Bayesian Hierarchical Modeling and Clustering of Malignant Cancer Diagnoses

S. Zhu, J. Ma, Z. Feng

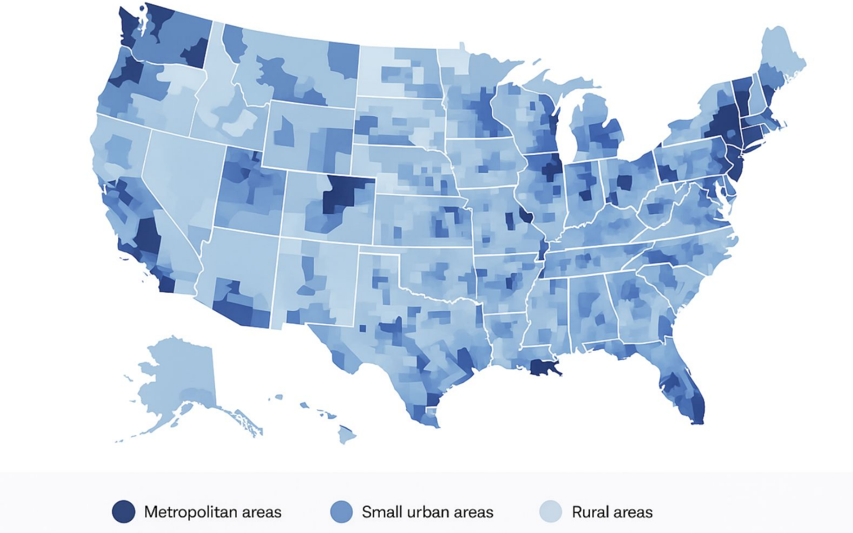
Johns Hopkins
University

Background

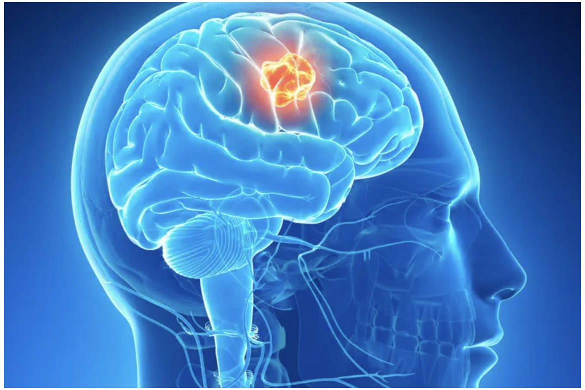
- Cancer remains a leading cause of mortality

- Patients come from different regions

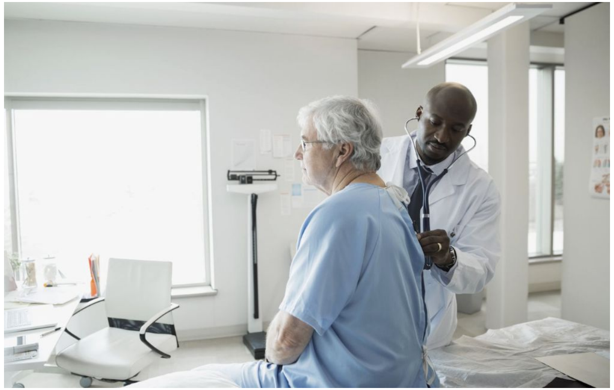
- Show different risks of late-stage tumors



Which level contributes most to late-stage diagnosis?



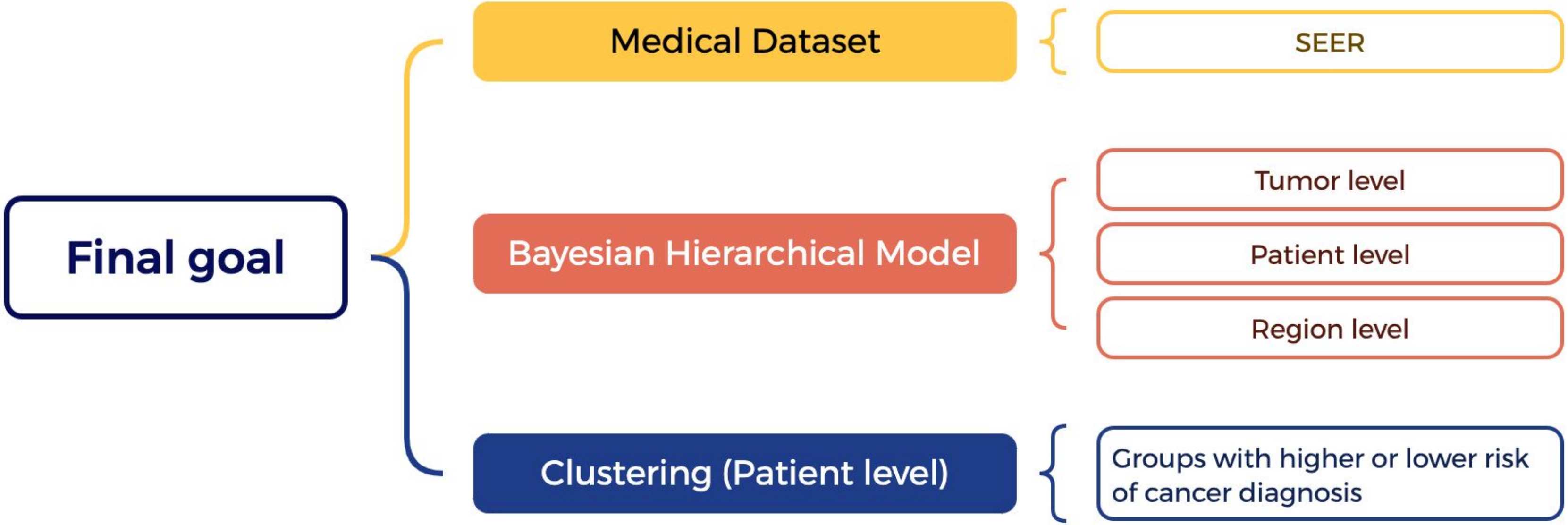
Tumor



Patient

Region

Final Goal



Final goal

Medical Dataset

SEER

Bayesian Hierarchical Model

Tumor level

Patient level

Region level

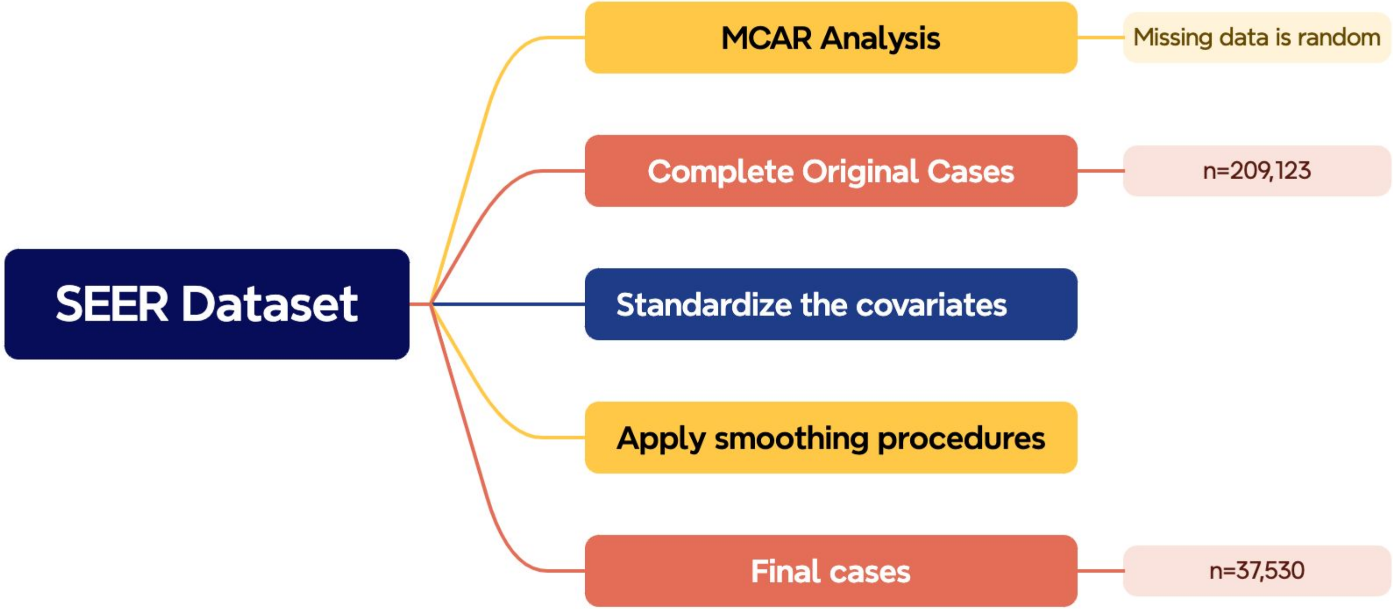
Clustering (Patient level)

Groups with higher or lower risk of cancer diagnosis

Dataset

SEER Dataset

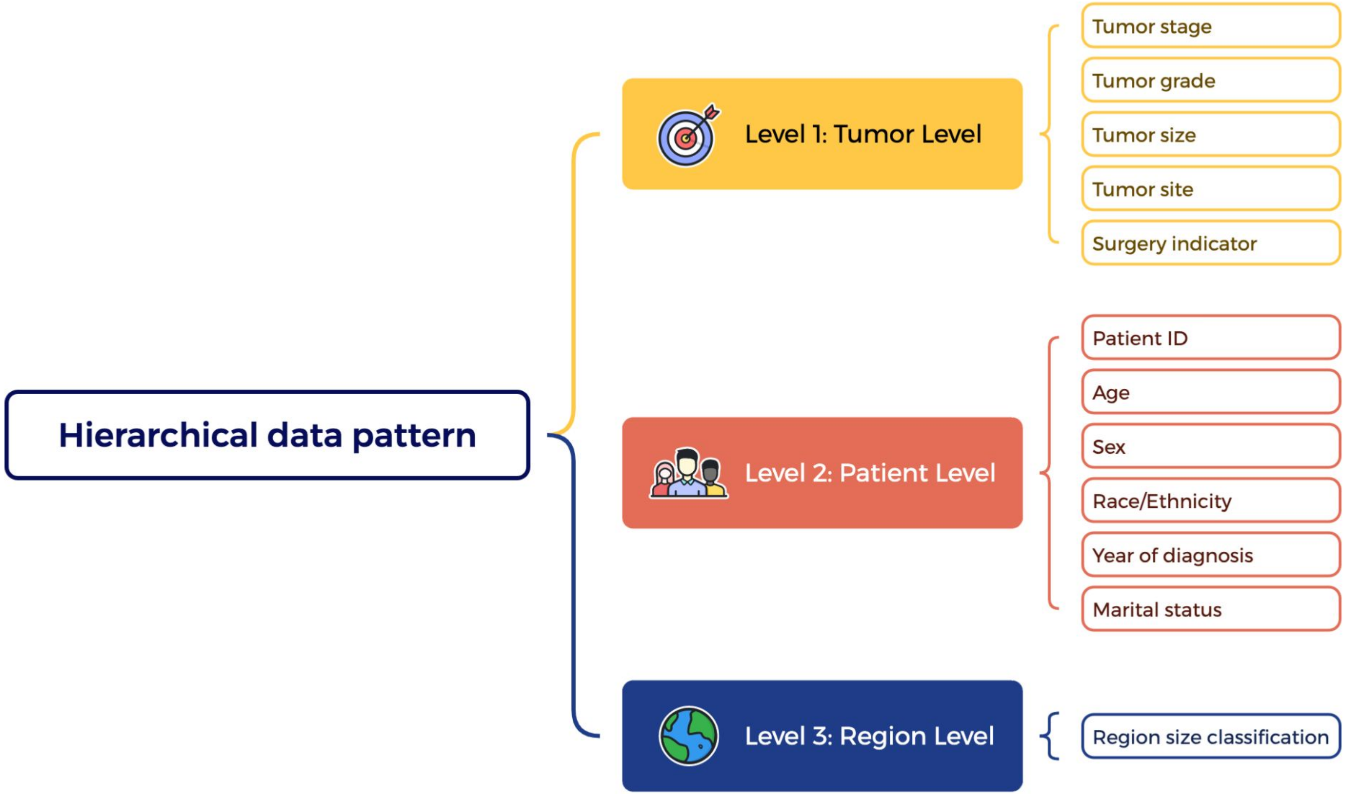
NATIONAL CANCER INSTITUTE
Surveillance, Epidemiology, and End Results Program



SEER Dataset

- MCAR Analysis — Missing data is random
- Complete Original Cases — n=209,123
- Standardize the covariates
- Apply smoothing procedures
- Final cases — n=37,530

Dataset Structure



Hierarchical data pattern

Level 1: Tumor Level

- Tumor stage
- Tumor grade
- Tumor size
- Tumor site
- Surgery indicator

Level 2: Patient Level

- Patient ID
- Age
- Sex
- Race/Ethnicity
- Year of diagnosis
- Marital status

Level 3: Region Level

- Region size classification

Model Specification

What we’re attempting to build on

- Lower SES status is associated with delayed cancer diagnosis.
- Effects persist even after accounting for tumor stage, indicating systemic inequalities in access and follow-up.
- Intersection between demographic and geographic factors causes further disparities.
- Much of prior work analyzes disparities one dimension at a time (e.g., rural vs urban, or SES, or race), Limitations highlight that this lack of multilevel modeling limits our ability to disentangle individual vs contextual effects.

Why Utilize a Hierarchical Model?

- Standard Regression Models mute intra-cluster correlation
- Tumor-Level:

$$\operatorname{logit}(p_{tpr}) = z_{tpr}^{\top}\beta_T + w_p^{\top}\beta_P + u_p + v_r,$$

- Patient-Level:

$$u_p \mid \sigma_u^2 \sim N(0, \sigma_u^2), \quad p = 1, \dots, P.$$

- Region-Level:

$$v_r \mid \mu_v, \sigma_v^2 \sim N(\mu_v, \sigma_v^2), \quad r = 1, \dots, R.$$

Prior Specification and Posterior Sampling

- Weakly-Informative Priors
  - Fixed Effects: Normal(0,4)
  - Intercept: Normal(0,25)
  - Group Std. Devs: t(3,0,2.5) (Gelman, 2006)

- No U Turn Sampling: A Ball rolling in the park
  - Implemented via `brms` which compiles to Stan and doesn’t require tuning

- Diagnostics
  - Scale Reduction Factors (Rhat)
  - Effective Sample Size
  - Variance Inflation Factor

Clustering

The Goal

- Cluster posterior random effects based on patient demographics (age, sex, race).

- Connecting the model results to risk stratification.

Bayesian Clustering Method

- Finding the center for each of the clusters

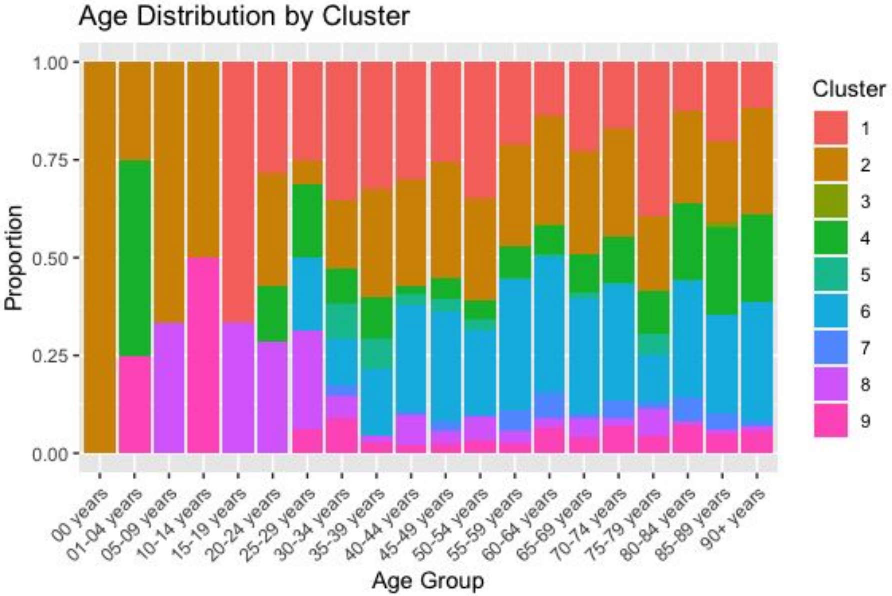
$$\bar{v}_i = \frac{1}{S}\sum_{s=1}^{S} v_i^{(s)}$$

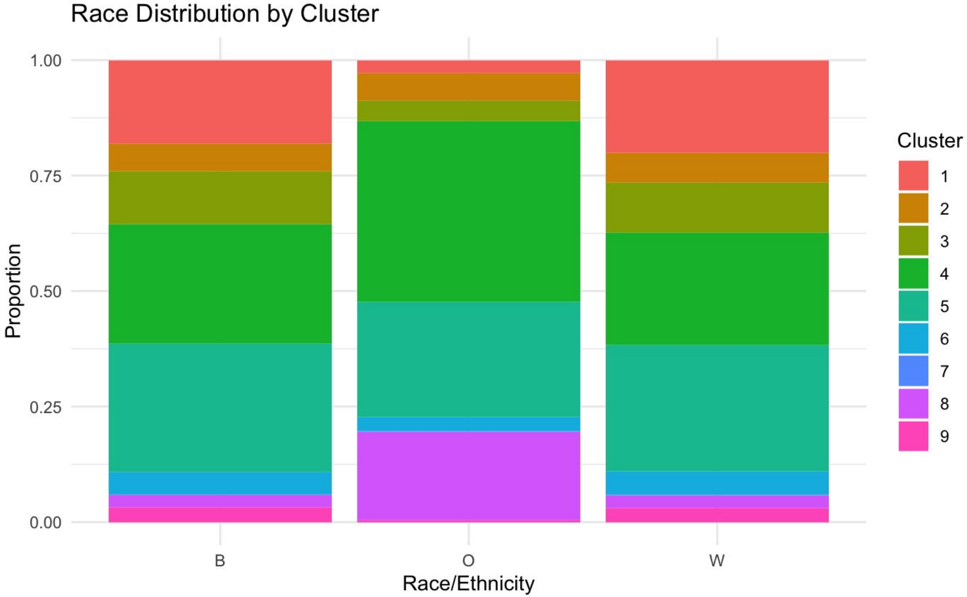
  - $V_i^{(s)}$ is the patient-level random effect for patient *i* in draw s
  - S is the number of samples
  - $\bar{v}_i$ is the posterior mean of that patient's random effect

- Gaussian Mixture Model

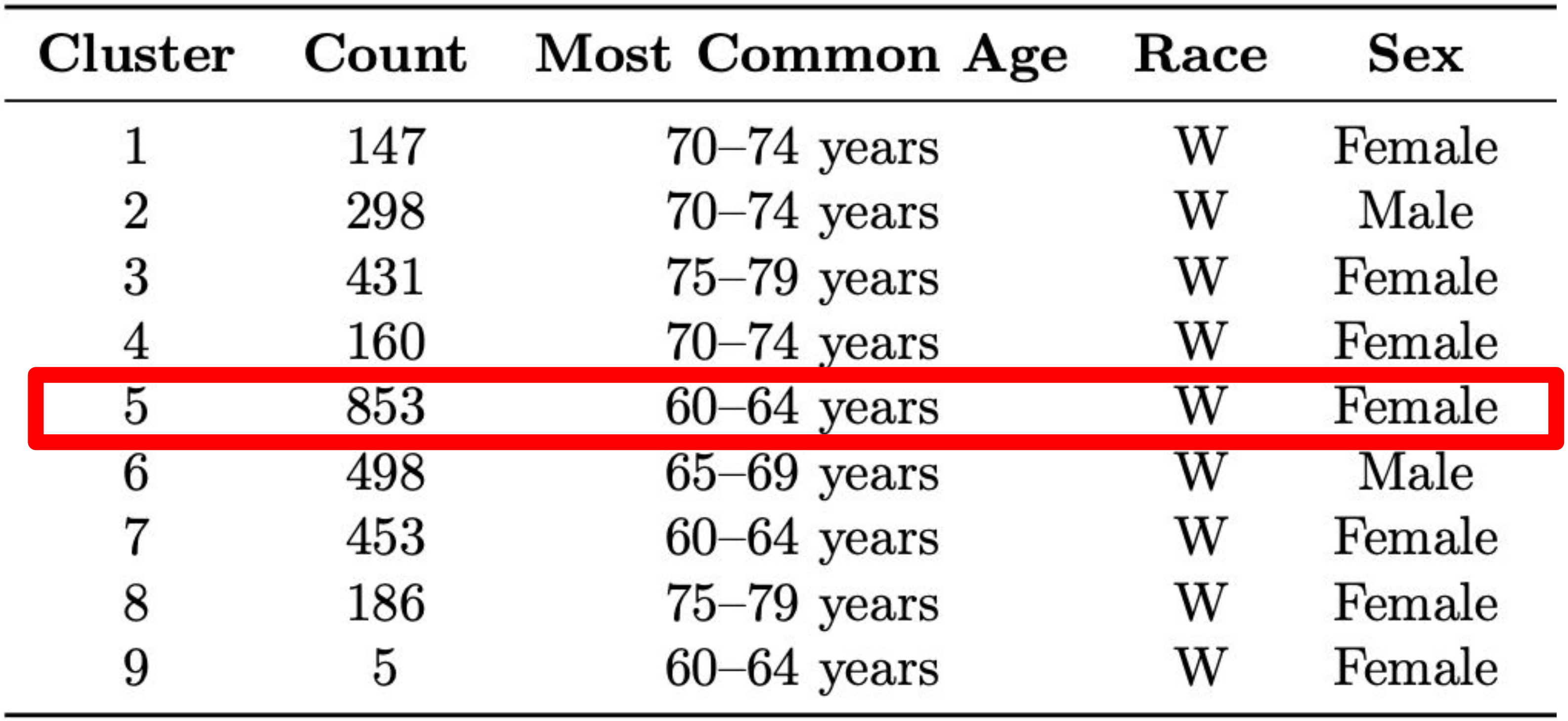
  - Fit a Bayesian Gaussian Mixture to $\{\bar{v}_i\}$:
    - $\bar{v}_i \,|\, z_i = k \sim N(m_k, s_k^2)$
    - $z_i \sim$ Categorical(ω)
    - $\omega \sim$ Dirichlet(α)

Cluster Visualization





Bayesian Clustering Result



Cluster	Count	Most Common Age	Race	Sex
1	147	70–74 years	W	Female
2	298	70–74 years	W	Male
3	431	75–79 years	W	Female
4	160	70–74 years	W	Female
5	853	60–64 years	W	Female
6	498	65–69 years	W	Male
7	453	60–64 years	W	Female
8	186	75–79 years	W	Female
9	5	60–64 years	W	Female

Discussion

Takeaways

- Demographics Drive Patterns:
  - Older white females dominated the largest group but showed wide tumor size variance, making them a high-burden subgroup.
  - Non-white older females had relatively large tumors despite small cluster size, indicating potential disparities in early detection or access.
  - The most common age group is 60-64, followed by 75-79 and 70-74

- Male Clusters Were Smaller but Informative:
  - Clusters with older white males were consistent in tumor size but warrant further investigation for under-detection or delayed diagnosis.

- Geography Uniform, Disparities Persist:
  - All clusters came from metropolitan areas, so risk variation is not geographic, but demographic.

References

1. Singh GK, Miller BA, Hankey BF, Edwards BK. Area Socioeconomic Variations in U.S. Cancer Incidence, Mortality, Stage, Treatment, and Survival, 1975–1999. NCI Cancer Surveillance Monograph Series, Number 4. Bethesda, MD: National Cancer Institute, 2003. NIH Publication No. 03-0000. Link

2. Mobley, Lee & Kuo, Tzy-Mey & Watson, Lisa & Brown, G.. (2012). Geographic Disparities in Late-Stage Cancer Diagnosis: Multilevel Factors and Spatial Interactions. Health & place. 18. 978-90. 10.1016/j.healthplace.2012.06.009. Link

3. Singh, Sarah, & Praveen Sridhar. "A narrative review of sociodemographic risk and disparities in screening, diagnosis, treatment, and outcomes of the most common extrathoracic malignancies in the United States." Journal of Thoracic Disease [Online], 13.6 (2021): 3827-3843. Web. 2 Dec. 2025 Link

4. Andrew Gelman. (2006). *Prior distributions for variance parameters in hierarchical models*. *Bayesian Analysis*, 1(3), 515–533

5. Hoffman & Gelman (2014). “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo.” *JMLR*.