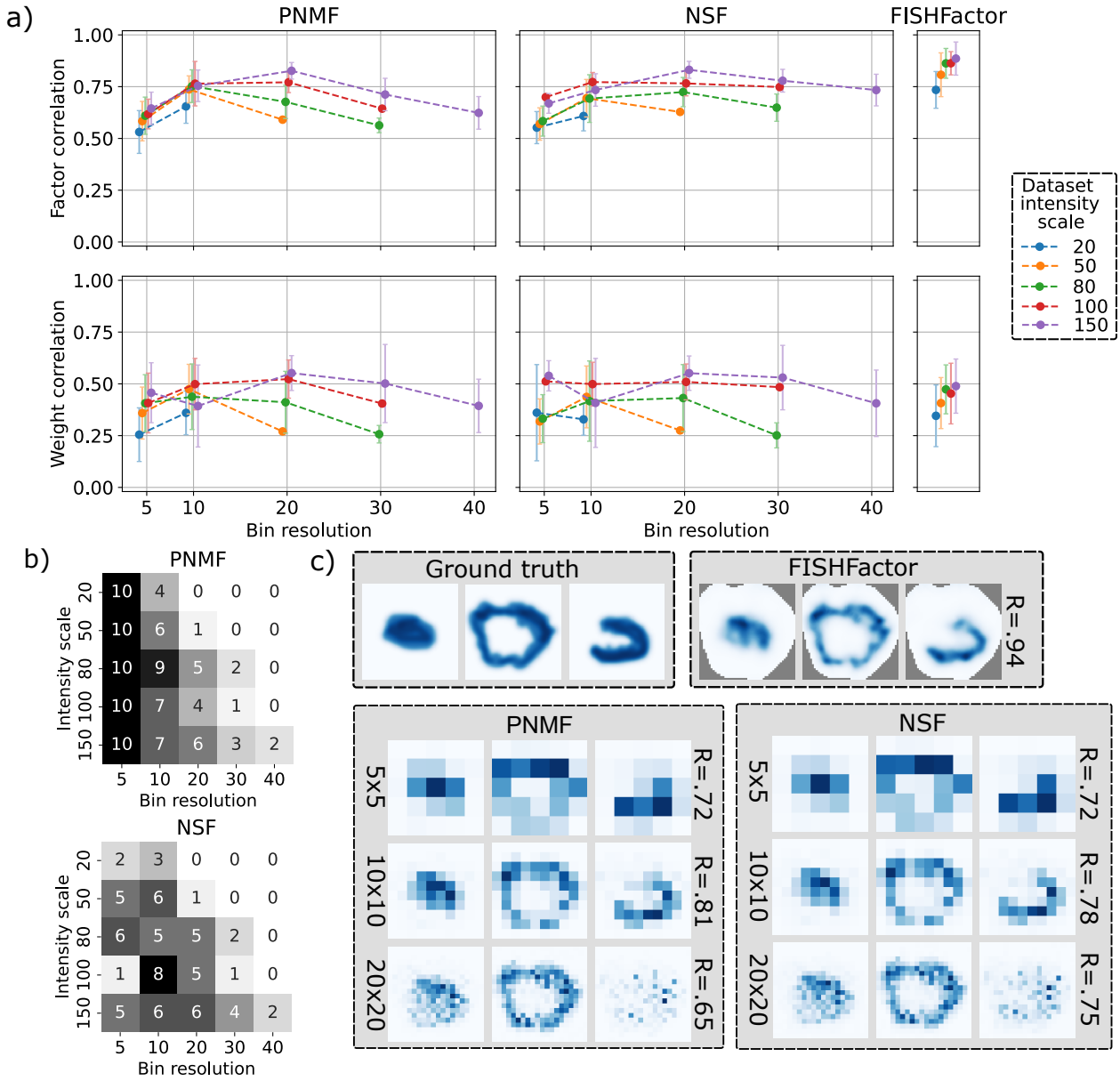

Fig. 3. Reconstruction of factors and weights with PNMF, NSF and FISHFactor in 10 simulated datasets with 5 different simulation intensity scales. **a)** Pearson correlation between inferred and ground truth values for the factors (top) and weights (bottom), *without* PNMF and NSF models that did not converge. Shown are average correlations across experiments. Error bars correspond to plus or minus one standard deviation. Jitter was added on the x-axis to improve visibility. **b)** Number of datasets out of 10 per intensity scale that PNMF and NSF converged on (without restarts). FISHFactor converged on all datasets, consistently achieving factor correlations > 0.6 (not shown). **c)** Ground truth and inferred factors in one experiment with dataset intensity scale $\mu = 80$, where PNMF and NSF only converged up to a bin resolution of 20×20. The Pearson correlation with the ground truth is stated for each factor set. This illustrates that in the case of convergence PNMF and NSF yield more noisy but otherwise comparable results to FISHFactor.

10, 20, 30 or 40 bins per spatial dimension. We then assessed the performance of all models in terms of their ability to recover the true simulated factor values and weights. As the resolution of inferred factors in PNMF and NSF is determined by the respective bin resolution, we first upsampled the factors to the simulated resolution by nearest-neighbor interpolation in order to be able to compute correlations.

Overall, our results indicate that FISHFactor infers accurate estimates of the true simulated latent variables, achieving as high or higher correlation values than PNMF or NSF when selecting the optimal binning resolution (Fig. 3a). In particular the inferred factors were more accurate in FISHFactor, showing a considerably higher degree of smoothness than PNMF and NSF (Fig. 3c). Moreover, we observed that PNMF and NSF run into numerical instabilities that result in aborted training for a substantial number of simulated settings (Fig. 3b, without restarts), especially when the intensity scale was low or the bin resolution was high, both leading to very sparse data matrices. Notably, similar problems were also observed when using the implementation of FISHFactor in combination with a Poisson likelihood, suggesting that they are intrinsic to the application of such models to sparse binned data. While reducing the bin resolution can help to mitigate these numerical issues, this results in a reduction of the spatial resolution. In contrast, the FISHFactor model achieved stable results on all datasets, while providing the full spatial resolution.

4.2 Application to cultured mouse embryonic fibroblasts from seqFISH+

We applied FISHFactor to a dataset comprising seqFISH+ measurements on a set of cultured mouse embryonic fibroblasts (NIH/3T3) [12]. Following the original publication, we focused on a set of 60 genes that were shown to be specifically expressed in three major subcellular spatial groups: protrusions, cytoplasm and nucleus/perinucleus.

Initially, we used these data to investigate the robustness of FISHFactor against data subsampling. We selected 20 cells at random and applied FISHFactor, PNMF and NSF to each of the cells individually (PNMF and NSF do not support multiple groups), with prior binning where necessary. PNMF and NSF only converged for a small set of of cells and bin widths (PNMF: 10/20 with 5x5 bins, 3/20 with 10x10 bins and 0/20 for higher resolution, NSF: 3/20 with 5x5 bins, 4/20 with 10x10 bins and 0/20 for higher resolution, all without restarts). Consistent with the results on simulated data, this implies that in order to obtain meaningful results using PNMF and NSF, the spatial binning resolution needs to be highly coarse-grained. Therefore we instead considered principal component analysis (PCA) on binned data as a comparison partner for FISHFactor, which always converged. Again we used the same bin resolutions as before and 3 factors or principal components. We evaluated the robustness of PCA and FISHFactor with respect to data subsampling by using the inferred factors and weights from the full datasets as a baseline and assessing their correlation to the values inferred from the subsampled data. This analysis revealed that FISHFactor was generally more stable than PCA (Fig. 4), in particular regarding the factors. As expected, PCA factors and weights could be better reconstructed when selecting low binning resolutions, where bins on average capture a larger number of points. However, this prevents finding high-resolution expression patterns.

Next, we applied FISHFactor to the same set of 20 cells jointly, in order to identify major subcellular expression patterns and cluster genes based on their localization within the cells. As a result the model yields cell-specific factors and one weight matrix that is shared across all cells. We found that based on the differences in the scale parameters of the factors, cells clustered into two distinct groups (Fig. 5a). In one group, the first factor was inactive, while in the other group it showed high activity at the cell borders. This suggests that this factor described protrusions, which may not be present in all cells. The other factors are active in all cells and outline the cytoplasm and nucleus of each cell (Fig. 5a). To identify spatially co-occurring groups of genes, we performed hierarchical clustering on the shared weight matrix and found three major clusters with high weights mostly on a single factor (Fig. 5b). When comparing the genes in the clusters to their assignment to spatial groups found in [12], we recovered an almost perfect agreement of our clusters with the known group labels (label colors in Fig. 5b). Notably, a clustering of genes and cells at the cellular level cannot reveal the distinct spatial groups of genes (Fig. 5c), which only become apparent when considering their subcellular localization.

Finally, we used the dataset to study the scalability of the method, finding an approximately linear computational complexity in the number of molecules per cell (Fig. 6a) and number of cells (Fig. 6b), demonstrating the feasibility of applying the method to larger data sets. The memory requirement has linear scaling in the number of data points per cell and in the number of factors, but remains approximately

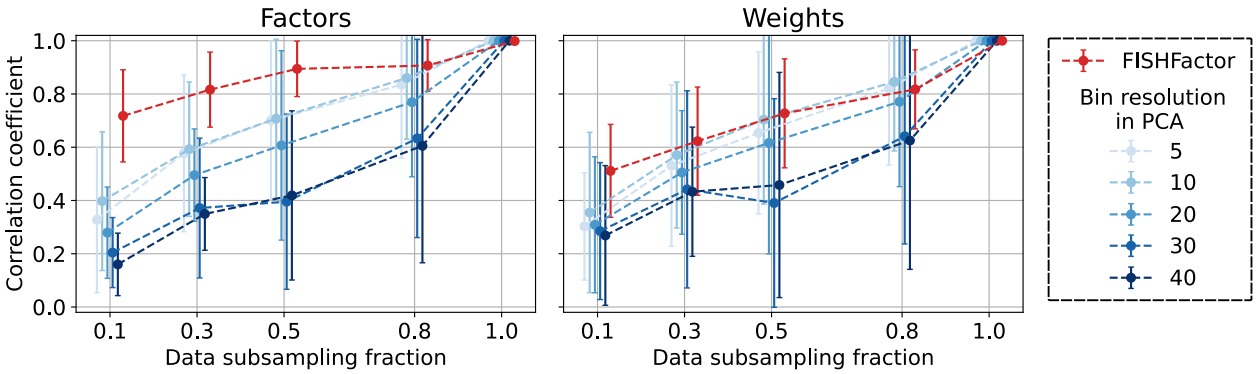

Fig. 4. Robustness of FISHFactor and PCA against data subsampling. FISHFactor (red) and PCA with varying bin sizes (blue colors) were applied to data from 60 genes in one out of 20 NIH/3T3 cells from [12] after subsampling various fractions of the data (x-axis). As a baseline for every model and bin resolution the corresponding inferred values on the full data for the respective cell were used. The lines show the average Pearson correlation of the inferred values to this baseline across the 20 cells with errorbars representing plus or minus one standard deviation. The points are slightly jittered on the x-axis for better visibility.

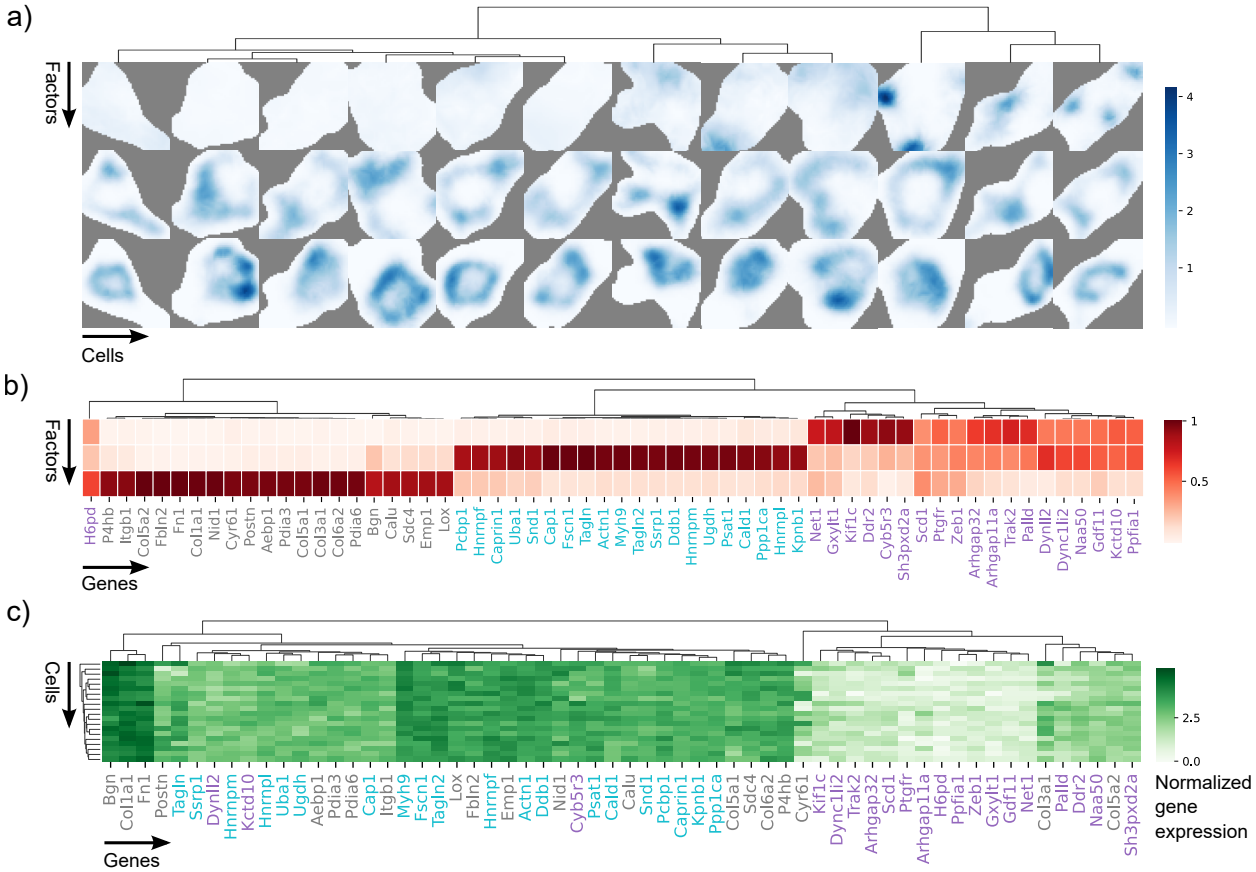

Fig. 5. Application of FISHFactor to coordinates of 60 genes in 20 NIH/3T3 cells from [12]. **a)** Inferred factors for the first 12 cells, with hierarchical clustering of the shown cells based on their scale parameters s_{mk}. **b)** Gene clustering based on inferred weights, with gene labels colored according to the spatial groups identified in [12]. **c)** Gene clustering at the cellular level based on normalized and log transformed gene counts per cell. Known spatial clusters of genes are less apparent compared to b).

constant in the number of groups. The peak memory requirement in the application to 20 cells was around 20 GB.

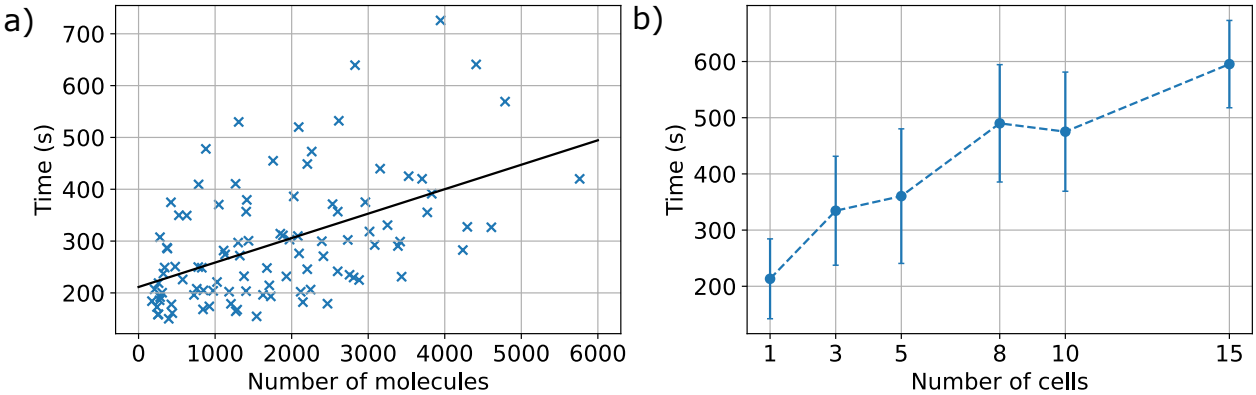

Fig. 6. Scalability of FISHFactor. **a)** Convergence time on individual NIH/3T3 cells ($M = 1$) for varying total number of molecules per cell. Points are based on the subsampled data as in Fig. 4 with stopping after 2000 epochs without improvement. **b)** Convergence time for different numbers of randomly selected NIH/3T3 cells ($M \geq 1$, x-axis), with stopping after 1000 epochs without improvement. Shown is the average convergence time across 10 random sets of cells of the given size, errorbars represent plus or minus one standard deviation.

5 Discussion

With FISHFactor we presented a spatial non-negative factor model for single-molecule resolved data and showed that using a Poisson point process likelihood is superior to applying existing methods to binned data, both in terms of numerical stability and the accuracy of the inferred variables. We found that FISHFactor is relatively robust to data subsampling, which makes it useful for data with sparsely expressed genes. FISHFactor could furthermore accurately identify major subcellular expression patterns and known spatial clusters of genes in a population of cells by jointly inferring a shared weight matrix and cell-specific factors. Future work could further increase scalability of the method, include an automatic choice of relevant genes or aim to reduce the sensitivity to the cell segmentation quality.

Code availability

Our code and experiments with used hyperparameters are freely available under https://github.com/bioFAM/FISHFactor.

Acknowledgments

The authors thank William Townes for his support in using PNMF and NSF, the Pyro and GPyTorch teams and communities for assistance in implementing FISHFactor, Vitalii Kleshchevnikov for helpful discussions on probabilistic programming and Jeongbin Park for sharing his expertise on spatial data.

B.V. was funded by the BMBF (COMPLS project MOFA no. 031L0171B). The Stegle research group was further supported by core funding from EMBL, the German Cancer Research Center and the European Commission (ERC project DECODE, 810296).

Competing interests

The authors have no competing interests.

Contributions

B.V., O.S. conceived the project.
F.W. implemented the model.
F.W. conducted the experiments.
F.W. generated the figures with input from B.V., O.S.
B.V., F.W. wrote the manuscript with input from O.S.
B.V., O.S. supervised the project.

References

1. Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., and Stegle, O. MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology 21*, 1 (2020), 1–17.
2. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. Multi-Omics Factor Analysis - a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology 14*, 6 (2018), e8124.
3. Berglund, E., Maaskola, J., Schultz, N., Friedrich, S., Marklund, M., Bergenstråhle, J., Tarish, F., Tanoglidi, A., Vickovic, S., Larsson, L., Salmén, F., Ogris, C., Wallenborg, K., Lagergren, J., Ståhl, P., Sonnhammer, E., Helleday, T., and Lundeberg, J. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nature Communications 9*, 1 (2018).
4. Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research 20*, 28 (2019), 1–6.
5. Brunet, J.-P., Tamayo, P., Golub, T. R., and Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U. S. A. 101*, 12 (2004), 4164–4169.
6. Buxbaum, A. R., Haimovich, G., and Singer, R. H. In the right place at the right time: Visualizing and understanding mRNA localization. *Nature Reviews Molecular Cell Biology 16*, 2 (2015), 95–109.
7. Byron, M. Y., Cunningham, J. P., Santhanam, G., Ryu, S. I., Shenoy, K. V., and Sahani, M. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In *Advances in neural information processing systems* (2009), pp. 1881–1888.
8. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science 348*, 6233 (2015), 1360–1363.
9. Chen, X., Sun, Y. C., Church, G. M., Lee, J. H., and Zador, A. M. Efficient in situ barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Research 46*, 4 (2018), 1–10.
10. Codeluppi, S., Borm, L. E., Zeisel, A., La Manno, G., van Lunteren, J. A., Svensson, C. I., and Linnarsson, S. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature Methods 15*, 11 (2018), 932–935.
11. Duncker, L., and Sahani, M. Temporal alignment and latent gaussian process factor inference in population spike trains. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2018), NIPS'18, Curran Associates Inc., p. 10466–10476.
12. Eng, C. H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G. C., and Cai, L. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature 568*, 7751 (2019), 235–239.
13. Eng, C. H. L., Shah, S., Thomassie, J., and Cai, L. Profiling the transcriptome with RNA SPOTs. *Nature Methods 14*, 12 (2017), 1153–1155.
14. Femino, A. M., Fay, F. S., Fogarty, K., and Singer, R. H. Visualization of Single RNA Transcripts in Situ. *Science 280*, 5363 (1998), 585–590.
15. Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems* (2018), S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc.
16. Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., and Nilsson, M. In situ sequencing for RNA analysis in preserved tissue and cells. *Nature Methods 10*, 9 (2013), 857–860.
17. Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Ferrante, T. C., Terry, R., Turczyk, B. M., Yang, J. L., Lee, H. S., Aach, J., Zhang, K., and Church, G. M. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nature Protocols 10*, 3 (2015), 442–458.
18. Lewis, P. A. W., and Shedler, G. S. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly 26*, 3 (1979), 403–413.
19. Littman, R., Hemminger, Z., Foreman, R., Arneson, D., Zhang, G., Gómez-Pinilla, F., Yang, X., and Wollman, R. Joint cell segmentation and cell type annotation for spatial transcriptomics. *Molecular Systems Biology 17*, 6 (2021), 1–15.
20. Lubeck, E., and Cai, L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nature Methods 9*, 7 (2012), 743–748.
21. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M., and Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods 11*, 4 (2014), 360–361.
22. Lyubimova, A., Itzkovitz, S., Junker, J. P., Fan, Z. P., Wu, X., and Van Oudenaarden, A. Single-molecule mRNA detection and counting in mammalian tissue. *Nature Protocols 8*, 9 (2013), 1743–1758.

23. Park, J., Choi, W., Tiesmeyer, S., Long, B., Borm, L. E., Garren, E., Nguyen, T. N., Tasic, B., Codeluppi, S., Graf, T., Schlesner, M., Stegle, O., Eils, R., and Ishaque, N. Cell segmentation-free inference of cell types from in situ transcriptomics data. *Nature Communications 12*, 1 (2021), 1–13.
24. Partel, G., and Wählby, C. Spage2vec: Unsupervised representation of localized spatial gene expression signatures. *FEBS Journal 288*, 6 (2021), 1859–1870.
25. Qian, X., Harris, K. D., Hauling, T., Nicoloutsopoulos, D., Muñoz-Manchado, A. B., Skene, N., Hjerling-Leffler, J., and Nilsson, M. Probabilistic cell typing enables fine mapping of closely related cell types in situ. *Nature Methods 17*, 1 (2020), 101–106.
26. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A., and Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods 5*, 10 (2008), 877–879.
27. Rao, A., Barkley, D., França, G. S., and Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature 596*, 7871 (2021), 211–220.
28. Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., and Macosko, E. Z. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science 363*, 6434 (2019), 1463–1467.
29. Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., and Frisén, J. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science 353*, 6294 (2016), 78–82.
30. Stein-O'Brien, G. L., Arora, R., Culhane, A. C., Favorov, A. V., Garmire, L. X., Greene, C. S., Goff, L. A., Li, Y., Ngom, A., Ochs, M. F., et al. Enter the matrix: factorization uncovers knowledge from omics. *Trends in Genetics 34*, 10 (2018), 790–805.
31. Townes, F. W., and Engelhardt, B. E. Nonnegative spatial factorization. *arXiv preprint arXiv:2110.06122* (2021).
32. Townes, F. W., Hicks, S. C., Aryee, M. J., and Irizarry, R. A. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology 20*, 1 (2019), 1–16.
33. Velten, B., Braunger, J. M., Arnol, D., Argelaguet, R., and Stegle, O. Identifying temporal and spatial patterns of variation from multi-modal data using MEFISTO. *bioRxiv* (2020).
34. Wang, X., Allen, W. E., Wright, M. A., Sylwestrak, E. L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., Nolan, G. P., Bava, F. A., and Deisseroth, K. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science 361*, 6400 (2018).
35. Witten, D. M., Tibshirani, R., and Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics 10*, 3 (2009), 515–534.
36. Xia, C., Fan, J., Emanuel, G., Hao, J., and Zhuang, X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proceedings of the National Academy of Sciences of the United States of America 116*, 39 (2019), 19490–19499.