Discussion of “Statistical disease mapping for heterogeneous neuroimaging studies”

Ben WU[1] and Jian KANG[2]*

[1]*Department of Mathematical Statistics, School of Statistics at Renmin University, Beijing, China*
[2]*Department of Biostatistics, University of Michigan, Ann Arbor, 48109, MI, U.S.A.*

Abstract: This is a discussion of Liu & Zhu (2021), which develops a novel statistical disease mapping framework for neuroimaging data analysis. *The Canadian Journal of Statistics* 49: 35–38; 2021 © 2021 Statistical Society of Canada

Résumé: C’est la discussion de Liu & Zhu (2021), qui propose un nouveau outil de cartographie pour les maladies basé sur des approches statistiques permettant l’analyse de données en neuroimagerie. *La revue canadienne de statistique* 49: 35–38; 2021 © 2021 Société statistique du Canada

We would like to congratulate the authors for their excellent work on developing a novel statistical disease mapping (SDM) framework that delineates imaging heterogeneity at both the individual and group levels. The proposed method is general and useful for different neuroimaging studies. The proposed SDM model-fitting procedure consists of two major components. The first component fits an individual-level image-on-scalar regression model based on a multivariate varying coefficient model (MVCM), where a hidden Markov random field model (HMRFM) estimates a set of voxel-wise latent diseased region indicators for each individual. The second component fits a spatial zero-inflated Poisson regression model (SZIPM) to characterize the disease map at the group level.

We thank the Editor for the opportunity to discuss this work. Our discussion will focus on the following four aspects: the Potts model for labelling diseased regions, alternative modelling strategies for the two model components, the computational complexity of the estimation method, and several possible future directions.

1. THE POTTS MODEL FOR LABELLING DISEASED REGIONS

In the HMRFM, for each individual, a Potts model is adopted to specify the joint distribution of the diseased region indicators across all voxels, where a spatial smoothness parameter and a definition of a neighbourhood both control the flexibility of estimated diseased regions in terms of region size and spatial distribution. In this work, the individual-specific diseased region indicators are assumed to have a common smoothness parameter across all individuals. As the authors have pointed out that “diseased regions can significantly vary across subjects and/or time in terms of their number, size, and location,” it is of interest to learn whether individual-specific smoothness parameters can increase flexibility in model fitting. It seems feasible to estimate

*Author to whom correspondence may be addressed.
E-mail: jiankang@umich.edu

individual-level smoothness parameters using a pseudo-likelihood approach. On a related note, it may be worth investigating the statistical efficiency of this pseudo-likelihood approach to estimation relative to the full-likelihood approach. In addition, one may also consider modelling individual-level smoothness parameters as random effects, that is,

$$p(\mathbf{b}_i|\tau_i) = \exp\left\{-U(\mathbf{b}_i)\tau_i - \log C(\tau_i)\right\}, \quad \tau_i \sim \text{Gamma}(a, b), \tag{1}$$

where $\mathbf{b}_i$ represents the voxel-wise disease indicators, τ_i represents the random smoothness parameter for individual i, and the definitions of $U(\cdot)$ and $C(\cdot)$ are as given for Equation (2) of the paper. Some relevant methods have been discussed for the problem of image reconstruction and segmentation (Storath et al., 2015), as well as for Bayesian hierarchical modelling (Song et al., 2020).

There are several alternative modelling strategies for detecting diseased regions. One simple strategy is to directly threshold the absolute value of the estimated spatially varying coefficients in the MVCM or the derived statistical parametric maps. For example, one may use a Z-statistic map to identify abnormal regions, as in other neuroimaging studies. One important issue is how to control the false discovery rate. Some related methods have been developed for the linear regression model. It is of interest to explore extensions to the MVCM. This method avoids modelling $\mathbf{b}_i$ in the MVCM.

To adopt a model-based approach, we may consider the soft-thresholded Gaussian process (Kang et al., 2018, STGP) for modelling the sparse and piecewise smooth, spatially varying functions in the paper. As a potential advantage, STGPs can select diseased regions and estimate voxel-specific effect sizes simultaneously, providing a more systematic modelling approach relative to the Potts model. The STGP may also be more flexible in modelling the spatial smoothness of the spatially varying functions as long as appropriate kernel functions are chosen.

2. ALTERNATIVE MODELLING STRATEGIES FOR BOTH MODEL COMPONENTS

To construct the group-level disease map, in the second component of the SDM framework, the estimates of the individual-level diseased region indicators $\mathbf{b}_i$ from the first component are collected as "raw data" to construct the "spatially-varying response variables" q_k and fit the disease regression SZIPM. Here, the spatial locations $\mathbf{s}_k$ are "predictors," and the spatially varying regression coefficients $\xi_\lambda(\cdot)$ and $\xi_\pi(\cdot)$ are used to represent the group-level disease map. We would like to point out that different distributional assumptions are made for the individual-level diseased region indicators $\mathbf{b}_i$ in the two model components.

In the first component, where $\mathbf{b}_i$ is assumed to follow a Potts model, the distribution of the estimator $\hat{\mathbf{b}}_i$ is completely determined by the data and this model assumption. Thus, the distribution of $q_k = \sum_{i=1}^n \hat{b}_i(\mathbf{s}_k)\zeta_i$ is also determined given the subgroup indicators $\zeta_i \in \{0, 1\}$. We wonder if this distribution is consistent with the SZIPM. What if we simply summarize the results from the first component? For example, for each voxel $\mathbf{s}_k$, we use the frequency (or the kernel-smoothed frequency) of subjects for which the voxel $\mathbf{s}_k$ is marked as being within a diseased region as an estimate of the group level probability of disease, that is,

$$f_n(\mathbf{s}_k) = \frac{q_k}{\sum_{i=1}^n \zeta_i} \text{ and } f_n(\mathbf{s}) = \sum_{k=1}^m K(\|\mathbf{s} - \mathbf{s}_k\|) f_n(\mathbf{s}_k), \tag{2}$$

where $K(\cdot)$ represents a kernel function. Can we define the group-level disease probability based on the limit of the frequencies, that is,

$$\Pr\left(\mathbf{s}_k \text{ belongs to the diseased region}\right) = \lim_{n\to\infty} f_n(\mathbf{s}_k)?$$

As an alternative modelling strategy, one may introduce the group-level disease indicators $\mathbf{b} = \{b(\mathbf{s}_k)\}$ to which the Potts model prior is assigned and specify the conditional probability mass of $b_i(\mathbf{s}_k)$ given $b(\mathbf{s}_k)$. For example, one simple choice is

$$\Pr\left(b_i(\mathbf{s}_k) = 1 | b(\mathbf{s}_k) = l\right) = \kappa_l(\mathbf{s}_k), \tag{3}$$

for $l = 0, 1$, where $\kappa_0(\cdot)$ $(\kappa_1(\cdot))$ represents the probability that the voxel $\mathbf{s}_k$ belongs to a diseased region for individual i given that $\mathbf{s}_k$ belongs to (does not belong to) a group-level diseased region. Under a Bayesian modelling framework, we can assign spatially dependent functional priors on $\kappa_l(\cdot)$, for example, the Gaussian process priors, and make posterior inferences on $\mathbf{b}$ and $\mathbf{b}_i$. Suppose we obtain the posterior samples of $\mathbf{b}_i$ and $\mathbf{b}$, denoted as $\{\mathbf{b}_i^h\}_{h=1,\ldots,H}$ and $\{\mathbf{b}^h\}_{h=1,\ldots,H}$, respectively. Then, the posterior probabilities of interest are given by

$$\widehat{\Pr}(\mathbf{s}_k \text{ belongs to the diseased region for individual } i) = \frac{1}{H}\sum_{h=1}^{H} I\{b_i^h(\mathbf{s}_k) = 1\}, \tag{4}$$

$$\widehat{\Pr}(\mathbf{s}_k \text{ belongs to the group level diseased region}) = \frac{1}{H}\sum_{h=1}^{H} I\{b^h(\mathbf{s}_k) = 1\}, \tag{5}$$

where $I(\cdot)$ is an indicator function.

On the other hand, in the second component, the author assumes that the group-level summation over $b_i(\mathbf{s}_k)$, that is, q_k, follows an SZIPM. In the alternative modelling strategy above, for the first component, can we specify a conditional distribution of $\mathbf{b}_i$ given q_k accordingly? For example, we may assume, with a specific zero-inflated probability π_k^0, that all of the $b_i(\mathbf{s}_k) = 0$ and, with probability $1 - \pi_k^0$, that each $b_i(\mathbf{s}_k)$ independently follows a Bernoulli distribution with probability π_k. Furthermore, we can link π_k and π_k^0 to q_k and impose spatial smoothness. In this model specification, the summation over $b_i(\mathbf{s}_k)$ well approximates the SZIPM as the Poisson distribution is a limiting case of the binomial distribution. However, it is unclear if this estimation procedure is still feasible from both the frequentist and Bayesian perspectives.

3. COMPUTATIONAL COMPLEXITY

The parameter estimation procedure includes three steps. The first step is weighted least squares (WLS) for the n_0 subjects considered normal controls. In a common high-dimensional neuroimaging problem, the number of voxels m is much larger than the number of individuals n or n_0, and the number of individuals is also usually larger than the number of covariates p. The number of imaging features J is moderate. Thus, the computational complexity of the WLS method is $O(Jmn_0)$. In the second step, an iterative algorithm is developed to estimate $\overline{\mathbf{B}}$ and $\mathbf{b}$. Denote by R the total number of iterations. From Equation (10) in the paper, the second step has a complexity of $O(RmnJ^2)$. The third step is an EM algorithm for SZIPM parameter estimation for which the complexity is negligible compared to the first two steps as the $\mathbf{b}_i$s have been summed over individuals. In summary, the total time complexity is of a linear order in the number of voxels and the number of individuals and is of a quadratic order in the number of imaging features. Thus, the computation can be scaled up to a large number of individuals for high-resolution images.

4. FUTURE DIRECTIONS

From this insightful work, some potential future directions can be pursued in mapping the heterogeneous diseased regions. First, it may be useful to combine the two model components with consistent model assumptions and theoretical guarantees. The key step is to model the group-level and individual-level diseased regions simultaneously. For example, we can borrow some ideas from factor analysis or independent component analysis and decompose $\mathbf{b}_i$ into group component(s) and individual component(s). Second, as we discussed before, Bayesian hierarchical models can be constructed to make posterior inferences on the parameters of interest, such as the probability that a specific voxel $\mathbf{s}_k$ belongs to the diseased region at the individual level. Bayesian models can be more flexible in specifying different activation region shapes and can be more accurate in quantifying the uncertainty of diseased region selection. A key challenge in developing Bayesian methods for SDM is an efficient posterior computation algorithm. Some existing, scalable MCMC algorithms and variational Bayesian methods can potentially be extended or modified for the proposed model. Third, it is also interesting to perform subgroup analysis to accommodate heterogeneity among individual images, especially when there is a lack of prior knowledge regarding group partitions. In particular, we may consider clustering individuals' brain activity patterns and/or associating brain activity with other covariates. This is a more complicated problem as we do not only summarize individual information into the group-level disease map but also determine individual group assignments for which a mixture model can be adopted. How to make a valid inference on the number of subgroups is also worth investigating.

BIBLIOGRAPHY

Kang, J., Reich, B. J., & Staicu, A.-M. (2018). Scalar-on-image regression via the soft-thresholded Gaussian process. *Biometrika*, 105, 165–184.

Liu, R. & Zhu, H. (2021). Statistical disease mapping for heterogeneous neuroimaging studies. *The Canadian Journal of Statistics*, 49, 10–34.

Song, Y., Zhou, X., Kang, J., Aung, M. T., Zhang, M., Zhao, W., Needham, B. L., Kardia, S. L., Liu, Y., Meeker, J. D., et al. (2020), "*Bayesian Hierarchical Models for High-dimensional Mediation Analysis with Coordinated Selection of Correlated Mediators*," arXiv preprint arXiv:2009.11409.

Storath, M., Weinmann, A., Frikel, J., & Unser, M. (2015). Joint image reconstruction and segmentation using the Potts model. *Inverse Problems*, 31, 025003.