


Objective

> BiocManager::install("nnSVG")

Weber et al. (2023) proposed nnSVG, a scalable approach to identify spatially variable genes (SVGs) in spatially-resolved transcriptomics data based on nearest-neighbor Gaussian processes (NNGP). $\{y(s); s \in \mathbb{R}^k\}$

Methods

$\mathbf{y}_g = (y_{g1}, \dots, y_{gN})^\top$, normalized and transformed (*logcounts*) gene expression levels, for $g = 1, \dots, G$ across the finite set of N spatial locations $\mathbf{S} = (\mathbf{s}_1^\top, \dots, \mathbf{s}_N^\top)^\top$. $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{b} + \boldsymbol{\epsilon}$, $\mathbf{b} \sim \mathcal{N}_N(\mathbf{0}, \sigma_s^2 \cdot \mathbf{K}) \perp \boldsymbol{\epsilon} \sim \mathcal{N}_N(\mathbf{0}, \tau^2 \cdot \mathbf{I})$

Response NNGP Model	$\mathbf{y} \sim \mathcal{N}_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	(1)
SpatialDE (spatial differential expression):	$\boldsymbol{\Sigma} = \sigma_s^2 \cdot \mathbf{K} + \tau^2 \cdot \mathbf{I}$	(2)
 Squared Exponential Kernel Function	$K_{ij} = \exp\left\{-d_{ij}^{\frac{2}{\alpha}} / \frac{2}{\alpha} l^{\frac{2}{\alpha}}\right\}$ for $i, j \in \{1, \dots, N\}$	(3)
nnSVG: a scalable approximation to Eq. 2, Stan	$\tilde{\boldsymbol{\Sigma}}^{-1} = (\mathbf{I} - \mathbf{L})^\top \mathbf{D}^{-1} (\mathbf{I} - \mathbf{L})$ <small>sparse</small>	(4)

l : Length scale (bandwidth) hyperparameter. $l \nearrow$, $K_{ij} \nearrow$, Smoothness \nearrow

How rapidly does the covariance decay as a function of the Euclidean norm $d_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|$?

σ_s^2 : Spatial component of variance in gene expression.

Identified highly variable genes (HVGs) using $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ that assumes $\mathcal{H}_0: \sigma_s^2 = 0$. (5)

τ^2 : Nugget, the non-spatial component of variance in gene expression.

$\sigma_s^2 / (\sigma_s^2 + \tau^2)$ defines the *effect size* per gene (Svensson et al., 2018).

Separability between $\boldsymbol{\epsilon}$ and latent \mathbf{b} (resolution; Allen et al., 2023, p. 1785; Shang & Zhou, 2022, p. 12).

- Fitted a separate model for each gene and obtained maximum likelihood estimates for l , σ_s^2 , and τ^2 using the fast optimization algorithms in the “BRISC” (bootstrap for rapid inference on spatial covariances, $N > 100$) R package.
- Performed a likelihood ratio (LR) test comparing the fitted model using (4) against the baseline model (5).
- Ranked genes by the estimated LR statistics and calculated multiple-testing (Benjamini-Hochberg) adjusted approximate p -values for statistical significance per gene using an asymptotic χ^2_2 -distribution. p. 9

Highlights

“nnSVG” v1.5.3+

- Identified genes that vary in expression continuously across the entire tissue ($\boldsymbol{\mu} = \boldsymbol{\mu}_g \cdot \mathbf{1}$, mostly) or within *a priori* defined spatial domains (regions of the tissue slide corresponding to anatomical features or tissue types; Fig. 2) by including the spatial domains as covariates within the model where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$; See also Tbl. 1;
- Used flexible gene-specific estimates of l_g within the models (one model per gene; See Fig. 1b; $\phi = l_g^{-1}$); (unlike SPARK-X which uses pooled l for all g and may lose sensitivity to spatial patterns);
- Scaled the computational complexity and runtime linearly with N , i.e., $\mathcal{O}(Nm^3)$ flops, where m is the number of NNs, See Fig. 3; and, thus, applied to large transcriptome-wide datasets, e.g., $N > 1,000$, high-throughput; (unlike SpatialDE and SpatialDE2 which scale cubically and quadratically). By default, $m = 10 \ll N$.

Limitations $y_i \sim \text{Pois}(\lambda_i)$ or $B(n_i, p_i)$, $\ln \lambda_i$ or $\text{logit}(p_i) = \mathbf{x}_i^T \boldsymbol{\beta} + \boxed{b_i + \epsilon_i}$, $\mathbf{b} = (b_1, \dots, b_N)^T \sim \mathcal{N}_N(\mathbf{0}, \sigma_s^2 \cdot \mathbf{K}) \perp \epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau^2)$

SpatialDE, SpatialDE2, and nnSVG typically transform raw count data into **normalized** data before DE analysis, which can result in a potential loss of power due to failing in accounting for the mean-variance relationship that exists in raw counts (Guo et al., 2023, p. 4). **Error: zeros** or low-expressed/quality. *SCTransform*.

nnSVG scaled better with the number of spatial locations than SpatialDE but was slower on datasets with **many genes** (Chen et al., 2024, p. 11). nnSVG managed to achieve linear time inference but was still the second slowest method. SpatialDE overtakes nnSVG at 10,000 cells but is faster for fewer cells in part because it runs on a **GPU** (Fig. 5; Jones et al., 2023, p. 7-8).

nnSVG and SOMDE (**self-organizing map**) did not identify any significant peaks (spatially variable peaks [**SVPs**]), indicating their limitations in capturing spatial variability in this context (Li et al., 2023, p. 15).

Discussion

1. Nonstationary spatial modeling based on the Matérn kernel.

$$K_{ij} = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} d_{ij}}{l} \right)^\nu \cdot K_\nu \left(\frac{\sqrt{2\nu} d_{ij}}{l} \right),$$

where $K_\nu(\cdot)$ is the modified Bessel function of the second kind, e.g., $\nu \rightarrow +\infty$ and $\nu = 1/2$.

Different kernels make different assumptions about smoothness, periodicity, and other properties.

Stationarity - statistical properties such as mean and variance are constant over space

Isotropy - these properties are uniform in all directions

2. High power of nnSVG in detecting biologically informative genes? Benchmarking.

Performance compared to HVGs (“*scrn*”), *deviance residuals* from a binomial regression model (“*scry*”), and Moran’s *I* (a measure of spatial autocorrelation; “*Rfast2*”).

p. 10

$$\text{Global } I = \frac{N}{\sum_{i=1}^N \sum_{j=1}^N w_{ij}} \frac{\sum_{i=1}^N \sum_{j=1}^N w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^N (y_i - \bar{y})^2}, \text{ where } w_{ij} \text{ are the spatial weights } (w_{ii} = 0).$$

3. The key idea of the NNGP: to model the data, we often don’t need to model the influence between all pairs of points; instead we can just consider a few of each point’s “neighbors” (Jones, 2021; Vecchia, 1988).

$$p(\mathbf{y}) = p(y_1, \dots, y_N) = p(y_1) \cdot p(y_2 | y_1) \cdot \dots \cdot p(y_N | y_1, \dots, y_{N-1}) = p(y_1) \prod_{i=2}^N p(y_i | \{y_j\}_{j=1}^{i-1}) \approx \prod_{i=1}^N p(y_i | \mathbf{y}_{N(s_i)})$$

Allen, C., Chang, Y., Neelon, B., Chang, W., Kim, H. J., Li, Z., Ma, Q., & Chung, D. (2023). A Bayesian multivariate mixture model for high throughput spatial transcriptomics. *Biometrics*, 79, 1775-1787. DOI: 10.1111/biom.13727

Chen, C., Kim, H. J., & Yang, P. (2024). Evaluating spatially variable gene detection methods for spatial transcriptomics data. *Genome Biology*, 25, 1-21. DOI: 10.1186/s13059-023-03145-y

Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111, 800-812. DOI: 10.1080/01621459.2015.1044091

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., & Banerjee, S. (2019). Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *Journal of Computational and Graphical Statistics*, 28, 401-414. DOI: 10.1080/10618600.2018.1537924

Guo, X., Ning, J., Chen, Y., Liu, G., Zhao, L., Fan, Y., & Sun, S. (2023). Recent advances in differential expression analysis for single-cell RNA-seq and spatially resolved transcriptomic studies. *Briefings in Functional Genomics*, 1-15. DOI: 10.1093/bfpg/eld011

Jones, A. (2021a). *Nearest Neighbor Gaussian Processes*. <https://andrewcharlesjones.github.io/journal/nnnp.html>

Jones, A. (2021b). *The Matérn Class of Covariance Functions*. <https://andrewcharlesjones.github.io/journal/matern-kernels.html>

Jones, D. C., Danaher, P., Kim, Y., Beechem, J. M., Gottardo, R., & Newell, E. W. (2023). An information theoretic approach to detecting spatially varying genes. *Cell Reports Methods*, 3, 1-12. DOI:10.1016/j.crmeth.2023.100507

Li, Z., Patel, Z. M., Song, D., Yan, G., Li, J. J., & Pinello, L. (2023). Benchmarking computational methods to identify spatially variable genes and peaks. *bioRxiv*, 1-32. DOI: 10.1101/2023.12.02.569717

Shang, L., & Zhou, X. (2022). Spatially aware dimension reduction for spatial transcriptomics. *Nature Communications*, 13, 1-22. DOI: 10.1038/s41467-022-34879-1

Svensson, V., Teichmann, S. A., & Stegle, O. (2018). SpatialDE: Identification of spatially variable genes. *Nature Methods*, 15, 343-346. DOI: 10.1038/nmeth.4636

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 50, 297-312. DOI: 10.1111/j.2517-6161.1988.tb01729.x

Weber, L. M. (2022). *Unsupervised Analyses of Spatially-Resolved Transcriptomics Data with {nnSVG} and R/Bioconductor*. https://youtu.be/C-3YQ_Tel3k and https://youtu.be/x7_BLWfrJXQ

Weber, L. M., Saha, A., Datta, A., Hansen, K. D., & Hicks, S. C. (2023). nnSVG for the scalable identification of spatially variable genes using nearest-neighbor Gaussian processes. *Nature Communications*, 14, 1-12. DOI: 10.1038/s41467-023-39748-z

Xia, B. (2020). <https://twitter.com/BoXia7/status/1261464021322137600>

Zhang, L. (2018). *Nearest Neighbor Gaussian Processes (NNGP) Based Models in Stan*. <https://mc-stan.org/users/documentation/case-studies/nnnp.html>

Future Work: (1) genes not independent; (2) Bayes factors