Statistical Methodology for Computed Tomography Scans of the Lungs

Sarah M. Ryan, MS

PhD Candidate
Department of Biostatistics and Informatics
Colorado School of Public Health
University of Colorado Anschutz Medical Campus

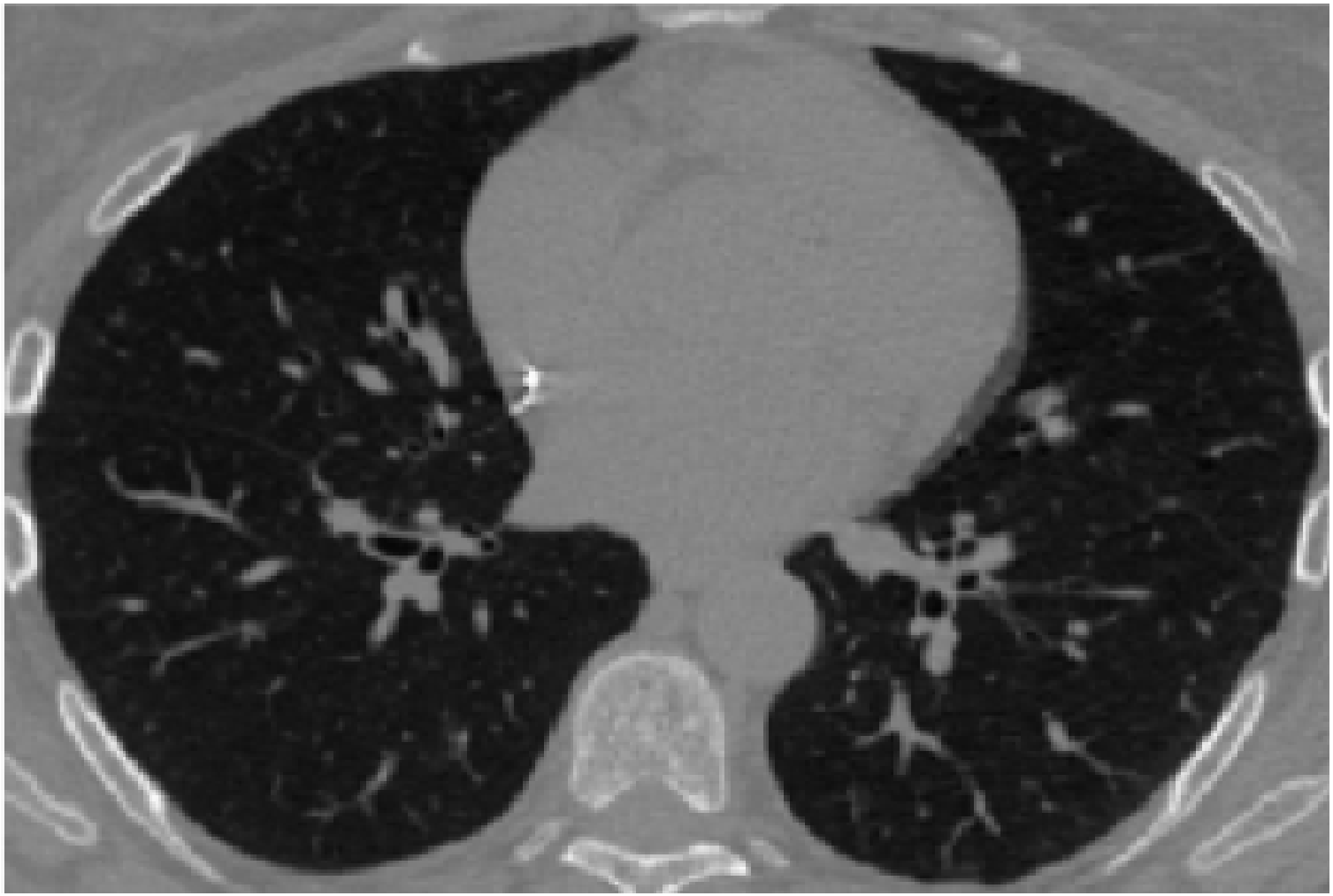
January 27, 2020

Computed Tomography (CT)

A computerized x-ray imaging procedure which generates cross-sectional images of the body that can be combined to form three-dimensional images

CT Scans of the Lung

Healthy Lung

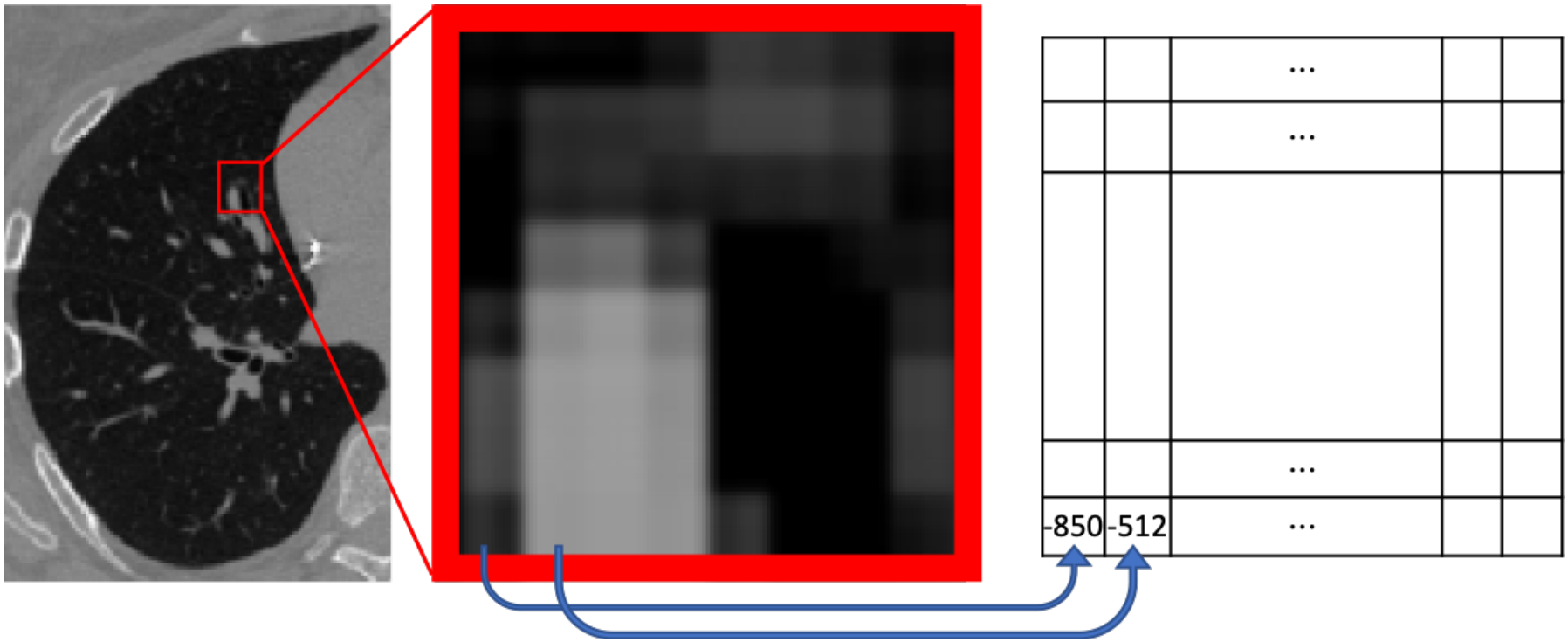


CT Scans of the Lung

Fibrotic Lung

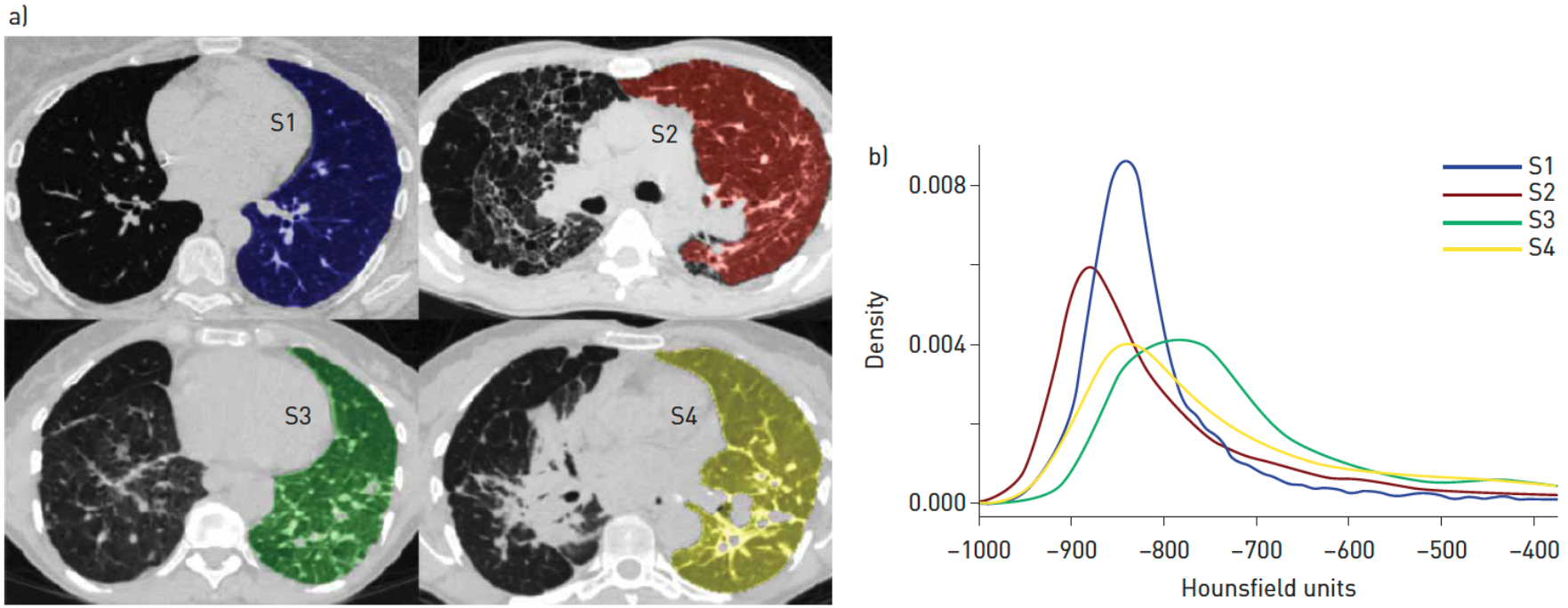
CT Scans are Data

Hounsfield Unit (HU): A measure of the radiodensity of a pixel from a CT scan



Existing Methods for Image Analysis

Radiomics: An emerging field in which large numbers of quantitative features are computed from medical images, providing a rapid, objective, and sensitive quantification of lung abnormalities [Ryan et al., 2019a]



colorado school of public health
UNIVERSITY OF COLORADO
COLORADO STATE UNIVERSITY
UNIVERSITY OF NORTHERN COLORADO

Clinical Question:

Where is disease commonly found in the lung?

Methodological Questions:

1. **How do we align spatial coordinates across scans when scans are different sizes and shapes?**
2. **After aligning spatial coordinates across scans, how do we identify significant areas of disease?**

colorado school of public health
UNIVERSITY OF COLORADO
COLORADO STATE UNIVERSITY
UNIVERSITY OF NORTHERN COLORADO

Clinical Question:

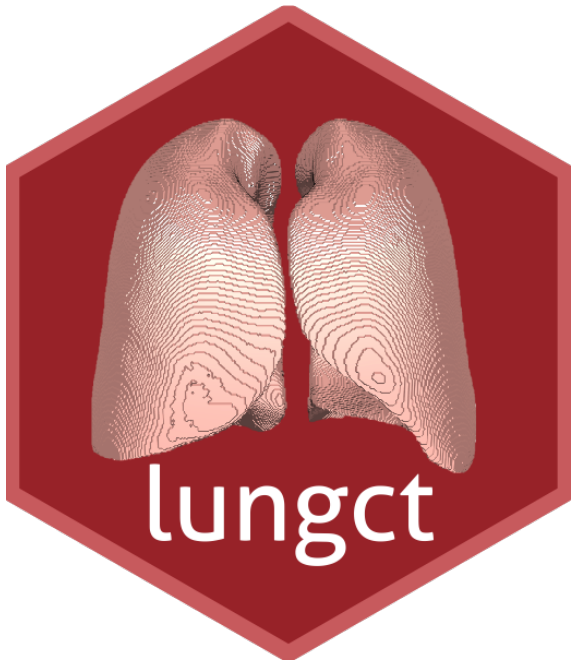
Where is disease commonly found in the lung?

Methodological Questions:

1. How do we align spatial coordinates across scans when scans are different sizes and shapes?
   - **Create a lung template**

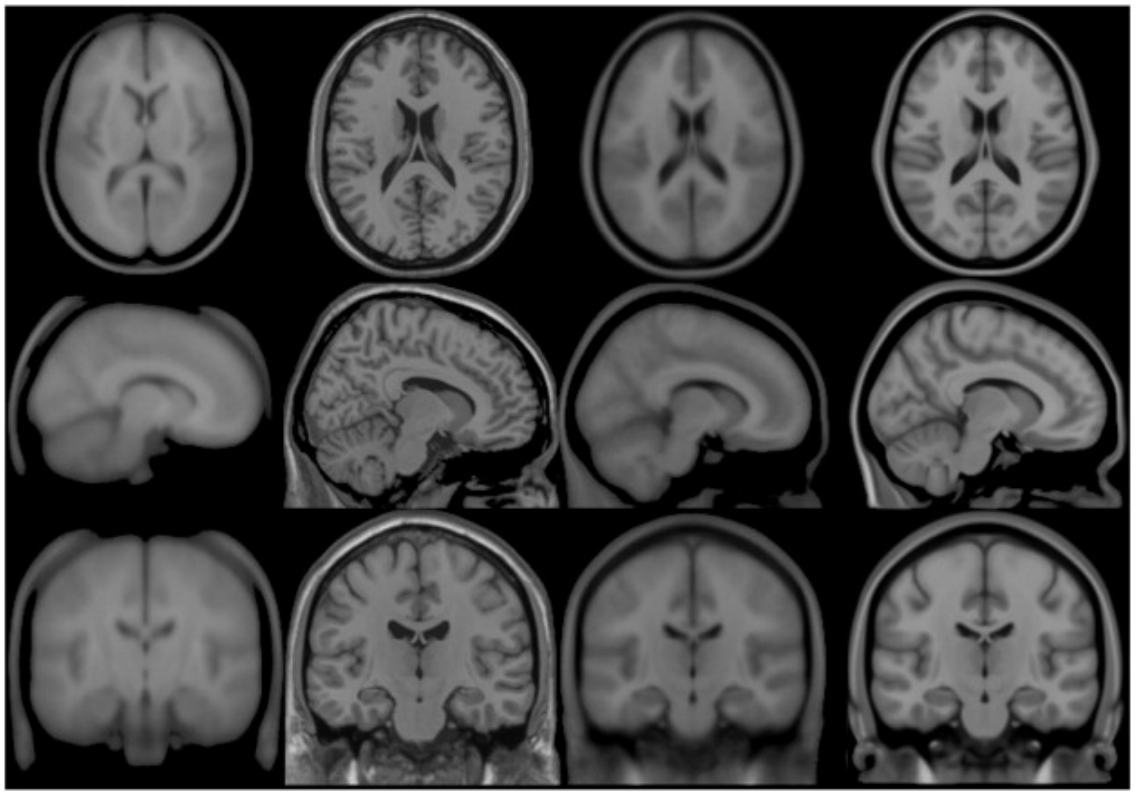
2. After aligning spatial coordinates across scans, how do we identify significant areas of disease?
   - **Develop a spatial model for whole-lung population-level inference**

Create a Lung Template *(with R!)*

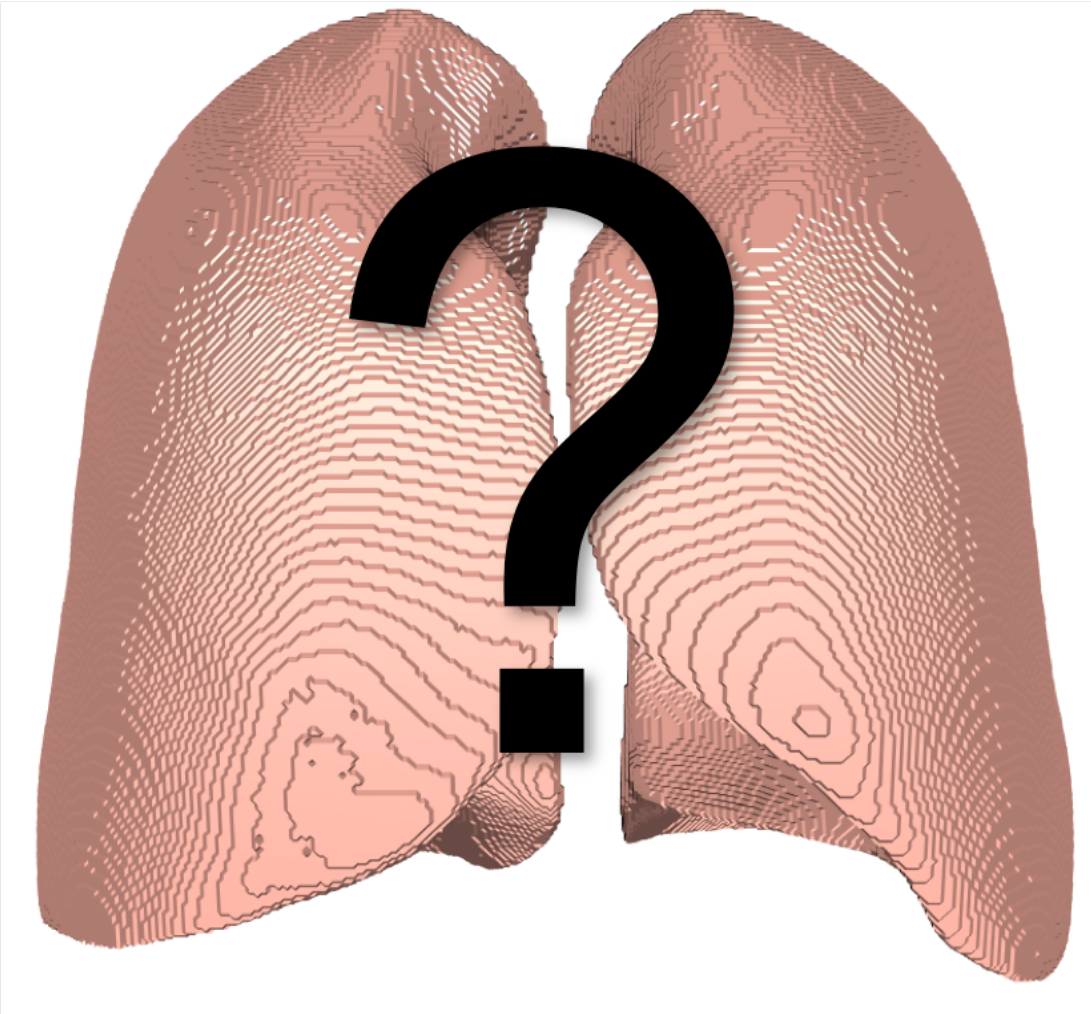


Templates

A **template** is a standardized 3D coordinate frame which allows researchers to combine data across subjects and/or studies [Evans et al., 2012].



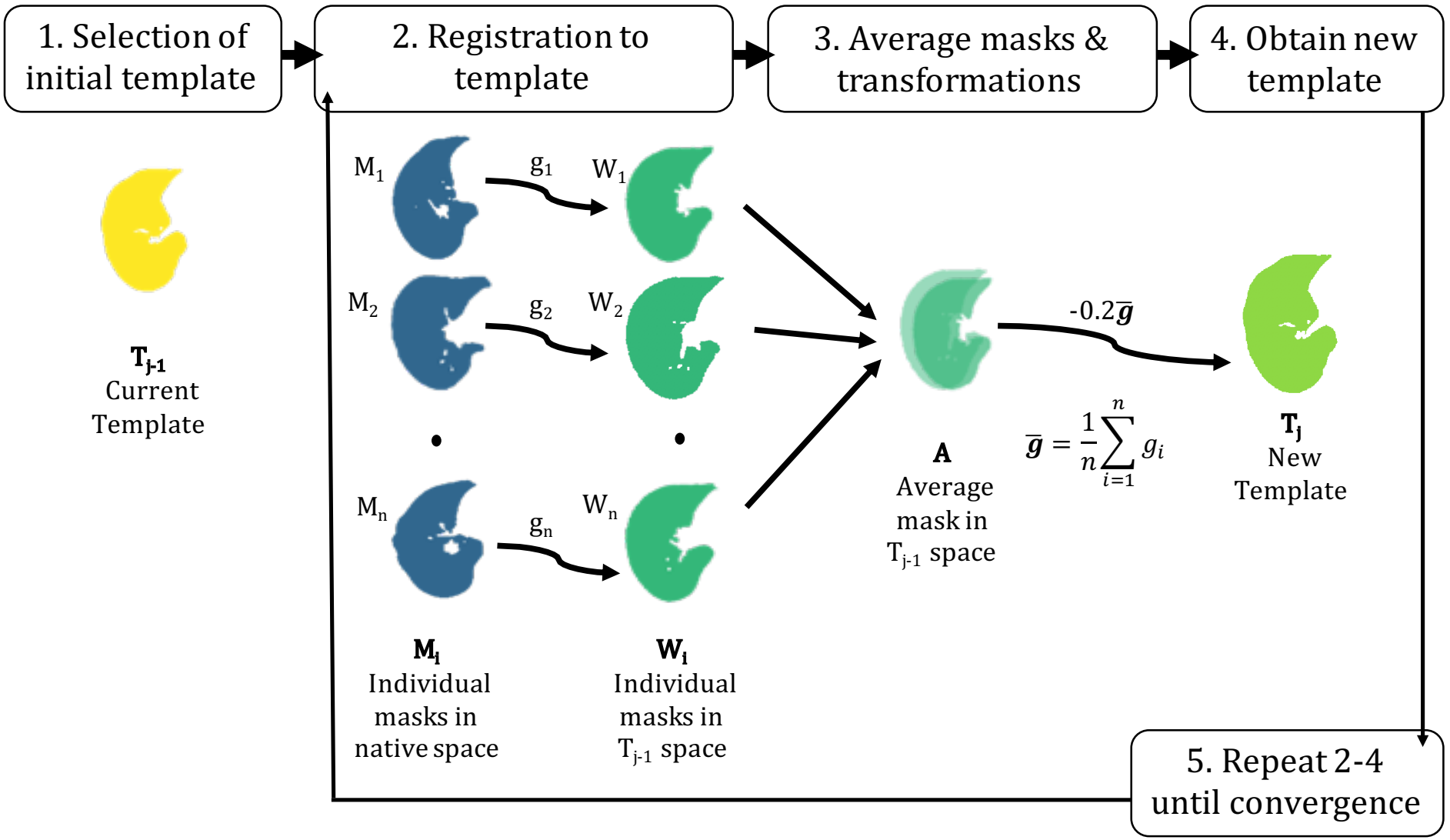
There is no standard lung!



Healthy Population

- **COPDGene** is a cross sectional prospective cohort of smokers with and without COPD, with a goal to perform epidemiological and genetic studies [Regan et al., 2011]
- A small (100) group of similar aged non-smokers without COPD were enrolled to provide a comparison point for CT scans in the aging lung
- We used $\mathbf{N = 62}$ adult subjects, with equal proportions of males/females, age range: 51-72 years.

Template Creation

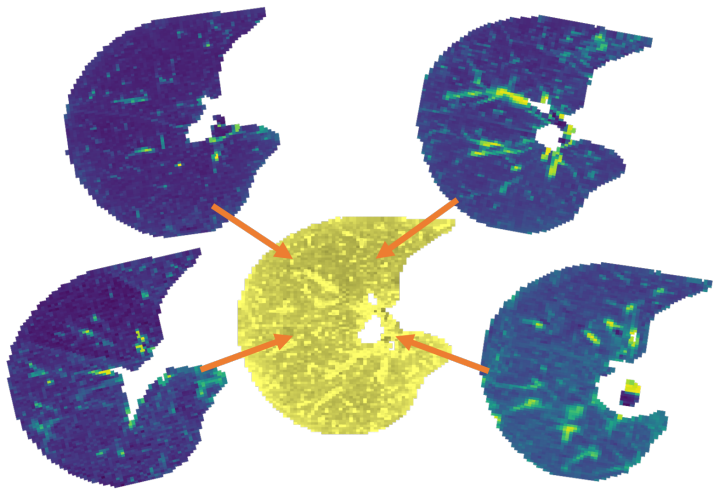


[Ryan et al., 2019b]

Notes on Registration

Registration is the procedure of transforming voxels from their original space into a common space

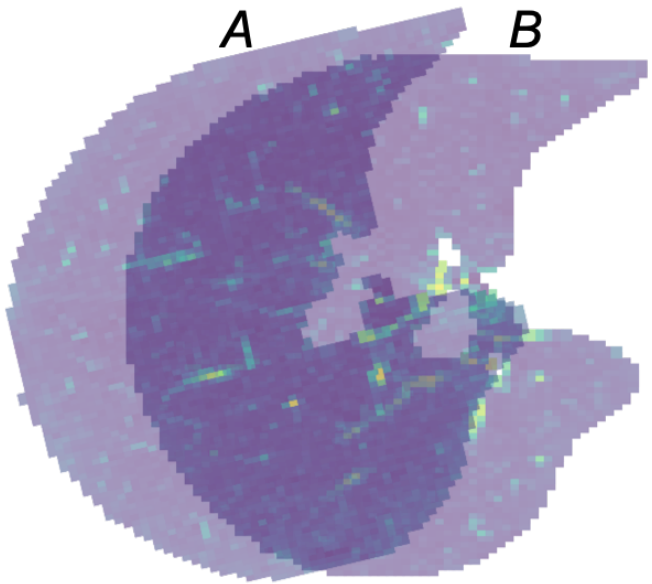
- Left and right lungs were registered separately to account for differences in lung shape and size
- Symmetric Normalization (SyN) non-linear registration was used due to its flexibility and success in EMPIRE10 Challenge [Avants et al., 2008]

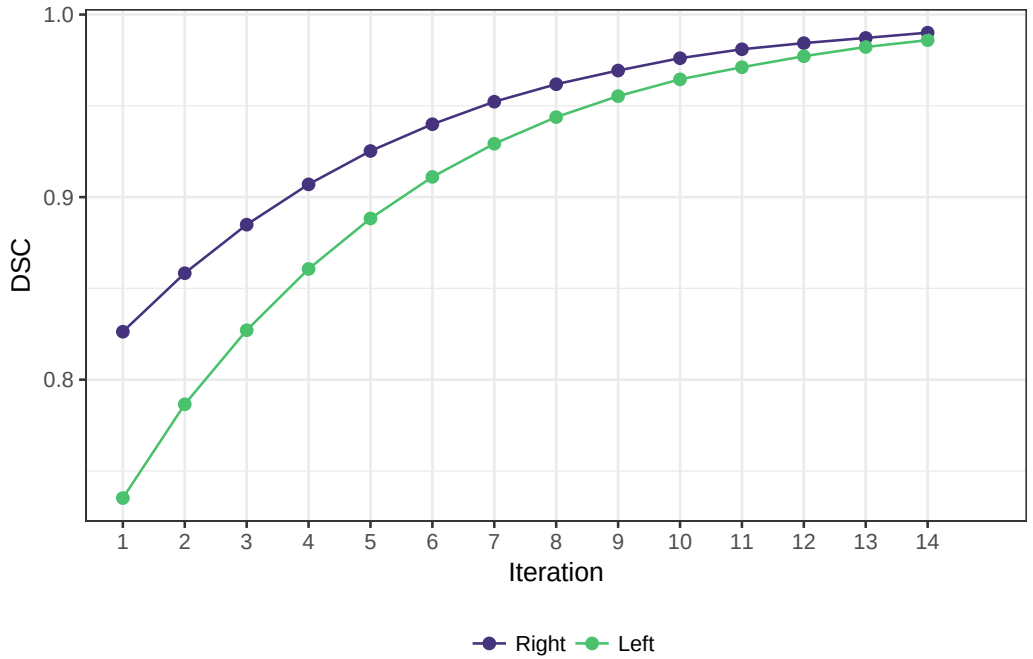


Convergence Criteria

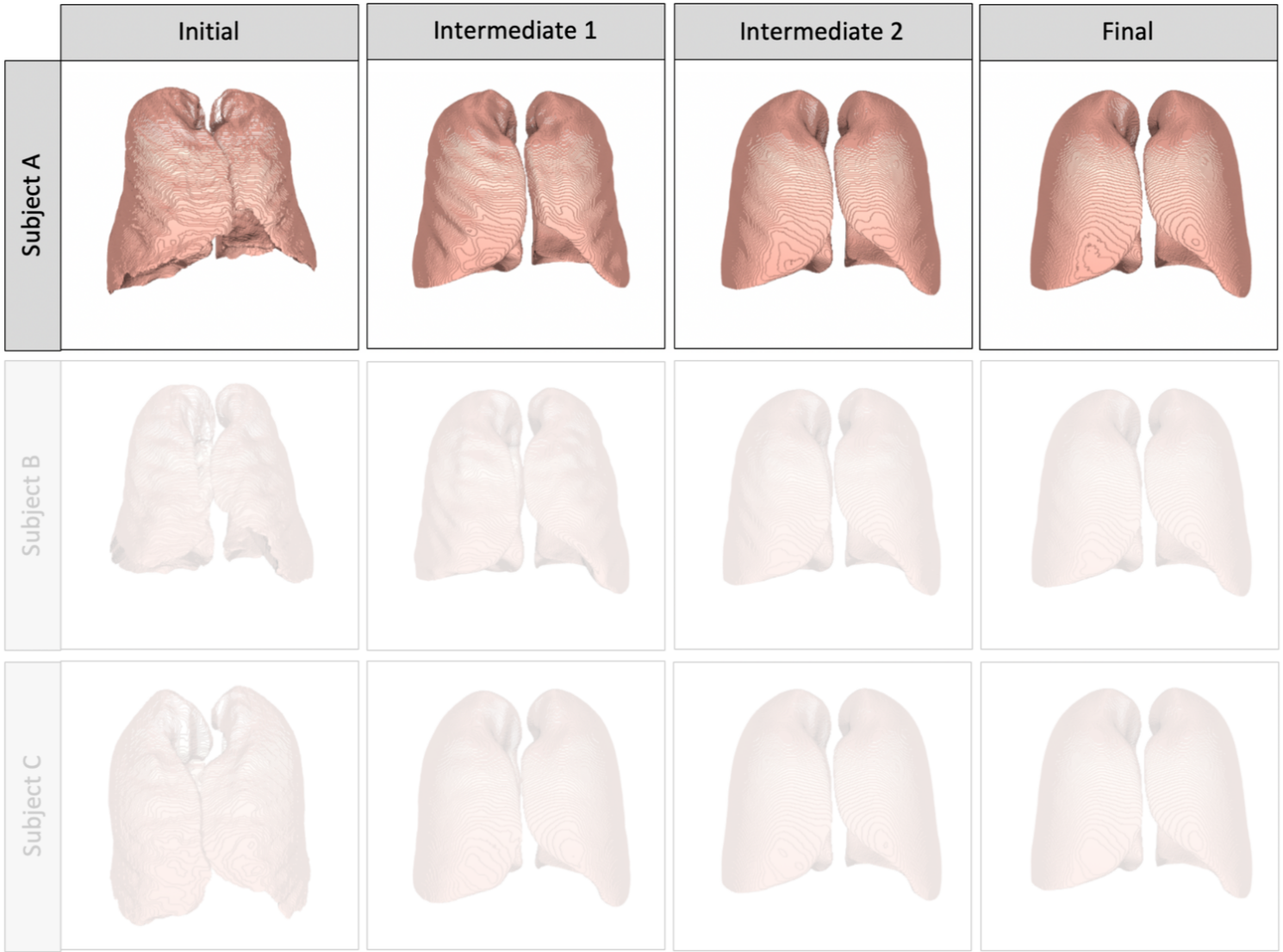
We define convergence using a dice similarity coefficient (DSC) between successive iterations of at least 0.99

$$DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

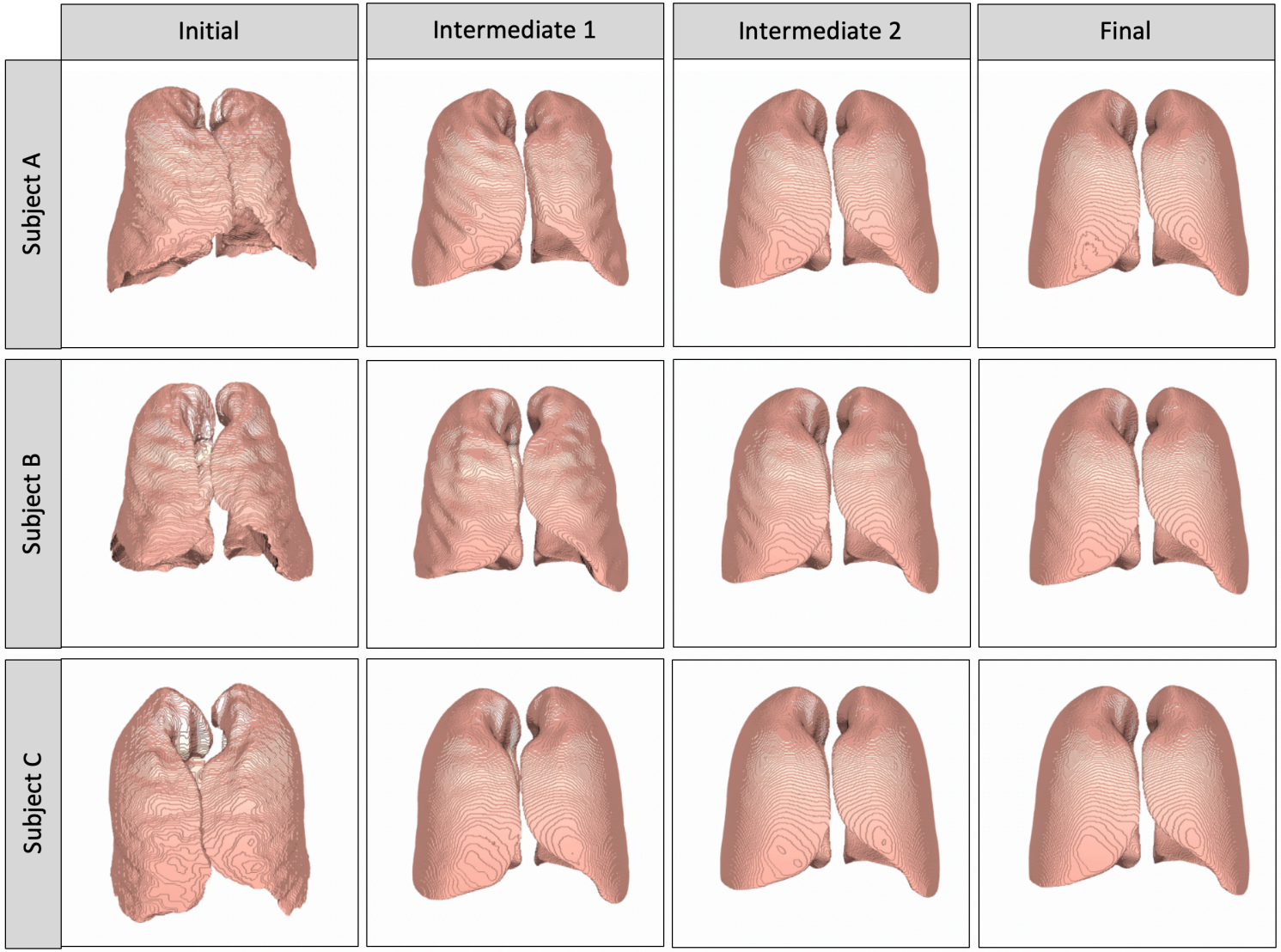




Convergence



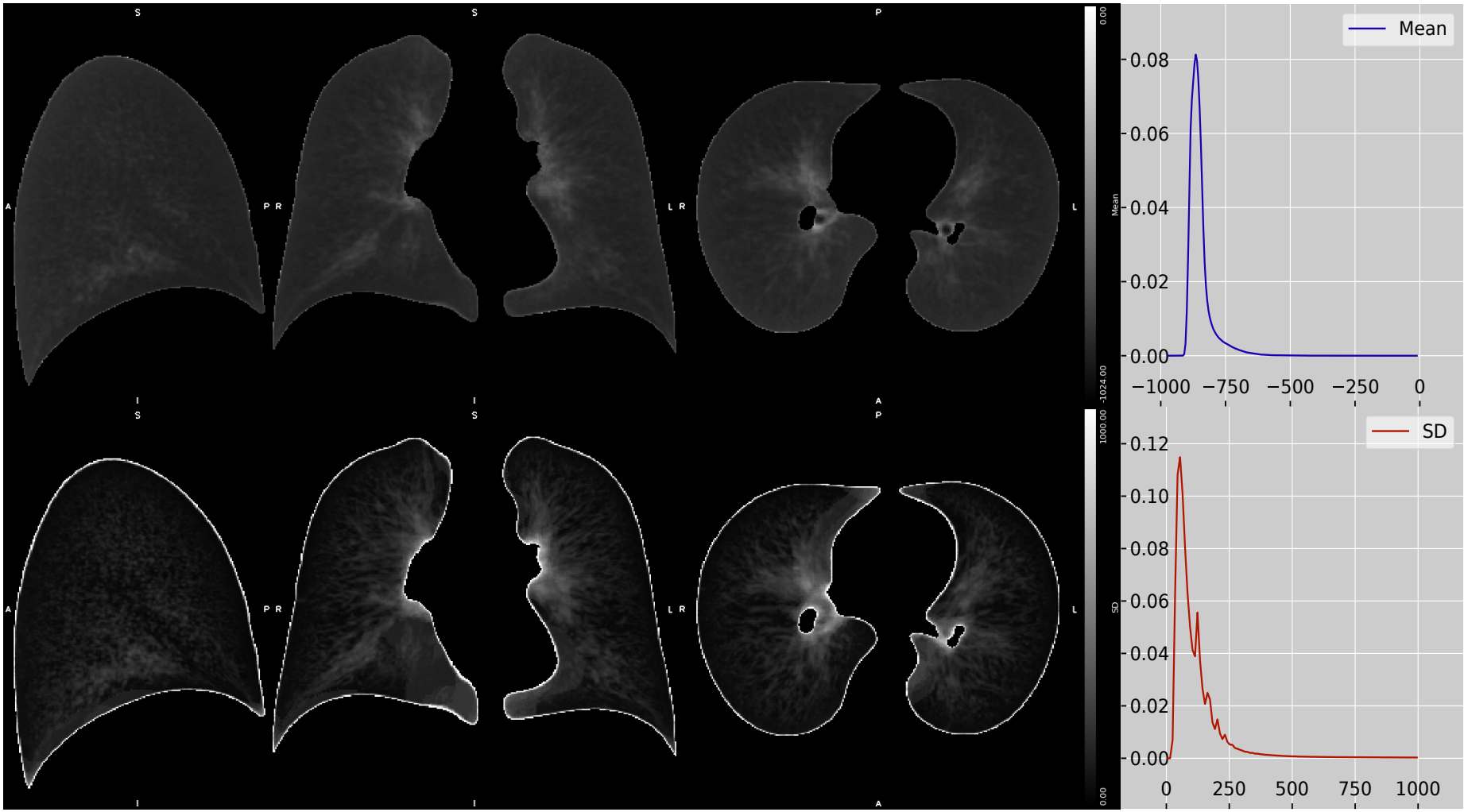
Sensitivity to Initial Template Choice



Healthy Lung Template Characteristics

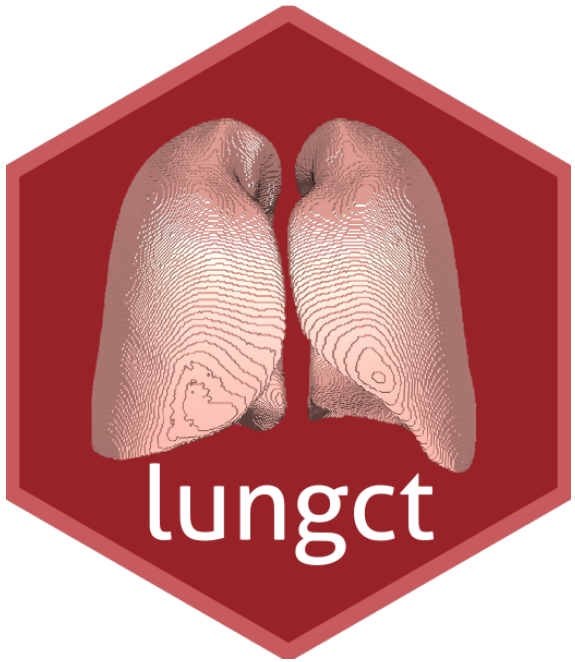
colorado school of public health
UNIVERSITY OF COLORADO
COLORADO STATE UNIVERSITY
UNIVERSITY OF NORTHERN COLORADO

Number of iterations: 14



Conclusions

1. We created the first publicly available standard lung template using healthy adults, which is available for download via *lungct* [Ryan et al., 2019b]
2. We develop a fully-automated and open-source image processing pipeline for lung CTs in R software

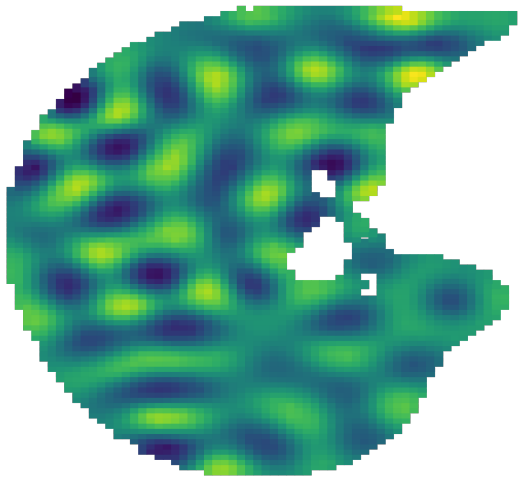


Where is disease commonly found in the lung?

Methodological Questions:

1. How do we align spatial coordinates across scans when scans are different sizes and shapes?
   - ~~**Create a lung template**~~
2. After aligning spatial coordinates across scans, how do we identify significant areas of disease?
   - **Develop a spatial model for whole-lung population-level inference**

Develop a Model for Whole-Lung Population-Level Inference



Classical Method for Imaging Data

Voxel-based morphometry (VBM) - A mass univariate statistical test where separate statistical models are fit on every voxel

- Fails to account for spatial correlation between voxels
- Arbitrary, post-hoc smoothing of data
- High false positive rate and low power depending on multiple comparison correction

Classical Method for Imaging Data

Voxel-based morphometry (VBM) - A mass univariate statistical test where separate statistical models are fit on every voxel

- Fails to account for spatial correlation between voxels
- Arbitrary, post-hoc smoothing of data
- High false positive rate and low power depending on multiple comparison correction

A spatial model for high-dimensional data is desired.

Spatial Models for High-Dimensional Data

- **Integrated Nested Laplace Approximation (INLA)**
  - Type of Gaussian Markov random field-based approximation
  - Provides computationally efficient estimation and inference for latent Gaussian models
  - Extended to neuroimaging fMRI data [Mejia et al., 2019]
  - Problem: Its memory consumption depends on N.
- **Eigenvector Spatial Filtering (ESF)**
  - A type of low rank approximation
  - Describes spatial variation using a linear combination of L basis functions where $L << N$
  - Related to Moran's I, a common spatial summary measure
  - A computationally-efficient and memory-free approach has been developed [Murakami and Griffith, 2019a]

Note: Not an extensive list of spatial modeling approaches for high-dimensional data

Eigenvector Spatial Filtering

ESF models spatial maps using eigenvectors associated with the Moran coefficient (MC) [Moran, 1950]:

$$\mathrm{MC}(\mathbf{y}) = \frac{N}{\mathbf{1}'\mathbf{C}\mathbf{1}} \frac{\mathbf{y}'\mathbf{M}\mathbf{C}\mathbf{M}\mathbf{y}}{\mathbf{y}'\mathbf{M}\mathbf{y}} \tag{1}$$

- $\mathbf{y}$ is a spatial response
- $\mathbf{C}$ is a spatial correlation matrix, such as the exponential
- $\mathbf{M}$ is a centering matrix, ensuring eigenvectors are orthogonal and uncorrelated
- $\mathbf{MCM}$ is decomposed into spatially orthogonal variables, or eigenvectors
- A subset of L eigenvectors, $\mathbf{E}$, where $L << N$, are selected

Murakami's ESF Model

The resulting eigenvector matrix, $\mathbf{E}$, can be used to construct spatially-varying coefficients (SVC) in a mixed-effects model [Murakami et al., 2017]:

$$\begin{aligned} \mathbf{y} &= \sum_{k=1}^{K} \mathbf{x}_k \circ \boldsymbol{\beta}_k^{SVC} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}\right) \\ \boldsymbol{\beta}_k^{SVC} &= \beta_k \mathbf{1} + \mathbf{E}\boldsymbol{\gamma}_k, \quad \boldsymbol{\gamma}_k \sim N\left(\mathbf{0}_L, \sigma_{k(\gamma)}^2 \boldsymbol{\Lambda}\left(\alpha_k\right)\right) \end{aligned} \tag{2}$$

- Parameters are estimated using restricted maximum likelihood (REML)
- Fits high-dimensional spatial data efficiently
- Shown to have less bias in the SVC estimates as compared to geographically-weighted regression [Murakami and Griffith, 2019a]

Our ESF Model for Imaging Data

Our spatial voxel-based morphometry, or **spVBM**, model extends the Murakami's ESF based SVC model by incorporating non-spatial random effects **Zb**. For a single subject, the model is:

$$\mathbf{y_i} = \sum_{k=1}^{K} x_{k,i} \circ \beta_{k,i}^{SVC} + \mathbf{Z_i b_i} + \boldsymbol{\varepsilon_i} \tag{3}$$

$$\beta_{k,i}^{SVC} = \beta_k \mathbf{1} + \mathbf{E_i} \gamma_k \tag{4}$$

$$\mathbf{b_i} \sim N(\mathbf{0}, D), \quad \boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \Sigma_i), \quad \boldsymbol{\gamma_k} \sim N\left(\mathbf{0}, \sigma_k^2 \boldsymbol{\Lambda}(\alpha_k)\right) \tag{5}$$

where $\mathbf{y_i}$ is the response vector, $x_{k,i}$ is the k^{th} covariate, $\beta_{k,i}^{SVC}$ is a vector comprised of the k^{th} spatially-varying coefficient, $\mathbf{Z_i}$ is a matrix of covariates for the random effects, $\mathbf{b_i}$ is the vector containing the random effects, and $\boldsymbol{\varepsilon_i}$ are the residuals. See [Murakami and Griffith, 2019b].

The spVBM Model: Estimation and Inference

The parameters are estimated in the following steps:

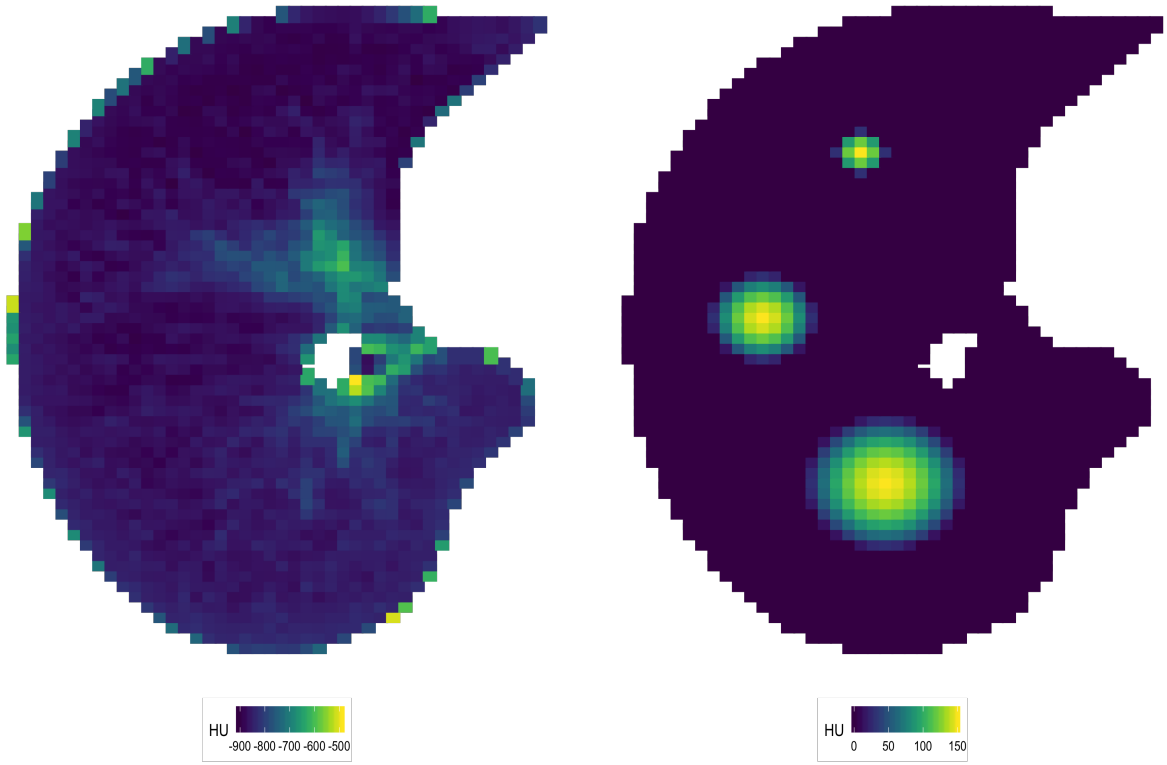
1. β are estimated using maximum likelihood and the normal equations
2. The variance parameters are sequentially estimated by maximizing the restricted log-likelihood

Inference:

- Interested in β_k^{SVC}
- Voxel-level null hypothesis that β_k^{SVC} at voxel j is equal to 0.
- Wald statistic: $W = \frac{(\beta_{k,j}^{SVC} - 0)^2}{Var(\beta_{k,j}^{SVC})}$

Simulation Study

- Simulate local regions of disease with a binary covariate to represent a case-control study
- Local regions based on small, medium and large disease regions using an Epanechnikov kernel
- Compare spVBM to VBM



Eigenvectors in the Lung

- Distance matrix: (N_u x N_u) dimension
- Eigenvector design matrix: (N_u x L) dimension

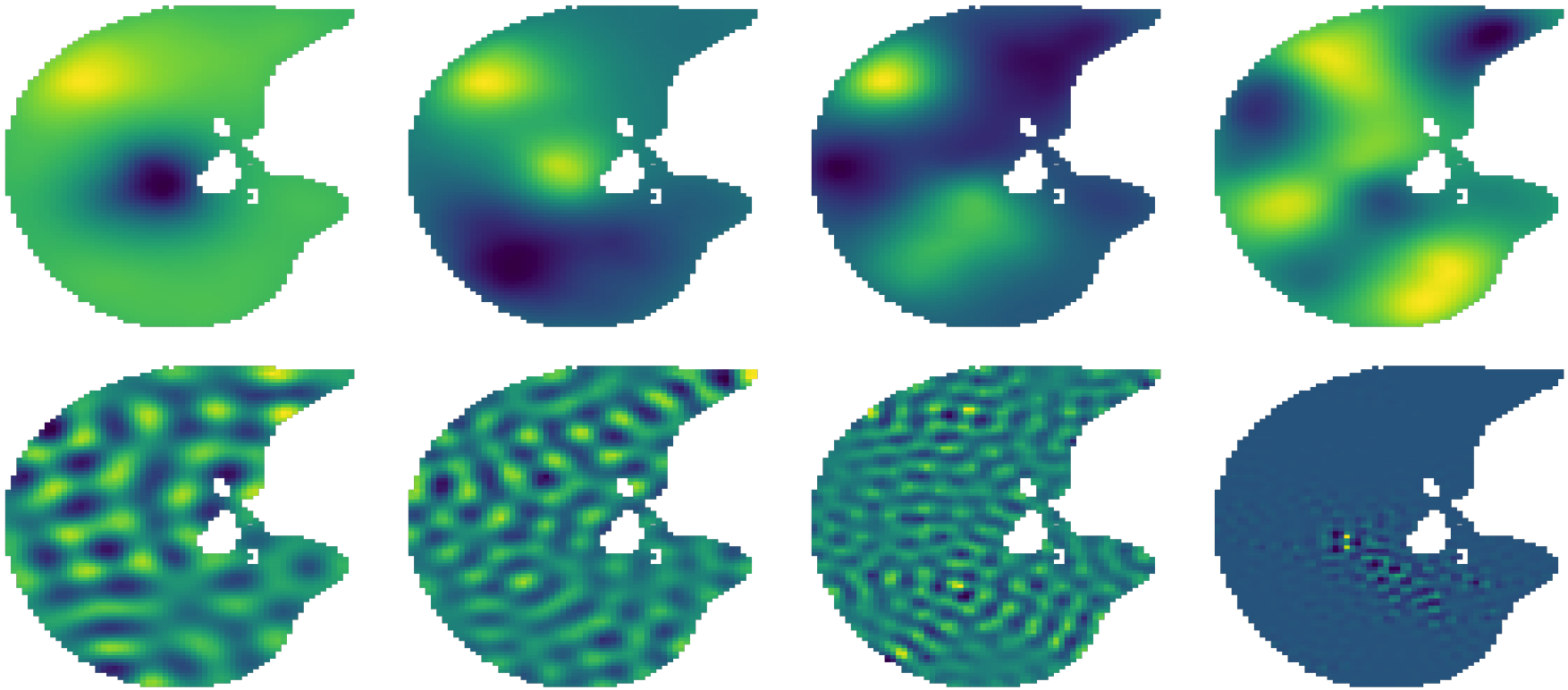
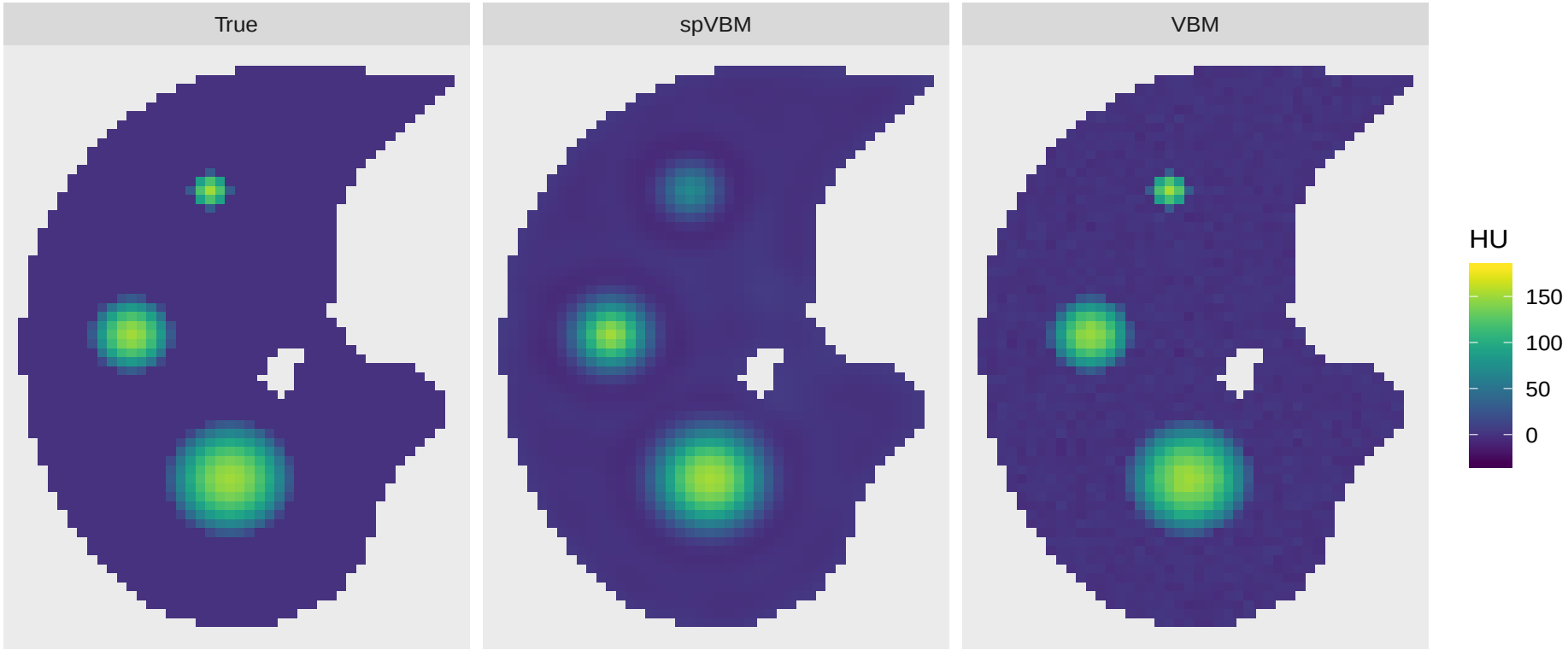
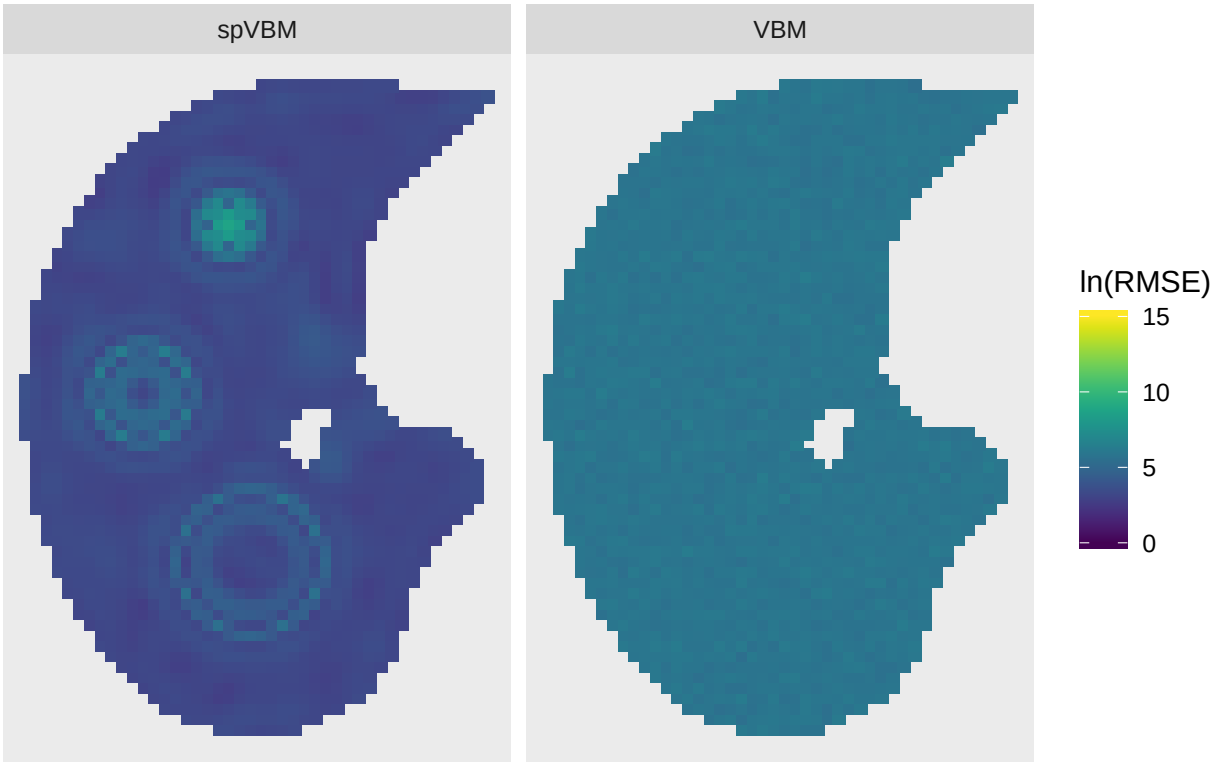


Figure: Moran eigenvectors (1, 2, 4, 8, 100, 200, 400, 800) based on the spatial correlation matrix from a 2D axial slice of the lung.


True
spVBM
VBM
HU
150
100
50
0

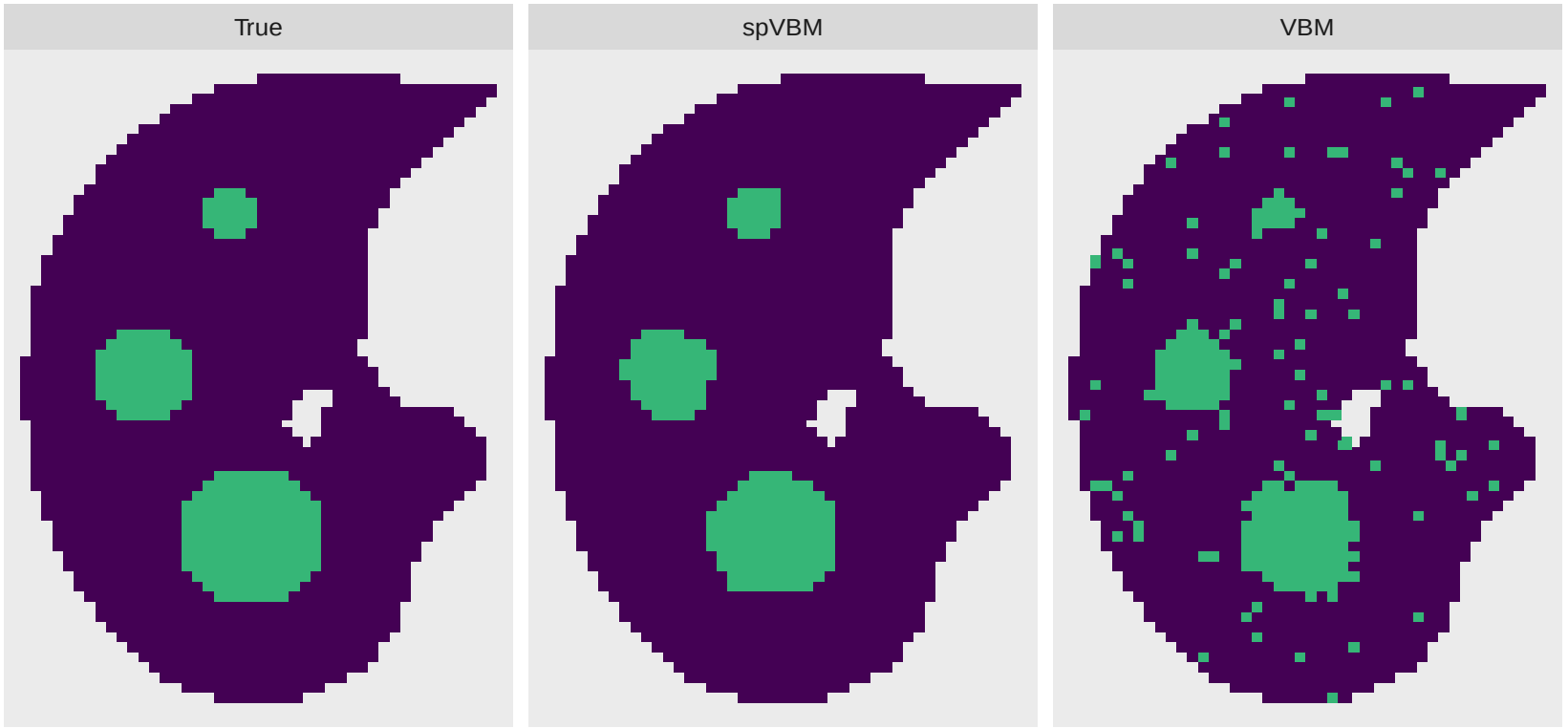
RMSE in $\hat{\beta}_1^{SVC}$



Inference for $\hat{\beta}_1^{SVC}$

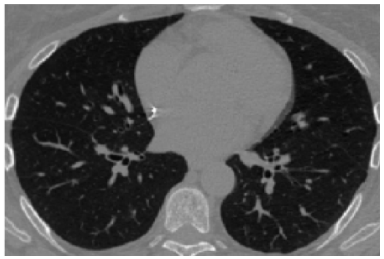
spVBM : FPR = 0.007, FNR = 0.105

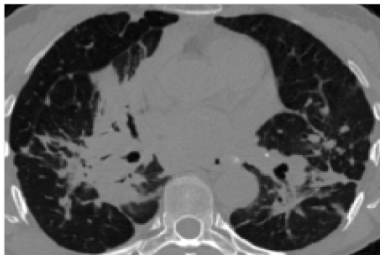
VBM : FPR = 0.050, FNR = 0.315



Application to Diseased Population

GRADS
GENOMIC RESEARCH IN ALPHA-1 ANTITRYPSIN DEFICIENCY AND SARCOIDOSIS



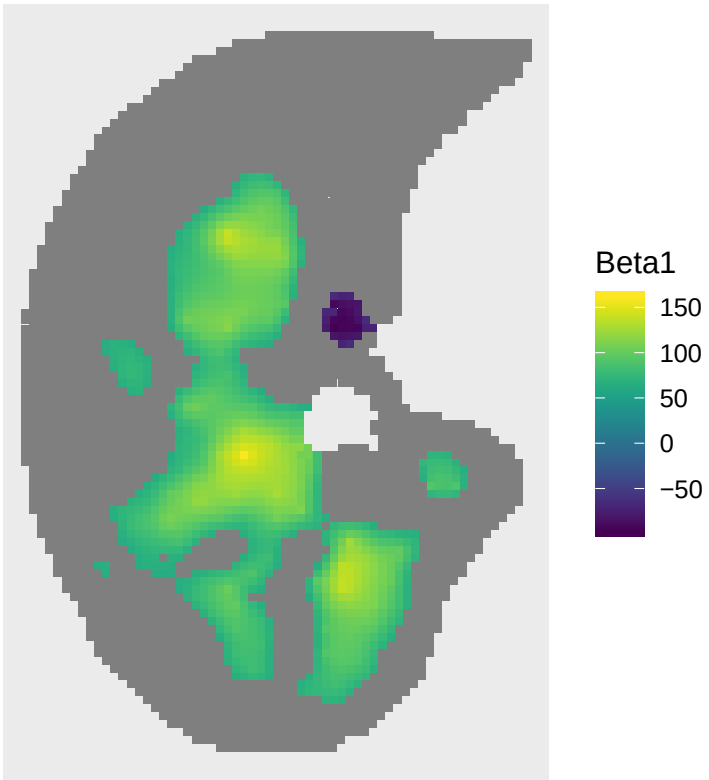


	Healthy	Fibrotic	P-value
Sample Size	13	29	
Male (%)	6 (46.2)	13 (44.8)	1.000
White (%)	11 (84.6)	17 (58.6)	0.194
Hispanic (%)	3 (23.1)	1 (3.4)	0.151
Age (mean (SD))	56.29 (8.87)	55.03 (7.69)	0.641
BMI (mean (SD))	33.09 (4.18)	28.83 (5.38)	0.016

Application to Diseased Population

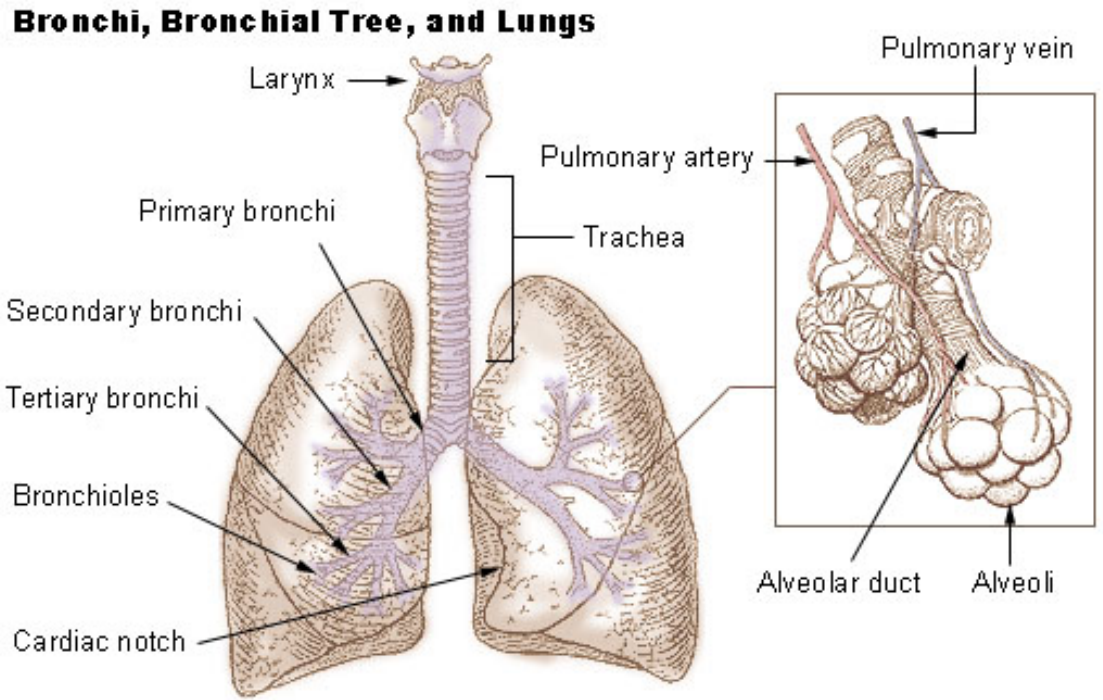
$$\mathbf{HU_i} = \beta_{0,i}^{SVC} + \beta_{1,i}^{SVC} * X_{fibrosis} + \beta_{2,i}^{SVC} * X_{BMI} + \mathbf{b_i} + \boldsymbol{\varepsilon_i}$$

$$\mathbf{b_i} \sim N\left(\mathbf{0}, \sigma_b^2 \mathbf{I}\right), \quad \boldsymbol{\varepsilon_i} \sim N\left(\mathbf{0}, \sigma_e^2 \mathbf{I}\right), \quad \boldsymbol{\gamma_k} \sim N\left(\mathbf{0}, \sigma_k^2 \boldsymbol{\Lambda}\left(\alpha_k\right)\right)$$



Compared to healthy patients, patients with fibrosis have areas of significantly higher intensity near the bronchi.

Application to Diseased Population



Compared to healthy patients, patients with fibrosis have areas of significantly higher intensity near the bronchi.

Conclusions

Clinical Question:

Where is disease commonly found in the lung?

Methodological Questions:

1. How do we align spatial coordinates across scans when scans are different sizes and shapes?
   - ~~**Create a lung template**~~
2. After aligning spatial coordinates across scans, how do we identify significant areas of disease?
   - ~~**Develop a spatial model for whole-lung population-level inference**~~

Future Directions

- Develop a spectrum of lung templates
  - Age-specific
  - Disease-specific
- Extend spVBM
  - To 3D data
  - To non-normal data
  - To different spatial correlations
  - Compare to Mejia's INLA method for imaging data
- Explore other imaging modalities and anatomical areas
  - fMRI
  - Animal studies

Collaborators

- Nichole Carlson, PhD, Colorado School of Public Health
- Tasha Fingerlin, PhD, National Jewish Health
- Debashis Ghosh, PhD, Colorado School of Public Health
- Lisa Maier, MD, MSPH, National Jewish Health
- John Muschelli, PhD, Johns Hopkins Bloomberg School of Public Health

Funding Sources

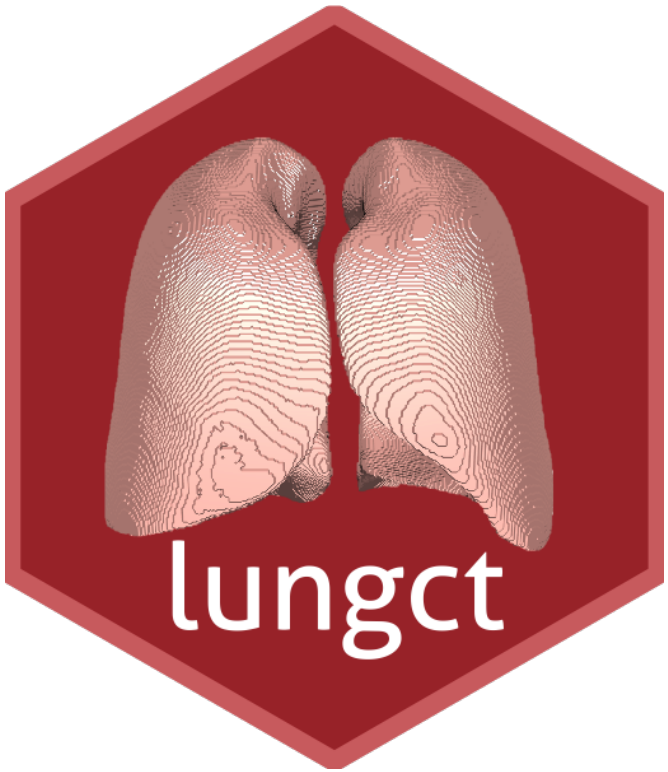
- National Institutes of Health (R01 HL114587; R01 HL142049; U01 HL112695)
- GRADS study (NIH grant U01 HL112707, U01 HL112707, U01 HL112694, U01 HL112695, U01 HL112696, U01 HL112702, U01 HL112708, U01 HL112711, U01 HL112712)
- COPDGene study (NIH grants U01 HL089856 and U01 HL089897, COPD Foundation)

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health.

Contact

colorado school of public health
UNIVERSITY OF COLORADO
COLORADO STATE UNIVERSITY
UNIVERSITY OF NORTHERN COLORADO

Email: Sarah.M.Ryan@cuanschutz.edu
Twitter: @SarahBiostats
GitHub: ryansar
website: www.SarahMRyan.com



References I

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008).
Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain.
Medical image analysis, 12(1):26–41.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014).
Fitting linear mixed-effects models using lme4.
arXiv preprint arXiv:1406.5823.

Evans, A. C., Janke, A. L., Collins, D. L., and Baillet, S. (2012).
Brain templates and atlases.
Neuroimage, 62(2):911–922.

Mejia, A. F., Yue, Y., Bolin, D., Lindgren, F., and Lindquist, M. A. (2019).
A bayesian general linear modeling approach to cortical surface fmri data analysis.
Journal of the American Statistical Association, pages 1–26.

Moran, P. A. (1950).
Notes on continuous stochastic phenomena.
Biometrika, 37(1/2):17–23.

Murakami, D. and Griffith, D. A. (2019a).
Eigenvector spatial filtering for large data sets: fixed and random effects approaches.
Geographical Analysis, 51(1):23–49.

Murakami, D. and Griffith, D. A. (2019b).
A memory-free spatial additive mixed modeling for big spatial data.
arXiv preprint arXiv:1907.11369.

References II

Murakami, D. and Griffith, D. A. (2019c).
Spatially varying coefficient modeling for large datasets: Eliminating n from spatial regressions.
Spatial Statistics, 30:39–64.

Murakami, D., Yoshida, T., Seya, H., Griffith, D. A., and Yamagata, Y. (2017).
A moran coefficient-based mixed effects approach to investigate spatially varying relationships.
Spatial Statistics, 19:68–89.

Regan, E. A., Hokanson, J. E., Murphy, J. R., Make, B., Lynch, D. A., Beaty, T. H., Curran-Everett, D., Silverman, E. K., and Crapo, J. D. (2011).
Genetic epidemiology of copd (copdgene) study design.
COPD: Journal of Chronic Obstructive Pulmonary Disease, 7(1):32–43.

Ryan, S. M., Fingerlin, T. E., Mroz, M., Barkes, B., Hamzeh, N., Maier, L. A., and Carlson, N. E. (2019a).
Radiomic measures from chest hrct associated with lung function in sarcoidosis.
European Respiratory Journal, page 1900371.

Ryan, S. M., Vestal, B., Carlson, N. E., and Muschelli, J. (2019b).
Template creation for high resolution computed tomography scans of the lung in r software.
Academic Radiology.

The spVBM Model: Multi-subject model

$$\mathbf{Y} = \sum_{k=1}^{K} X_k \circ \beta_k^{SVC} + \mathbf{Zb} + \varepsilon \tag{6}$$

$$\beta_k^{SVC} = \beta_k \mathbf{1} + \mathbf{E}\gamma_k \tag{7}$$

where $\mathbf{Y} = [\mathbf{y_1} \ldots \mathbf{y_N}]^T$, $X_k = [x_{k,1} \ldots x_{k,N}]^T$, $\mathbf{b} = [\mathbf{b_1} \ldots \mathbf{b_N}]^T$, $\varepsilon = [\varepsilon_1 \ldots \varepsilon_N]^T$ and $\mathbf{E} = [\mathbf{E_1} \ldots \mathbf{E_N}]^T$, and where $\mathbf{Z}$, is the block diagonal matrix with blocks $\mathbf{Z_i}$ on the diagonals and zeros elsewhere. The dimension of $\mathbf{Y}$ is equal to $n = \sum_{i=1}^{N} n_i$.

The spVBM Model: Reparameterizatior

Separating the global $\mathbf{X}\beta$ and spatial terms $\tilde{\mathbf{E}}\tilde{\gamma}$:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \tilde{\mathbf{E}}\tilde{\gamma} + \mathbf{Z}\mathbf{b} + \varepsilon, \quad \varepsilon \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \\ \tilde{\mathbf{E}} &= [\mathbf{X}_1 \circ \mathbf{E} \cdots \mathbf{X}_K \circ \mathbf{E}] \end{aligned} \tag{8}$$

Obtaining common variance terms as in [Bates et al., 2014]:

$$\mathbf{Y} = \mathbf{X}\beta + \tilde{\mathbf{E}}\tilde{\mathbf{V}}(\Theta)\tilde{\mathbf{u}} + \mathbf{Z}\Omega(\Phi)\mathbf{w} + \varepsilon, \quad \tilde{\mathbf{u}}, \mathbf{w}, \varepsilon \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$$

$$\tilde{\mathbf{V}}(\Theta) = \begin{bmatrix} \mathbf{V}\left(\theta_1\right) & & \\ & \ddots & \\ & & \mathbf{V}\left(\theta_K\right) \end{bmatrix}, \quad \Omega(\Phi) = \begin{bmatrix} \Omega\left(\phi_1\right) & & \\ & \ddots & \\ & & \Omega\left(\phi_H\right) \end{bmatrix} \tag{9}$$

The spVBM Model: Estimation

1. $\boldsymbol{\beta}, \tilde{\mathbf{u}}, \mathbf{w}$ are estimated using maximum likelihood and the normal equations;

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\tilde{\mathbf{u}}} \\ \hat{\mathbf{w}} \end{bmatrix} = \mathbf{P}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \tilde{\mathbf{V}}(\boldsymbol{\Theta})\tilde{\mathbf{E}}'\mathbf{y} \\ \boldsymbol{\Omega}(\boldsymbol{\Phi})\mathbf{Z}'\mathbf{y} \end{bmatrix} \tag{10}$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\tilde{\mathbf{E}}\tilde{\mathbf{V}}(\boldsymbol{\Theta}) & \mathbf{X}'\mathbf{Z}\boldsymbol{\Omega}(\boldsymbol{\Phi}) \\ \tilde{\mathbf{V}}(\boldsymbol{\Theta})\tilde{\mathbf{E}}'\mathbf{X} & \tilde{\mathbf{V}}(\boldsymbol{\Theta})\tilde{\mathbf{E}}'\tilde{\mathbf{E}}\tilde{\mathbf{V}}(\boldsymbol{\Theta}) + \mathbf{I} & \tilde{\mathbf{V}}(\boldsymbol{\Theta})\tilde{\mathbf{E}}'\mathbf{Z}\boldsymbol{\Omega}(\boldsymbol{\Phi}) \\ \boldsymbol{\Omega}(\boldsymbol{\Phi})\mathbf{Z}'\mathbf{X} & \boldsymbol{\Omega}(\boldsymbol{\Phi})\mathbf{Z}'\tilde{\mathbf{E}}\tilde{\mathbf{V}}(\boldsymbol{\Theta}) & \boldsymbol{\Omega}(\boldsymbol{\Phi})\mathbf{Z}'\mathbf{Z}\boldsymbol{\Omega}(\boldsymbol{\Phi}) + \mathbf{I} \end{bmatrix} \tag{11}$$

The spVBM Model: Estimation

2. The variance parameters are estimated by maximizing the restricted log-likelihood

$$\text{loglik}_R(\boldsymbol{\Theta}, \boldsymbol{\Phi}) = -\frac{1}{2}\log|\mathbf{P}| - \frac{N-K}{2}\left(1 + \log\left(\frac{2\pi\tilde{d}(\boldsymbol{\Theta})}{N-K}\right)\right) \quad (12)$$

$$\tilde{d}(\boldsymbol{\Theta}, \boldsymbol{\Phi}) = \min_{\beta, \tilde{\mathbf{u}}, \mathbf{w}} \|\mathbf{Y} - \mathbf{X}\beta - \tilde{\mathbf{E}}\tilde{\mathbf{V}}(\boldsymbol{\Theta})\tilde{\mathbf{u}} - \mathbf{Z}\boldsymbol{\Omega}(\boldsymbol{\Phi})\mathbf{w}\|^2 + \|\tilde{\mathbf{u}}\|^2 + \|\mathbf{w}\|^2 \quad (13)$$

The spVBM Model: Estimation

3. The residual variance is estimated below, where n is the dimension of Y and K is the number of fixed effects:

$$\hat{\sigma}^2 = \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \tilde{\mathbf{E}}\mathbf{V}(\boldsymbol{\Theta})\tilde{\mathbf{u}} - \mathbf{Z}\boldsymbol{\Omega}(\boldsymbol{\Phi})\mathbf{w}\|^2}{n - K} \tag{14}$$

To perform statistical inference on the β_k^{SVC}, we obtain the covariance matrix of β_k^{SVC} following [Murakami and Griffith, 2019c], where $\hat{\mathbf{P}}_k^*$ is a subset of $\mathbf{P}^{-1}$, only including rows and columns associated with x_k:

$$Var(\hat{\beta}_k^{SVC}) = \hat{\sigma}^2 \begin{bmatrix} 1 & \mathbf{E}\mathbf{V}(\hat{\theta}_k) \end{bmatrix} \hat{\mathbf{P}}_k^* \begin{bmatrix} \mathbf{1}^T \\ \mathbf{V}(\hat{\theta}_k)\mathbf{E}^T \end{bmatrix} \tag{15}$$