Contributed Discussion

Andrea Sottosanti*, Davide Risso†, and Cristian Castiglione‡

We first want to congratulate the authors for the impressive work, which consists in a non-parametric method for estimating the spatial dependence structure of both stationary and non-stationary fields. For convenience, we refer to their method as `NPVecchia`. We are convinced that their work represents a notable advance in spatial statistics and brings a powerful and flexible analysis tool into many real-data problems. Nevertheless, to better understand which domains of application could benefit of such innovation, some open issues should be further discussed.

Due to the absence of an explicit model for the underlying continuous spatial field, we are concerned about the possibility of using `NPVecchia` for performing common operations in spatial data analysis, such as spatial interpolation and prediction. This limitation would restrict the range of applications to contexts of in-sample analysis, where prediction over unobserved sites is not the main goal. Moreover, geo-spatial data are frequently observed upon a collection of sites distributed extremely irregularly over the space and so the distances between different points can vary considerably. On the contrary, Kidd and Katzfuss (2021) implicitly assume an almost uniform distribution of the observed locations.

It therefore appears that `NPVecchia` is appropriate for applications whose data are characterized by roughly equally-spaced sites, with many observations per site, and whose main goal is not the prediction over unobserved locations. Based on these considerations, we believe that `NPVecchia` would be ideal for modelling the data processed by a new, groundbreaking class of technologies for DNA sequencing, called *spatial transcriptomics* (*s.t.*). For the substantial contributions that *s.t.* is carrying into the study of biological organisms, it was named *method of the year 2020* (Marx, 2021). The 10X Visium sequencing platform (Rao et al., 2020), one among several *s.t.* technologies, collects the cells of a tissue sample through a grid of equally spaced spots on the surface of a chip. The transcriptome is sequenced within each spot, where a few neighbour cells are collected. The output of the procedure is the expression of thousands of genes within each spot, together with the coordinates of the spots. Figure 1 is an example of a human prostate cancer tissue sample processed with 10X Visium.[1]

The growing popularity of *s.t.* has allowed researchers to identify the so-called *spatially expressed* (*s.e.*) genes, i.e., genes that show spatial variation patterns across the tissue. Discovering and comprehending the functions of *s.e.* genes is of great scientific interest and might lead to new insights and discoveries of specific biological processes.

*Department of Statistical Sciences, University of Padova, via C. Battisti 241, Padova, Italy, andrea.sottosanti@unipd.it

†Department of Statistical Sciences, University of Padova, via C. Battisti 241, Padova, Italy, davide.risso@unipd.it

‡Department of Statistical Sciences, University of Padova, via C. Battisti 241, Padova, Italy, cristian.castiglione@phd.unipd.it

[1] `https://www.10xgenomics.com/resources/datasets`.

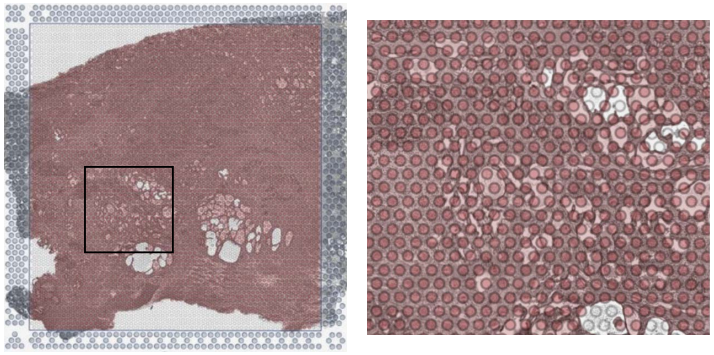

Figure 1: Left: human prostate cancer sample analysed with the 10X Visium platform. The tissue covers a total of 4,371 spots. Right: detail of the left figure corresponding to the black square. The spots on the chip are visible as circles over the whole surface.

Svensson et al. (2018) and Sun et al. (2020) tackle this research question as a statistical hypothesis testing problem where, for each gene, the presence of spatial variation patterns is tested. Both these methods assume a stationary field and express the dependency across the spots using parametric spatial correlation functions. To overcome these limitations and perform an accurate inferential process, we discuss a possible use of the idea of Kidd and Katzfuss (2021) into the analysis of *s.t.* experiments, with the aim of improving the discovery of *s.e.* genes.

Let $\mathbf{Y} = (\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(N)})^T$ be an $N \times n$ experiment matrix, where $\mathbf{y}^{(\ell)}$ is the expression of gene ℓ over the n observational sites (spots) with spatial coordinates $\mathbf{s}_1, \dots, \mathbf{s}_n$. We assume that the data have been centered and pre-processed in such a way that $y_i^{(\ell)} \in \mathbb{R}$ and the histogram of each $\mathbf{y}^{(\ell)}$ is approximately symmetric. Then, we assume the following model:

$$\mathbf{y}^{(\ell)} | \mathbf{f}^{(\ell)}, \lambda_\ell^2, \sigma_\varepsilon^2 \sim \mathcal{N}_n(\mathbf{f}^{(\ell)}, \lambda_\ell^2 \mathbf{I}_n + \sigma_\varepsilon^2 \mathbf{I}_n), \quad \mathbf{f}^{(\ell)} | \tau_\ell^2, \boldsymbol{\Sigma} \sim \mathcal{N}_n(\mathbf{0}_n, \tau_\ell^2 \boldsymbol{\Sigma}), \tag{1}$$

where $\mathbf{f}^{(\ell)}$ is the gene-specific spatial field with marginal variance τ_ℓ^2 and common covariance matrix $\boldsymbol{\Sigma}$, while λ_ℓ^2 and σ_ε^2 are the variances of idiosyncratic error terms. We assume the prior structure on the precision matrix $\boldsymbol{\Sigma}^{-1}$ proposed by Kidd and Katzfuss (2021), and non-informative priors for the variance parameters λ_ℓ^2 and σ_ε^2 as suggested by Gelman (2006). Last, taking inspiration from the recent literature on shrinkage priors and on the extraction of sparse signals (Bhadra et al., 2019), we propose to consider a prior model for τ_ℓ^2 that performs an aggressive shrinkage toward 0 if no spatial patterns arise, while leaving a high level of flexibility when the genes show a significant amount of spatial correlation. Within this framework, an interesting choice with optimal theoretical properties is the Horseshoe prior (Carvalho et al., 2010), corresponding to a hierarchical Half-Cauchy distribution on the standard deviation parameter τ_ℓ.

Formula (1) can be seen as a generalized, Bayesian version of the SpatialDE model proposed by Svensson et al. (2018), where all the unknown parameters, including the

spatial covariance matrix $\mathbf{\Sigma}$, are inferred directly from the data using, for example, a Gibbs sampling algorithm as described in Section 2.8 of Kidd and Katzfuss (2021). Thanks to the shrinkage imposed on τ_ℓ^2 through its a priori setup, the *s.e.* genes can be determined by evaluating the posterior distribution of $\delta_\ell = \tau_\ell^2/(\tau_\ell^2 + \lambda_\ell^2)$, that is the percentage of spatial variability specific of gene ℓ. For example, one may define an operating rule based on some threshold conditions, classifying as *s.e.* only those genes which have $\mathbb{P}(\delta_\ell \geq t|\mathbf{Y}) \geq p$ for t close to 0 and p close to 1.

Although we see a lot of promise in applying the work of Kidd and Katzfuss (2021) to the problem of identifying *s.e.* genes, it remains an open question whether irregularities on the edges and within the surface of tissues, as the one that appears in Figure 1 (left), could somehow affect the estimate of $\mathbf{\Sigma}$.

Several generalizations of the model in Formula (1) could be explored. First, it is often of clinical interest to evaluate biological processes common to a cohort of patients. Hence, the model could be extended to identify *s.e.* genes by simultaneously evaluating multiple tissue samples. Second, since *s.t.* raw data are highly variable, possibly zero-inflated counts, a Poisson or Negative Binomial extension could be considered, similarly to what has been done by Sun et al. (2020).

References

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2019). "Lasso meets horseshoe: a survey." *Statistical Science*, 34(3): 405–427. MR4017521. doi: https://doi.org/10.1214/19-STS700. 338

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 338

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515–533. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 338

Kidd, B. and Katzfuss, M. (2021). "Bayesian nonstationary and nonparametric covariance estimation for large spatial data." *Bayesian Analysis*. 337, 338, 339

Marx, V. (2021). "Method of the Year 2020: spatially resolved transcriptomics." *Nature Methods*, 18: 9–14. 337

Rao, N., Clark, S., and Habern, O. (2020). "Bridging genomics and tissue pathology." *Genetic Engineering & Biotechnology News*, 40(2): 50–51. 337

Sun, S., Zhu, J., and Zhou, X. (2020). "Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies." *Nature Methods*, 17(2): 193–200. 338, 339

Svensson, V., Teichmann, S. A., and Stegle, O. (2018). "SpatialDE: identification of spatially variable genes." *Nature Methods*, 15(5): 343–346. 338