UAB LIBRARIES
The University of Alabama at Birmingham

University of Alabama at Birmingham
UAB Digital Commons

All ETDs from UAB

UAB Theses & Dissertations

2020

Incorporating Spatial Structure into Bayesian Variable Selection Using Spike-and-Slab Priors with Application to Imaging Data

Justin Leach
University of Alabama at Birmingham

Follow this and additional works at: https://digitalcommons.library.uab.edu/etd-collection

 Part of the Public Health Commons

Recommended Citation

Leach, Justin, "Incorporating Spatial Structure into Bayesian Variable Selection Using Spike-and-Slab Priors with Application to Imaging Data" (2020). *All ETDs from UAB*. 837.
https://digitalcommons.library.uab.edu/etd-collection/837

This content has been accepted for inclusion by an authorized administrator of the UAB Digital Commons, and is provided as a free open access item. All inquiries regarding this item or the UAB Digital Commons should be directed to the UAB Libraries Office of Scholarly Communication.

INCORPORATING SPATIAL STRUCTURE INTO BAYESIAN VARIABLE SELECTION USING SPIKE-AND-SLAB PRIORS WITH APPLICATION TO IMAGING DATA

by

JUSTIN M. LEACH

INMACULADA ABAN, COMMITTEE CHAIR
LLOYD J. EDWARDS
RAJESH K. KANA
KRISTINA VISSCHER
NENGJUN YI

A DISSERTATION

Submitted to the graduate faculty of the University of Alabama at Birmingham,
in partial fullfillment of the requirement for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

2020

Copyright by
Justin M. Leach
2020

INCORPORATING SPATIAL STRUCTURE INTO BAYESIAN VARIABLE SELECTION USING SPIKE-AND-SLAB PRIORS WITH APPLICATION TO IMAGING DATA

JUSTIN M. LEACH

BIOSTATISTICS

ABSTRACT

Many applications in neuroscience, psychology, medicine, and public health require collecting and analyzing imaging data. While fine resolution images may appear continuous to the naked eye, they are in fact made up of measurements at discrete locations, either pixels in 2D or voxels in 3D. These data may be treated as the scientific outcomes of interest, or as predictors of outcomes of interest. Here, the concern is with the latter situation, which involves converting images into formats amenable to a linear modeling framework; i.e., each subject's image is converted into a vector, and used to model the subject's (scalar) outcome of interest.

Imaging data can complicate statistical analyses because using images as predictors can make traditional linear models invalid, or non-identifiable, for two reasons. First, images usually contain as many or more pixels or voxels, i.e., predictors, as subjects. Second, the measurements taken from images are usually highly correlated. Both situations violate the assumptions of traditional linear models. In contrast, Bayesian linear models can circumvent these issues but require principled approaches to selecting prior distributions. The primary goal of this work is to explore and develop a class of priors that is relevant to modeling scalar outcomes using images.

The outline of this work is as follows. After a review of literature, the first paper develops and presents an `R` package, `sim2Dpredictr`, for simulating data in a situation where images are used to model scalar outcomes; this package is freely available on CRAN. The second paper extends the spike-and-slab lasso prior by generalizing it to the spike-and-slab elastic net, and by incorporating spatial structure explicitly into variable

selection using Intrinsic Autoregressions (IAR) as priors; an `R` package to fit the models, `ssnet`, is available on github. We evaluate the proposed models with a simulation study by generating data with `sim2Dpredictr`. The final paper demonstrates the practical utility of the methods by applying them to classification problems using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Finally, we summarize the contributions of the dissertation, and discuss future research directions.

Keywords: Bayesian GLM, Bayesian Variable Selection, EM Algorithm, Image Analysis, Spike-and-Slab Lasso, Elastic Net

DEDICATION

This dissertation is dedicated to my immediate family. To my parents, Ken and Tamie, for their love and support, and for making me do math when I'd have preferred to play my Gameboy. To my brother, Jared, my oldest friend and closest confidant, always ready to join me in pondering the mysteries of the universe. To my wife, Megan, who has endured living with and supporting a stressed out graduate student for no less than half a decade, and yet has not run out of either patience or love. And to the little one on the way, already pushing me to finish up with work already so I can go out and play. I love you all.

ACKNOWLEDGMENTS

I would like to express my gratitude to my committee members, Drs. Lloyd Edwards, Rajesh Kana, Kristina Visscher, and Nengjun Yi for their support and suggestions that helped improve this project.

I would also be remiss if I did not thank the UAB Department of Biostatistics. During my time here many professors have opened their doors to offer advice and support. While certainly not exhaustive, I want to specifically thank Drs. Suzanne Judd, David Redden, Lloyd Edwards, Leanne Long, Erika Austin, Byron Jaegar, and Charles Katholi, each of whom listened, dispensed advice and wisdom, and in general kept me on my feet.

There are too many of my fellow students to thank, and I will inevitably do some a disservice by omitting their names here. At that risk and no particular order, I would nevertheless like to thank Holly Hartman, Amanda Pendegraft, Allison Fialkowski, Nichole Pompey, Boyi Guo, Rouba Chahine, Anastasia Hartzes, and Tarrant McPherson.

Finally, I would like to thank my advisor and mentor, Dr. Inmaculada Aban. Several years ago, I showed up at her office interested in imaging statistics, and we set off into a wilderness of research in neuroscience, neuroimaging, and spatial statistics. As with most dissertations, there were many times when it felt as though progress was impossible, which is a vivid illusion, but an illusion nonetheless. Setbacks are inevitable, and in these times Dr. Aban always redirected my attention to the things I could control, which allowed me to continue progressing, eventually moving past each setback.

The lone wolf dies. I have been fortunate not to be a lone wolf. I apologize to anyone I have neglected here; my sins of omission are a reflection of my memory's fallibility, not the weight of your importance. Thank you all.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table *Page*

sim2Dpredictr: AN R PACKAGE FOR SIMULATING SCALAR OUTCOMES WITH SPATIALLY DEPENDENT PREDICTORS

INCORPORATING SPATIAL STRUCTURE INTO INCLUSION PROBABILITIES FOR BAYESIAN VARIABLE SELECTION

THE SPIKE-AND-SLAB ELASTIC NET AS A CLASSIFICATION TOOL IN ALZHEIMER'S DISEASE

LIST OF FIGURES

Figure *Page*

LITERATURE REVIEW

sim2Dpredictr: AN R PACKAGE FOR SIMULATING SCALAR OUTCOMES WITH SPATIALLY DEPENDENT PREDICTORS

INCORPORATING SPATIAL STRUCTURE INTO INCLUSION PROBABILITIES FOR BAYESIAN VARIABLE SELECTION

THE SPIKE-AND-SLAB ELASTIC NET AS A CLASSIFICATION TOOL IN ALZHEIMER'S DISEASE

LIST OF ABBREVIATIONS

AUC	Area Under the ROC Curve
AD	Alzheimer's Disease
BhGLM	Bayesian Hierarchical Generalized Linear Model
BHM	Bayesian Hierarchical Model
CAR	Conditional Autoregression
CN	Cognitive Normal
EEG	Electroencephalography
EM	Expectation Maximization
EM-IWLS	Expectation Maximization Iterative Weighted Least Squares
FDR	False Discovery Rate
FDP	False Discovery Proportion
fMRI	Functional Magnetic Resonance Imaging
FWER	Family-Wise Error Rate
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
GMRF	Gauss-Markov Random Field
IAR	Intrinsic Autoregression
IWLS	Iterative Weighted Least Squares
LM	Linear Model
LMM	Linear Mixed Model
MAE	Mean Absolute Error

MCI	Mild Cognitive Impairment
MEG	Magnetoencephalography
MLE	Maximum Likelihood Estimation
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
MTP	Multiple Testing Procedure
MVN	Multivariate Normal
NPV	Negative Predictive Value
OLS	Ordinary Least Squares
PET	Positron Emission Tomography
PFE	Per-Family Error Rate
PCE	Per-Comparison Error Rate
pFDR	Positive False Discovery Rate
PPV	Positive Predictive Value
REML	Restricted Maximum Likelihood
RFT	Random Field Theory
ROC	Receiver Operating Characteristic Curve
SPM	Statistical Parametric Map
SSL	Spike-and-Slab LASSO
VLSM	Voxel-Based Lesion-Symptom Mapping

LITERATURE REVIEW

1 Introducing the Problem

Consider the problem where a random sample of n subjects has a scalar outcome Y_i, $i = 1, \ldots, n$, and predictors, or independent variables, consist of values at many spatial locations. For example, suppose each subject has a scalar outcome and a 50×50 image. Further, consider two wrinkles: Wrinkle 1 is that the subject images have spatial correlation (association). That is, for example, the value of the image at location (i, j) is 1 is likely not independent of the value of the image at location $(i, j + 1)$. An example is displayed in Figure 1, where the green pixels correspond to value 1 and the white pixels to value 0. While the data are binary in this case, it is clear that 0/1 values are not uniformly distributed in the two-dimensional space, and zeros are more likely to be surrounded by zeros than ones and vice versa. Wrinkle 2 is that the area a good model should identify as associated with the outcome is usually clustered, rather than disperse; an example is displayed in Figure 1, where the orange pixels correspond to the area that is actually associated with an outcome and the white pixels to the area that is not associated with the outcome. The question is how to best use the images like that in Figure 1 to identify the orange area in Figure 1. While the image in Figure 1 is presented as a matrix of zeros and ones, the problem at hand also encompasses matrices or arrays with non-independent entries made up of any real numbers. This document aims to explore solutions to the above problem and hopefully find a (as much as possible) unique approach to improving or expanding the existing methods.

The rest of the document is organized as follows: Section 2 introduces and briefly expounds upon the classical linear model framework. The general linear model and its

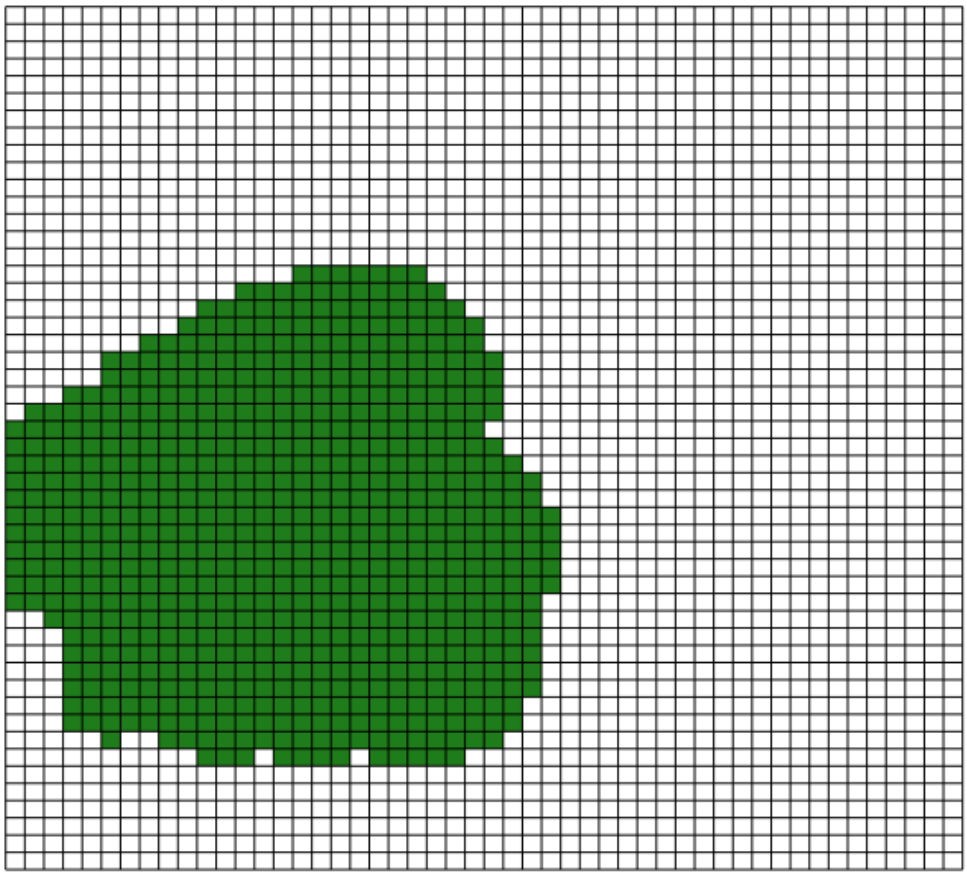

Figure 1: Example of a subject image matrix. Each box corresponds a location in the subject image matrix. In this example, locations that are green indicate data value 1 and locations that are white indicate data value 0; i.e. this is a matrix of spatially related binary predictors.

assumptions are elucidated and approaches for handling analyses when two key assumptions have been violated is discussed. Finally, an answer is given to the question of whether (in general) these models help with the problem described above.[1] Section 3 introduces Bayesian statistics by addressing its origin in the form of Bayes Theorem, and comparing

[1]It should come as no surprise that the answer is generally “No", or at the very least, not without some severe and possibly dubious adjustments.

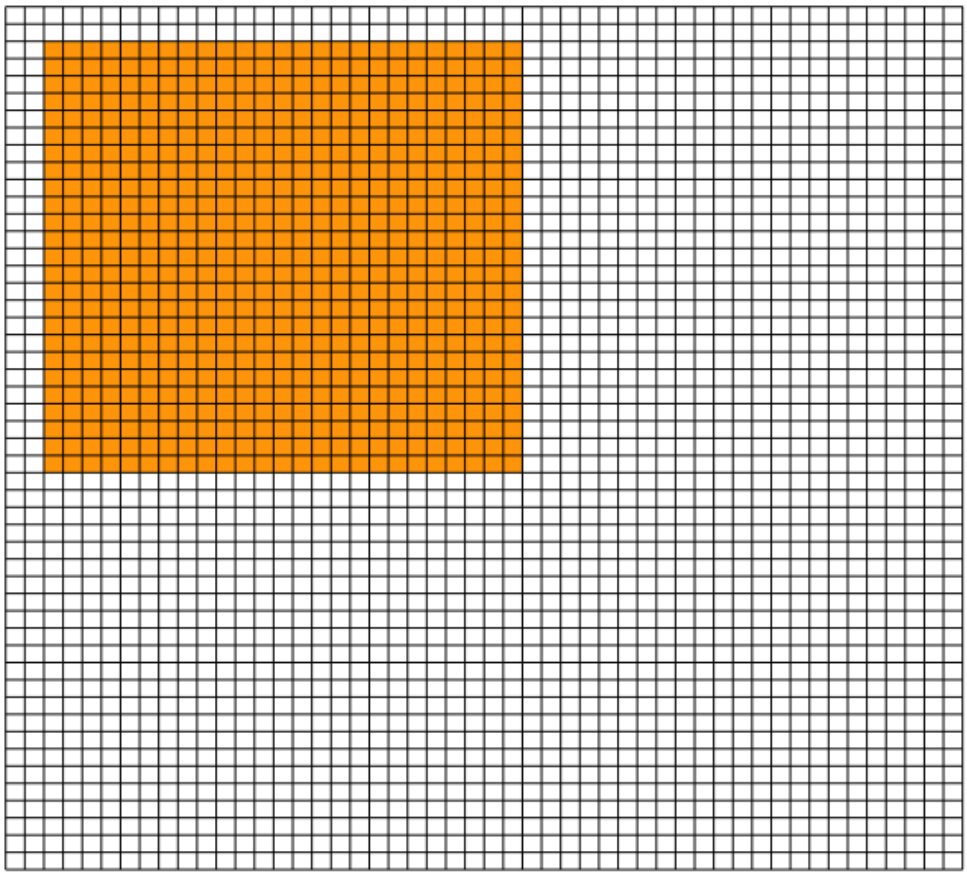

Figure 2: Example of a spatially clustered parameters. Each box corresponds to a location in the subject image matrix. The locations that are colored orange are locations that are truly associated with the outcome of interest. Those that are white are not truly associated with the outcome of interest.

and contrasting the differences in Classical vs. Bayesian estimation. We then discuss the Bayesian approach to linear modeling and discuss how the inclusion of prior distributions in the linear model setting can provide specific benefits in certain situations. Section 4 introduces the issue of multiple testing in statistics and distinguishes multiple testing from multiple comparisons. The discussion includes the viability of hypothesis testing in the

classical paradigm and the use of multiple testing procedures to control errors and discuss the relationship between multiple testing issues and Bayesian statistics. Section 5 describes a method from genetic research that employs Bayesian GLM's with shrinkage priors, and which holds promise for possibly being adapted to spatial data. This section discusses the general estimation framework, shrinkage priors and their purpose, and multiple testing in this particular Bayesian framework. Section 6 explores a recently developed method for Bayesian variable selection that combines the strengths of spike-and-slab priors with the LASSO, and discusses an adaptation of this approach to genetics.[2] This section also includes a brief discussion regarding the concept of penalized likelihoods, and spike-and-slab priors more broadly. Section 7 discusses how to convert an image into a format appropriate for a linear model and (attempts to) answer questions like "How does one incorporate spatial information into the Bayesian GLM's discussed in previous sections?", "How would one go about estimating the necessary parameters?" Section 8 provides an overview of some useful spatial statistics that will aid us in our task. Note also that the ultimate goal is to generalize the above problem in magnetic resonance images (MRI) applications, and so a brief overview on neuroimaging and MRI in particular is found in Appendix A.

2 Linear Models

2.1 Basic Linear Models

Linear models have become nearly ubiquitous in research, particularly when one desires to model how multiple variables, viewed together, affect a single outcome. That is, for $i = 1, \ldots, n$ subjects with outcomes Y_i and predictors $X_{i1}, \ldots, X_{iJ}$, employ the following model:[3]

$$Y_i = \beta_0 + X_{i1}\beta_1 + \cdots + X_{iJ}\beta_J + \epsilon_i \tag{2.1.1}$$

[2] This a mathematical discipline and so no one should be surprised that this approach was christened "The Spike-and-Slab LASSO".

[3] Unless otherwise specified, the information in this section can be found in Myers and Milton (1998), but basically any reputable linear models text will do just fine.

where $\beta_0, \ldots, \beta_J$ are parameters associated with the variables of interest and e_i is the random error. This model can be expressed in matrix form by stacking the elements of equation (2.1.1) to obtain the $n \times 1$ outcome vector $\boldsymbol{Y}$, the $n \times (J+1)$ design vector $\boldsymbol{X}$, a $(J+1) \times 1$ parameter and the $n \times 1$ vector of random errors $\boldsymbol{\epsilon}$:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.1.2}$$

This is the form of a general linear model. The ordinary least squares (OLS) estimate, i.e. minimizing the sum of squared errors[4] of $\boldsymbol{\beta}$ can be shown to be:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{2.1.3}$$

where $\boldsymbol{y}$ are the observed outcomes.[5] Since the model assumes the random error has mean 0 and variance $\sigma^2 \boldsymbol{I}$, it is only necessary to estimate σ^2. The sample variance is given as follows and can be shown to be unbiased for σ^2:

$$s^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n - (J+1)} \tag{2.1.4}$$

The presence of the inverse in equation (2.1.3) means that $\boldsymbol{X}$ must be full rank in order for $(\boldsymbol{X'X})^{-1}$ to exist; this implies linearly independent columns in $\boldsymbol{X}$. In practice this does not imply that there exists *no correlation* between these columns, only that no columns are a function of any collection of other columns. Generally, the correlation between predictors is non-zero, if small. However, larger magnitude correlations can ultimately cause problems, sometimes with identifiability and sometimes with interpreting parameter estimates. Predictors can be confounding, or in some cases, measures of (nearly) the same thing. Consider a model to study the association between lung function and footrace times:

[4] In practice, for observed data the sum of squared residuals is minimized.

[5] Following at least some people's convention, capital letters shall herein refer to random variables and lower case letters to observed values.

including data on both 400m time and 500m time are likely to interfere with each other since both are essentially measuring the same kind of fitness. Alternatively, consider modeling lung function with 400m and smoking status; smoking status likely affects 400m and lung function, and therefore one might expect a sizable correlation between these predictors, which could muddle interpretation of parameters or cause the model to be non-identifiable.[6]

Otherwise, this approach requires only that the random errors $\boldsymbol{\epsilon}$ have mean $\mathbf{0}$ and variance $\sigma^2\boldsymbol{I}$, i.e. they are independent with common variance, in order for the estimator $\hat{\boldsymbol{\beta}}$ to be a *Best Linear Unbiased Estimator (BLUE)* for $\boldsymbol{\beta}$.[7] However, further application, e.g. inference, interval estimation, and maximum likelihood proofs require one to further assume normality of the random errors. Normal random errors imply that the outcome is also assumed to be normally distributed with mean $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ and variance $\sigma^2\boldsymbol{I}$.

2.2 Generalized Linear Models (GLM)

Generalized linear models arise from the case where outcomes/errors are not normally distributed and/or do not have independent errors, but mostly the former, as a Generalized Least Squares estimate can be derived for non-independent errors.[8] GLM's are generally appropriate for data that belong to an exponential family, and they do not model the outcome explicitly, but rather model a function of the expectation, called the *link function*. The link function transforms the expectation to obtain a linear model:

$$h(\mathrm{E}(Y_i|X_{i1},\ldots,X_{iJ})) = \beta_0 + X_{i1}\beta_1 + \cdots + X_{iJ}\beta_J \tag{2.2.1}$$

[6] A related but different topic is the interaction between predictors in their association with an outcome. For example, suppose lung function is the outcome, and age and smoking status (Yes or No) as predictors. It could be the case the association between lung function and age is different depending on whether the subject smokes or not. If in equation (2.1.1), $X_{i1} = \begin{cases} 1 & \text{smoker} \\ 0 & \text{non-smoker} \end{cases}$ and $X_{i2} =$ subject age, then an interaction term is given by $X_{i3} = X_{i1} * X_{i2}$, and has parameter β_3. However, this does not immediately imply anything about the relationship between smoking status and age.

[7] The is shown via proving the Gauss-Markov Theorem. See Myers and Milton (1998).

[8] The problem here is that if you do not know the form and magnitude of the correlation between errors, then estimation is difficult and ultimately requires an approach different from that found in most introductory linear models textbooks. See McCulloch et al. (2008) and section 2.3.

where $h(\cdot)$ is the link function and the right hand side (RHS) is called the *linear predictor*, often denoted η_i. Link functions differ from transformations in that link functions are functions of the theoretical expectation, whereas a transformation is a function on the raw data itself. Many of the most useful distributions in the exponential family have a relationship between the mean and variance; that is, the variance is actually a function of the mean. This is true for some of the most commonly used distributions such as the Binomial, Poisson, and Gamma distributions.[9] In one sense this simplifies the process in that it (often) eliminates the need to estimate the variance separately, although it is possible to estimate an over-dispersion parameter, commonly given as ϕ;[10] however, when it comes time to maximize the likelihood, the relationship between the mean and variance prevents the existence of an analytic solution.[11] The method of *iterative weighted least squares (IWLS)* is an (obviously) iterative approach to solving for $\boldsymbol{\beta}$ to obtain $\hat{\boldsymbol{\beta}}$ that sidesteps the analytic issues (McCulloch et al., 2008). This approach also has the added benefit of asymptotic normality for the parameter estimates, allowing for Wald Tests to be used inference.[12] Note that ultimately similar concerns about the design matrix $\boldsymbol{X}$ arise in the case of the GLM because one must still calculate an inverse that is a function of $\boldsymbol{X}$, which implies that $\boldsymbol{X}$ must be full rank. Ultimately, GLM's are most commonly employed to address estimation when the outcome of interest is binary[13] or count data.[14]

[9] See Appendix C.

[10] The technical use of this parameter is apparently not well agreed upon. For some, it is a dispersion parameter and thus naturally 1 for distributions such as the Binomial and Poisson distributions, but naturally needs to be estimated for distributions such as the Gamma distribution. Others prefer to label ϕ as "over-dispersion" so it is always something estimated in addition to that which is standard to a theoretical distribution.

[11] i.e. it is impossible to isolate $\hat{\boldsymbol{\beta}}$ as in equation (2.1.3). Rather, the result is an equation with $\boldsymbol{\beta}$ on both sides with no (analytic/algebraic) way to get it by itself on one side of the equation.

[12] Other inference approaches are available; e.g. Likelihood Ratio Test. In addition, IWLS is not the only approach to estimation. See McCulloch et al. (2008), chapter 5.

[13] e.g. Consider a model where the outcome is pass/fail on some test and the predictors are college major and whether or not the subject took a Red Bull beforehand.

[14] e.g. Consider a model where the outcome is the number of hospital visits within a year and the predictors are alcohol consumption and age.

2.3 Correlated Errors

Normally distributed data

Section 2.1 briefly addressed non-normal random errors, and often this situation implies a functional relationship between the mean and variance. While the reader is referred to a textbook, McCulloch et al. (2008), for further details, it was claimed that parameter estimates could be obtained via IWLS. Another common occurrence is correlated random errors, which can arise when the same outcome is measured multiple times for a subject, or alternatively, when subjects within some cluster can be assumed to have a correlation in outcomes/random errors. When the outcome is normally distributed, the model is an extension of equation (2.1.1) into a multivariate case:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon_i} \tag{2.3.1}$$

where $\boldsymbol{Y}_i$ is a $n_i \times 1$ vector of observations for subject/cluster i, with $n_i \times (J+1)$ design matrix $\boldsymbol{X}_i$, $(J+1) \times 1$ parameter vector $\hat{\boldsymbol{\beta}}$, and $n_i \times 1$ random error vector $\boldsymbol{\epsilon}_i$. Now, assume that $\boldsymbol{\epsilon}_i$ is normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_i$, which may not be equal to a scalar multiplied by an identity matrix. Stacking equation (2.3.1) across all $i = 1, \ldots, n$ obtains equation (2.1.2) with the difference being that the variance of $\boldsymbol{\epsilon}$ is no longer $\sigma^2\boldsymbol{I}$, but rather a block diagonal consisting of the $\boldsymbol{\epsilon}_i$ from each subject, denoted $\boldsymbol{\Sigma}$. A correlation structure can be assumed for $\boldsymbol{\Sigma}_i$, which may then be estimated and used for inference. It is also common to partition the error into random effects and within subject error which results in a model format known as a *Linear Mixed Model (LMM)*. In the case of longitudinal data, this partitioning allows that there are population level effects (the "fixed effects" $\boldsymbol{\beta}$), but that individuals or clusters actually have their own deviation from this population level effect (the so-called random effect), from which they will vary due either to measurement error and/or random error. Whatever approach one chooses (partitioning error or not), estimation runs into similar issues as the GLM case in that the estimating of the

(co)variance results in analytically intractable situations, requiring numerical methods for obtaining solutions. Additionally, while estimation in the GLM case still relies on Maximum Likelihood (and technically so can estimation here), an alternative approach known as *Restricted (Residual) Maximum Likelihood (REML)* transforms the model to contain none of the fixed effects and then performs maximum likelihood estimation to obtain (co)variance estimates which may then be used to estimate a solution for $\hat{\beta}$; this process reduces, but does not eliminate, the bias in variance estimation that results when using maximum likelihood.[15]

As an example (and hopefully somewhat intuitive at that), consider a situation where each subject runs 400 meters as fast as possible on 4 separate Saturdays in July. Subjects are randomized to take either Red Bull, Beer, Coffee, or Water before their sessions (the same beverage each time). The fixed effect accounts for population trajectories, i.e. for each beverage group, on average did those subjects improve (or not) and by how much? The random effects accounts for the fact that difference subjects are likely to individual specific differences in trajectory of (hopefully) improvement. The within subject error accounts for the fact that there is likely additional variation not accounted for by individual differences. That is, there is some variation about a subject's time that may be tied to neither individual fitness nor beverage.

Non-Normal data with correlated errors

Non-normal data tends to be even more difficult to handle when errors, and thus outcomes, are not independent. This is generally beyond the scope of this document, but similar to LMM's, one may attempt to model random effects or not, i.e. partition the random error or not. If the random error is not partitioned it is common to employ Generalized Estimating Equations, which depend on a kind of quasi-likelihood to specify the relationship

[15] See McCulloch et al. (2008) for theoretical details. Essentially, the problem is that the ML solution estimates the variance components as if we know β, but we have estimated β, which introduces additional variation that ML ignores. The beauty of REML is that the variance components are estimated without respect to particular values of β, which alleviates the bias introduced by pretending we know β. In practice, this does not often noticeably change the estimate $\hat{\beta}$, but since variance estimates can be different, the results of hypothesis tests can change noticeably.

between mean and variance, and which employs an ad-hoc addition of a working correlation matrix to account for dependence between observations (Liang and Zeger, 1986). One may alternatively specify a *Generalized Linear Mixed Model (GLMM)*, which combines the LMM and GLM approaches described above; however, estimation methods are typically more complicated and there is no general agreement about the best optimization approach (Tuerlinckx et al., 2006).[16]

2.4 Application to the Question at Hand

None of the models as described will be able to handle the general data structure described in Section 1. Why? The first issue is that in many situations the design matrix will be non-identifiable, either because the number of predictors in the image will exceed the number of subjects ($J > n$) or because the associations between the predictors may be relatively strong in many cases. A second point is that the frameworks for handling correlations discussed in this section are related to correlations between errors, which in reality arise due to the correlated nature of the collected outcomes. These approaches cannot immediately address image data unless the image is used as the outcome. Possible practical goals in thinking about such a question might be using MRI images to predict disease progression or to determine what areas of the brain are associated with some behavioral outcome. However, a key difference between this situation is that the spatial structure of the predictors and their parameters actually contain information about how likely it is that a location is associated with the outcome of interest. If it is known that none of the locations surrounding a particular location are associated with the outcome then that provides some information about whether this particular location is associated with the outcome. Ultimately, the intention is to use the available spatial information in inference. This will be the ultimate task of this work; however, in order to address this task, we must first review Bayesian approaches to statistics, as well as some aspects of established spatial statistics.

[16] Despite this difficulty, one should not be dissuaded from fitting the model one deems most appropriate. Read Tuerlinckx (2006) or consult your friendly neighborhood statistician.

3 Bayesian Approaches to Modeling

3.1 Basics Bayes

Bayesian statistics is built upon exploiting the rules of conditional probability. Following Casella and Berger, consider the definition of the conditional probability of event A given event B from some set of possible events S:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{3.1.1}$$

It is straightforward to see that $\Pr(A \cap B) = \Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A)$, which may be incorporated into equation (3.1.1) to find:

$$\Pr(A|B) = \frac{\Pr(A|B)\Pr(A)}{\Pr(B)} \tag{3.1.2}$$

The generalization of this insight is called *Bayes' Theorem*. Let $A_1, A_2, \ldots$ be a partition of the sample space, S and B be any set. Then for each $i = 1, 2, \ldots,$

$$\Pr(A_i|B) = \frac{\Pr(B|A_i)\Pr(A_i)}{\sum_{j=1}^{\infty} \Pr(B|A_j)P(A_j)} \tag{3.1.3}$$

Note that in Bayesian modeling equation (3.1.3) can be generalized to the continuous case such that the denominator contains an integral rather than a summation, and the probabilities are replaced by probability densities.

3.2 Classical vs. Bayesian Estimation

Bayesian philosophy, hypothesis testing, and specification of priors

In its simplest form, the disagreement between Frequentists and Bayesians is best understood by thinking about the assumptions placed on the parameters, $\boldsymbol{\theta}$, associated with a particular distribution. Frequentists assume that this parameter vector contains *constants*, which is estimated, and whose estimates shall have some variation requiring consideration;

however, if the sample size went to infinity, the variation around these parameters would disappear, resulting in the true and exact parameter values instead of an estimate. To be clear, in frequentist approaches it is the parameter estimates that vary, not the parameters themselves. The label "frequentist" is better understood through traditional probability exercises, say a coin flip. The assumption is that if a coin is fair, then in a finite sample the proportion of heads likely differ from the proportion of tails, but the longer the series of coin flips the closer these proportions will get to 1/2, and an infinite series of coin flips would produce 1/2 exactly. The parameter, in this case the probability of a heads (or equivalently, a tails for a fair coin), is determined by the relative frequency of flips that ends on heads.

The Bayesian approach fundamentally rejects the idea that the parameters are constants, with only their estimation varying on account of finite sample sizes. Bayesians argue that parameters of interest, $\boldsymbol{\theta}$, themselves have distributions with constant parameters, called *hyperparameters*;[17] that is, $\boldsymbol{\theta}$ is not a constant, but rather a random variable with a distribution of its own.[18] Bayesians prefer to think about probabilities less as frequencies and more as approaches for measuring and quantifying "uncertainty", particularly in cases where the frequency framework requires mental gymnastics. For example, Jeffreys and Berger describe probability in Bayesian terms as a "measure of plausibility of a hypothesis or proposition" (Jeffreys and Berger, 1992); e.g., given some seismic data, what is the probability of an earthquake occurring? A frequentist might counter by asking, given this seismic data, how often does an earthquake follow? However, interest is not always focused on the large sample qualities of how often an earthquake follows, but rather on the probability that an earthquake occurs *in this particular case*.

[17](Depending on the depth of hierarchy, it may take a few levels to reach the constant parameters.

[18]Some authors, e.g., Greenland (2006), argue that it is incorrect to say that "parameters are treated as fixed by the frequentist and as random by the Bayesian". Nevertheless, mathematically speaking, this is exactly the case; Bayesians estimate a parameter's posterior distribution, while frequentists find a point estimate for a parameter. What this distinction means for interpretation in practice may vary. In Greenland's subjective approach, the uncertainty is always around a parameter's value, and not a "property" of the parameter, but the mathematics do not confine us to this particular interpretation.

In the simplest application of Bayesian probability, there exists a set of mutually exclusive hypotheses, each of which is assigned a prior probability, collect data, and then update the probabilities that each hypothesis is true; i.e., for i hypotheses H_i, and data $\mathbf{y}$ there is the following application of Bayes theorem:

$$\Pr(H_i|\mathbf{y}) = \frac{\Pr(\mathbf{y}|H_i)\Pr(H_i)}{\Pr(\mathbf{y})} \tag{3.2.1}$$

This differs from frequentist interpretations because a frequentist would say that the true hypothesis has probability 1 after data have been collected, while all others have probability 0; the Bayesian, however, is giving each hypothesis a "probability" that attempts to quantify the uncertainty about which hypothesis is true.[19]

However, simple framework of equation (3.2.1) is a rather poor description of (many) modern applications of Bayesian statistics and has received much criticism.[20] Gelman and Shalizi (2013) explain in detail the common inadequate and inappropriate conceptions that surround Bayesian methodology. Often, Bayesian statistics is described as assigning a prior distribution, which incorporates the "prior beliefs" of the researchers into the model, which may then be updated when new data is collected. The idea is then that over time the updating shall converge upon the "true hypothesis" or true distribution of some parameters $\boldsymbol{\theta}$. However, as is the case with other types of models, the model is virtually always false in the strict sense and therefore to some degree misspecified. The degree to which the model is misspecified determines whether the model may result in useful inference, or not.

[19] This particular application lines up fairly well with the interpretation of Greenland (2006), who we maligned in a previous footnote. Here we assume that one of the hypotheses is true, we just do not know which one. Our point is not that such an interpretation is always incorrect, but that it is not always implied by using Bayesian statistics.

[20] See for example Mayo (1996), who argues strenuously against Bayesian approaches to hypothesis testing, with particular focus on philosophy of science (Mayo, 1996). Many of the mathematical justifications and arguments for Bayesian applications in experiment and hypothesis testing can be found in Berger and Wolpert (1988), but this is largely beyond the scope of this document.

A key objection to using equation (3.2.1) in practice is that it is possible for the prior distribution to *not include all possible truths*. As in frequentist approaches, models can be checked for goodness of fit, often by simulating from the posterior distribution to determine whether the model would be likely to produce the observed data. Additionally, the role of the prior distribution as representing the researchers' "prior beliefs" is often misleading. The prior is more similar to a regularization device, which causes fitted models to be less sensitive to various details in the data. Prior distributions, in the course of model checking, are as testable as any other aspect of the data. A better outlook is to take the perspective that " 'the model', for a Bayesian, is the combination of the prior distribution and the likelihood, each of which represents some compromise among scientific knowledge, mathematical convenience, and computational tractability" (Gelman and Shalizi, 2013). The general point is that Bayesian models are not necessarily about "updating" one's subjective beliefs, but rather they are flexible tools for modeling uncertainty, which have assumptions that should be stated and checked for accuracy.

Bayesian approaches are often useful in cases where classical approaches fail in some way, but where the data still contain useful information; e.g., see Gelman et al. (2008), whose weakly informative priors can help address the issue of complete separation in logistic regression. That is not to say that there are not circumstances where equation (3.2.1) can be a useful approach or that no one advocates the use of prior distributions as a way to incorporate prior beliefs, but only that in many practical applications a more nuanced approach is available. For example, Greenland (2006) shows how relatively simple applications of "subjective" Bayesian statistics can be useful in epidemiology by exploring how in many cases prior distributions can be represented directly as prior data, which can inform reasonable prior distribution choices, and can aid interpretation of the strength of results. Furthermore, Greenland discusses how the assumptions of many traditional methods can fail in practice, and how Bayesian approaches can force a reckoning with the assumptions researchers make in every day analysis. It may even be possible to evaluate

model performance in somewhat of a traditional framework, e.g., in simulation studies does a procedure obtain the correct answer with reasonable frequency? For example, if the expression of some gene is associated with a particular disease, does the model tend to discover that fact?

Estimation approaches

Classical, or Frequentist, approaches to estimation typically assume that a random sample of observations is drawn from some distribution, from which the *likelihood function* is constructed: $L(\boldsymbol{\theta}|\mathbf{y})$, which is the joint distribution of the data, $\mathbf{y}$, as function of some set of parameters $\boldsymbol{\theta}$, given $\boldsymbol{y}$. In the case of independent and identically distributed (i.i.d.) random variables this is simply the product of the individual random variables; i.e. for a sample of n i.i.d. random variables with probability distribution function $f(y|\boldsymbol{\theta})$, the likelihood is given by:

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}) \tag{3.2.2}$$

One then maximizes the likelihood as a function of $\boldsymbol{\theta}$, given the data $\mathbf{y}$, in order to obtain an estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$. This approach essentially tells us what the most likely parameter values are, given the collected data. This is the essence of classical *maximum likelihood estimation*, which assumes there exist unique constant values for all entries in $\boldsymbol{\theta}$; the uncertainty is baked in the fact that we take a finite sample, and so we will not have $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$.[21]

Bayesian statistics still use the likelihood function, but the addition of a prior distribution changes the situation: the goal is now to estimate the *distribution* of $\boldsymbol{\theta}$ given the data $\mathbf{y}$; i.e. the assumption is that $\boldsymbol{\theta}$ is a vector containing random variables rather than constants. What follows are some key ideas and concepts taken from Gelman et al.

[21] It is not implied that maximum likelihood estimation guarantees that estimates are unbiased, i.e., $\mathrm{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$, or consistent, i.e., that as the sample size goes to infinity $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}$. This is often the case, but one must go through the proofs to be sure.

(2013). Estimation is achieved by beginning with Bayes' Theorem. Adopting the Bayesian convention of $p(\cdot)$ to denote the probability density, the posterior distribution is given by:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta}, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{3.2.3}$$

The whole of Bayesian statistics arises from this relatively simple formulation.[22] The density function $p(\boldsymbol{\theta})$ is called the *prior distribution*. This may incorporate prior information about the parameters or may be assigned to be "non-informative" or flat. The *posterior distribution* is given by $p(\boldsymbol{\theta}|\mathbf{y})$, and is the "updated" distribution of the parameters given the observed data. While the notation has changed, notice that $p(\mathbf{y}|\boldsymbol{\theta})$ is called the *likelihood* as it actually has the same form as equation (3.1.3). Lastly, $p(\mathbf{y}) = \sum_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})$ is the marginal distribution of $\mathbf{y}$, and it is often treated as normalizing constant that does not require attention; when distributions are continuous rather than discrete the summation is replaced by an integral. Therefore, one often deals with the *unnormalized* posterior density, as indicated by equation (3.2.3).

In general, the posterior distribution can be thought of as a compromise between the information contained in the data and prior information (Gelman et al., 2013). Therefore, the form of prior can matter to inference. If no prior information is available or trustworthy, then the prior distribution should not affect inference; assigning flat or non-informative priors is not always an easy task, especially for parameters who can take on any real value. Consider that if the parameter space spans all $\mathbb{R}$, then assigning a uniform prior will result in an *improper distribution*, i.e., its pmf/pdf will not sum/integrate to one. These improper priors may be used, but one must take care to assure that the posterior distribution is proper.

In some cases, the distributional form of prior has an obvious choice, whether or not there exists information on its parameter values. For example, with the Binomial distribution,

[22] This does not imply that the resulting wing of statistics is simple; many posterior distributions lack closed form and of those distributions that have closed form, the derivations are often quite nasty.

whose parameter is a probability, it makes sense to use the beta distribution as a prior, since its support is continuous and bounded between 0 and 1. An important concept relating to prior specification is that of *conjugacy*, where the posterior distribution has the same parametric form as its prior. This applies to the binomial model with a beta prior: both the prior and posterior distribution of θ have a beta distribution and the beta prior distribution is a *conjugate family* to the binomial likelihood. In general, any distributions that can be expressed in the form of an exponential family have *natural conjugate prior distributions*. It is also the exponential families that typically have closed form solutions for their posterior distributions.

A further difference between Bayesian and Frequentist approaches is that of Hierarchical Models, where some confusion can arise. Hierarchical models arise in the general scenario where $Y|N \sim f_Y(N, \theta)$ where in addition $N \sim f_N(\lambda)$. Note that in some cases N may be a vector rather than a scalar, but for simplicity, let N, θ, λ be scalar quantities. This situation can arise without Bayesian philosophy and where the marginal distribution of Y can be calculated. A (fully) Bayesian approach would make two key changes. First, the frequentist account can allow N to vary, but θ to remain an unknown constant; the Bayesian account would require that θ take a prior distribution. Secondly, λ would not be treated as a constant, but also have a prior distribution. Traditional examples of this setup often involve Y being the number of surviving offspring in some group and N being the number of offspring in that group; even without Bayesian accounting, one can see that both of these quantities would be random variables in practice. The Bayesian says that additionally, for any parameters, an additional layer of variability, or uncertainty, should be modeled. For example, if θ is the probability of survival and λ is the mean number of offspring, a Bayesian would reject the notion that this is or could be understood as a constant value.

4 Multiple Testing Procedures (MTP)

When one simultaneously performs a large number of hypothesis tests, typically at least some of these will result in rejecting the null hypothesis, even in cases where *none* of the tests should have been rejected. Obviously, it is ideal to include no false positives, and at least desirable to minimize the number of false positives included. Making matter worse, the probabilities of obtaining false positives are less and less intuitive as the number of tests arise. Multiple testing procedures are an attempt to quantify the expected number of false positives under various (joint) null hypothesis formulations, and to correct the results to minimize the probabilities that false positives are included in the results. In this section we review statistical hypothesis testing and discuss several approaches to minimizing false positives.

4.1 Hypothesis Testing Review

A return to the idea of the hypothesis test itself is necessary in order to understand what purpose may be served by correcting for multiple hypothesis tests. Traditional hypothesis testing is an outgrowth of classical (frequentist) statistics that assumes some distribution f, which has parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. In data collection the distribution from which the data are drawn is often, though not always, clear. Parameter values for the distribution are then estimated and compared to some hypothesized parameter values to determine whether these differ significantly from each other. In the case of a single parameter, this consists of a null hypothesis $H_0 : \theta = \theta_0$ and an alternative hypothesis, generally one of the following:

1. 2-sided: $H_A : \theta \neq \theta_0$
2. 1-sided: $H_A : \theta > \theta_0$ or $H_A : \theta < \theta_0$

These can also be generalized to test multiple parameters at once. The essence of hypothesis testing is to *assume the null hypothesis is true*, i.e., determine what the world would/should

look like if the null hypothesis were true, and then evaluate whether the results are consistent with that hypothesis/worldview. If yes, then the test fails to reject the null hypothesis, but notably has not *shown* there is no difference, but rather have said that at the very least the results are not that inconsistent with what the world would look like if the null hypothesis were true. If no, then the test rejects the null hypothesis in favor of the specified, usually composite, alternative.

Generally, judgments regarding rejection proceed by first determining the distribution of a test statistic under H_0 and then calculating the probability of sampling a value as or more extreme than the one observed; this is the infamous p-value. Before testing one specifies a value, α, and rejects H_0 when the p-value is less than α; that is, if the probability of drawing a value as or more extreme under H_0 is sufficiently low, then reject H_0. One sets α based an acceptable type I error rate: that is, the rate of false positives, or false rejections of the null hypothesis, that is deemed acceptable. While the form of null distribution will vary and in some cases become quite complicated, the basic idea is always essentially to compare the sample estimates with some hypothesized value. Even if the hypothesized values are true, sometimes the tests incorrectly determine that they are not the true values and it is necessary to make some judgment on the likelihood of this having occurred, at least in the long run. The other key error to avoid is failing to reject H_0 when it is false; these are called Type II errors, or false negatives. At a given sample size, one may not arbitrarily lower false negative rates without increasing false positive rates, and vice versa. This language is often confusing and results in incorrect interpretations of the p-value. It is important to understand that false positive/negative rates are properties of the procedure rather than the sample; that is, if $\alpha = 0.05$, then a procedure that incorrectly rejects H_0 5% of the time given that H_0 is true is deemed acceptable. This is not equivalent to a statement about the probability that a particular sample has incorrectly rejected H_0, i.e. the p-value is not the probability that the result is a false positive. This latter probability is either 0 or 1, after sampling, whereas the probability of drawing a sample that yields an incorrect rejection is 5%. In general, a

procedure is good if when H_0 is true, it says so, and when H_0 is false it says so. The error rates are properties of a particular hypothesis testing procedure in two different conditions; given that the procedure could reasonably encounter both situations, it is necessary to tweak the procedure so that it performs relatively well in either circumstance.

4.2 The Logic of Multiple Testing Procedures

Multiple testing and multiple comparisons situations arise when multiple hypothesis tests are conducted simultaneously, or may be treated as such. Consider m independent hypothesis tests, each of which is tested at level α. By the logic described in the previous section, it is clear that $\alpha * m$ tests are expected to be false rejections when all m tests are truly null; e.g., for 20 tests where the null is true for each, on average 1 hypothesis test erroneously rejected. Conversely, the probability of obtaining at least 1 false positive when all m are true is 1 minus the probability of not rejecting any tests falsely, which under independence is given by $1 - (1 - \alpha)^m$; e.g., for 20 truly null tests, the probability of obtaining at least 1 false positive is $1 - (0.95)^{20} = 64.15\%$.[23] Thus, as the number of independent (and true null) tests, m, increases the probability of including at least one false positive approaches one.

Families of tests and error rates

Multiple testing focuses on *families of tests*. A useful definition of "family" is given by Hochberg and Tamhane (1987) as "[a]ny collection of inferences for which it is meaningful to take into account some combined measure of errors". This idea will be revisited in discussing *when* multiple comparisons procedures are necessary or useful. Families of tests may consist of hypothesis tests related to multiple related outcomes (e.g., a clinical trial with multiple endpoints), multiple categories within linear models (e.g., did any of number of treatments outperform the placebo), or pairwise comparisons between many

[23] Strictly speaking, the probability is based on taking repeated samples such that if in each sample the 20 tests are independent then $\sim 64\%$ of samples will contain at least one false positive.

sets of categories (e.g., how do each of multiple treatments compare to each other). The key insight is that families should consist of tests that make sense to be viewed together.

Error rates can be quantified in several ways, but for now, we focus on type I errors, or false positives, since this is the error rate that increases in the presence of multiple tests. How these procedures affect the type II error, or false negative rate, will be discussed throughout.[24] When there is a family of tests the *error rate* should be controlled, but what is meant by such a term? In the case of a single test, false positives are controlled by selecting a sufficiently low significance threshold, α, so that the null hypothesis is rejected in only $\alpha\%$ of samples where the null hypothesis is true.

The most basic error rate to control is the *family-wise error rate (FWER)*, which is the probability of at obtaining least one false positive in family of tests. For some family of tests, $\mathcal{F}$, and multiple comparison procedure, $\mathcal{P}$, the formal definition of FWER is as follows (Hochberg and Tamhane, 1987):

$$\text{FWER}(\mathcal{F}, \mathcal{P}) = \Pr(M(\mathcal{F}, \mathcal{P}) > 0) \tag{4.2.1}$$

where $M(\mathcal{F}, \mathcal{P})$ is the number of wrong inferences, most commonly treated as the number of false positives.[25] Other alternatives include:

1. *Per-family error rate (PFE)*, which is the expected number of false positives:

$$\text{PFE}(\mathcal{F}, \mathcal{P}) = \mathrm{E}(M(\mathcal{F}, \mathcal{P})) \tag{4.2.2}$$

[24] Note that the power of a test to detect a particular deviation from the null hypothesis is one minus the false negative rate. In many works, including this one, it is ultimately more intuitive to talk about power and depending on the context, one or the other term is employed.

[25] Strictly speaking, the number of wrong inferences could involve false negatives, but this is not a standard interpretation of the FWER.

2. *Per-comparison error (PCE) rate*, which is the ratio of the PFE to the cardinality of $\mathcal{F}$, $N(\mathcal{F})$:

$$\text{PCE}(\mathcal{F}, \mathcal{P}) = \frac{\text{E}(M(\mathcal{F}, \mathcal{P}))}{N(\mathcal{F})} \tag{4.2.3}$$

Note that this error rate is what is controlled in classical hypothesis testing with no MTP/MCP correction.

3. *False Discovery Rate (FDR):* The FDR, as well as methods for its control, was introduced by Benjamani and Hochberg and is the expectation of the proportion of incorrectly rejected null hypotheses (Benjamini and Hochberg, 1995). Define the random variable $\mathbf{Q} = \frac{\mathbf{V}}{\mathbf{V}+\mathbf{S}}$ where $\mathbf{V}$ = the number of incorrect rejections and $\mathbf{S}$ is the number of correctly rejected null hypotheses. Then the FDR, Q_e is $\text{E}[\mathbf{Q}]$:

$$Q_e = \text{E}[\mathbf{Q}] = \text{E}\left[\frac{\mathbf{V}}{\mathbf{V}+\mathbf{S}}\right] = \text{E}\left[\frac{\mathbf{V}}{\mathbf{R}}\right] \tag{4.2.4}$$

where $\mathbf{R} = \mathbf{V} + \mathbf{S}$. In common parlance, the FDR is the expected proportion of false positives in a set of rejected hypotheses.

It is clear that PCE $\leq$ FWER $\leq$ PFE. Infinite families are theoretically possible, but finite families are nearly always sufficient in practical application. Under mutually independent inferences and error rate α, PCE $= \alpha$, PFE $= \alpha N(\mathcal{F})$, and

$$\text{FWER} = 1 - (1-\alpha)^{N(\mathcal{F})} \tag{4.2.5}$$

Note that $(1-\alpha)^{N(\mathcal{F})}$ is the probability of obtaining no false positives.

4.3 Controlling Error Rates

Family-wise error control

Following Nichols and Hayasaka (2003), suppose a subset of test statistics, $T = \{T_i\}$, is obtained for $i \in \mathcal{V} = \{1, \ldots, V\}$ which is used to test the family of hypotheses

$$H_i = \begin{cases} 1 & \text{Alternative Hypothesis is True} \\ 0 & \text{Null Hypothesis is True} \end{cases}$$

That is, for each of V tests, the null hypothesis is either true ($H_i = 0$) or not ($H_i = 1$); the task is to identify whether $H_i = 0$ or $H_i = 1$. Let H_0 be the case where all V null hypotheses are true, i.e. $H_i = 0, \forall i = 1, \ldots, V$. An α-level test for test i is to reject when $\Pr\{T_i \geq u | H_i = 0\} \leq \alpha$ for rejection threshold u. FWER is controlled in one of two senses:

1. **Weak** control requires control under the complete null hypothesis; that is, when the null hypothesis is true for the entire family of hypotheses:

$$\Pr\left(\bigcup_{i \in \mathcal{V}} \{T_i \geq u\} | H_0\right) \leq \alpha \tag{4.3.1}$$

   This means that if there is in reality no tests for which the null hypothesis is false, then the probability of at least 1 false positive is less than or equal to α.

2. **Strong** control requires control for the false positives when the null hypothesis is true for any subset of the family of hypotheses, $\mathcal{V}_0 \subset \mathcal{V}$:

$$\Pr\left(\bigcup_{i \in \mathcal{V}_0} \{T_i \geq u\} | H_i = 0, i \in \mathcal{V}_0\right) \leq \alpha \tag{4.3.2}$$

   This means that if in reality the null hypothesis is true for some tests and false for other tests within a family of tests, then the rate of false positives in the subset where

the null hypothesis is true is less than or equal to α. It makes no explicit statement about whether a procedure correctly identifies the tests where the null hypothesis is truly false; i.e. in the extreme example, reject no tests at any level, which strongly controls the FWER at $\alpha = 0$, but fails to identify a single true positive.

The commonly termed "omnibus" test, e.g. the test of model significance compared to an intercept only model in a regression, exhibits weak control, but strongly controlled tests allow for rejections on individual tests, which is often exactly what is desired. That is, weak control allows rejecting at least 1 test, but cannot necessarily say which one(s), while strong control allows specification of which one(s) should be rejected.

The FWER is closely associated with the distribution of the maximum test statistic, $M_T = \max_i T_i$, with the idea being that at least one test shall exceed a threshold if and only if M_T exceeds the threshold:

$$\bigcup_i \{T_i \geq u\} = \{M_t \geq u\}$$

In order to control the FWER at level α, define $\mu_{\alpha_0} = F^{-1}_{M_T|H_0}(1-\alpha)$, which defines the $100(1-\alpha)$ percentile of the distribution of the maximum test statistic under the complete null hypothesis, i.e., when the null hypothesis is true for all tests in the family. This results in weak control of the FWER:

$$\begin{aligned} \Pr\left(\bigcup_i \{T_i \geq u_\alpha \middle| H_0\}\right) &= \Pr(M_T \geq u_\alpha | H_0) \\ &= 1 - F_{M_T|H_0}(u_\alpha) = \alpha \end{aligned}$$

Note also that u_α has strong control of the FWER, assuming pivotality, which holds for a family of tests if the null distribution of a subset of tests, $\mathcal{V}_0 \subset \mathcal{V}$, is not dependent on the state of other null hypotheses. That is, we have the following:

$$\Pr\left(\bigcup_{i\in\mathcal{V}_0}\{T_i \geq u_\alpha\}\middle| H_i = 0, i \in \mathcal{V}_0\}\right) = \Pr\left(\bigcup_{i\in\mathcal{V}_0}\{T_i \geq u_\alpha\}|H_0\right)$$
$$\leq \Pr\left(\bigcup_{i\in\mathcal{V}}\{T_i \geq u_\alpha\} \geq u_\alpha|H_0\right) = \alpha$$

The maximal test statistic is an important part of permutation testing in neuroimaging, and more generally in classical approaches to solving spatial problems similar to that presented in section 1.

The Bonferroni correction

One of the most common multiple testing procedures is the Bonferroni correction. This correction is based on the the heuristic "Union-Intersection" Method (Hochberg and Tamhane, 1987). In such case, for a family of V tests the complete null hypothesis, $H_0 = \cap_{i=1}^{V} H_{0i}$, which hypothesizes that the simultaneous truth of V hypotheses. This may be compared to the alternative $H_1 = \cup_{i=1}^{V} H_{1i}$, which says that if any one of the V hypotheses is false then one broadly rejects H_0.

For such a test to have level α,[26] using appropriate test statistics, $T = \{T_i\}$ for $i \in \mathcal{V} = \{1, \ldots, V\}$, it must be the case that:

$$\Pr_{H_0}\{|T_i| > \xi_i, \text{for some } i = 1, \ldots, V\} \leq \alpha \tag{4.3.3}$$

It is common to specify $\xi_i = \xi$ for all i, which implies that individual rejection levels, $\alpha_i = \Pr_{H_{0i}}\{|T_i| > \xi_i\}$, are also the same. Following Hochberg and Tamhane (1987), the Bonferroni inequality is produced from the first order approximation of Boole's formula.

[26] From Casella and Berger (2002), a test is of *level* α, $\alpha \in [0, 1]$, if the test's power function, $\beta(\theta)$, satisfies $\sup_{\theta\in\theta_0} \beta(\theta) \leq \alpha$, where $\theta \in \theta_0$ define some parameter values under the null hypothesis. A *size* α test changes the inequality to an equality and so if a test is a size α test then it is a level α test, but not necessarily vice versa. In the case of a single test, the size of the test essentially specifies the probability that one would find a test statistic whose absolute value is greater than some possible value, under the null hypothesis. This is generalized to family of tests here.

For some random events $\mathcal{E}_i, i = 1, \ldots, k$, Boole's Formula is given by

$$1 - \Pr\left(\bigcap_{i=1}^{k} \mathcal{E}_i\right) = \Pr\left(\bigcup_{i=1}^{k} \mathcal{E}_i^c\right)$$
$$= \sum_{i=1}^{k} \Pr(\mathcal{E}_i^c) - \sum_{i<j} \Pr(\mathcal{E}_i^c \cap \mathcal{E}_j) + \cdots + (-1)^{k-1} \Pr\left(\bigcap_{i=1}^{k} \mathcal{E}_i^c\right) \quad (4.3.4)$$

where $\mathcal{E}_i^c$ is the complement of $\mathcal{E}_i$. The first order approximation is called the *Bonferroni inequality*:

$$\Pr\left(\bigcap_{i=1}^{k} \mathcal{E}_i\right) \geq 1 - \sum_{i=1}^{k} \Pr(\mathcal{E}_i^c) \quad (4.3.5)$$

Equivalently, this inequality is more commonly represented as follows:

$$\Pr\left(\bigcup_{i=1}^{k} \mathcal{E}_i\right) \leq \sum_{i=1}^{k} \Pr(\mathcal{E}_i) \quad (4.3.6)$$

and is sometimes referred to as Boole's inequality; e.g. in Casella and Berger, which may be consulted for a proof of the inequality represented in 4.3.6 (Casella and Berger, 2002). The Bonferroni procedure simply instructs one to choose a desired α and divide the p-value for each individual tests by the number of tests. The Bonferroni inequality can be used to show that this procedure controls the FWER at level α. If $\mathcal{E}_i$ are p-values, p_i, for V tests where there are V_0 true null hypotheses, and FWER control is desired at level α, and the null hypothesis is stated in terms of the p_i rather the test statistics, then:

$$\Pr\left(\bigcup_{i=1}^{V_0} \{p_i \leq \alpha/V\}\right) \leq \sum_{i=1}^{V_0} \Pr(\{p_i \leq \alpha/V\}) = V_0(\alpha/V) \leq V(\alpha/V) = \alpha \quad (4.3.7)$$

which implies strong control of the FWER. However, the Bonferroni procedure is conservative, particularly for non-independent tests; i.e., when the rejection or failed rejection of a test contains information about whether another test is likely to reject or fail to reject the null hypothesis. In such cases the FWER is typically controlled well below the desired level.

Bonferroni-type procedures

Bonferroni-type procedures include step-up and step-down tests. Both approaches order p-values or test statistics from smallest to largest and compute a significance threshold for each test based upon the number of tests and the rank of each p-value. Step-down procedures begin with the smallest p-value and if it is less than its threshold move to the next test, continuing until reaching a test that does not reject the null hypothesis. Step-up tests reverse this process and begin with the largest p-value; if this value is above its threshold, then continue to the next largest p-value and so on until reaching a test with a p-value below its threshold, at which point one rejects the rest of the tests of lower rank. One of the most common applications of a step-down procedure is the Holm step-down test (Holm, 1979). The process is to order the p-values and associated hypotheses as $p_{(1)}, \ldots, p_{(V)}$ and $H_{(1)}, \ldots, H_{(V)}$, respectively, then reject all $H_{(i)}$ where $i \leq j$ when

$$p_{(j)} \leq \frac{\alpha}{V - j + 1} \tag{4.3.8}$$

The Hochberg step-up test uses the same threshold as in 4.3.8, essentially reversing the process and thereby increasing the power of the procedure (Hochberg, 1988). These and related measures have either no assumptions or weak assumptions on dependence among tests, and in doing so take no account of spatial structure or image smoothness, but are not invalidated by the presence of dependence among tests. It can be shown that both these procedures have strong FWER control (Hochberg, 1988).

FDR control

As shown by Benjamini and Hochberg, the FDR is equivalent to the FWER when all null hypotheses are true (Benjamini and Hochberg, 1995); i.e. FWER control in the weak sense, as described above. Otherwise, if some subset of tests $\mathcal{V}_0 \subset \mathcal{V}$ are truly null, then the FDR$\leq$FWER, but the FDR has greater potential for increases in power, i.e. better control

of type II errors. The FDR arose out the idea that there are cases where the FWER is too conservative a goal; it may often be acceptable to allow a greater number of false positives in order to gain some power, so long as there exists some quantification of this trade-off. Further, if a large number of rejections is expected and the interpretation of the analysis will not be (greatly) affected by a single false positive, then FDR may be a better choice. Benjamini and Hochberg considered several alternative formulations of the FDR before focusing on the above formulation. However, the list of these discussed in Benjamini and Hochberg (1995) is below for completeness:

1. One may desire control at the level of individual realizations, i.e. for a particular test in a family, but when all hypotheses are true, then $\boldsymbol{V}/\boldsymbol{R} = 1$ for even a single incorrect rejection; i.e. for a single test, if there is a false rejection then the number of rejections is $\boldsymbol{R} = 1$ and the number of false rejections is $\boldsymbol{V} = 1$. Benjamini and Hochberg show that $P(\mathbf{R} > 0)\,\mathrm{E}(\mathbf{V}/\mathbf{R}|\mathbf{R} > 0)$ may be controlled in this case, but not $\mathrm{E}(\mathbf{V}/\mathbf{R}|\mathbf{R} > 0)$.

2. Another alternative mentioned by Benjamini and Hochberg is found in Soric (1989), which seeks to control $Q' = \mathrm{E}[\mathbf{V}]/r$ and runs into problems with control when all null hypotheses are true.

3. Lastly, consider $\mathrm{E}[\mathbf{V}]/\,\mathrm{E}[\mathbf{R}]$, which also has trouble when all hypotheses are true.

Setting aside these formulations, Benjamini and Hochberg's procedure for controlling $\boldsymbol{Q} = E[\boldsymbol{V}/\boldsymbol{R}]$ is as follows: Consider V hypotheses $H_i, i = 1, \ldots, V$ with corresponding p-values, $p_i, i = 1, \ldots, V$, which may be ordered $p_{(1)} \leq \cdots \leq p_{(V)}$ with hypotheses $H_{(i)}$ corresponding to $p_{(i)}$. The procedure is variant of the Bonferroni multiple testing procedure, where if k is defined by:

$$k = \max\left\{p_{(i)} \leq \frac{i}{V}q\right\} \tag{4.3.9}$$

then reject all $H_{(i)}$, for $i = 1, \ldots, k$. Benjamini and Hochberg show that the procedure controls the FDR at q for independent test statistics and any configuration false null hypotheses, although independence is not strictly required. It is also shown that this procedure chooses α to maximize the number of rejections at that level, $r(\alpha)$, subject to $\alpha V / r(\alpha) \leq q^*$, which allows simultaneous maximization of $\mathbf{R}$ and FDR control, post-experiment. The method is shown to be more powerful than a number of FWER controlling methods, with power improving with the number of non-null hypotheses.

In addition Benjamini and Yekutieli show that equation (4.3.9) is valid when the joint distribution of $p_{(1)} < \cdots < p_{(V)}$ (or the test statistics from which they came) have positive regression dependency on each one from a subset (PRDS), where the subset consists of the null hypotheses that are true (Benjamini and Yekutieli, 2001). A classic and widely used example of PRDS is in data arising the multivariate normal distribution with non-negative correlation terms; however, non-negative correlation does not imply PRDS in general. When PRDS does not hold, Benjamini and Yekutieli (2001) show that one may replace q in equation (4.3.9):

$$q^* = \frac{q}{\sum_{i=1}^{V} \frac{1}{i}} \tag{4.3.10}$$

and control the FDR at $\leq \frac{V_0}{V} q \leq q$, where V_0 is the number of true null hypotheses. This adjustment tends to be more conservative than the former adjustment.

Additionally, it would be remiss to neglect recent alternative interpretations in the use of the FDR to control false positive rates. One of the more influential metrics is the *positive false discovery rate (pFDR)*, which is defined as the expected proportion of false positives in the set of rejections, given that the set of rejections is not empty:

$$pFDR = E\left[\frac{\boldsymbol{V}}{\boldsymbol{R}} \middle| \boldsymbol{R} > 0\right] \tag{4.3.11}$$

where $\boldsymbol{V}$ is the number of incorrect rejections and $\boldsymbol{R}$ is the total number of rejections. This approach was introduced by Storey (2003), who shows that the pFDR has a Bayesian interpretation as a posterior probability. Another common approach is the FDX, or tail probability of the false discovery proportion (FDP):

$$tFDP(c) = \Pr(FDP > c) \tag{4.3.12}$$

which quantifies the probability that the FDP in a sample of tests exceeds c (Genovese and Wasserman, 2006). While the weight of the literature in multiple testing procedures could crush a mere mortal, Farcomeni (2008) provides an excellent overview of multiple testing issues with particular focus on developments in FDR interpretation and application.

4.4 Multiple Testing in Neuroimaging

The data structure

Images consist of discrete units of measurement, *pixels* (2D) or *voxels* (3D). The literature in neuroimaging statistics has long had to address multiple comparisons issues and is useful in understanding classical/frequentist approaches in situations where the number of parameters/predictors greatly exceeds the number of subjects. Neuroimaging data often consists of structural MRI, with single intensity value measurements within each voxel, or a functional MRI (fMRI), where a time series of intensities is associated with each voxel.[27] While neuroimaging research is diverse in its statistical and mathematical methodology, many research questions reduce to whether some brain region is associated with a cognitive behavior, physiological response, or disease. Multivariate methods are often computationally prohibitive, when they are possible at all; the number of data points is often in the hundreds of thousands, which in practice always outnumbers the sample size, causing identifiability issues.

[27] By no means is MRI the only modality, and even with MRI there is a wide range of options for both obtaining and processing the images; however, such discussion is beyond the current section.

A popular method designed to handle such large datasets is the "mass-univariate" approach, which models each voxel separately. The result is a map of test statistics used to perform voxel-level inference; Karl Friston introduced and pioneered this approach in neuroimaging within the Statistical Parametric Mapping context, which has since become ubiquitous in application and methodological development (Friston et al., 1991, 1994; Worsley and Friston, 1995; Worsley et al., 1996) . In functional neuroimaging, one typically seeks to answer whether a particular voxel is "activated", although many modern modern questions in neuroscience, psychology, and medicine are far more complicated (Ombao et al., 2017). In the fMRI context, "activation" detection involves inferring neuronal activity by detecting changes in blood flow. In structural MRI, higher-resolution images show a snapshot of the brain, rather than change over time. At least some research questions using structural MRI involve extracting meaningful inference from a large number of voxels such that the mass-univariate approach can applied; for example, Voxel-based lesion-symptom mapping (VLSM) variations on the mass-univariate approach to determine whether lesion status is associated with a particular cognitive decline (Bates et al., 2003). This approach allows for so-called "omnibus" tests, i.e. "is there activation anywhere in the brain?", as well as tests at individual voxels or subsets of voxels. However, while this approach is enticing in its simplicity, performing so many tests inflates false positive rates; i.e., even if the brain is nowhere activated, and one tests at the $\alpha = 0.05$ level, then on average 5% of "activated" voxels are expected to be spurious.[28]

However, there is another problem. Recall from section 4.3 that the Bonferroni procedure is not a size α test if only a subset of tests is null, but rather is a level α test that is permitted to control false positives at a rate equal to or below the specified α. Most images exhibit spatial structure and hence pixel/voxel values and/or their associated test statistics are not independent, leading the Bonferroni procedure to be more conservative than desired.

[28] It is reasonable to doubt that a living subject could have activation in no area of the brain, but the general point remains: with so many tests one should expect a significant number of false positives in areas not associated with the outcome or exposure of interest.

In particular, if one voxel is “active" then this tells provides information about the likelihood of finding other activated voxels nearby. Therefore, confidence in research findings require addressing the multiple testing problem while accounting for, or at least circumventing, the inherent spatial structure so as to avoid excessive conservatism in false positive control.

Nichols and Hayasaka (2003) provide a general FWER-control overview for neuroimaging contexts they identify three traditional approaches to FWER control in neuroimaging:

1. Bonferroni-type Procedures
2. Random Field Procedures
3. Nonparametric methods, specifically permutation tests

Bonferroni-type procedures were covered in section 4.3 and are not re-covered here, but note that their application tends to be conservative (Nichols and Hayasaka, 2003). FDR control is briefly reexamined.

Random field theory

Keith Worsley and Karl Friston played a large role in bringing Random Field Theory (RFT) into application for neuroimaging as a method to obtain valid inference by treating the Statistical Parametric Maps as lattice approximations of random fields. Worsley et al. (1992) introduced the methodology and subsequently, through collaboration with Karl Friston, these researchers combined forces to marry SPM and RFT (Worsley et al., 1996).

The RFT approach is typically based on Gaussian random field, which should have 0 mean and unit variance under the null hypothesis. Regions with values above some threshold, u, are known as excursion sets, from which one may calculate the *Euler characteristic*, a topological measure. A formal treatment is beyond the scope of this work, but an intuitive explanation is that the method counts the number of connected suprathreshold regions or “clusters”, minus the number of “holes”, plus the number of “hollows”. Sufficiently high

threshold typically avoid holes and hollows and so the Euler characteristic will simply be the number of clusters. An accessible discussion can be found in either Nichols and Hayasaka (2003) or Lindquist and Meija (2015). Those desiring all the gory mathematical details should consult Worsley et al. (1992).

Resampling approaches

Resampling approaches follow RFT in that they seek to approximate the upper tail of the distribution of the maximum test statistic, $F_{M_T}(t)$, but resampling avoids stringent assumptions required of parametric methods by empirically estimating the distribution of the maximum test statistic. Resampling procedures generally consist of either permutations, which resamples the data without replacement, or bootstrapping, which resamples residuals with replacement; both approaches construct a null distribution. Nichols and Hayasaka (2003) differentiate between *randomization tests*, which justifies resampling by the random selection inherent to a particular experimental design, and *permutation tests*, which simply requires exchangeability under the null hypothesis. The difference is more philosophical than practical; randomization tests explicitly refer to the sample at hand, while permutation tests assume a population distribution so that inference may be made.

The permutation test as a concept deserves further elaboration. A permutation test permutes labels, or equivalently data, and calculates a test statistic for each permutation. This process assumes that labels are exchangeable under the null hypothesis. A set of labels is *exchangeable* if the distribution of the statistic is invariant to labeling. The classic permutation testing example is not related to multiple testing issues, but was intended to address smaller sample sizes or situations where standard parametric assumptions appear to fail or seem dubious. A classic example is the permutation test version of the 2-sample t-test. The observed t-statistic is calculated as usual, but the t-distribution is not used to obtain a p-value. Rather, the group labels are permuted about the subjects, with each subject keeping their outcome value; equivalently, subjects could retain their group label

and permute their outcome values. For each permutation the t-statistic is re-calculated, and this set of permutation t-statistics creates an empirical null distribution, which is used to calculate a p-value. If the number of possible permutations is too large, it is common to sample a few thousand possibilities as an approximation to the complete empirical null distribution. This approach is related to the bootstrap, but a more detailed discussion is beyond our scope here. See Berry et al. (2002) or Long et al. (2007) for readable review papers. Efron and Tibshirani (1993) or Davison and Hinkley (1997) are useful textbooks on the bootstrap.

However, for images the process is slightly more complicated. The key insight is that *entire images* are resampled. This insight is motivated by following Nichols and Holmes (2001). If one assumes no autocorrelation, then permutation tests are permitted within each voxel, assuming a time series is associated with each voxel location. For instance, if the question pertains to whether particular voxels are associated with a specific motor task, then it would suffice to find all subject label permutations, or else sample from the possible permutations when the total is prohibitively large, and calculate or approximate the relevant p-value. However, with an image, the voxels are not independent, but rather have a spatial structure. If one permutes subject labels at each voxel then the results will be invalid. Rather, the subject labels are permuted about entire images, which preserves spatial structure. Then calculating the distribution of the maximum test statistic is straightforward and one rejects the hypothesis of no activation for voxels whose p-value is less than the upper $\alpha\%$ of the distribution of the maximum test statistic. This approach has strong control over the FWER; a formal proof is found in Holmes et al. (1996). This approach is well-founded for most structural MRI or PET, where there is no times series and no autocorrelation, respectively. Functional MRI time series, however, have temporal autocorrelation, run afoul of the exchangeability restriction, and inflated false positives can result if this violation is not addressed. One approach is decorrelate the images and then resample; this involves estimating autocorrelation in the residuals, decorrelating the residuals, resampling, and

finally recorrelating the resampled residuals (Nichols and Hayasaka, 2003). Lindquist and Meija (2015) note that permutation tests have found greater use in group-level analysis where exchangeability holds between subjects, but this is a less common approach for subject-level data, given the trouble with temporal auto-correlation. Although details are not given, presumably a simple variant on group level analysis is to calculate test statistics for each individual and use permutations of group labels to make comparisons. Furthermore, the approach is valid for various structural images, since each subject has one observation per voxel, and therefore no temporal autocorrelation.

FDR in neuroimaging

The seminal paper for FDR use in neuroimaging is Genovese et al. (2002); as described above, the idea is to control the proportion of *rejected hypotheses* that are falsely rejected. It turns out that the initial procedures proposed for FDR control can be appropriate for spatial data, and neuroimaging data in particular. The traditional Benjamini-Hochberg approach in equation (4.3.9) holds under Gaussian noise with nonnegative correlation between voxels; even if one felt uncomfortable with those assumptions, the modifications made by Benjamini and Yekutieli (2001) would be valid (i.e., equation (4.3.10)). A key benefit to this approach is that it avoids arbitrary thresholds and adapts to the signal strength. Depending on the group analysis method employed, this approach also allows for the thresholds to vary by subject. However, in application it is important to bear in mind that the FDR is based on large sample theory in the sense that the approach controls FDR on average over many many replications, so for a given data set the observed false discovery proportion (FDP) may not be below the desired q.[29] Relatedly, the FDR alone fails to tell us the exact proportion of truly activated voxels detected; this would require the *false non-discovery rate (FNR)*, i.e. the proportion of active voxels a method failed to label as such, which is 0 if

[29] Consider the similarity to confidence intervals where a given interval may not contain the "true" parameter; e.g. a 95% CI will contain the "true" value in 95% of samples, but this does not mean that a given interval has a 95% chance of containing the "true" parameter.

all voxels are labeled active. FDR is more conservative as correlations increase and thus more powerful for un-smoothed data, this in contrast to random field theory, which are more conservative for un-smoothed data.

4.5 On the Validity of Hypothesis Testing and MTP's

Both traditional hypothesis testing and multiple testing procedures have variously come under fire. This section is a response to several of the more pervasive criticisms in the literature.

The relative merits of p-values and hypothesis testing

Misperceptions regarding statistical procedures are not particularly new. It is practically a right of passage for young statisticians to deliver a sermons to students, colleagues, or both about p-values and their interpretation. Inevitably, after so many years have passed and old habits resurface, statisticians must once again deliver such sermons in publications, many of which are quite readable, e.g., Best et al. (2016) or Greenland et al. (2016). Periodically, the misunderstanding turns into outright hostility, as when the Journal of Basic and Applied Psychology outright banned p-values (Trafimow, 2014; Trafimow and Marks, 2015), which then prompts a response from various outlets, including statisticians (Wasserstein and Lazar, 2016). It is largely our concern here to defend multiple testing procedures, but it seems this requires defending the hypothesis testing framework itself.[30]

In defense of hypothesis testing and p-values

Hypothesis testing is based upon specifying a (null) hypothesis for a given vector of unknown parameters, $\boldsymbol{\theta}$, and may or may not involve assuming an underlying distribution:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \tag{4.5.1}$$

[30] Which for better or worse involves an explanation beyond a well-placed name-drop.

One then determines whether a given sample provides evidence to reject this hypothesis, typically by calculating some test statistic, from which one derives a p-value; that is, the probability of obtaining a test statistic of equal or more extreme value. For example, suppose a statistical test results in some test statistic, t. If this is assumed to be a realization from the random variable representing the distribution possible test statistics under the null hypothesis, T, then the p-value is $\Pr(T > |t|)$ for a two-sided test. If this p-value is below a pre-specified significance threshold, α, then one rejects H_0; otherwise, one fails to reject H_0.[31]

There are some obvious limitations to such a procedure's usefulness. For one, it must be the case that the null hypothesis is plausible, but one typically *knows it* to be false, at least in the strict sense. Especially when hypotheses involve differences between groups, which they often do, most researchers would suspect scientific malpractice if a paper claimed the difference between some groups was *exactly zero*. In addition, p-values are themselves random variables and subject to variation, which makes assuming that a similar p-value would be obtained in a subsequent study rather dubious, especially for small sample sizes. If the strict null hypothesis is nearly always false and p-values fluctuate (sometimes wildly, it seems) across samples, then why not abandon the approach entirely?[32]

One often ignored detail is that research hypotheses and statistical hypotheses typically differ. Bradley (1968) lends some relevant insights in a text that is as useful for thinking about the limitations of the central limit theorem in practice and hypothesis testing in general as it is about its nominal subject (Distribution-free, or non-parametric, statistical inference); that is, what to do when distributional assumptions about our data

[31] It may be argued that such a process is an unholy union between Neyman-Pearson hypothesis testing and Fisher's p-values (see Berger (2003) for an interesting perspective on the conflicting views of Fishern Neyman, and Jeffreys). Fair enough. Nevertheless, opinions and interpretations notwithstanding, there is a theoretical connection between these methods that can be reduced to straight-up mathematics, which may be seen in such texts as Casella and Berger (2002).

[32] It may seem odd that a strident defense of traditional hypothesis testing is given here, since this dissertation will ultimately focus on Bayesian methodologies. For now it shall suffice to point out that the author is what one might call a statistical pluralist; i.e., different practical situations may call for different statistical approaches, and reasonable people may sometimes disagree about which is best.

run amok. There are two levels of null hypothesis, each of which have their corresponding assumptions. One level is the null hypothesis in which the experimenter is actually interested, for example something like “Is there a *meaningful/scientific/practical difference* between or across groups?" To this question there is often, but not always, an answer properly summarized as “Yes” or “No”. The other level pertains to how to test that hypothesis, which is the domain of the statistical hypothesis. The statistical hypothesis converts the research hypothesis into a form that can quantify and answer the question at hand and may or may not involve distributional assumptions. The statistical formulation has a certain exactness that the real world typically lacks, but this does not mean it is not, or cannot be a useful tool in some circumstances. Note that this argument more formally favors some testing criteria and need not depend explicitly upon a p-value to be a valid approach. Rather, it is a reminder that statistical methods are in practice tools to answer scientific questions in the face of uncertainty. While the formal mathematics may be rigorous, in practice we are approximating reality.[33]

A deeper issue is that studies sample data, which involves inherent uncertainty, no matter the measure. While Trafimow (an erstwhile editor for the Journal of Basic and Applied Psychology) may dislike p-values because they cannot reliably tell us the truth, given that they are subject to randomness, it no less true that an effect size can be due to chance sampling variation, especially in small samples, and in fact this is exactly the kind of thing that probabilistic modeling is supposed to *protect against*, however imperfectly; different samples yield different results no matter the methodology, and incorrect inferences in either direction are possible. In addition, if the assumptions about a sample are incorrect, then no amount statistical witchcraft will tell you that you have in fact not discovered what you think you have discovered, and attacking statistical methods is far easier than addressing deficits in study design.[34]

[33]We simply cannot afford to dig further here, as we are treading dangerously close to finding ourselves deep into the philosophy of science.

[34]This is not necessarily to attack (or exonerate) Psychology; study design is difficult and incorrect sampling

The task is to *quantify uncertainty* relating to the employed procedure, for no excorcist can cast out uncertainty from scientific studies. It is true that effect sizes, variances, interval estimation, etc. are important considerations and capable researchers will consider the data in a holistic manner so as to achieve some glimpse of scientific relevance and accuracy. There is often a certain ridiculousness to the celebration/dejection dichotomy when a test (fails to) fall(s) below the sacred $\alpha = 0.05$ level. Given a large enough sample size, significant differences can be obtained between means for any separation, no matter how small; this is why one should consider whether the observed difference is clinically meaningful and appropriate application of any statistical method(s) requires caution and the patience to fully examine and describe the qualities of the data. Scientists should be able to recognize cases where the significant difference is not scientifically relevant in both their own and others' work.

There are other approaches to testing and reporting results. Hubbard and Lindsay (2008) argue against p-values and in favor of estimation via confidence intervals. For a given α value, hypothesis test results will be the same, but confidence intervals can help avoid some pitfalls of p-values, despite being subject to the same sampling variation issues as p-values. Interval estimation often provides an easier outlet for determining the practical import of a significant difference, allows for a more robust interpretation of the results than a p-value alone, and can provide information regarding the quality of the obtained estimate. Perhaps the most useful aspect of the confidence interval is that it provides a measure of precision: one can easily see the range of "plausible" values. Consider the following 95% CI for an odds ratio: [1.01, 1.02]. This interval does not contain 1, but it is very narrow and just barely over 1; we may doubt whether this "significant" effect is meaningful, depending on the context. In contrast, a 95% CI [0.97, 1.87] is not statistically significant, but it is also not very precise, perhaps making it difficult to argue convincingly about the effect size.

assumptions finds its way into many studies involving humans. The scientific community as a whole, and its sub-disciplines, has a responsibility to seriously address these issues. The Basic and Applied Psychology situation is simply a well-publicized example in what the author considers misguided thinking.

In conclusion, even in abandoning the hypothesis testing framework entirely, some subjectivity remains in assigning increased value to some differences over others. Which of these are large enough and how small a variance about those estimates should give provide confidence in the obtained results? In many cases, one must judge whether or not a study confirms a researcher's intuitions, and this may require an answer: Yes or No? For example, does a drug outperform placebo or not? There are grey areas and obtaining extreme samples is possible, but for better or worse a substantial amount of research requires making decisions in categorical manner, which requires a principled approach in decision making that should go beyond eye-ball-testing descriptive statistics in order to have any relevance. A balance between the need for clear-cut answers and broad holistic data inspection is necessary, and these two goals are not mutually exclusive, especially when one is transparent about which analyses were performed. Blind p-value worship may be anathema, but concerns regarding hypothesis testing do not invalidate all inferential frequentist statistical methods and do not mean that hypothesis testing can never be useful. Researchers should rectify themselves to understanding that these are tools, based on assumptions that are at best approximate, seek to understand how and when they fail, and embrace uncertainty.

A case study in confusion about multiple testing procedures

As it so happens, controversy over hypothesis testing also spreads into multiple testing procedures. For example, Rothman (1990) argues against multiple testing procedures in epidemiology, largely by winning subsequent battles against two strawmen. The first is "Chance not only can cause unusual findings in principle, but it does cause many or most such findings". While he delves into an interesting discussion of what the term "chance" means, he largely misses what a (bio)statistician is talking about:[35] The discussion relates to distributions of numbers and the chance involved in drawing (at least approximately) "randomly" from these distributions. Random sampling involves no funny pseudo-philosophical view

[35] At least this one, anyway.

that the values measured in some study subjects arise from anything other than deterministic processes, which he rightly believes would obviate the whole of empirical research. Chance is best understood in this context as randomness, or variability, that arises due to taking finite samples. Random (or, in fact, non-random) sampling processes can result in samples that are not indicative of the truth or representative of their underlying distributions/populations. If a population is finite and one samples the entire population then there are no assumptions involving chance or randomness in parameter estimation or differences between estimates; on then possesses a census.

In some models, further considerations, such as random error, can also play a role. Consider the unlikely scenario where one assembles the entirety of the American population and subjects them to a 400 meter time trial; there is thus no sampling variation due to sampling Americans, because they are all present. However, there are variations in run times within an individual, and even if each person runs only once there is no certainty that this is truly the fastest time possible or even that the time is representative of the individual's running ability.[36] Therefore, testing group differences in time will still be subject to randomness in the sense that individuals have variation about the time they can run. In a simpler sense, one could sample 100 Americans and find group differences in run time by "chance", in that one could by "chance" obtain a sample that did not adequately represent the groups, but this does not imply that the run times were causeless, only that random or otherwise unexplained variation could account for why there existed a difference between groups in the sample, and that the difference may then not be endemic to group membership. The important point here is that due to finite samples, it is possible to be severely mislead about causes and effects.

[36]This example highlights an issue with the idea of truly taking a census, and what it would mean to do so. One could imagine that an individual's 400m run times are themselves a random variable from which to sample, so that a reasonable interpretation is that one first samples from a population of random variables (the individual runner) and then sample from that random variable (the run times for that runner).

Rothman's second goal is to refute that "No one would want to earmark for further investigation something caused by chance." While the bulk of this section continues to scrutinize definitions of chance, the fact remains that there is no conceivable reason that anyone *should* want to waste time and money researching on a difference that arose due to sampling variation and is not actually representative of the "true" underlying distribution, if that is the case; time and attention, like samples, are finite and precious. However, this complaint is as much about the kind of research one conducts as it is about the philosophy of hypothesis testing. Hypothesis tests are always based on an arbitrary threshold, as is in fact the "eyeball test", which if hypothesis testing is microwaving an egg, then the "eyeball test" is on the order of microwaving a fork.[37] Thresholds can vary based upon study goals and especially in early research phases more or less stringent thresholds may be required depending on the consequences of false positives or negatives; sample size considerations may also affect what is considered a reasonable threshold. Moderate findings are surely worth examining further in many cases, but if the "truth" is that there is no scientifically or medically relevant difference present in a study then beating the dead horse is as useless as it ever was.[38]

Rothman also criticizes the manner in which multiple testing procedures depend on the validity of the "universal null hypothesis", defined as the hypothesis that "all associations we observe (in a given body of data) reflect only random variation", and its likelihood of being true. However, multiple testing/comparison procedures are not in general based only on the plausibility of all possible hypotheses in an experiment or data set being true; much of the theory derived and discussed in the literature lends attention to strong control of the FWER; that is, ensuring that the FWER is controlled when any subset of hypotheses are truly null. Likewise, methods for controlling FDR are specifically designed

[37]The author is being hyperbolic - do not under any circumstances attempt to empirically discover the effects of microwaving either object. The requisite studies have been performed and are available here.

[38]The author wishes to clarify that no individual should ever beat a horse, living or dead. There are better outlets available for relieving stress.

to account for situations where some hypotheses are truly null and others are not. While early work and introductions focus on something like the "universal null hypothesis", the bulk of the literature is far more broad. A more appropriate critique would be that many FWER methods are in general too conservative when the universal null hypothesis is not a reasonable assumption. Correctly interpreted, this critique would mean advocating for improved methods for controlling false positives, but would not obviate the underlying intuitions about why multiple testing procedures are necessary in general.

Finally, perhaps the most egregious oversight is that Rothman avoids discussing families of tests. While some subjectivity is involved in deciding what constitutes a family, this distinction is more relative than arbitrary and should correspond reasonably to the real world. This (somewhat) obviates concerns corresponding to "slippery slope" logic. Reasonable thinking on what constitutes a family of tests can largely prevent one from feeling obligated to extend multiple testing to every test on a dataset. An important subtext is that multiple comparisons procedures and hypothesis testing procedures in general should be viewed as modeling approaches with various large sample qualities. Neither process is designed to ensure certainty about a particular question within a single dataset, but rather speak to how well the approaches work upon multiple applications or else to provide information about how different a result is from what one would expect if there were no difference in the population (Mayo, 1996).

The reason that multiple testing becomes an issue is that the possible errors the testing procedure can make change depending upon the number of tests. Many researchers are bothered by penalizations for "peeking" at large numbers of tests; how could more information be a bad thing? However, the concern is not actually that in a particular sample the probability of a wrong inference on a particular test is increased, but rather that if one adopts a "peeking procedure" in every dataset, or as a procedure in general, without any correction for multiple testing then a large proportion of studies will report a spurious

result; an uninhibited peeking procedure basically ensures that one finds the false positives and reports them as interesting findings. The analysis procedure itself will produce false positives with higher frequency because there are more ways to obtain a false positive.[39] Multiple testing procedures are designed to improve the reliability and replicability of the scientific process as a whole by addressing the nature of procedures when applied repeatedly. While not specifically addressing multiple testing procedures, Chapter 9 of Deborah Mayo's book *Error and the Growth of Experimental Knowledge*, provides an extended discussion in a philosophy of science context regarding why one should care about the changing error characteristics of procedures as more tests are conducted (Mayo, 1996).

The intention is not to embarrass Rothman, insinuate that his arguments may not have been somewhat refined since 1990, or even summarize the literature rebutting his arguments.[40] However, this paper still appears to carry some influence, and it seems pertinent to understand why these seemingly reasonable arguments go astray. Furthermore, the multiple testing problems are even trickier in more complex data and addressing these objections helps identify misconceptions and thus think more clearly about the problems at hand. To summarize:

1. *Hypothesis testing is an inexact procedure*, and can often be based on the quantifying qualitative questions; e.g. Is there a *meaningful* gender difference in reaction to a therapy? Research hypotheses and statistical hypotheses are related, but rarely exactly the same.

2. *P-values can be useful tools for hypothesis testing, but they should never be severed from a holistic perspective on the data.* At the very least, sample sizes, effect sizes, in-

[39] At the risk of belaboring the point, consider a simple example. If one tests a single truly null hypothesis in repeated samples at level 0.05 then in the long run 0.05% of the time a spurious significant result is found. If one tests 3 independent truly null hypotheses in repeated samples, each at the 0.05 level, then any combination of the 3 tests could be spurious and so $1 - (1 - 0.05)^3 \approx 14\%$ of the samples will contain at least 1 false positive.

[40] In any case, the idea that a tenured professor would be all that concerned about a graduate student's criticism seems to be a null hypothesis no one should count on.

terval estimation, and the potential consequences of false positives and false negatives should all play a role in how results are reported. Blind trust in p-values is the wide path that should be avoided, but it may not (yet) be necessary to entirely abandon the tool.

3. *Multiple testing procedures require at least some defensibly defined family of tests in order to be interpretable or useful.* Families should make sense in qualitative terms. This prevents many inappropriate wanderings down "slippery slopes". Families of tests can correspond to populations which could possibly be modeled by common distribution (it is then the researcher's job to determine whether this is the case). However, this is not always the case. Multiple testing can arise in cases where the various tests involve outcomes with different measures or scales and may not be reasonably considered as having a common distribution under the null hypothesis; e.g. consider studying differences in human brains based on imaging studies. Perhaps a study considers changes in volume (sMRI), BOLD activity (fMRI), and electrical activity (EEG). The outcomes could not reasonably be expected to have a common distribution, but yet the three tests constitute a family. In simpler settings, if height does not differ by sexual orientation, say 3 main groups heterosexual, homosexual, and bisexual, then this means that it is reasonable to treat the 3 groups as being drawn from a common distribution, or at least treated as having a common mean if the variances are drastically different, and one may say that sexual orientation does not yield any useful information about height. It is less clear why sexual orientation and childhood diet (e.g. omnivore, vegetarian/vegan, or pescatarian) would constitute a family of tests.

4. *Hypothesis testing, p-values, and statistics in general can be conceived of as attempts to quantify or handle uncertainty.* Modern research are uses what amounts to approximate methodology to answer scientific questions with real answers, but

which are subject to variability, due at the very least to finite sampling from distributions/populations, to say nothing about other concerns like measurement error. Deeper discussions on the definition of chance and randomness are left to philosophers and quantum physicists.

5. *Hypothesis testing and multiple testing procedures are designed to address the large sample properties of statistical tests.* It is important to bear in mind that for the body of scientific research to be valid then the procedures employed must possess certain qualities over many repeated applications. It is not always possible to generate high confidence in the results within a particular sample, but concern should be with how methods perform in the long run because it is those properties that validate the scientific process.[41]

Multiple testing controversies in neuroimaging

As previously described, statistical controversies come out of the woodwork in various forms and with various frequencies over time, but they often manifest similar concerns, misconceptions, or philosophical bents. Previous sections discussed the arbitrary nature of p-value thresholds, or in a spirit of transparency, thresholds in general. There is no perfect threshold and a quick foray into any mathematical statistics textbook shall introduce the reader to the tension between the ability to avoid false positives and false negatives simultaneously; as a testing procedure improves one error type, the other becomes worse (Casella and Berger, 2002).[42]

Lieberman and Cunningham (2009) provide an apparently common view of the tension between power and false positive control, in both their legitimate and exaggerated

[41] Note that strictly speaking it is not necessary to take a repeated applications approach to justify hypothesis tests, but ultimately one must bear in mind that a single sample can only be so informative.

[42] Note that this tension is *within procedures*; e.g., suppose procedure 1 has 40% power at the $\alpha = 0.05$ level. If α is lowered to 0.01, then procedure 1's power must decrease. This does not mean that another procedure for the same scientific question, say procedure 2, could not have 80% power at the $\alpha = 0.01$ level. This is an occasional task in mathematical statistics: to find the procedure that maximizes the power for a given α level.

claims. To start with their legitimate concerns, it is quite defensible to worry that too much concern is given to false positives, especially given the state of many sub-disciplines in neuroscience; the field is relatively young and so researches may reasonably desire to avoid overconfidence in any direction. Especially given the "fundable" sample size in many imaging studies, there is concern that only large effects, e.g., motor effects, shall be found while more moderate effects, e.g., what is expected in cognitive processes, will be missed. Any hope of combining results into meta-analyses is certainly frustrated by the fact that papers with useful non-significant information are rarely published, and modern science desperately needs an avenue for publishing well-conducted science that fails to pass pre-specified decision thresholds; publication bias skews the literature and robs science of useful results (Joober et al., 2012). However, despite these legitimate concerns, what follows is another case-study in misunderstanding the application of statistical methods.

One of the authors' key complaints is that the behavioral science standards to which cognitive neuroscience aspires to emulate are not required to correct for all their tests, and in addition the old "gold standard" was $p < 0.001$, which was then adjusted to account for spatial extent so that $p < 0.005$ and 10 voxel extent was deemed acceptable. The authors present some simulations showing fairly reasonable results while decrying a perceived demand for certainty. Perhaps some in their field demand certainty, but as discussed, statistics is rather the endeavor to account for the exact opposite: uncertainty lives and may even be immortal. While, generally speaking, scientists, and statisticians in particular, should not be called to account for jokes that fall flat, a tired hyperbole is to ask whether entire journals should adjust for all tests in their journal; this "joke" betrays a misunderstanding about the underlying philosophy. For one, this ignores the meaning of multiple tests; multiple testing procedures are relevant when a group of tests can be reasonably regarded as a family. It is also little consolation that simulations appear to reasonably control false positives; the multiple testing procedure philosophy is not to find the perfect p-value threshold, but rather to help researchers understand and quantify the level

at which the employed procedure controls false positives. There is no need to retroactively apply a witch hunt to papers published with less than informative thresholds; a reasonable balance is available.

In the same journal issue, Bennett et al. (2009) provide a more nuanced view and lament the opposite problem: that too many researchers see no need to account for multiple testing. They make an astute observation that is often lost amidst more pure statistical theoretical terms: multiple testing procedures should not only to control false positives, but provide information on their rate under various assumptions and conditions. It is important to understand the directionality of the process: one often needs to quantify a procedure's over the long run, *so that* informed decisions about acceptable error rates are possible in various contexts; at the very least one should understand how likely it is that a procedure errs in its inference: a "principled correction... definitively identifies for the reader the probability or proportion of false positives that could be expected in the reported results." While arbitrary thresholds are unavoidable in hypothesis testing, the underlying idea of multiple testing says that a threshold does not mean the same thing in terms of error control at different numbers of tests. A chief concern with arbitrarily lowering the thresholds as a solution is that $p < 0.001$ is not really testing 10,000 and 6,000 voxels at the same level.

Methodology ranges from more to less conservative, and in practice more liberal thresholds should sometimes be employed. False positives cannot be avoided in all studies unless one rejects no hypotheses and so the balance between the false positives and false negatives is one that must be considered, but this should be done by thinking about the level at which one actually wishes to test and employing a multiple testing procedure to get there. If examination shows that a false negative rate is to high for a given level of significance, then this should be stated explicitly and thresholds assigned accordingly; the answer is not to ignore a known statistical issue or to address the issue with another arbitrary threshold.

The trouble with FWER control

There has been an element of polemic to this section, but this should not imply an insensitivity to the difficulty researchers face in practice. Problems arise in attempting FWER control, and especially when very large numbers of tests are necessary it does become impractical, especially at something like the voxel level. Nichols and Hayasaka (2003) note that most Bonferroni methods are far too conservative to be acceptable, especially with non-independent data, and RFT is also fairly conservative, if much improved over Bonferroni. Permutation tests have the distinction of least conservative, at least in practice according to simulations and experience by Nichols and Hayasaka (2003).

It is no mystery why controlling the FWER is excessive in some cases: to control the FWER is to control the probability of *one or more false positives*. In many cases, a few false positives would not change the interpretation of the data and controlling the FWER comes at the cost of missing true positives without much added protection against spurious results at a relevant scientific level; that is, often somewhere in between a conservative test and no test is desired, but one must still choose such a test in a principled way. An example is Mirman et al. (2018), who strive for some balance in their study of Voxel-based lesion-symptom mapping (VLSM). Mirman and colleagues are interested in the association between lesion status at a voxel and cognitive deficit severity in structural MRI studies. It is unlikely that a single voxel difference would be meaningful in this context and so a different approach is needed to achieve a reasonable balance; they address multiple testing issues in a creative way that avoids the extremes of FWER control or no correction, and provide insight into the limitations of several common approaches; e.g. cluster-based thresholds and FDR-correction.

Clusters are defined as a contiguous subset of voxels that all exceed some threshold of statistical significance; researchers often set minimum cluster sizes to prevent excessive type I errors. This approach is typically based on the idea that significantly activated brain

activity is likely to have a spatial extent greater than a single or small number of voxels (Lieberman and Cunningham, 2009). While setting a minimum cluster size accounts for spatial dependence and that activation of any sort likely has a spatial extent, choosing the appropriate spatial extent should not be arbitrary. A common approach is to choose a primary threshold for significance along with a secondary threshold of voxel extent; e.g. as mentioned above $p < 0.001$ and a minimum of 10 contiguous voxels (Lieberman and Cunningham, 2009). However, using liberal primary thresholds along will give rise to larger clusters and provide the illusion of robustness (Lindquist and Meija, 2015).

While none can escape the arbitrary nature of the primary threshold, Mirman et al. (2018) escape further obfuscation by following Nichols and Holmes (2001) with a non-parametric procedure for hypothesis testing. After applying the primary threshold to the data, they compute the size of the largest supra-threshold voxel cluster; this is essentially the test statistic. The permutation test comes into play when the behavioral data, i.e. the outcome of interest, is permuted and analysis re-conducted many times to build a null distribution to which the observed maximum cluster size is compared. For example, significant clusters are those whose size exceed 95% the null distribution cluster sizes. While this methodology is more nuanced than setting arbitrary cluster sizes, simulations showed that permissive thresholds, 0.05 and 0.01, resulted in clusters too large to be anatomically meaningful, while less permissive thresholds, 0.001, 0.0001, result in clusters that extended beyond the bounds of the known significant area. The cluster-based method fails to avoid false positives in the areas where they are the most likely, thus finding problems with even more sophisticated cluster size methods.

The next approach is to apply the concept of *k-FWER control*, based on work by Romano and Wolf (2007), who described methodology to control the probability of $> k$ false positives. When applied to permutations, this approach is fairly easy to apply in that it is the same as traditional permutation except that rather than finding the maximum test

statistic and creating a null distribution about it, one takes the v highest test statistic from each permutation; e.g. if $v = 5$, then the null distribution is created from the 5^{th} highest test statistic from each image. This process is christened *continuous permutation-based FWER* since it allows $v > 1$. While similar to FDR, the difference is that FDR controls the expected proportion of false positives in the sample of rejected voxels, generally denoted q, while a generalized version of FWER still controls the rate of a particular number of false positives. Nominal q-values can be computed with this approach to compare methodology, and determine which approaches actually do best in controlling FDR. In smaller samples, traditional FDR-control methods were more variable than k-FWER, where direct FDR-control becomes anti-conservative and the voxel dependence exaggerates the skew toward smaller p-values in the presence of true signal; ultimately, continuous FWER tended to outperform direct FDR-control in VLSM simulation studies, in terms of avoiding false positives and detecting true signal, especially at sample sizes typical in those studies. The authors believe that the inherent flexible thresholds of this methodology allows more ability to describe the data and report multiple thresholds, thus avoiding dichotomous significance statements, which should not be a problem so long as practitioners are transparent about their choices.

Summary

This section presented a summary of the literature with respect to methodology for handling multiple testing in neuroimaging, but also discussed controversies with respect to p-values, hypothesis testing, and multiple testing procedure applications; it is hoped that the reader will have been disabused of at least a few misconceptions in the literature. Statistics is not an exercise in perfection, but rather the attempt to quantify inevitable uncertainty in research questions. The application thereof requires some combination of mathematical precision and artful application; a great scientist once said: "Sometimes science is more art than science, Morty. A lot of people don't get that." Roiland (2014). Defending the

usefulness of traditional methods is not to imply that they cannot or should not be improved, nor does it imply that novel frameworks could not prove more useful than older ones. However, many attacks on classical hypothesis testing and multiple testing procedures are either misguided or represent partial truths; furthermore, multiple authors have provided more nuanced approaches to its implementation. It is on such a foundation that better, more reliable, science can be conducted.

4.6 Bayesian Perspectives on Multiple Testing

Given Bayesian statistics' hesitancy about hypothesis testing in general, it may be surprising to discover that there exists a literature on Bayesian approaches to multiple testing problems. Nevertheless, such a literature exists, and we provide some exploration thereof in the current section.

The traditional problem

In many cases an alternative framework for statistical modeling is the younger sibling of frequentist statistics, Bayesian statistics, which is not particularly new, but required significant computational advances to become viable in complicated practical situations. As such, the Bayesian paradigm may ultimately play an increasing role in science. However, Bayesian statistics requires a different stance towards multiple testing, and arguments for or against multiple testing procedures thus differ from those in classical statistics. Berry and Hochberg (1999) are followed throughout in motivating the problem, although their example is generalized to k groups rather than two. Consider k random variables Y_i, $i = 1, \ldots, k$. After sampling, one obtains vectors of observations $\boldsymbol{y}_i$, $i = 1, \ldots, k$, whose length may not be equal among groups. Let the k parameters representing means be denoted as $\theta_1, \ldots, \theta_k$.[43]

[43] It also makes sense to think of the θ_i as differences; i.e., suppose θ_1 is a reference group mean and $\theta_2, \ldots, \theta_k$ represent the respective differences between groups $2, \ldots, k$ and group 1, respectively. There is also nothing stopping the generalization of θ_i from a scalar to vector, but for now we take a simpler approach.

The posterior distribution is formulated as follows:

$$p(\theta_1, \ldots, \theta_k | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) \propto p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k | \theta_1, \ldots, \theta_k) p(\theta_1, \ldots, \theta_k) \tag{4.6.1}$$

This is the posterior distribution, but with the data partitioned according to some relevant groups. Assuming that the observations are conditionally independent, then

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k | \theta_1, \ldots, \theta_k) = p(\boldsymbol{y}_1 | \theta_1) \ldots p(\boldsymbol{y}_k | \theta_k)$$

Then if one assumes the θ_i are independent and condition on $\boldsymbol{y}_i$, the θ_i are also independent in posterior distribution:

$$p(\theta_1, \ldots, \theta_k | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) \propto p(\boldsymbol{y}_1 | \theta_1) p(\theta_1) \ldots p(\boldsymbol{y}_k | \theta_k) p(\theta_k) \propto p(\theta_1 | \boldsymbol{y}_1) \cdots p(\theta_k | \boldsymbol{y}_k) \tag{4.6.2}$$

Consider carefully what equation (4.6.2) implies: If the θ_i are independent then there is no need for multiple comparisons because information about θ_i, obtained via $\boldsymbol{y}_i$, provides *no information about* θ_j, when $i \neq j$. Notice that this insight refers to the parameters, θ_i, and not the data, $\boldsymbol{y}_i$. Therefore, when the parameters are independent of each other there is no need for adjustments for multiplicities when comparing means.[44]

Extension to hierarchical models

However, very few studies allow for such a strong assumption a priori, especially in hierarchical models, where one assumes that parameters are drawn from a common distribution; e.g., should a shepherd assume that knowledge about the survival probability of

[44] In some sense this is analogous to the concept of families of tests. For example, suppose we want to know about differences in height across many nationalities. The set of pairwise tests is a coherent family of tests, and in some sense the distributions of heights for each nationality is still measuring humans. If an alien species landed in Norway and measured the height of 100 Norwegians, this information about human height would tell them something about what to expect if they then flew down to Australia to also measure height. In contrast, this data would give them no information about mean circumference of pine trees across the fifty states of the USA, which would be a separate family.

lambs in one flock tells him/her/them absolutely nothing about the survival probability in another flock? This is unlikely, because even if a large reason for the difference in survival probability between two flocks is largely due to the shepherd's competence, there will be other measured or unmeasured factors that flocks may share in common that may affect these survival probabilities. However, while independence is rarely a reasonable assumption in these circumstances, exchangeability may or may not be a valid assumption; that is, $\theta_1, \ldots, \theta_k$ may be considered a random sample drawn from a population distribution, say with mean θ_0 and variance τ^2.[45] In the case of hierarchical models the posterior distribution for θ_i depends on data $\boldsymbol{y}_i$, but it also depends on the other data $\boldsymbol{y}_j$, $i \neq j$, because the θ_i all give information about the population mean, θ_0, which itself gives information about the θ_i; the process thus "shrinks" the θ_i in each other's direction.

Berry and Hochberg (1999) illustrate this discussion by example where the θ_i are assumed to arise from a normal distribution with hyperparameters assigned by a Dirichlet process prior distribution.[46] Without getting too technical, the posterior distribution of θ_0 and/or τ^2 given the data $\boldsymbol{y_i}$ is a mixture of Dirichlet processes, which implies positive probability to $\theta_i = \theta_j$, $i \neq j$, so that one may find posterior probabilities that any combination of the θ_i are equal to each other. Berry and Hochberg also argue that posterior probabilities regarding whether some $\theta_i = \theta_j$ should account for estimates of other parameters rather than be calculated in isolation.

An application to DNA microarray data

While ultimately the focus of this work will turn to a generalized Hierarchical modeling approach, it is helpful to consider the work of Scott and Berger (2006). Analysis of DNA microarray data tests many thousands of genes for "activation", i.e., over- or under-

[45] A truly Bayesian approach would likely assign vectors of hyperparameters to describe the prior distributions of θ_0 and τ^2.

[46] The details are beyond the scope of this text, note that the Dirichlet distribution is a multivariate generalization of the beta distribution. As the beta distribution is the conjugate prior to the binomial distribution, so is the Dirichlet distribution to the multinomial distribution. A Dirichlet process is essentially a generalization of the Dirichlet distribution (Teh, 2011).

expression in the presence of stimuli. A basic approach to such modeling assumes that data on expression for M genes, $\boldsymbol{y} = (y_1, \ldots, y_M)$, arise from independent normal distributions, i.e., $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$; this implies that the y_i are modeled as a "true mean" μ_i plus some measurement error with variance σ^2.[47] Researchers then need to determine which genes are active ($\mu_i \neq 0$) or inactive ($\mu_i = 0$). Define $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$ and $\gamma_i = \begin{cases} 0 & \text{if } \mu_i = 0 \\ 1 & \text{if } \mu_i \neq 0 \end{cases}$, so that the full likelihood is

$$f(\boldsymbol{y}|\sigma^2, \gamma, \boldsymbol{\mu}) = \prod_{j=1}^{M} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_j - \gamma_j \mu_j)^2}{2\sigma^2} \right) \right] \tag{4.6.3}$$

The following inference is of interest:

1. The posterior probability that $\mu_i = 0$ given the data, i.e. that gene i is inactive.
2. The marginal posterior density of μ_i, given $\mu_i \neq 0$.
3. The joint posterior distribution of the key hyperparameters.

Scott and Berger derive the relevant theory, introduce importance sampling for computation, discuss posterior summaries, consider the implications of various prior distributions for the probability that a particular gene is inactive, and discuss the application of decision theory to the problem; these details are beyond the scope of this work, and here the relevance of their approach is that one can see both the posterior probability that a gene is inactive as well as a prediction for what the distribution would look like, given the mean was non-zero. The relevant derivations depend upon a common prior probability that the $\mu_i = 0$ (are inactive), and that the μ_i are zero or not *independently*, i.e., whether or not a particular gene is active does not depend on the active/inactive status any other genes.

[47] While Scott and Berger call this measurement error, it is not clear that some genes might be activated due to process not necessarily related to the given stimulus, so it is conceivable to this author that errors may encompass more than measurement error.

The question relevant to the current endeavor is whether, when, and to what extent do Bayesian methods automatically correct for multiple testing. Scott and Berger provide some guidance and determine that under this particular set of distributional assumptions some accounting of multiple comparisons occurs. In simulations, where 10 "signal" points and n "noise" points were generated, at least some natural penalty seems to be imposed in that the probabilities that a given gene is non-zero decrease as the number of "noise" observations gets larger. This statistical phenomena arises from the fact that the posterior distribution of the common probability of a gene being inactive nears one as the number of noise observations grows, with the result that there is less evidence that a gene is active if it occurs in a sea of noise; this is related to the Occam's razor effect in Bayesian variable selection, whereby models are penalized for complexity. However, there is a distinction between the Occam's razor effect and multiple testing corrections; this is more clearly expounded upon in Scott and Berger (2010), to which we now turn.

Scott and Berger return: When and how does multiplicity correction occur?

Scott and Berger continued their research into Bayesian multiple comparisons corrections and the result was a 2010 tome exploring the topic in the context of variable selection as applied by Empirical-Bayes and Full-Bayes methodologies (Scott and Berger, 2010).[48] The paper is worth a full read, especially for those interested in discrepancies and convergences between Empirical-Bayes and Full-Bayes methodology, but its relevance this work is their contributions to understanding how and when Bayesian methodology automatically makes corrections for multiple testing.

Multiple testing correction in a Bayesian framework occurs through prior specification; while this fact was noted briefly in their previous work, the insight is more thoroughly

[48] Briefly, Empirical-Bayesian procedures treat hyperparameters in prior distributions as quantities to be estimated from the data. Full Bayesian approaches assign the prior and sample from its distribution. In the context of the cited paper, the priors of concern are the probabilities that any parameter is 0, and these arise from prior probabilities that a model under consideration is the right one. Empirical-Bayes would estimate these probabilities from the data while full Bayes would assign the priors.

investigated and clearly explicated in their later paper. In the context of variable selection, these priors relate to the prior probabilities that certain parameters are or are not zero, which arise from prior probabilities that particular models are the correct one. However, an important distinction must be made between multiple testing correction and the so-called "Occam's Razor" effect: The Occam's razor effect is essentially a penalty for *more complex models*, and results from integrating the likelihood over parameter spaces of higher dimension, producing relatively more dispersed predictive distributions. That this is no multiplicity correction can be seen intuitively by noting that Bayes Factors are invariant to the number of other models under consideration.[49]

Not all prior specifications result in multiplicity adjustment; most importantly, assigning a common prior probability of inclusion of 1/2 for each parameter provides no multiple testing adjustment. One may subjectively adjust this prior to address beliefs about the proportion of included variables in the correct model, but there is no prior that can be set independently of the number of variables under consideration and still implement multiple testing corrections. Note that Empirical-Bayes approaches confront multiple testing by the fact that, given a fixed number of true inclusion variables, the estimated common probability of inclusion shall approach 0 as the number of variables under consideration increases. Using a uniform prior for probability of inclusion results in marginal inclusion probabilities of 1/2, but this approach in not equivalent to simply setting the inclusion probabilities to be 1/2 since the probability is not divided among models in the same way.[50] The lesson is that in variable selection problems, multiple testing correction occurs via the prior inclusion probabilities, but the degree thereof is connected to the data.

[49] From Gelman et al. (2013), the Bayes factor is defined as the ratio of marginal densities under 2 models in question; i.e. for data $\boldsymbol{y}$, and 2 models H_1, H_2 with respective parameter vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, the Bayes Factor is $\frac{p(\boldsymbol{y}|H_2)}{p(\boldsymbol{y}|H_2)} = \frac{\int p(\boldsymbol{\theta}_2|H_2)p(\boldsymbol{y}|\boldsymbol{\theta}_2,H_1)\,d\boldsymbol{\theta}_2}{\int p(\boldsymbol{\theta}_1|H_1)p(\boldsymbol{y}|\boldsymbol{\theta}_1,H_1)\,d\boldsymbol{\theta}_1}$.

[50] The uniform prior is implemented via a beta distribution with both parameters equal to 1.

Summary

Multiple testing is of no concern for the Bayesian in the simple example where the joint posterior distribution can be factored into a product of marginal posterior distributions, arising from independence of parameters. However, this is a poor assumption for most data, and in particular is both philosophically and mathematically incoherent for hierarchical models. An accounting for multiple testing in this scenario can be achieved by considering all combinations of hypotheses containing the hypothesis of interest; e.g. if a shepherd wants to know whether 2 flocks have equal survival probabilities for its lambs, then he should consider all possible combinations of equality/inequality that contain this hypothesis. In more complicated data, this kind of approach is perhaps infeasible, and while there exists a penalty on model complexity inherent to Bayesian methodology, this is not equivalent to multiple testing correction; such correction is tied to the choice of priors. While it is not possible to go into detail here, note that Scott and Berger's work provide both theoretical and simulation-based justification for these arguments.

5 Bayesian Shrinkage Priors and Generalized Linear Models

5.1 Bayesian Generalized Linear Models

While linear models are commonly associated with classical statistics, they are also ubiquitous in Bayesian applications; the difference is, of course, that Bayesian applications explicitly incorporate prior distributions for the parameters in the models. One common approach for including the priors is as "additional data points", where a data point is added for each parameter and the prior variance is included in the covariance matrix (Gelman et al., 2008, 2013). This approach allows for fitting ill-posed models, as well as incorporating prior distributions as prior information or tuning parameters. This section considers a particular variation on this approach that was initially proposed for genetic research.

5.2 *The Basic GLM*

In general, one begins with a vector of outcomes $\boldsymbol{y} = (y_1, \ldots, y_n)$ which can be divided into $K \geq 1$ groups, $G_k, k = 1, \ldots, K$ where the K^{th} group contains $J_k > 1$ variables; in Yi et al. (2014) these are genetic predictor variables (Yi et al., 2014).[51] Non-grouped covariates may also be collected and incorporated into the analysis. The form of the generalized linear model is as follows:

$$h(\mathrm{E}(y_i|\boldsymbol{X}_i)) = \beta_0 + \sum_{j=1}^{J_0} x_{ij}^c \beta_j^c + \sum_{k=1}^{K} \sum_{j \in G_k}^{J_k} x_{ij}^g \beta_j^g = \boldsymbol{X}_i \boldsymbol{\beta}, \quad i = 1, \ldots, n \tag{5.2.1}$$

where h is the link function, n is the number of individuals, β_0 is the intercept, x_{ij}^c, x_{ij}^g are observed covariates and genetic variables, respectively, the coefficients β_j^c, β_j^g are the non-genetic and genetic effects, respectively, $j \in G_k$ indicates membership in group j, $\boldsymbol{X}_i$ is an $1 \times J$ vector containing all variables, $\boldsymbol{\beta}$ is a $J \times 1$ vector containing all coefficients as well as the intercept; i.e. $\boldsymbol{X}_i = (1, x_{i1}, \ldots, x_{iJ})$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_J)'$, and $J = \sum_{k=0}^{K} J_k$ is the total number of variables.[52]

Note that the term $\sum_{k=1}^{K} \sum_{j \in G_k}^{J_k} x_{ij}^g \beta_j^g$ can be adapted to also contain parameters for at the group level (Yi et al., 2011):

$$\sum_{k=1}^{K} g_k \sum_{j \in G_k}^{J_k} x_{ij}^g \beta_j^g \tag{5.2.2}$$

where g_k is the *group effect* and α_j are the *weights* of the j variants. Hierarchical prior distributions are necessary here in order to obtain an identifiable model. Group effects are particularly useful when the the effects of particular variants may be small or moderate but yield a significant cumulative effect. This approach includes as special cases several other methods in the literature regarding rare variants; further details are found in Yi et al. (2011).

[51] As a look ahead, these effects can be mapped from genes to pixels or voxels.

[52] As this a general formulation, note that the vector $\boldsymbol{X}_i$ could be extended to include multiple observations for a given subject.

The data distribution, or likelihood, is defined as

$$p(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, \phi) = \prod_{i=1}^{n} p(y_i|\boldsymbol{X}_i\boldsymbol{\beta}, \phi) \tag{5.2.3}$$

where this distribution can generally be allowed to take the form of any distribution that can be expressed as an exponential family, and ϕ is the dispersion parameter. For some distributions, e.g. Poisson, this parameter is not necessary and thus $\phi = 1$.

Generalized linear models are usually fit via iterated weighted least squares, which proceeds by constructing the pseudo-response z_i and pseudo-weight w_i for each data point y_i, given the current estimates $(\hat{\boldsymbol{\beta}}, \hat{\phi})$:[53]

$$z_i = \hat{\eta}_i - \frac{L'(y_i|\hat{\eta}_i)}{L''(y_i|\hat{\eta}_i)}$$
$$w_i = -L''(y_i|\hat{\eta}_i)$$

and the likelihood is thus approximated by the weighted normal likelihood:

$$p(y_i|\boldsymbol{X}_i\boldsymbol{\beta}, \phi) \approx \mathcal{N}(z_i|\boldsymbol{X}_i\boldsymbol{\beta}, w_i^{-1}\phi)$$

where $\hat{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta}$, $L(y_i|\hat{\eta}_i) = \log p(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, \phi = 1)$, $L'(y_i|\eta_i) = \frac{\partial L(y_i|\eta_i)}{\partial \eta_i}$, and $L''(y_i|\eta_i) = \frac{\partial^2 L(y_i|\eta_i)}{\partial \eta_i^2}$. An iterative process then updates the parameters $(\boldsymbol{\beta}, \phi)$ until reaching convergence. Normal linear regression has $z_i = y_i$ and $w_i = 1$, eliminating the need for iterations.

5.3 Hierarchical Models

Previous sections discussed hierarchical models in both the frequentist and Bayesian perspectives, but following Yi et al. (2014), more details are explored. While classical generalized linear models with either too many or correlated coefficients can be nonidentifiable, Bayesian inference has been used to circumvent such issues.

[53] See for example McCulloch et al. (2008).

Let the coefficients be normally distributed with 0 mean and unique variance, τ_j^2:

$$\beta_j|\tau_j^2 \sim \mathcal{N}(0, \tau_j^2), \quad j = 1, \ldots, J \tag{5.3.1}$$

Let the intercept β_0 and dispersion ϕ have noninformative priors; e.g. $p(\beta_0|\tau_0^2) = \mathcal{N}(0, \tau_0^2)$ for τ_0^2 large and $p(\log \phi) \propto 1$. Given τ_j^2, the conditional posterior distribution for β is approximated by $\mathcal{N}_J(\hat{\boldsymbol{\beta}}, \mathrm{var}(\hat{\boldsymbol{\beta}}))$ where

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X'\Sigma_z^{-1}X} + \boldsymbol{\Sigma_\beta})^{-1}\boldsymbol{X'\Sigma_z^{-1}z} \\
\mathrm{var}(\hat{\boldsymbol{\beta}}) &= (\boldsymbol{X'\Sigma_z^{-1}X} + \boldsymbol{\Sigma_\beta})^{-1}\phi \\
\Sigma_z &= \mathrm{diag}(w_1^{-1}, \ldots, w_n^{-1}) \\
\Sigma_{\boldsymbol{\beta}} &= \mathrm{diag}(\tau_0^2/\phi, \ldots, \tau_j^2/\phi)
\end{aligned}$$

Coefficients with prior variance equal to zero are shrunk to precisely zero, so that no infinities shall be involved in calculations of the above estimates; alternatively, the prior variance could be set to a very low value with very similar results. When $\phi \neq 1$, it is updated by

$$\hat{\phi} = (\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\Sigma_z^{-1}(\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n$$

which results in well defined $\hat{\boldsymbol{\beta}}$ with finite variance, no matter whether the data have high dimension or are highly correlated.

5.4 Shrinkage Priors

In general, shrinkage priors consist of using the variance parameter(s) in a prior distribution to adaptively shrink parameters that are not important to the outcome of interest, while leaving important parameters relatively unchanged; this is a particular variant of Bayesian variable selection. In the previous section, the β_j would be subject to shrinkage on account of the τ_j values, and the adaptive nature of the process is tied to the hyperprior

specifications of the τ_j. Yi and Ma (2012) discuss the use of shrinkage priors as well as the limitations of other penalization approaches for handling correlation and/or high dimensional data, and the competing approaches described therein tend to have one or more of the following limitations:

1. Unique tuning parameters applied to all coefficients, which may result in over- or under-shrinking of coefficients varying in importance.

2. Methods with varying penalty parameters have thus far tended to depend too heavily on the quality of initial estimates.

3. Inadequate ability to include or handle hierarchical structures.

The variance parameters τ_j^2 determine parameter shrinkage whereby $\tau_j^2 = 0$ results β_j shrinking to 0, $\tau_j^2 = \infty$ does not shrink β_j at all, thus it contributes 1 degree of freedom. It is worth explaining why this is the case. Consider (4.3.7) to see that if $\tau_j^2 = 0$ then $\beta_j = 0$ has *no variance* and is thus a constant. If $\tau_j = \infty$, or in practice is simply very large, then the distribution of β_j is so disperse that any value of β_j is as believable as any other and there is little reason to pull the estimate towards the prior mean.

The τ_j^2 are assumed unknown and have variances themselves. Yi and Ma recommend two priors: the half-Cauchy and exponential priors. Modeling the half-Cauchy prior requires expression hierarchically as follows:

$$\tau_j^2|s_j \sim \text{Inv-}\chi^2(\nu, s_j^2), \quad s_j^2|b_{k[j]} \sim \text{Gamma}(a, b_{k[j]}), \quad p(\log b_k) \propto 1 \tag{5.4.1}$$

Alternatively, an exponential prior for τ_j^2 is given by

$$\tau_j^2|s_j \sim \text{Exponential}(s_j^2/2), \quad s_j|b_{k[j]} \sim \text{Gamma}(a, b_{k[j]}) \tag{5.4.2}$$

where $k[j]$ indexes the group k within which the j^{th} group resides.[54] Yi and Ma (2012) discuss further details regarding selection of the hyperparameters for equations 5.4.1 and 5.4.2. The exponential approach leads to higher power and quicker convergence and is therefore the general recommendation. The important detail is that the group specific b_k allow for pooling information within groups as well as different shrinkage among groups.

5.5 Computation

Computational implementation combines the Expectation-Maximization (EM) algorithm with the iterated weighted least squares (IWLS) algorithm (EM-IWLS algorithm) where the τ_j^2 and their hyperparameters are treated as missing data so that $(\boldsymbol{\beta}, \phi)$ are updated by averaging over the missing data. The general process as described by Yi and Ma (2012) is briefly summarized as follows:

1. Start with a crude parameter estimate.
2. For some number of iterations:
   (a) E-step: Take conditional expectations with respect to the parameters τ_j^2, s_j, b_k.
   (b) M-step: Calculate the pseudo-data and pseudo-weights, update the β estimates via the augmented weighted normal linear regression, and update ϕ if necessary.

At convergence the estimates $(\hat{\boldsymbol{\beta}}, \hat{\phi})$ and $\mathrm{var}(\hat{\boldsymbol{\beta}})$ obtained. P-values may be obtained for hypotheses $H_0 : \beta_j = 0$ via the test statistic $\frac{\hat{\beta}_j}{\sqrt{\mathrm{var}(\hat{\beta}_j)}}$, which follows an approximate standard normal when ϕ is not included or t-distribution with n degrees of freedom when ϕ is included.[55]

[54] Modeling s_j rather than s_j^2 allows easier computation of the posterior distribution.

[55] It may seem odd for a Bayesian method to use "traditional" p-values, but in practice the approach appears to produce reasonable results, especially when combined with the hierarchical Bonferroni procedure described below. See Yi et al. (2014).

5.6 *Effective Number of Effects*

The *effective number of parameters* or *degrees of freedom* is an indication of Bayesian hierarchical model complexity, defined by the posterior mean of deviance minus the deviance at the posterior mean or mode (Gelman et al., 2013):

$$\rho = \bar{D}(\boldsymbol{\theta}) - D(\hat{\boldsymbol{\theta}}) \tag{5.6.1}$$

where $\boldsymbol{\theta}$ includes all parameters, $D(\boldsymbol{\theta}) = -2\log\{p(\boldsymbol{y}|\boldsymbol{\theta})\}$, $\bar{D}(\boldsymbol{\theta})$ is the posterior mean of $D(\boldsymbol{\theta})$, and $\hat{\boldsymbol{\theta}}$ is the posterior mean or mode of $\boldsymbol{\theta}$. While this is generally obtained via posterior simulation, Yi et al. (2014) approximately estimate equation (5.6.1) for any subset of parameters conditional on the estimates of the other parameters from the EM-IWLS algorithm.

Since the generalized linear likelihood is approximated by the weighted normal likelihood, the expectation of D is taken with respect to the conditional posterior distribution of genetic effects $p(\boldsymbol{\beta}^g|\beta_0, \boldsymbol{\beta}^c, \phi, \boldsymbol{\tau}^2)$ where $\boldsymbol{\beta}^g, \boldsymbol{\beta}^c, \boldsymbol{\tau}^2$ are the vectors of all genetic effects, covariate effects, and variances, respectively, so that the conditional posterior distribution is approximated by:

$$p(\boldsymbol{\beta}^g|\beta_0, \boldsymbol{\beta}^c, \phi, \boldsymbol{\tau}^2) \sim \mathcal{N}(\hat{\boldsymbol{\beta}}^g, \boldsymbol{V}_g)$$

where $\boldsymbol{V}_g = (\boldsymbol{X}_g'\boldsymbol{\Sigma}_z^{-1}\phi^{-1}\boldsymbol{X}_g + \boldsymbol{\Sigma}_{\beta^g}^{-1})^{-1}$, $\boldsymbol{X}_g$ is the design matrix of $\boldsymbol{\beta}^g$, and $\boldsymbol{\Sigma}_{\beta^g}$ is a diagonal matrix consisting of the τ_j^2 for genetic effects. Thus, $\bar{D} = D(\hat{\boldsymbol{\beta}}) + \mathrm{Tr}(\boldsymbol{X}_g'\boldsymbol{\Sigma}_z^{-1}\phi^{-1}\boldsymbol{X}_g\boldsymbol{V}_g)$ so that the effective number of genetic effects is

$$\rho = \mathrm{Tr}(\boldsymbol{X}_g'\boldsymbol{\Sigma}_z^{-1}\phi^{-1}\boldsymbol{X}_g\boldsymbol{V}_g) = J_g - \mathrm{Tr}(\boldsymbol{\Sigma}_{\beta^g}^{-1}\boldsymbol{V}_g) \tag{5.6.2}$$

where $J_g = \sum_{k=1}^{K} J_k$ is the total number of genetic effects; this has the interpretation as being the effective number of tests in a hypothesis testing framework, which shall be more

obvious when the hierarchical Bonferroni procedure is described in the next section. It is straightforward to adjust equation (5.6.2) to determine the effective number of effects within any subset of coefficients as well as the effective number of coefficients:

$$\rho = (J+1) - \text{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{V})$$

where $J+1$ is the total number of coefficients, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \text{diag}(\tau_0^2, \ldots, \tau_j^2)$, and $V = (\boldsymbol{X'\Sigma_z^{-1}}\phi^{-1}\boldsymbol{X} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1}$. It is clear that $0 \leq \rho \leq J_g$ so that $\text{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{V})$ represents the reduction in the parameters and depends directly on the τ_j^2.

5.7 Hierarchical Bonferroni

Hierarchical modeling results in dependencies among parameters and lowers the number of independent tests, which makes traditional Bonferroni too conservative; however, the effective number of tests can be used to construct the *hierarchical Bonferroni correction*:

$$p_j' = \min(1, J_\rho \cdot p_j) \tag{5.7.1}$$

where $J_\rho = \max(\rho, (\rho + 0.05 \cdot J_g)/2)$ and p_j and p_j' are the adjusted and initial p-values for $H_0 : \beta_j = 0$, respectively.

6 The Spike-and Slab Lasso (SSL)

A relatively new development is the Spike-and-Slab Lasso, developed by Ročková and George (2018) and adapted for genetic research by Tang et al. (2017) and Tang et al. (2018). This section describes these advances, as well as first covering background necessary for full comprehension. Most of the background information is covered to some extent in these cited works, but some topics deserve further description, given the relatively unlimited nature of this document.[56]

[56] At least for now. It is the author's hope that the document remain finite in both space and time.

6.1 Background: Lasso and Penalized Likelihood in Brief

Before tackling the spike-and-slab lasso, it is helpful to understand the "regular" lasso. Here we provide an overview of the lasso, as well as a brief discussion of penalized likelihood methods in general.

Lasso 1996

Robert Tibshirani originally developed the lasso as a biased estimation approach for variable selection in linear models of the kind discussed in section 2.1. OLS estimates are sometimes unsatisfactory in that while they are unbiased they can still have large variance, which may result in poor prediction accuracy. Interpretation is often obfuscated by large numbers of variables and a smaller subset of those exhibiting the strongest effects is often desired. lasso was developed to address concerns in two other approaches: Ridge regression and subset selection. Ridge regression adds a penalty to the likelihood and ultimately shrinks coefficient values from their OLS values. Subset selection is a discrete process of removing "unimportant" variables. The "Least Absolute Shrinkage and Selection Operator" (lasso) approach was designed to keep the better aspects of Ridge Regression and Subset selection, by shrinking coefficients (like ridge regression), some just a little and some to zero (unlike ridge regression). The fact that some coefficients are shrunk to zero effectively removes them from the model, performing a kind of variable selection. The lasso is given by:

$$\hat{\boldsymbol{\beta}} = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{J} x_{ij}\beta_j \right)^2 \right\}, \quad \text{subject to} \sum_{j=1}^{J} |\beta_j| \leq t \tag{6.1.1}$$

with all definitions the same as section 2.1.[57] Tibshirani goes on to show the form of the lasso estimates, algorithms for obtaining said estimates, and compares it to several competing methods, all of which can be found in Tibshirani (1996). The insight here is that this estimation approach can both improve the variance of estimates as well as (at least some

[57] Note that the difference between equation (6.1.1) and Ridge regression is that ridge regression is subject to $\sum_{j=1}^{J} \beta_j^2 \leq t$.

of the time) reduce "unimportant" variables to exactly zero, performing automatic variable selection and in general avoiding the need for traditional hypothesis testing. A final benefit is that the lasso also has a Bayesian interpretation as applying double exponential priors to the β_j, which makes it an attractive approach for both frequentists and Bayesians alike (Park et al., 2012).

Penalized likelihoods

The lasso method maximizes a *penalized likelihood*. Standard maximum likelihood maximizes the likelihood equation with respect to some parameter vector $\boldsymbol{\theta}$; penalized likelihoods add a term to the objective function (likelihood) to "penalize" various undesirable possibilities. For example, smoothing splines penalize roughness in estimation, and thereby avoid overfitting. The general form of penalized optimization is given by:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \{L(\boldsymbol{\theta}|\boldsymbol{y}) + \lambda \operatorname{pen}(\boldsymbol{\theta})\} \tag{6.1.2}$$

where $L(\cdot)$ is the likelihood, $\boldsymbol{y}$ is a vector of observations, λ is a tuning parameter, and $\operatorname{pen}(\boldsymbol{\theta})$ is a penalty function. Equation 6.1.2 can be generalized and re-expressed in this form as follows:

$$\hat{\boldsymbol{\beta}} = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{J} x_{ij}\beta_j \right)^2 + \lambda \operatorname{pen}(\boldsymbol{\beta}) \right\} \tag{6.1.3}$$

where, for example $\operatorname{pen}(\beta) = \sum_{j=1}^{J} \beta_j^2$ for ridge regression and $\operatorname{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{J} |\beta_j|$ for lasso. The nail through the foot is the tuning parameter λ, which needs to be selected by the analyst. Once again, it is beyond the scope of this document to examine this topic in greater detail, but note that several methods have been developed to avoid arbitrary λ selection, most commonly some form of cross-validation. Various adaptations of generalized cross validation or k-fold cross validation are common approaches to choosing λ (Craven and Wahba, 1979; Efron and Tibshirani, 1993; Golub et al., 1979; Hastie et al., 2009).

Note finally that the form penalized likelihood employed in Rockeva and George primarily addresses the case of the linear model in equation (2.1.2) and gives the following penalized estimated:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^J}{\arg\max} \left\{ -\frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \text{pen}_\lambda(\boldsymbol{\beta}) \right\} \tag{6.1.4}$$

where they have chosen to simply index the penalty function by the parameter λ. It is important to realize that equations 6.1.3 and 6.1.4 represent the same basic process; that is, for a given penalty function the results would not differ between them.[58]

6.2 The Spike-and-Slab Lasso: Theory

Spike-and-slab priors

Spike-and-slab priors are fundamentally mixture priors, i.e., priors consisting of mixture distributions. This prior formulation allows more flexibility in modeling/estimation compared to traditional conjugate priors (Ročková and George, 2018). Spike-and-slab priors handle sparse data by creating a mixture containing a "slab" distribution, which has a wider variance to model the effects of relatively larger magnitude, and a "spike" distribution, which has a narrow variance to model effects that are relatively unimportant. The general idea being that the slab distribution models parameters that should be included in the model, and the spike distribution models parameters that should be excluded from the model.

Following the notation in Ročková and George (2018), the spike-and-slab prior is given by:

$$\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \prod_{j=1}^{J} \left[\gamma_j \psi_1(\beta_j) + (1-\gamma_j)\psi_0(\beta_j)\right], \quad \boldsymbol{\gamma} \sim \pi(\boldsymbol{\gamma}) \tag{6.2.1}$$

[58] Key differences are: 1. The former notation actually performs a minimization and the latter a maximization, which is accounted for by the choice of whether to include minus signs. 2. Presumably, using the notation pen_λ allows for somewhat greater flexibility, but generally the chosen form still consists of multiplying a penalty parameter by a penalty function. 3. The latter notation is in matrix, rather than scalar, form.

where the slab distribution, $\psi_1(\beta)$, is diffuse to account for "signal" or large effects, and the spike distribution, $\psi_0(\beta)$ is narrow to account for "noise" or small effects, depending on the context. These are given double exponential distributions in Ročková and George (2018):[59]

$$\psi_1(\beta) = \frac{\lambda_1}{2} e^{-\lambda_1|\beta|}$$
$$\psi_0(\beta) = \frac{\lambda_0}{2} e^{-\lambda_0|\beta|}$$

where generally $\lambda_0 >> \lambda_1$.[60]

While other priors for $\boldsymbol{\gamma}$ are possible, the one that concerns us here is that appropriate for subset selection, i.e., the goal in mind is to separate signal from noise. Thus, prior distribution for γ is Binomial:

$$\pi(\boldsymbol{\gamma}|\theta) = \prod_{j=1}^{J} \theta^{\gamma_j}(1-\theta)^{1-\gamma_j}, \quad \theta \sim \pi(\theta) \tag{6.2.2}$$

where $\theta = P(\gamma_j = 1|\theta)$ denotes prior beliefs about the fraction of the β_j that are signal rather than noise.[61]

The SSL prior

The SSL prior builds a penalty from equations 6.2.1 and 6.2.2:

$$\pi(\boldsymbol{\beta}|\theta) = \prod_{j=1}^{J} [\theta\psi_1(\beta_j) + (1-\theta)\psi_0(\beta_j)] \tag{6.2.3}$$

[59] These are also known as Laplace distributions, depending on which textbook you cut your teeth. The author has no allegiance to either.

[60] Note that as λ increases the distribution becomes sharper, and as λ decreases the distribution becomes more diffuse. Thus, large λ_0 produces a spiky distribution and small λ_1 produces a flatter distribution comparatively, hence the names.

[61] This may appear to be quite arbitrary, but it may often be the case that the research question on hand gives some indication of some appropriate ranges. For example, if the case that the predictors are pixels or voxels from an MRI image and the outcome is some behavioral score, then coarser scores should employ large θ compared to finer scores, because coarse scores should be expected to be associated with larger sub-volumes of the brain, while finer measures should be expected to be associated with smaller sub-volumes.

Equation 6.2.3 has margined out $\boldsymbol{\gamma}$ in equation (6.2.2) and may be incorporated as a penalty directly into equation (6.1.4). This approach is ultimately a compromise between best subset selection and the lasso, and collapses into each under certain circumstances. Best subset selection is obtained when $\psi_0(\beta_j) = I(\beta_j = 0)$ (i.e. $\lambda_0 \to \infty$) and $\psi_1(\beta_j) \propto c > 0$ (i.e. $\lambda_1 \to 0$), while assigning $\psi_1(\beta_j) = \psi_0(\beta_j)$ is equivalent to the traditional lasso described above. However, a fully Bayesian approach treats θ as random as well, assigning it a prior distribution; while the mathematics may become more difficult as a result, which is mostly left to Ročková and George (2018), there are several benefits, including but not limited to greater ability to handle sparsity, potentially automatic multiple testing adjustments, and avoiding the need for cross-validation with respect to θ.

Coordinate-wise optimization for fitting the SSL

Ročková and George (2018) recommend coordinate-wise optimization for fitting the SSL model. Most penalties in the literature are separable, but not the SSL penalty. However, the similarity between approaches for the traditional Bayesian lasso and the SSL, along with some theory shown in the original SSL paper, allows Rockeva and George to develop a coordinate ascent algorithm.

6.3 The Spike-and-Slab Lasso Generalized Linear Model

Building off Ročková and George (2018), Tang et al. (2017) apply the spike-and-slab lasso to generalized linear models in a Bayesian framework of the GLMs in section 2.2; i.e., the responses should able to be treated as arising from a distribution in the exponential family, where the linear predictor is given by $\eta_i = \beta_0 + \sum_{j=1}^{J} x_{ij}\beta_j = h(\mathrm{E}(Y_i|X_i))$ and $h(\cdot)$ is a suitable link function. Thus, the data distribution (likelihood) is given as:

$$p(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, \phi) = \prod_{i=1}^{n} p(y_i|\boldsymbol{X}_i\boldsymbol{\beta}, \phi) \tag{6.3.1}$$

where $\boldsymbol{y}$ is the $n \times 1$ response vector, $\boldsymbol{X}$ is the $n \times J$ design matrix, $\boldsymbol{\beta}$ is the $(J+1) \times 1$ parameter vector, and ϕ is the dispersion parameter. We have previously mentioned the inability of this kind of model to handle situations where $J > n$ (non-identifiability), or more generally when it is not the case that $n >> J$ (over-fitting). The lasso and Bayesian applications thereof have provided work-arounds to these problems by fitting lasso GLM's with the cyclic coordinate descent algorithm (Friedman et al., 2010). However, as mentioned in section 6.2, the lasso has only one parameter, λ, and so cannot apply varying levels of shrinkage to different parameters. Therefore, the benefits described in section 6.2 for the general linear model case would apply to generalized linear models as well, if there were an algorithm to fit the model.

The EM coordinate descent algorithm

In order to avoid confusion, notation is given to match Tang et al. (2017). Define $\lambda_1 = 1/s_1$ and $\lambda_0 = 1/s_0$, so that

$$\beta_j | \gamma_j, s_0, s_1 \sim \frac{1}{2S_j} e^{-|\beta_j|/S_j} \tag{6.3.2}$$

where β_j and γ_j are interpreted as in section 6.2 and $S_j = \begin{cases} s_1, & \gamma_j = 1 \\ s_0, & \gamma_j = 0 \end{cases}$. The algorithm developed combines the Expectation Maximization (EM) algorithm with the previously cited coordinate descent algorithm and is aptly named the "EM Coordinate Descent Algorithm". The process is as follows:

1. First, set up the log joint posterior distribution of $(\boldsymbol{\beta}, \phi, \boldsymbol{\gamma}, \theta)$:

$$\begin{aligned}\log p(\boldsymbol{\beta}, \phi, \boldsymbol{\gamma}, \theta|\boldsymbol{y}) &= \log p(\boldsymbol{y}|\boldsymbol{\beta}, \phi) + \sum_{j=1}^{J} \log p(\beta_j|S_j) + \sum_{j=1}^{J} \log p(\gamma_j|\theta) + \log p(\theta) \\ &\propto \ell(\boldsymbol{\beta}, \phi) - \sum_{j=1}^{J} \frac{1}{S_j}|\beta_j| + \sum_{j=1}^{J} (\gamma_j \log(\theta) + (1-\gamma_j)\log(1-\theta))\end{aligned} \tag{6.3.3}$$

   Note that the algorithm needs starting values $\boldsymbol{\beta}^0, \phi^0, \theta^0$. In many cases $\phi = 1$ as a base assumption, but $\phi^0 = 1$ may still be an appropriate starting point if one wishes or needs to estimate (over)dispersion. Similarly, $\boldsymbol{\beta}^0$ should often begin as a vector of zeroes, absent strong prior knowledge. It is perhaps θ^0 that is most interesting to the author, but discussions of its choice are deferred until later. It is not clear that a default starting value is appropriate for all situations.

2. *E-Step:* The values for the γ_j are unknown, so the algorithm treats them as "missing data" and averages over their posterior distribution to estimate $(\boldsymbol{\beta}, \phi, \theta)$. The E-step takes the expectation of equation (6.3.3) with respect to conditional posterior distribution of the γ_j, which is derived as:

$$\begin{aligned}p_j &= p(\gamma_j = 1|\beta_j, \theta, \boldsymbol{y}) \\ &= \frac{p(\beta_j|\gamma_j = 1, s_1)p(\gamma_j = 1, \theta)}{p(\beta_j|\gamma_j = 0, s_0)p(\gamma_j = 0, \theta) + p(\beta_j|\gamma_j = 1, s_1)p(\gamma_j = 1, \theta)}\end{aligned} \tag{6.3.4}$$

   where $\theta = p(\gamma_j = 1, \theta)$, $1 - \theta = p(\gamma_j = 0, \theta)$, and $p(\beta_j|\gamma_j = k, s_k)$, $k = 0, 1$ are given by equation (6.3.2). On account equation (6.3.4), it can be shown that the conditional posterior expectation of S_j^{-1} is as follows:

$$\begin{aligned}\mathrm{E}(S_j^{-1}|\beta_j) &= E\left(\frac{1}{(1-\gamma_j)s_0 + \gamma_j s_1}\middle|\beta_j\right) \\ &= \frac{1-p_j}{s_0} + \frac{p_j}{s_1}\end{aligned} \tag{6.3.5}$$

Larger estimates for β_j result correspondingly larger estimates for p_j, S_j, so that coefficients are not uniformly shrunk.

3. *M-Step:* Update $(\boldsymbol{\beta}, \phi, \theta)$ by maximizing equation (6.3.3) while using the posterior expectations from the E-step for γ_j, S_j^{-1}. The parameters $(\boldsymbol{\beta}, \phi)$ and θ are updated separately since they are included in separate terms in equation (6.3.3). The following respective terms are then maximized with respect to the appropriate parameters:

$$Q_1(\boldsymbol{\beta}, \phi) = \ell(\boldsymbol{\beta}, \phi) - \sum_{j=1}^{J} \frac{1}{S_j} |\beta_j| \tag{6.3.6}$$

$$Q_2(\theta) = \sum_{j=1}^{J} [\gamma_j \log(\theta) + (1 - \gamma_j) \log(1 - \theta)] \tag{6.3.7}$$

The second term in equation (6.3.6) functions as a lasso penalty with $1/S_j$ as the penalty parameter. Recall that a Bayesian lasso GLM is appropriately maximized via the cyclic coordinate descent algorithm; it is so here. $Q_2(\theta)$ is maximized by $\hat{\theta} = \frac{1}{J} \sum_{j=1}^{J} p_j$.

4. Ultimately, a convergence criterion is necessary. Tang et al. (2017) use

$$\frac{|d^{(t)} - d^{(t-1)}|}{0.1 + |d^{(t)}|} < \epsilon$$

where $d^{(t)} = -2 \log \ell(\beta^{(t)}, \phi^{(t)})$ is the deviance estimate at iteration t and ϵ is suitably small.

6.4 Selecting Shrinkage Parameters

The general strategy for selecting shrinkage parameters is to fix s_1, or equivalently λ_1, and perform analysis using a sequence of values for s_0, or equivalently λ_0. The formulation used by Ročková and George (2018) involves selecting a sequence $\{\lambda_0^1 < \cdots < \lambda_0^L\}$, where λ_0^1 is near or equal to λ_1 while preserving $\lambda_0 \geq \lambda_1$. Then an appropriate algorithm for fitting the SSL is run with λ_0^1 and starting point $\boldsymbol{\beta}^0 = \mathbf{0}$ such that an estimate for $\boldsymbol{\beta}$ obtained. This

estimate for $\boldsymbol{\beta}$ is then used as the starting point for re-running the algorithm with λ_0^2. This process is continued until some L is reached where the solution is no longer affected by an increase in λ_0.

The formulation used by Tang et al. (2017) is similar, except that since they use S_j, the sequence is given as $\{s_0^1 < \cdots < s_0^L\}$ where $s_0^1 > 0$ and $s_0^L < s_1$. From this sequence of models one choose the best fitting model with any of the following measures:

1. *Deviance:* A goodness-of-fit measure appropriate for GLM's and given by:

$$d = -2\sum_{i=1}^{n} \log p(y_i|X_i\hat{\boldsymbol{\beta}}, \hat{\phi}) \tag{6.4.1}$$

2. *Mean Squared Error (MSE):*

$$MSE = \frac{1}{n}(y_i - \hat{y}_i)^2 \tag{6.4.2}$$

3. *AUC* and *Misclassification:* For logistic regression additional measures may be appropriate: The area under the ROC curve (AUC) and misclassification, the latter of which is defined as follows:

$$\frac{1}{n}\sum_{i=1}^{n} I(|y_i - \hat{y}_i| > 0.5) \tag{6.4.3}$$

where $I(|y_i - \hat{y}_i| > 0.5) = \begin{cases} 1, & |y_i - \hat{y}_i| > 0.5 \\ 0, & |y_i - \hat{y}_i| \leq 0.5 \end{cases}$.

4. *Pre-validated Linear Predictor Analysis:* This approach uses the estimated linear predictor $\hat{\eta}_i$ as a continuous covariate in a univariate GLM and then examines relevant statistics to assess predictive performance.

7 Modeling Scalar Outcomes with Images as Predictors

In section 1 the problem of interest was briefly introduced, but the discussed lacked a more precise mathematical framework for modeling beyond saying something like “How can one determine what aspects of this image affect some scalar outcome?" In the preceding sections many variations on the linear model were discussed, and linear models are often the most natural framework for handling such questions. In this section the basic form of the model is introduced before discussing its relation to the various estimation approaches discussed above.

Assume that there is some $n \times 1$ vector of outcomes $\boldsymbol{Y}$, where n is the number of subjects. Each subject has an accompanying image, which may consist any real numbers; i.e. an $A \times B$ matrix given by:

$$\boldsymbol{M}_i = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1B} \\ m_{21} & m_{22} & \dots & m_{2B} \\ \vdots & \vdots & \ddots & \vdots \\ m_{A1} & m_{A2} & \dots & m_{AB} \end{bmatrix} = \{m_{ab}\}_{A \times B} \tag{7.0.1}$$

where $m_{ab} \in \mathbb{R}$. While a matrix cannot be included in a linear model as stated in equation (2.1.1), the matrix $\boldsymbol{M}_i$ may be converted into a vector in a principled way by concatenating the rows of $\boldsymbol{M}_i$:

$$\boldsymbol{X}_i' = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iJ} \end{bmatrix} = \begin{bmatrix} m_{11} \\ m_{12} \\ \vdots \\ m_{1B} \\ m_{21} \\ \vdots \\ m_{2B} \\ \vdots \\ m_{AB} \end{bmatrix} \tag{7.0.2}$$

where $J = A * B$. In such a manner one can stack the $\boldsymbol{X}_i$ to create an $n \times J$ "design matrix" $\boldsymbol{X}$, which can be included in either a general linear model as in equation (2.1.2), or a generalized linear model as in equation (2.2.1).

However, any image with a something resembling fine resolution is likely to result in $J >> n$, since samples sizes are often in the hundreds or low thousands and images often contain hundreds of thousands of pixels (2D) or voxels (3D). Thus, as mentioned in section 2, traditional methods are usually poor candidates for fitting such a model. However, as mentioned in section 3, and explored more fully in sections 5 and 6, Bayesian approaches can handle models where $J > n$, which mean the approaches discussed above are candidates for answering our question from section 1. An additional wrinkle is how to account for the correlation structure inherent in an image, which is one of the main goals of this work. We discuss how to incorporate spatial structure into variable selection in Chapter **??**, where we extend the spike-and-slab lasso to incorporate spatial structure into prior probabilities of model inclusion, but first we provide an overview of some relevant spatial statistical concepts and outline how these might be used to extend the spike-and-slab lasso GLM.

8 Spatial Statistics Overview

8.1 Introduction

The multivariate Normal distribution is often a useful tool for modeling spatial dependence, but in many cases both interpretation and computation are aided by using a conditional framework rather than the traditional multivariate specification. In Section 8.2 we define the Markov Random Field and the Gauss Markov Random Field. In Section 8.3 we introduce the CAR model and provide an intuitive introduction and comparison to its improper cousin, the intrinsic autoregression (IAR). Section 8.4 describes the formal structure of Intrinsic GMRF's (IGMRF's), and how IAR's arise within such framework. Section 8.5 shows how we can use GMRF's to model probabilities in a spatial framework.

8.2 Gauss Markov Random Fields (GMRF's)

A common and usually reasonable framework for modeling spatial correlation is a Gaussian Markov Random Field (GMRF), because the multivariate Normal distribution easily and intuitively incorporates spatial structure.[62] We provide a definition of a Markov Random Field from Cressie and Wikle (2011), before going on the discuss GMRF's.

Markov Random Fields (MRF)

Suppose we have spatial locations $j = 1, \ldots, J$, and the distribution for a random variable $X(\boldsymbol{s}_j)$ at location $\boldsymbol{s}_j$ is defined conditional only on points within a *neighborhood*, $N(\boldsymbol{s}_j)$, of $\boldsymbol{s}_j$. That is, if $\boldsymbol{X}_{(-j)}$ consists of the random variables for all other locations $\boldsymbol{s}_k$, $j \neq k$, then:

$$p(X(\boldsymbol{s}_j)|\boldsymbol{X}_{(-j)}) = p(X(\boldsymbol{s}_j)|\boldsymbol{X}(N(\boldsymbol{s}_j))) \tag{8.2.1}$$

where $\boldsymbol{X}(N(\boldsymbol{s}_j)) \equiv (X(\boldsymbol{u})|\boldsymbol{u} \in N(\boldsymbol{s}_j))^\top$.

[62] Note that Markov Random Fields (MRF) are not confined to the Normal distribution, even though we limit our focus to such distributions here; see Cressie and Wikle (2011), chapter 4.2.2 for a more general treatment of MRF's.

Definition of GMRF

GMRF's confine equation (8.2.1) to the case where $\boldsymbol{X}(\boldsymbol{s}_j)$ is normally distributed, and often use graph notation to facilitate interpretation. Following Rue and Held (2005), a random vector $\boldsymbol{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ is a GMRF with respect to graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and symmetric positive definite precision matrix $\boldsymbol{Q}$ if and only if its density has the form

$$p(\boldsymbol{x}) = (2\pi)^{-n/2}|\boldsymbol{Q}|^{1/2} \exp\Big\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{\mu})\Big\} \tag{8.2.2}$$

and $Q_{ij} \neq 0 \iff \{i, j\} \in \mathcal{E}$ for all $i \neq j$. The precision matrix $\boldsymbol{Q}$ is equal to the inverse of the covariance matrix; i.e., $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$. The marginal mean and precision are

$$\mathrm{E}(x_i|\boldsymbol{x}_{-i}) = \mu_i - \frac{1}{Q_{ii}} \sum_{j:j\sim i} Q_{ij}(x_j - \mu_j) \tag{8.2.3}$$

$$Prec(x_i|\boldsymbol{x}_{-i}) = Q_{ii} \tag{8.2.4}$$

where $j \sim i$ indicates that x_i and x_j are neighbors, and the correlation between x_i and x_j given $\boldsymbol{x}_{-ij}$ is

$$Corr(x_i, x_j|\boldsymbol{x}_{-ij}) = -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}, \quad i \neq j \tag{8.2.5}$$

The beauty of the precision matrix

Standard multivariate statistics usually specify models with the covariance matrix, so what is gained by specifying a conditional model in terms of the precision matrix? For one, the interpretation of the precision matrix is in terms of conditional relationships, while the covariance matrix is interpreted in terms of marginal relationships; i.e., the diagonal of the precision matrix contains precisions conditional on all other variables, and the off-diagonal contains conditional correlation information (after appropriate scaling). This means that

$Q_{ij} = 0$ implies the x_i and x_j are independent, conditional on the rest of the variables. In contrast, the diagonal of the covariance matrix holds marginal variances, while the off-diagonal holds marginal covariances. Thus, if our interest is in conditional relationships, then the precision matrix may better suit our purposes.

Additionally, while the covariance matrix is generally completely dense, the precision matrix is often sparse, which lends computational benefits and aids interpretation, since the conditional expectation in equation (8.2.3) explicitly shows how a realization of x_i arises as function of its surrounding neighbors. For example, consider an autoregressive process of order one, abbreviated AR(1), for $t = 1, \ldots, n$ and unit variance:

$$x_t = \phi x_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, 1), \quad |\phi| < 1 \tag{8.2.6}$$

The covariance matrix for the AR(1) process will be completely dense, and on its own will not help us understand the relationship between random variables x_i and x_j. That is,

$$\boldsymbol{\Sigma} = \frac{1}{1-\phi^2} \begin{bmatrix} 1 & \phi & \phi^2 & \ldots & \phi^{n-1} \\ \phi & 1 & \phi & \ldots & \phi^{n-2} \\ \phi^2 & \phi & 1 & \ldots & \phi^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \phi^{n-3} & \ldots & 1 \end{bmatrix} \tag{8.2.7}$$

However, the precision matrix will be sparse, specifically tridiagonal:

$$\boldsymbol{Q} = \begin{bmatrix} 1 & -\phi & & & & & \\ -\phi & 1+\phi^2 & -\phi & & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & & -\phi & 1+\phi^2 & -\phi \\ & & & & & -\phi & 1 \end{bmatrix} \tag{8.2.8}$$

Equation (8.2.8) clearly shows us the defining dependence structure of the AR(1) process: that x_t is conditionally dependent on the realization before and after it (i.e., x_{t-1} and x_{t+1}). An AR(1) process is reasonably well understood simply by considering equation (8.2.6), but in more complex situations we may not have easily interpretable representations and specifying and interpreting useful models will be aided by using the precision matrix and conditional mean.

A final clarification: two variables can be correlated when they are not neighbors; in fact, this is usually the case in temporal and spatial settings. This is made clear by comparing equations (8.2.7) and (8.2.8). The AR(1) model has non-zero correlation between all variables, even if this correlation is negligible after some temporal or spatial distance. It is the dense nature of the covariance matrix that makes interpretation difficult; we can tell whether variables are neighbors just by looking inside $\boldsymbol{Q}$, but $\boldsymbol{\Sigma}$ only tells us whether the correlation between variables is strong or not.

8.3 Conditional Autoregressions (CAR's)

Conditional autoregressions (CAR) can be thought of as a particular generalization of standard autoregressions in one dimension, and are used extensively in spatial statistical research. Here we provide a brief overview, and connect them to GMRF's.

Notation and basic properties

As alluded to in section 8.2, GMRF's are often best understood through their conditional properties rather than explicitly through the precision/covariance matrices, $\boldsymbol{Q}$ and $\boldsymbol{\Sigma}$. The resulting specifications are *conditional autoregressions* (CAR), which are slightly reparameterized versions of equations (8.2.3) and (8.2.4) (Rue and Held, 2005; Cressie and Wikle, 2011). Rather than explicitly specifying the Q_{ij} within $\boldsymbol{Q}$, we adopt general

parameters $\{\beta_{ij},\, i \neq j\}$ into the conditional mean and precision as

$$\mathrm{E}(x_i|\boldsymbol{x}_{-i}) = \mu_i - \sum_{j:j\sim i} \beta_{ij}(x_j - \mu_j) \tag{8.3.1}$$

$$Prec(x_i|\boldsymbol{x}_{-i}) = \tau_i \tag{8.3.2}$$

where we require that

1. $\beta_{ij} \neq 0 \leftrightarrow \beta_{ji} \neq 0$,
2. $Q_{ii} = \tau_i$ and $Q_{ij} = \tau_i\beta_{ij}$, and
3. symmetric $\boldsymbol{Q}$, which implies $\tau_i\beta_{ij} = \tau_j\beta_{ji}$.

This parameterization results in a unique joint distribution that is multivariate Normal (Besag, 1974; Rue and Held, 2005).

While it is often easier to work with the full conditionals, it is also important to verify that this specification results in a unique joint distribution; this is achieved via Brook's Lemma, which allows us to obtain the following joint multivariate normal distribution (Besag, 1974; Banerjee et al., 2015):

$$p(x_1, \ldots, x_n) \propto \exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \mathbf{D}^{-1}(\mathbf{I} - \mathbf{B})(\boldsymbol{x} - \boldsymbol{\mu}) \right\} \tag{8.3.3}$$

where $\boldsymbol{B} = \{\beta_{ij}\}$ and $\boldsymbol{D} = \mathrm{diag}\left(\frac{1}{\tau_i}\right)$, which implies that we can specify the precision and covariance matrices as $\boldsymbol{Q} = \boldsymbol{D}^{-1}(\boldsymbol{I} - \boldsymbol{B})$ and $\boldsymbol{\Sigma} = (\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{D}$, respectively.[63]

In many cases it is useful to specify neighbor relations with a *proximity* or *adjacency* matrix, $\boldsymbol{W} = \{w_{ij}\}$ (Banerjee et al., 2015; Brown et al., 2014). When $w_{ij} \neq 0$, then x_i and x_j are neighbors ($i \sim j$). In most cases $w_{ii} = 0$, and when $w_{ij} = 1$ for $i \sim j$, W contains only zeroes and ones, so that if x_i and x_j are neighbors then $w_{ij} = 1$ and otherwise $w_{ij} = 0$.

[63] Note that $\boldsymbol{I}$ is a $J \times J$ identity matrix.

In more complicated cases, e.g., irregular lattices, the w_{ij} may be weights for neighbors based on some distance between them. Generally speaking, the w_{ij} determine the structure and weights of spatial connection between locations.

An informal introduction to intrinsic autoregressions (IAR's)

Intrinsic GMRF's (IGMRF) are GMRF's that have precision matrices less than full rank and thus improper joint distributions, while intrinsic autoregressions (IAR) are a particular variant of IGMRF applied to CAR models (Besag and Kooperberg, 1995; Rue and Held, 2005). It turns out that IAR's have some practical use, and are especially flexible as prior distributions in spatial applications. We discuss both formally in section 8.4; however, it is often difficult to interpret IGMRF's, and thus before discussing theoretical details pertaining to CAR and IAR models we relate two common temporal models to CAR and IAR frameworks to motivate their interpretation.

We first introduced the autoregressive process of order one, i.e., AR(1), in Section 8.2. The AR(1) process is most intuitively represented by equation (8.2.6), but for a given sample size, n, and assuming that ϵ_t has mean 0, an AR(1) can be specified as a multivariate normal distribution with mean 0, variance σ^2, and correlation between x_s and x_t, $corr(x_s, x_t) = \phi^{|s-t|}$. The AR(1) is process is *stationary*, which implies that the mean is constant (here zero), and whose covariance is dependent only on the distance between x_s and x_t, not their specific values; hence, the distribution of the process is not dependent on t.[64]

A process related to the AR(1) is the random walk, which is obtained from equation (8.2.6) when $\phi = 1$; i.e., for $t = 1, 2, \ldots$

$$x_t = x_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, 1) \tag{8.3.4}$$

[64] Stationarity is formally divided into *strong* and *weak* forms. Strong stationarity implies that the (finite) joint distribution of the process at one location (temporal, spatial, or otherwise) is the same as that of "lagged" locations; i.e., the joint distribution is the same everywhere. Weak stationarity simply requires finite variance, constant mean, and covariance that only depends on some distance measure between locations (spatial, temporal, or otherwise). See, e.g., Diggle (1990) or Cressie and Wikle (2011) for details.

Note that strictly speaking both the AR(1) and random walk processes can have "noise" with non-zero mean and non-unit variance; i.e., $\epsilon_t \sim \mathcal{N}(\mu_w, \sigma_w^2)$. In such case it can be shown that mean, variance, and covariance of the process in equation (8.3.4) are all functions of t; i.e. $\mathrm{E}(x_t) = x_0 + t\mu_w$, $\mathrm{var}(x_t) = t\sigma_w^2$, and $cov(x_t, x_s) = \min(t, s)\sigma_w^2$. This means that the random walk process is *not stationary*. Thus, while the AR(1) and random walk processes are similar in initial structure, they will exhibit different behaviors; for example, the AR(1) process will vary about the same mean no matter the value of t, but the random walk is better understood as varying about a "local" mean determined by its neighbors; e.g., compare the behavior of an AR(1) process to a random walk for $n = 2,000$ in figure 3. The AR(1) process dances around its mean of zero, while the random walk wanders all over the place, and does not vary around any specific mean, even though both processes have noise $\epsilon_t \sim \mathcal{N}(0, 1)$. Note also that most AR(1) simulations look fairly similar to each other, while random walk simulations vary wildly in their appearance, which is consistent with the process not having a constant mean.

The spatial analog to the AR(1) process is the CAR model described in section 8.3, both of which are stationary processes (Diggle, 1990; Rue and Held, 2005).[65] While natural ordering is no longer possible when the dimension is greater than 1, the correlation structure of a CAR model decays in the same fashion as the AR(1) process, but in space rather than time, and its mean and covariance is not dependent on location, although in finite lattices the variances may depend on the distance of a point from the boundary (Besag and Kooperberg, 1995). In contrast, the spatial analog to the random walk is an *intrinsic autoregression (IAR)*. IAR's, like random walks, are not stationary. In particular, neither process has a constant mean about which the process varies; rather, the process varies about "local" levels, rather than a global level. Also, like the random walk, while the process itself may not be stationary, the increments are stationary (Künsch, 1987). That is, the distribution of $x_j - x_i$

[65] See Diggle (1990) for a broader discussion of the AR(1) process, and note that Rue and Held (2005) (Chapter 2.6, p. 72) define conditional autoregression using by placing the restrictions on the form of spectral density function of a stationary Gaussian process.

Figure 3: AR(1) versus random walk. Both processes have noise $\epsilon_t \sim \mathcal{N}(0, 1)$. Notice how the AR(1) process clearly varies about a constant mean, in this case zero. In contrast the random walk jumps around and only "locally" resembles the AR(1) process.

is stationary for $i < j$; e.g., if $\Delta x_i \sim \mathcal{N}(0, \sigma_w^2)$, $i = 1, \ldots, n$, then we have:

$$x_j - x_i \sim \mathcal{N}(0, (j-i)\sigma_w^2), \quad i < j \tag{8.3.5}$$

8.4 Formal Descriptions of IGMRF's and IAR's

Having provided the intuition behind IGMRF's and IAR's via their analog in the random walk, we now proceed to formal definitions of these processes.

IGMRF definition

In defining and discussing and IGMRF's and IAR's, we primarily follow Rue and Held (2005) for formal definitions and Banerjee et al. (2015) for model interpretations; we depend to a lesser extent on primary sources Besag and Kooperberg (1995) and Künsch

(1987).[66] If $\boldsymbol{Q}$ is a symmetric positive semi-definite matrix of rank $n-k > 0$, then $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ is an *improper GMRF* of rank $n-k$ with parameters $(\boldsymbol{\mu}, \boldsymbol{Q})$[67], if its density is

$$p(\boldsymbol{x}) = (2\pi)^{-(n-k)/2}(|\boldsymbol{Q}|^*)^{1/2} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu}) \right\} \tag{8.4.1}$$

where $|\cdot|^*$ is a generalized determinant given by the product of non-zero eigenvalues; $\boldsymbol{x}$ is an improper GMRF with respect to the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where

$$Q_{ij} \neq 0 \iff \{i,j\} \in \mathcal{E}, \quad \forall i \neq j$$

and $\mathcal{V}$ is a set of vertices and $\mathcal{E}$ is a set of edges.

An *IGMRF of first order* is simply an IGMRF of rank $n-1$ with the linear constraint $\boldsymbol{Q}\mathbf{1} = \mathbf{0}$, which is to say that the $\sum_j Q_{ij} = 0$, $\forall i$. For example, when $\boldsymbol{\mu} = \mathbf{0}$ we have

$$\mathrm{E}(x_i | \boldsymbol{x}_{-i}) = -\frac{1}{Q_{ii}} \sum_{j:j\sim i} Q_{ij} x_j$$

where $-\sum_{j:j\sim i} \frac{Q_{ij}}{Q_{ii}} = 1$, which implies that the the conditional mean of x_i is a weighted mean of its neighbors with no overall level (mean); this is true in general for IGMRF's, and is one of the model's primary benefits in applications where it may not be realistic to assume a constant *global* mean. Recall the comparison between AR(1) and random walk processes; since the AR(1) process assumes a constant mean and the random walk does not, they may not be appropriate models for the same applications. Note that not all IGMRF's are analogous to the (first order) random walk, in the same way that GMRF's are more broad

[66] Rather confusingly, definitions and notation differ depending on the reference. We mean to say that the definitions and notations in this work correspond to Rue and Held (2005), but this is not to imply any favoritism towards any particular setup; we simply take the approach that it is simpler to accept the existing notation within a textbook, rather than assemble our own from many papers and texts.

[67] Strictly speaking, the mean and precision no longer exist, although the interpretation of the parameters in this framework is similar.

than autoregressive models.[68] Thus, IAR models are related to the broad class of IGMRF as CAR models are related to the broad class of GMRF.

IAR definition

Rue and Held (2005) obtain a basic IAR model on a two dimensional lattice by first considering the distribution of the "increments", $x_i - x_j$ for some locations i and j:

$$x_i - x_j \sim \mathcal{N}\left(0, \frac{1}{w_{ij}\tau}\right) \tag{8.4.2}$$

where the w_{ij} are the weights from the proximity matrix, $\boldsymbol{W}$, and which results in the following joint density:

$$p(\boldsymbol{x}) \propto \tau^{(n-1)/2} \exp\left\{-\frac{\tau}{2}\sum_{j:j\sim i} w_{ij}(x_i - x_j)^2\right\} \tag{8.4.3}$$

The elements of the "precision" matrix, $\boldsymbol{Q}$ are now

$$Q_{ij} = \tau \begin{cases} \sum_{k:k\sim i} w_{ik} & i = j, \\ -w_{ij} & i \sim j, \\ 0 & \text{otherwise}, \end{cases} \tag{8.4.4}$$

and the respective expectation and precision are

$$\mathrm{E}(x_i|\boldsymbol{x}_{-i}, \tau) = \frac{\sum_{j:j\sim i} x_j w_{ij}}{\sum_{j:j\sim i} w_{ij}} \tag{8.4.5}$$

$$Prec(x_i|\boldsymbol{x}_{-i}, \tau) = \tau \sum_{j:j\sim i} w_{ij}. \tag{8.4.6}$$

When the locations are equally spaced (i.e., a regular lattice), it reasonable to set $w_{ij} = 1$,

[68] For example, CAR models are stationary processes by definition (see Chapter 2.6 in Rue and Held (2005)), but GMRF's can be proper and non-stationary. For example, we can easily imagine a GMRF on a two-dimensional lattice that has constant mean vector $\boldsymbol{\mu}$, but where the correlation between two random variables in the lattice is dependent on their location therein.

which results in $\sum_{k:k\sim i} w_{ik} = n_i$, where n_i is the number of neighbors to location i. Thus, equation (8.4.4) becomes

$$Q_{ij} = \tau \begin{cases} n_i & i = j, \\ -1 & i \sim j, \\ 0 & \text{otherwise}, \end{cases} \tag{8.4.7}$$

which implies that the conditional distribution of x_i is

$$x_i | \boldsymbol{x}_{-i}, \tau \sim \mathcal{N}\left(\frac{1}{n_i} \sum_{j:j\sim i} x_j, \frac{1}{n_i \tau} \right). \tag{8.4.8}$$

Thus, we see how the conditional distribution is explicitly defined as varying about a local mean, since the conditional expectation of any x_j is defined as the mean of its neighbors. When variable specific precision is desired, we replace τ with τ_i in equation (8.4.8).[69] Note that while we never explicitly specified that $\boldsymbol{Q}\mathbf{1} = \mathbf{0}$, this is in fact the case. Additionally, we may express IAR's by their (improper) "joint distribution", i.e.,

$$p(x_1, \ldots, x_n) \propto \exp\left\{ -\frac{\tau_i}{2} \boldsymbol{x}^\top (\boldsymbol{D_w} - \boldsymbol{W}) \boldsymbol{x} \right\} \tag{8.4.9}$$

where $\boldsymbol{W} = \{w_{ij}\}$ is the proximity matrix as before, $\boldsymbol{D}_w = \text{diag}(w_{i+})$, and $w_{i+} = \sum_j w_{ij}$; obviously, $\boldsymbol{Q} = \boldsymbol{D_w} - \boldsymbol{W}$.

A unified framework for CAR and IAR models

As described in Banerjee et al. (2015), if it becomes necessary or desirable to make an IAR proper while maintaining at least some of its other benefits, then a *propriety parameter*, ρ, can be included such that $\boldsymbol{Q} = \boldsymbol{D}_w - \rho \boldsymbol{W}$. This will induce a proper distribution when $\rho \in (1/\lambda_{(1)}, 1/\lambda_{(n)})$ where $\lambda_{(1)} < \cdots < \lambda_{(n)}$ are the ordered eigenvalues

[69] It is also possible to include some "location" parameter, $\boldsymbol{\mu} = \{\mu_i\}$, as is included in the definition of an IGMRF (i.e., equation (8.4.1)). However, in practice usually $\boldsymbol{\mu} = \mathbf{0}$.

of $\boldsymbol{D}_w^{-1/2}\boldsymbol{W}\boldsymbol{D}_w^{-1/2}$. Additionally, if we scale $\boldsymbol{W}$ by w_{i+}, i.e., $\widetilde{\boldsymbol{W}} \equiv \text{diag}(1/w_{i+})\boldsymbol{W}$, then $\boldsymbol{Q} = M^{-1}(I - \alpha\widetilde{\boldsymbol{W}})$, where M is diagonal, and if $|\alpha| < 1$, then $(I - \alpha\widetilde{\boldsymbol{W}})$ is nonsingular. Including ρ in this manner leaves the conditional precision unchanged, but the conditional mean in equation (8.4.5) becomes

$$\text{E}(x_i|\boldsymbol{x}_{-i}, \tau) = \rho \times \frac{\sum_{j:j\sim i} x_j w_{ij}}{\sum_{j:j\sim i} w_{ij}}, \tag{8.4.10}$$

which shows that whereas an IAR models the locations as varying about the (weighted) mean of the neighbors, including a propriety parameter alters the model so that the locations now vary about a *proportion* of the (weighted) mean of the neighbors. At least one issue with this formulation is that including the propriety parameter can significantly affect strength of association; when using the scaled adjacency matrix, α must often be very close to 1 to obtain reasonable correlations, which is is at least one reason for preferring the improper IAR model to the proper CAR model (Besag and Kooperberg, 1995; Banerjee et al., 2015).

8.5 Using CAR's to Model Prior Probabilities

Suppose we want to model the probability that a parameter should be included in a statistical model, but the parameters arise in a spatial framework and are reasonably assumed to exhibit spatial clustering. This issue can be addressed indirectly by imposing a GMRF as a prior for the probabilities of inclusion; however, probabilities are bounded at 0 and 1, so we must transform the prior inclusion probabilities, θ_j, in order to use a GMRF to model the spatial dependence among them. While $\theta_j \in [0, 1]$, $\psi_j = \text{logit}(\theta_j) = \log \frac{\theta_j}{1-\theta_j} \in (-\infty, \infty)$, which has the same support as the (multivariate) Normal distribution. After modeling the ψ_j, we can find $\theta_j = \text{logit}^{-1}(\psi_j)$.

Section 8.4 revealed that a propriety parameter unifies the framework and interpretation of CAR and IAR models. We thus specify the conditional prior distributions as follows:

$$p(\psi_j|\psi_i,\tau) = \mathcal{N}\left(\alpha \sum_{j:j\sim i} \frac{\psi_i}{n_j}, (\tau^2 n_j)^{-1}\right) \tag{8.5.1}$$

where n_j is the number of neighbors for location j, and the $\sum_{j:j\sim i} w_{ij} = n_j$ because we specify $w_{ij} = 1$ if $j \sim i$ and 0 otherwise. We further specify a common precision parameter τ so that the precision $\tau^2 n_j$ only varies based on the number of neighbors. The joint distribution of this model then simplifies to the following (Jin et al., 2005; Banerjee et al., 2015):

$$\boldsymbol{\psi} = \mathcal{N}\left(\mathbf{0}, [\tau^2(\boldsymbol{D} - \alpha \boldsymbol{W})]^{-1}\right) \tag{8.5.2}$$

where $\boldsymbol{D} = \text{diag}(n_j)$ and $\boldsymbol{W} = \{w_{ij}\}$ where $w_{ij} = \begin{cases} 1 & j \sim i \\ 0 & \text{otherwise} \end{cases}$.

This framework will allow us to set $\alpha = 1$ so that we may obtain the IAR model, set α manually, or impose a prior distribution on α while still conserving the same basic interpretation of the model. Note that in this case the IAR is likely the preferable model for several reasons. First, the CAR model assumes mean 0 and is often employed to model spatial random effects. However, we are modeling prior probabilities, which cannot be reasonably assumed to vary about 1/2.[70] Second, it is not clear that the best way to model prior probabilities is to assume a common mean at all; the IAR allows far more flexibility in the structure of the prior probabilities and does not impose a common mean, which is likely to align better with data acquired in the real world. Third, additional parameters would be needed to make the proper CAR flexible enough to handle a wide variety of spatial

[70] If we assume ψ_j varies about 0, then θ_j varies about 1/2, since $\text{logit}(1/2) = 0$.

structures; in particular, we would need to model the common mean. This process would add more subjectivity and complexity to the model, and thus would open the process to risks of overfitting and/or other errors. Finally, as mentioned above, the CAR model is often incapable of imposing strong correlations, and so especially in cases where spatial correlation is likely to strong, the IAR would be preferred.

9 Dissertation Outline

So far this work has focused primarily on statistical concerns already existing in the literature, and while the introduction and preceding subsection provides a general idea of the problem at hand, one may reasonably wonder where all this leads. This section provides an outline of the proposed research direction and contains an informal (to the extent possible) description regarding the goals of the dissertation.

Ultimately, we will need to be able to evaluate the performance of models. In order to do so one must probe at the methodology, and simulations often provide a useful framework for understanding the strengths and weaknesses of a model. The first paper presents an `R` package, `sim2Dpredictr`, to simulate scalar outcomes using spatially dependent data. This package is available for download on CRAN. This paper has been submitted to the *Journal of Statistical Software*.

We presented two Bayesian methods, both of which were designed to shrink many variables to zero and perform automatic variable selection. The approach presented in section 5 uses an EM-IWLS algorithm to fit a Bayesian GLM, and in this context prior distributions can be set to shrink many variables to zero, performing variable selection; in addition, the hierarchical Bonferroni procedure could be used to correct for multiple comparisons. However, in order to use spatial information in variable selection it would be necessary to directly model the correlation among the parameters, β_j, which would require inverting a large covariance matrix when the number of predictors is large; this is computationally undesirable. The approach presented in Section 6 uses spike-and-slab priors

to model parameters as drawn from a mixture of distributions, one each for parameters that should and should not be selected, respectively. Similar to the former approach, double exponential priors are used to shrink some parameter estimates to zero and leave others nearer to their initial estimates. Spatial information can be incorporated into variable selection by modeling dependence among probabilities of inclusion, which avoids the need to invert a large covariance matrix, and is more directly related to the task at hand: variable selection. Using the latter methodology prevents us from directly applying any traditional multiple testing procedures, but the computational and interpretative benefits of the spike-and-slab lasso point to it being the more fruitful avenue for research. This path will shift focus away from multiple testing procedures towards a more general concern about model generalizeability; i.e., having selected a model, how well does it work when applied to independent data?

In the second paper we describe how to extend the spike-and-slab lasso GLM by placing IAR priors on the probabilities of inclusion, discuss how to incorporate a more general shrinkage approach known as the elastic net, and present a simulation study to explore the strengths and weaknesses of the model, with particular focus on the model's ability to generalize to independent data. The development version of an `R` package to fit the model, `ssnet`, is available on github.

In the third paper we explore how the model performs in real-world data, by exploring the prediction/classification potential of the model using data from the Alzheimer's Disease Neuroimaging Initiative (ANDI). This chapter is similar in spirit to the second paper, in that we will still be concerned with generalizeability. However, it differs in that we shift focus towards the classification potential of the algorithm, since (Bayesian) GLM's can be also be used to classify subjects by using logistic regression with a classification rule, which allow us to probe the method's potential usefulness in identifying disease status using imaging data.

sim2Dpredictr: AN R PACKAGE FOR SIMULATING SCALAR OUTCOMES WITH SPATIALLY DEPENDENT PREDICTORS

JUSTIN M. LEACH AND INMACULADA ABAN

Submitted to the *Journal of Statistical Software*

Format adapted for dissertation

1 Introduction

Images are ubiquitous in statistical analysis, both as outcomes of interest and predictors, and as with other areas of statistical application, model evaluation involving images can require statistical simulations. High resolution images often appear continuous to the naked eye, but in reality they consist of discrete locations containing intensity values (pixels in 2D, voxels in 3D) that can be treated as random variables in appropriate contexts. When images are used to predict or model outcomes, the statistical problem is one of large-dimensional variable selection; e.g., presumably only a subset of brain locations' intensity values are associated with or reliably predict cognitive ability or future disease development, and good methods should find those locations without including too many false positives and/or provide reliable prediction on independent data. The development and evaluation of variable selection methods should begin in controlled and relatively simple modeling/simulation frameworks where relative strengths and deficiencies are more easily probed; if all goes well, testing will include gradually more complicated (e.g., real world) settings. `sim2Dpredictr` provides tools for simulating spatially dependent data (predictors) with specific dependence structures; these "predictors" are then used to generate scalar outcomes. The simulated data can then be used to test the performance of variable selection models; the flowchart in Figure 1 depicts the basic framework for such a simulation process.

For continuous intensity values, random fields are a useful starting place, and there are several useful and well-documented tools available for simulating and modeling random fields in `R`, particularly the `R` packages `RandomFields` and `gstat`. These tools generally are designed with geostatistical-type problems in mind, where the outcome of interest is the image itself, and the goal is to model and/or understand the process underlying the spatial

data (Schlather et al., 2015; Gräler et al., 2016; Pebesma, 2004); similarly, `neuRosim` provides many tools for simulations in fMRI contexts (Welvaert et al., 2011).

In contrast, our primary interest is in how spatial dependence among predictors affects variable selection methodology, particularly in a (generalized) linear model (LM/GLM) framework. In this setting the primary focus is on how spatially structured correlation within the design matrix and clustering of non-zero parameters affects variable selection methodology. Continuous predictors are often simulated using Gauss-Markov Random Fields (GMRF), with a focus on their relation to the multivariate Normal distribution (MVN) Rue (2001). However, the learning curve needed to use and manipulate such tools can be steep, and if the primary concern is strength of correlation, then it is often more intuitive to simulate directly within the MVN framework. Additionally,many applications use images that have been binarized, e.g., lesion maps in the brain (Mirman et al., 2018); flexible simulation of clusters in binary maps is not well-suited to a traditional GMRF or MVN simulation framework. `sim2Dpredictr` provides image simulations that focus on specific correlation strengths for continuous images following a simple random field or MVN framework, and allows flexible simulation of binary maps via the Boolean Model, which can generate irregularly shaped random sets in multiple dimensions (Cressie and Wikle, 2011). `sim2Dpredictr` also contains the requisite tools to simulate full data sets; we provide a function to automate generation of spatially clustered parameter vectors, and subject images and outcomes can be generated within a single function.

In Section 2, we briefly overview the basic theoretical framework for simulation in the LM/GLM realm. In Section 3, we introduce the relevant simulation functions and describe their capabilities. In Section 4, we follow through a practical application of the presented functions. Finally, in Section 5, we provide a brief summary and discussion.

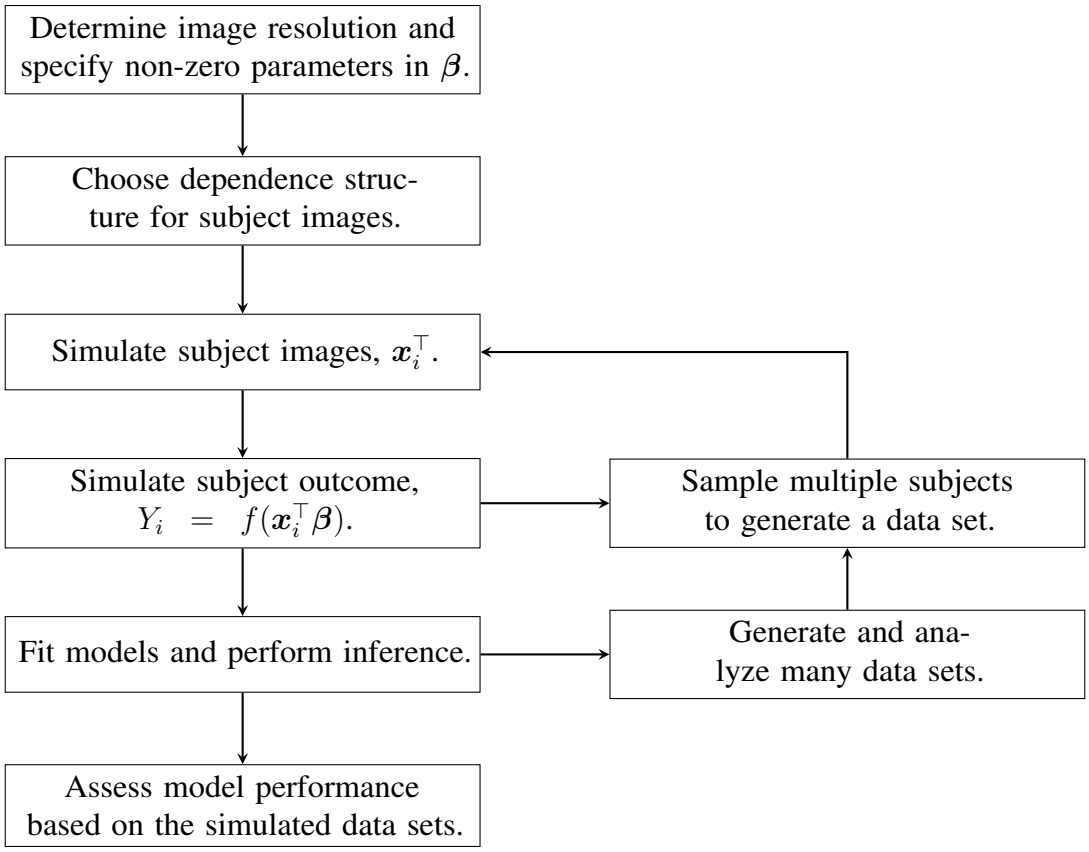

Figure 1: This flowchart describes the basic workflow. Although in many cases simulating a dataset can be made more efficient than subject-by-subject generation, this format makes the intuition behind sampling multiple subjects clear.

2 Theory Overview

2.1 Generating Scalar Outcomes with (Generalized) Linear Models

Suppose we want to model a scalar outcome using spatially dependent predictors in two dimensions; i.e., an $A \times B$ lattice resulting in $J = A * B$ predictors. If each univariate outcome, Y_i, is Normally distributed and the n outcomes are independent and identically distributed, then a linear model for the i^{th} subject has the form:

$$Y_i = \beta_0 + X_{i1}\beta_1 + \cdots + X_{iJ}\beta_J + \epsilon_i \tag{2.1.1}$$

where X_{ij} is the pixel intensity value at the j^{th} location, β_j, $j = 1, \ldots, J$, is the parameter corresponding to X_{ij}, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is the Normally distributed random error. These

equations can be represented in matrix formulation. Given a non-zero intercept, β_0, we obtain:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.1.2}$$

where $\boldsymbol{Y}$ is the $n \times 1$ outcome vector, $\boldsymbol{X}$ is an $n \times (J+1)$ design matrix, $\boldsymbol{\beta}$ is a $(J+1) \times 1$ parameter vector, and $\boldsymbol{\epsilon}$ is the random error vector. The formulation in Equation 2.1.2 can be used to directly simulate subject outcomes. Fortunately, once generated, correlation in $\boldsymbol{X}$ is only an impediment to model estimation, not outcome generation; given $\boldsymbol{X}$ and $\boldsymbol{\beta}$, one only needs to simulate $\boldsymbol{\epsilon}$ to successfully generate $\boldsymbol{Y}$.

Non-Normal outcomes are often represented in the generalized linear model (GLM) framework, which is most commonly used to model binary or count data:

$$g(\mathrm{E}(Y_i)) = \beta_0 + X_{i1}\beta_1 + \cdots + X_{iJ}\beta_J \tag{2.1.3}$$

where $g(\cdot)$ is an appropriate link function for modeling the expectation of the outcome. When operating in this framework the expectation is obtained and then the relevant distribution can be used to draw the subject-specific outcome values, although as we shall see some forms of Non-Normal data can be easily obtained by transforming or thresholding draws using Equation 2.1.2.

In order to test a new method for estimation and/or hypothesis testing/variable selection, we need to be able to simulate a data set under these specifications (i.e., using either Equation 2.1.2 or Equation 2.1.3), which requires the following steps:

1. Construct a parameter vector, $\boldsymbol{\beta}$. In most spatial settings non-zero parameters are clustered together, and the ordering of $\boldsymbol{\beta}$ should correspond appropriately to spatial locations.

2. Simulate images and convert them into design vectors, $\boldsymbol{x}_i^\top = [X_{i1}, \ldots, X_{iJ}]$, for each subject according to desired distributional and dependence structures.

3. For continuous outcomes, select a random error variance; i.e., if $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, then select a value for σ^2.

4. Use steps 1-3 to generate the outcomes, Y_i.

This process is easily repeated to generate many data sets, which allows for evaluation of variable selection methods. In the following sections we demonstrate how the functions in `sim2Dpredictr` can make this workflow relatively painless, but first we overview the basic theory needed to simulate the subject-specific images, $\boldsymbol{x}_i^\top$.

2.2 Generating Spatially Dependent Design Vectors

Continuous predictors: Drawing from the multivariate Normal distribution

Spatial dependence is often incorporated into models via Markov Random Fields, which are characterized by local dependence structure; i.e., the distribution of a random variable $X(\boldsymbol{s}_j)$ with (location) index $\boldsymbol{s}_j$ is specified based entirely by other random variables that are within a neighborhood $N(\boldsymbol{s}_j)$ of $X(\boldsymbol{s}_j)$ (Cressie and Wikle, 2011). Gaussian Markov Random Fields (GMRF) are Markov Random Fields with a Normal joint distribution characterized by conditional correlation structure; i.e., $X(\boldsymbol{s}_j)$ is *conditionally* independent of random variables outside its neighborhood; note that this does not imply marginal independence. An intuitive case of GMRF is the Conditional Autoregression (CAR), which can be simplified into a particularly intuitive form (Banerjee et al., 2015):

$$p(X_j|X_i, \tau_j, \alpha) = \mathcal{N}\left(\alpha \sum_{j:j\sim i} \frac{X_i}{n_j}, (\tau_j^2 n_j)^{-1}\right) \tag{2.2.1}$$

where $j \sim i$ indicates X_i and X_j are neighbors, τ_j is the precision (inverse variance) parameter, n_j is the number of neighbors for the j^{th} variable, and $\alpha \in [0, 1)$ controls the level of dependence. This formulation allows for interpreting the mean of X_j as varying

about a proportion of the mean of its neighbors. Note that while $\alpha = 0$ implies independence, $\alpha = 1$ results in an improper distribution that cannot be used to generate data, and in general α is not interpretable as a correlation (Besag and Kooperberg, 1995; Jin et al., 2005). The joint distribution is then multivariate Normal:

$$\boldsymbol{X} = \mathcal{N}\left(\mathbf{0}, [\boldsymbol{T}(\boldsymbol{D} - \alpha \boldsymbol{W})]^{-1}\right) \tag{2.2.2}$$

where $\boldsymbol{T} = \text{diag}(\tau_j^2)$, $\boldsymbol{D} = \text{diag}(n_j)$ and $\boldsymbol{W} = \{w_{ij}\}$ where $w_{ij} = \begin{cases} 1 & j \sim i \\ 0 & \text{otherwise} \end{cases}$.

Since GMRF's are a particular manifestation of MVN's, methods for simulating MVN's are applicable to simulating GMRF's. Generating draws from an MVN is remarkably straightforward. Following Ripley (1987), we may conjure MVN draws by transforming a set of Standard Normal draws to have non-zero mean, non-unit variance, and a covariance structure; we need do only the following:

1. Specify a mean vector for the MVN, $\boldsymbol{\mu}$.
2. Specify a covariance matrix, $\boldsymbol{\Sigma}$, for the MVN.
3. Obtain the Cholesky factor of the covariance matrix; i.e., the lower triangular matrix $\boldsymbol{L}$ in $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$.

Given that a set of random variables, $\boldsymbol{Z} = Z_1, \ldots, Z_J$ are drawn from the Standard Normal distribution, it can be shown that:

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2.2.3}$$

In `R`, standard Normal draws are obtained via `rnorm()`, and the Cholesky factor is obtained with `chol()`. In addition, the function `mvrnorm()` in `MASS` draws from MVN using eigendecomposition, which is more stable but slower for larger dimensions.

When the dimension of the MVN is large and the conditional dependence structure is sparse, then the precision matrix $\boldsymbol{Q}$ is sparse. We can exploit the sparseness of $\boldsymbol{Q}$ to make the Cholesky decomposion more efficient; e.g., we can quickly draw observations from CAR models by permuting $\boldsymbol{Q}$ to be a band matrix before taking the Cholesky decomposition Rue (2001). The process is slightly different from traditional MVN sampling since we take the Cholesky decomposition of the precision matrix rather than the covariance matrix:

1. As before, specify a mean vector for the MVN, $\boldsymbol{\mu}$.
2. Specify the precision matrix, $\boldsymbol{Q}$, for the MVN.
3. Obtain the Cholesky factor of the precision matrix matrix; i.e., the lower triangular matrix $\boldsymbol{Q}$ in $\boldsymbol{Q} = \boldsymbol{L}_Q \boldsymbol{L}_Q^\top$.
4. Since we have taken the Cholesky decomposion of $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$, we cannot directly draw from an MVN as before, and must solve $\boldsymbol{L}^\top \boldsymbol{X} = \boldsymbol{Z}$ to obtain draws from the MVN. However, this is easily solved via back substitution (Rue, 2001).

The function `chol.spam()` in the `R` package `spam` can permute the labels to obtain a band matrix and the base `R` function `backsolve()` can be used to obtain the draws from the MVN with `rmvnorm.prec()` (Furrer and Sain, 2010). The more challenging step is to explicitly specify correlation structure in $\boldsymbol{\Sigma}$ or $\boldsymbol{Q}$, which we implement in Section 3.1.

Binary predictors: Generating binary maps via the Boolean method

In general, the Boolean Model simulates a random set by "centering independent realizations of a compact random set S in $\mathbb{R}^d$ at each point of a realization $\{\boldsymbol{u}_i\}$ of a homogeneous Poisson point process Z_0 in $\mathbb{R}^d$, and taking the union of all the centered compact sets." It is easy to generate binary maps by thresholding continuous data, but such approaches constrain our freedom to control the size and shape of the clusters of 1's and 0's. The Boolean Method generates random sets in a more precise manner and then maps the sets from continuous space onto a lattice space to create a binary map (Cressie and Wikle,

2011). This approach allows the user to exercise more control over the the binary maps while still incorporating randomness into their creation. The Boolean model proceeds as follows (Cressie and Wikle, 2011):

1. Take a realization $\{\boldsymbol{u}_i\}$ from a homogeneous Poisson point process Z_0 in $\mathbb{R}^d$.
2. Take independent realizations $\{S_i\}$ from a compact random set S in $\mathbb{R}^d$.
3. Center the realizations from S at the "events" in $\{\boldsymbol{u}_i\}$.
4. The union of the compact random sets is then itself a random set.

This random set can easily be discretized, i.e., mapped onto a lattice, to create a binary image. Clustering tendencies in the images are controlled by altering the parameters that generate the random sets $\{S_i\}$ and draw from the homogeneous Poisson point process Z_0; further flexibility is obtained by extending this framework to include inhomogeneous Poisson point processes, e.g., in Cressie and Wikle (2011), such an extension is used to model tumor growth.

3 Simulation Functions

3.1 Building Blocks

While specifying the precision matrix in a CAR model can often result in a sparse matrix that enables fast draws from a MVN, it is often more intuitive to think about dependence structures in terms of correlation functions, which require us to specify a correlation matrix. `sim2Dpredictr` has several correlation functions built-in to allow easy specification and interpretation of a covariance matrix. We discuss these in this section.

Constructing a correlation matrix

Most reasonable spatial correlation models decay as a function of distance; an obvious option is a generalization of an auto-regressive of order one (AR(1)) correlation structure that is common in time-series analysis, of which many spatial models are extensions

(Diggle, 1990; Cressie and Wikle, 2011). Given correlation parameter ρ and suitable distance measure d, the correlation between two locations is given by:

$$\rho(d) = \rho^d \tag{3.1.1}$$

In many texts, e.g., Cressie and Wikle (2011), the correlation function in Equation (3.1.1) is introduced in the context of time-series analysis, given as $\rho^{|d|}$, and assumes $d = 0, \pm 1, \pm 2, \ldots$. However, in this application we require only that $d \geq 0$, since d represents spatial distance and not forward or backward in time.

The AR(1) structure is easily manipulated, but coarse; greater flexibility in decay can be obtained using an exponential correlation model in which increasing $\phi > 0$ results in steeper decay (Diggle et al., 2002):

$$\rho(d) = \exp(-\phi d) \tag{3.1.2}$$

A Gaussian correlation model is the same as the exponential model, except that the distance, d, is squared (Diggle et al., 2002):

$$\rho(d) = \exp(-\phi d^2) \tag{3.1.3}$$

In the rare case that an image with a Compound Symmetric correlation structure is desired, the correlation between any two points is given a single parameter and $\rho(d) = \rho$. One can further restrict non-zero correlations to be above some specified level by creating a neighborhood about each point to define marginal non-zero correlations. Simple approaches to creating this neighborhood are circular regions defined by a radius, r, or rectangular regions defined by respective width and height, w and h. If one simply desires to keep non-zero correlations above a certain level, then that level can be set explicitly; e.g., if a correlation is below 0.01, then set it to zero. This last approach can often have the effect of

making a covariance matrix relatively sparse without losing significant correlations in the image.

Each subject's image is initially treated as a matrix and its coordinates are given by the (row, column) position of a location. Thus, $X_{\ell,m}$ denotes the variable at the ℓ^{th} row and m^{th} column. For example, a subject's realized value for a variable with coordinates $(2, 3)$ (i.e., $X_{2,3} = x_{2,3}$)) is stored in the second row, third column. Similarly, the (Euclidean) distance d and radius r are calculated using this coordinate system; i.e., the distance between $X_{\ell,m}$ and $X_{u,v}$ is given by:

$$d(X_{\ell,m}, X_{u,v}) = \sqrt{(\ell - u)^2 + (m - v)^2} \tag{3.1.4}$$

Thus, then the correlation between locations $X_{\ell,m}$ and $X_{u,v}$ is $\rho(d(X_{\ell,m}, X_{u,v}))$; as mentioned above, we may threshold this value to avoid very small non-zero correlations and improve computational efficiency.

However, while this notation is useful for understanding how correlation between predictors is constructed, we need to index predictors such that the subscript corresponds to predictors in Equation (2.1.1); i.e., $X_{\ell,m}$ needs to be converted to X_{ij}, where (ℓ, m) denotes a location in two-dimensional space, while i denotes the subject index, and j indicates the predictor index in Equation (2.1.1). We assign these indices "by row"; e.g., for a 4×4 image and dropping the subject index i, the variable indices for the subject design vectors are organized as follows:

```
##      [,1]   [,2]   [,3]   [,4]
## [1,] "X_1"  "X_2"  "X_3"  "X_4"
## [2,] "X_5"  "X_6"  "X_7"  "X_8"
## [3,] "X_9"  "X_10" "X_11" "X_12"
## [4,] "X_13" "X_14" "X_15" "X_16"
```

For example, we see from matrix above that $X_{4,3}$ in two-dimensional space becomes X_{15} in Equation (2.1.1). These image vector indices will correspond to the row/columns of the correlation matrix; e.g., the correlation between X_1 and X_8 will be housed in cells $(1, 8)$ and $(8, 1)$ of the correlation matrix.

For example, we can use `correlation_builder()` to construct the correlation matrix for a 4×4 image with AR(1) correlation structure (i.e., using Equation (3.1.1) with $\rho = 0.25$, and correlations < 0.01 are set to 0):

```
S <- sim2Dpredictr::correlation_builder(
        corr.structure = "ar1",
        im.res = c(15, 15),
        rho = 0.25, corr.min = 0.01)
```

Note that `corr.structure = c("ar1", "exponential", "CS", "gaussian")` determines the correlation structure, `im.res` denotes the number of rows and columns in the image (i.e., image resolution), respectively, `rho` is the correlation parameter, ρ, and `corr.min` determines the smallest allowable correlation. When using either exponential or Gaussian correlation functions, `phi` denotes ϕ in Equations (3.1.2) and (3.1.3). See documentation for additional, if less commonly necessary, arguments.

Constructing a precision matrix

Alternatively, one may prefer to specify conditional dependence via a precision matrix, especially for finer image resolutions where the conditional dependence structure is sparse, so that the sparseness can be exploited to speed up computation. The `precision_builder()` constructs precision matrices. For example, we can construct a precision matrix for a 15×15 image, with unit precision parameter, and propriety parameter $\alpha = 0.75$:

```
Q.ar1 <- sim2Dpredictr::precision_builder(
        im.res = c(15, 15),
        tau = 1, alpha = 0.75,
        neighborhood = "ar1")
```

Here `tau` defines the precision parameters τ_j from Equation (2.2.1) and has length equal to 1 for common precision or equal to `prod(im.res)` for variable specific precisions. Note that since the precision depends on the number of neighbors, τ_j is not equal to precision, but rather the precision is proportional to τ_j. The propriety parameter α is defined by `alpha` and determines the strength of dependence, although again recall that α is not directly interpretable as a correlation. The `neighborhood` argument determines which other variables are neighbors. The default is `"ar1"`, which defines locations directly above, below, right, and left of a variable to be its neighbors. More complicated neighborhoods can be created by using `neighborhood = "round"`, which defines the neighborhood about a location as all other variables with a radius `r`:

```
Q.rnd <- sim2Dpredictr::precision_builder(
        im.res = c(15, 15),
        tau = 1, alpha = 0.75,
         neighborhood = "round", r = 3)
```

Rectangular neighborhoods can be constructed specifying `neighborhood = "rectangle"` and `w` and `h` to specify the width and height of the neighborhood, respectively:

```
Q.rec <- sim2Dpredictr::precision_builder(
        im.res = c(15, 15),
        tau = 1, alpha = 0.75,
        neighborhood = "rectangle",
```

```
         w = 3, h = 4)
```

Note that some situations may call for a proximity matrix, $\boldsymbol{W}$, that is not binary; e.g., if the neighborhoods are larger we may want a proximity matrix that inversely assigns weights to neighbors based on distance between them. This is possible by selecting `weight = "distance"`, which assigns neighbors weights equal to the inverse of Euclidean distance times a constant, which is specified by `phi`. The default is `weight = "binary"` to obtain the adjacency matrix.

Constructing a parameter vector

Like correlation matrix construction, manual specifying clustered non-zero parameter locations and their values in $\boldsymbol{\beta}$ is a time-consuming and error-prone process. The `beta_builder()` function handles this process automatically, and (obviously) assumes the same indexing structure as for the predictor values; i.e by row. For a 4×4 image the spatial structure of the parameter indices is:

```
##       [,1]      [,2]      [,3]      [,4]
## [1,] "beta_1"  "beta_2"  "beta_3"  "beta_4"
## [2,] "beta_5"  "beta_6"  "beta_7"  "beta_8"
## [3,] "beta_9"  "beta_10" "beta_11" "beta_12"
## [4,] "beta_13" "beta_14" "beta_15" "beta_16"
```

In most cases only a subset of locations will have non-zero parameter values, and the non-zero parameter values need not equal each other. The `beta_builder()` function offers several approaches to specify the locations and values for the non-zero parameters.

The argument `index.type = c("manual", "rectangle", "ellipse", "decay")` determines how non-zero locations are assigned. All four of these require specification of the arguments `row.index` and `col.index`, but the details on how to use these arguments varies. Manual entry uses `row.index` and `col.index` as vectors of row

and column coordinates, respectively, to specify non-zero locations; e.g., if `row.index = c(1, 3, 4)` and `col.index = c(4, 2, 4)`, then the parameters at locations $(1, 4)$, $(3, 2)$, and $(4, 4)$ are non-zero and all other locations have parameters equal to zero.

However, in many spatial contexts the non-zero parameters will be clustered, and more efficient assignment is helpful. When `index.type = "rectangle"`, `row.index` and `col.index` are strictly increasing sequences of whole numbers that define the rows and columns of the rectangular region of non-zero parameters; e.g., if `row.index = 1:3` and `col.index = 2:5`, then the rectangular region spanning the first through third rows and second through fifth columns will have non-zero parameters.

Both elliptical (`index.type = "ellipse"`) and decay index types use `row.index` and `col.index` as coordinates for the center of the area of non-zero parameter values in β. Elliptical neighborhoods additionally use the arguments `w` and `h` define the width and height of the ellipse, respectively; e.g., if `row.index = 2` and `col.index = 5` and `w = 1` and `h = 3`, then the area of non-zero parameters will be defined as the area within an ellipse of width 1, height 3, and center $(2, 5)$. Finally, `index.type = "decay"` treats the center coordinate as a "peak" parameter value, and the parameter values decay with distance from the peak according to either an exponential or Gaussian function of distance, as in `correlation_builder()`; the only difference between the decay here and in the correlation function is that the decay functions in Equations (3.1.2) and (3.1.3) are multiplied by the "peak" parameter value, so that the decay is from the peak value rather than one.

Values for the non-zero parameters are given by `B.values`; if `B.values` is a scalar, then all non-zero parameters are assigned that value unless `index.type = "decay"` is specified, in which case it is the "peak" parameter value; otherwise parameter values are assigned by row in the same way as the parameter indices. For example, in a 4×4 image, if manually chosen locations $(1, 1)$, $(1, 2)$, and $(4, 4)$ are non-zero and

`B.values = c(3, -1, 1/4)`, then $\beta_1 = 3$, $\beta_2 = -1$, and $\beta_{16} = 1/4$. However, a 32×32 image with an elliptical non-zero parameter region, as seen in Figure 2, is probably a more reasonable example. The code necessary to generate the non-zero parameters in Figure 2 is as follows:

```
Bex <- sim2Dpredictr::beta_builder(
        row.index = 8, col.index = 11,
        im.res = c(32, 32),
        B0 = 16, B.values = 3,
        index.type = "ellipse",
        h = 8, w = 8,
        output.indices = FALSE)
```

Additionally, one draw parameters from a distribution with `bayesian = TRUE`, which then treats the argument `B.values` as a vector of means. Distribution options are specified with `bayesian.dist = c"gaussian", "uniform"` to draw from a Normal or Uniform distribution, respectively. The scale of the distribution is specifed by the list `bayesian.scale`. To specify unique spreads the first element of the list is `"unique"` and the second element is a vector of standard deviations (Normal) or widths (Uniform). Otherwise, the first element is `"binary"` and the second element is a vector of length 3 that specifies the scale for the intercept, non-zero (important) parameters, and zero (non-important) parameters, respectively. For example, we can convert the framework described above where non-zero parameters were set to 3. These now have Normal distributions with means equal to 3 and standard deviation 1, while the "zero" parameters are drawn from a Normal distribution with standard deviation 1/3. Setting the scale to 0 has the effect of drawing some parameters from a distribution, but not others. Here, the intercept is set equal 16, declining to draw its value randomly by setting its scale to 0:

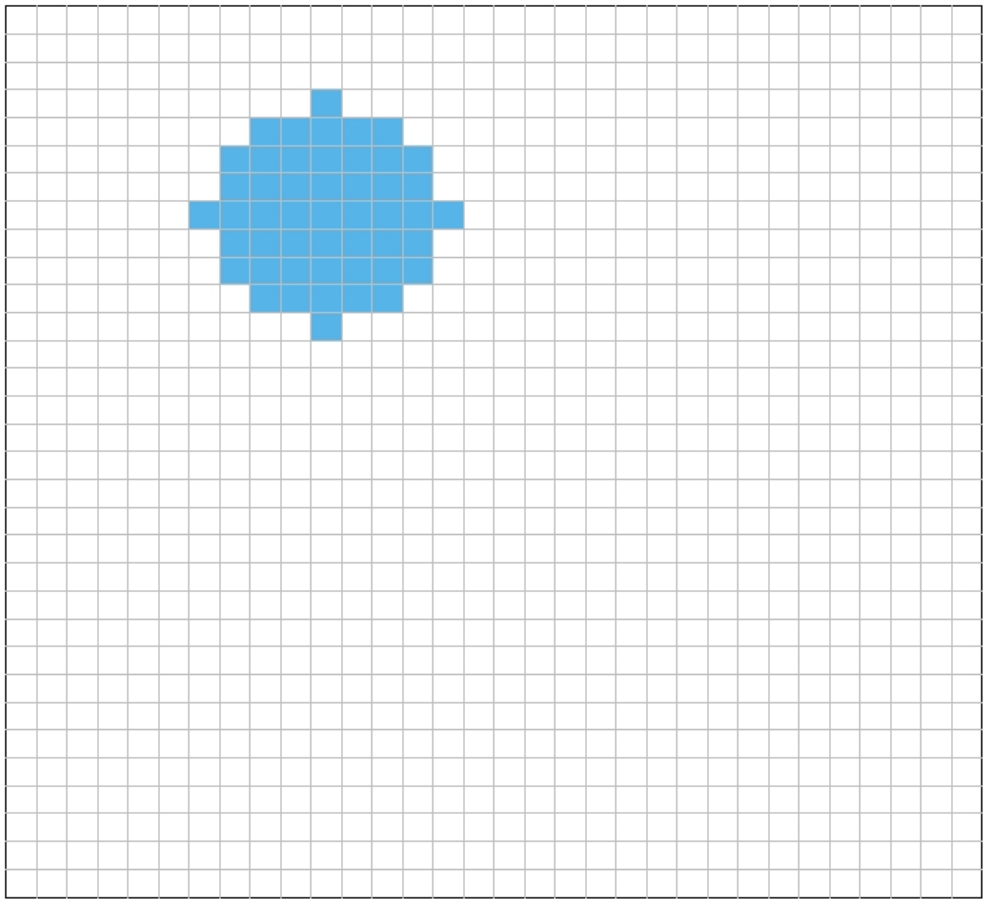

Figure 2: The blue pixels denote spatial locations with non-zero parameter values; i.e., $\beta_j \neq 0$.

```
Bex.bayes <- sim2Dpredictr::beta_builder(
        row.index = 8, col.index = 11,
        im.res = c(32, 32),
        B0 = 16, B.values = 3,
        index.type = "ellipse",
        h = 8, w = 8,
```

```
        bayesian = TRUE, bayesian.dist = "gaussian",
        bayesian.scale = list("binary", c(0, 1, 1/3)),
        output.indices = FALSE)
```

We can visualize what this situation looks like in Figure 3, where darker colored pixels are the ones drawn from the "important" distribution. Note that the default output of `beta_builder()` is a list whose first element is a vector of index values for non-zero parameters and whose second element is the desired parameter vector; this encourages the user to double-check that the parameter vector generated in fact corresponds to the desired locations, but can be disabled with `output.indices = FALSE`.

3.2 Drawing Spatially Dependent Continuous Predictors From an MVN

Cholesky decomposition

We noted in Section 2 that GMRF's are MVN's, and that simulating draws from an MVN requires the Cholesky decomposition of the distribution's covariance or precision matrix. The base `R` function `chol()` is sufficient in many cases, but for large sparse matrices `chol_spam()` from the `R` package `spam` uses more efficient methods; if the generated matrix has many zeros, then it is recommended to use `spam` (Furrer and Sain, 2010). `chol_s2Dp()` constructs the desired covariance or precision matrix by specifying one of `matrix.type = c("cov", "prec")`, respectively, and takes the Cholesky decomposition of that matrix; it contains the same arguments as `correlation_builder()` and `precision_builder()`, plus additional arguments for Cholesky decomposition options. The argument `use.spam = c(TRUE, FALSE)` determines whether base `R` or `spam` is used to take the Cholesky decomposition, `triangle = c("upper", "lower")` determines whether to output the Cholesky factor as lower triangular (i.e., $\boldsymbol{L}$ in Equation 2.2.3) or upper triangular (i.e., $\boldsymbol{L}^\top$ in Equation 2.2.3), and `sigma` or `tau` specifies the standard deviations or precision parameters for each variable. If `sigma` or `tau` is scalar, then common variance or precision parameter is assumed; otherwise, `sigma` and

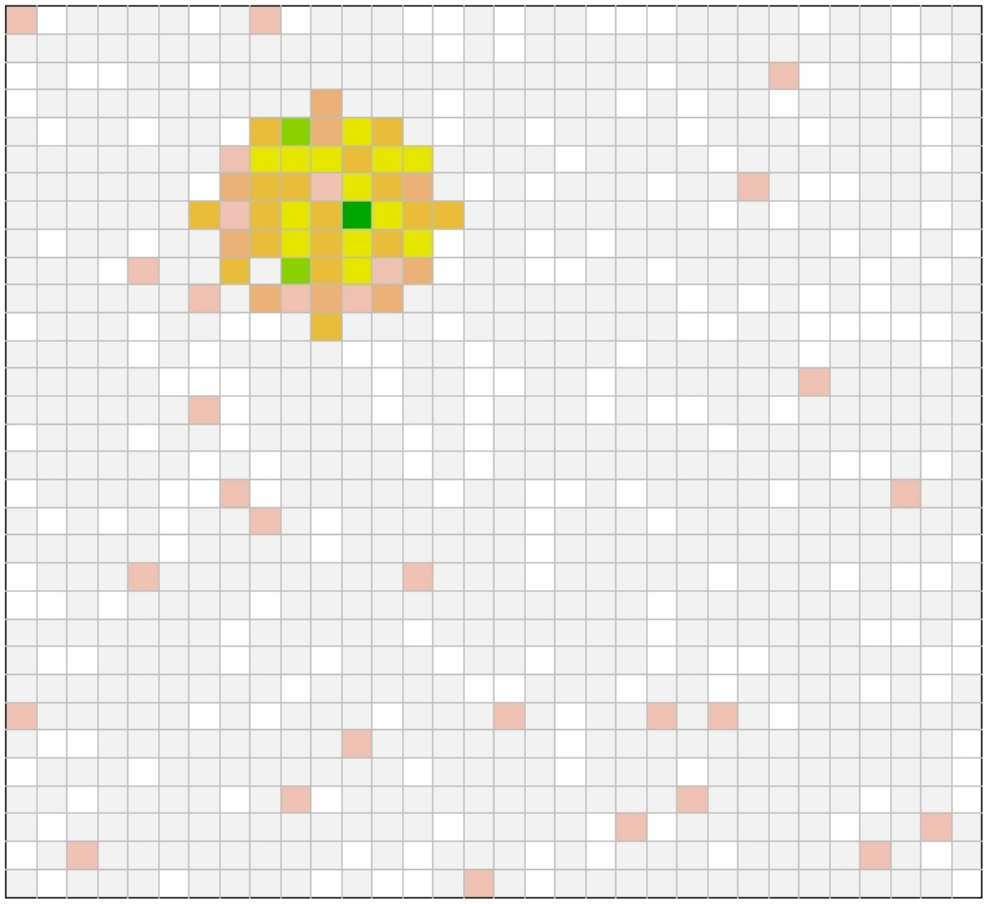

Figure 3: Darker pixels denote spatial locations with larger parameter values.

`tau` should be the same length as the number of parameters and contain parameter values for each variable by row, to match the indexing structure in Sections 3.1 and 3.1. Selecting `return.cov = TRUE` or `return.prec = TRUE` will tell to return a list whose first element is the covariance or precision matrix and whose second entry is the Cholesky factor of that matrix; this is the default setting.

Predictor generation (MVN)

Armed with the Cholesky factor, generating draws from an MVN requires only a vector of standard Normal draws $Z_1, \ldots, Z_J$ and vector of variable means, $\boldsymbol{\mu}$. `sim_MVN_X` generates `N` draws from a spatially correlated MVN given a Cholesky factor (specified by either `L` (lower triangular) or `R` (upper triangular)) and vector of means, `mu`. Like `sigma` in `chol_s2Dp`, if `mu` is a single scalar value then a common mean is assumed; otherwise, `mu` should be the same length as the number of parameters and contain means for each variable according the indexing structure in Sections 3.1 and 3.1. When using `spam`, we incorporate the functions, `rmvnorm.spam()` and `rmvnorm.prec()` to generate draws using covariance or precision matrices, respectively. These functions require us to include the covariance or precision matrix, but they also allow us to specify the Cholesky factor so that time is not wasted recomputing the Cholesky factor for successive calls. It does not matter whether `L` or `R` is used to specify the Cholesky factor, but whichever is specified must correspond correctly to a lower or upper triangular matrix, respectively; i.e., if `L` is specified then the matrix should be lower triangular. The arguments `S` and `Q` take the covariance or precision matrix, respectively. Figure 4 displays the first image from a sample of $N = 30$ 32×32 standardized images that have AR(1) correlation structure with $\rho = 0.9$ and minimum correlation $= 0.01$. The code necessary to produce Figure 4 is as follows:

```
set.seed(26378)
Lex <- sim2Dpredictr::chol_s2Dp(
        corr.structure = "ar1", matrix.type = "cov",
        im.res = c(32, 32), use.spam = TRUE,
        rho = 0.9, sigma = 1, corr.min = 0.01,
        triangle = "lower")

X <- sim2Dpredictr::sim_MVN_X(
        N = 30, mu = 0,
```

```
        L = Lex$L, S = Lex$S)


rotate = function(x){
  t(apply(x, 2, rev))
}
image(matrix(X[1, ], byrow = TRUE, nrow = 32), axes = FALSE)
box()
```

3.3 Binary Images Via the Boolean Model

Why not threshold GMRF's?

While perhaps less common, some applications require spatial dependence for predictors whose values are not continuous; e.g., lesion maps in structural brain images are binary, that is, the values are either 0 (no lesion at the location) or 1 (yes lesion at the location) (Bates et al., 2003; Mirman et al., 2018). In this section we present the Boolean model as a useful approach, but one may wonder why simply thresholding GMRF's is so undesirable as to require using the Boolean model. The answer is that the Boolean model is much more flexible and allows greater control over the size and shape of the clusters in the image, while thresholding GMRF's leaves us at the mercy of unconstrained and often unpredictable clustering patterns with potentially no resemblance to practical applications in mind.

Nevertheless, we have included arguments in `sim_MVN_X()` that allow for thresholding GMRF's to create categorical images; these are specified using `X.categorical = TRUE`, `X.category.type = c("manual", "percentile")`, and `X.num.categories`. `X.num.categories` determines the number of categories to generate. A vector of manual thresholds is specified with `X.manual.thresh`. Alternately, one can use `X.percentiles` to manually specify percentiles for thresholds; otherwise equally spaced percentiles are created using the number

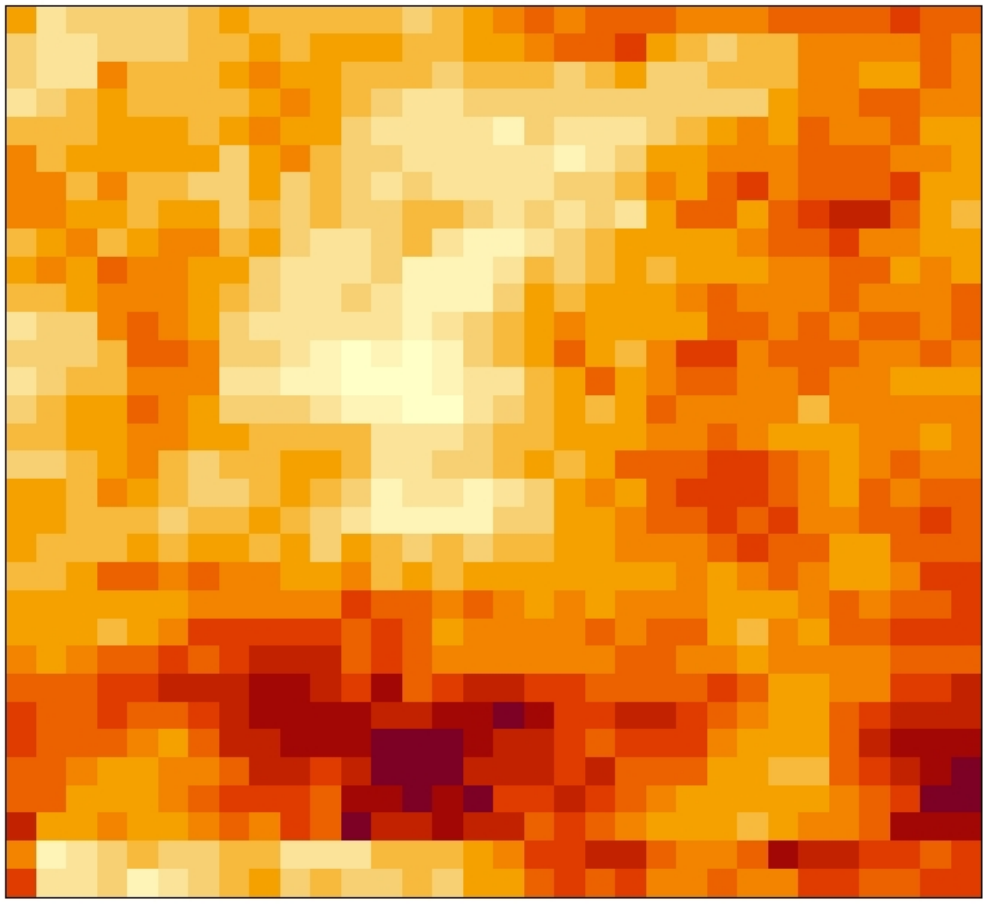

Figure 4: Here lies the simulated "image" for subject # 1. Looping over `X` by row with `image()` will allow for visualizing each of the 30 subjects' simulated images in succession. Note that `Lex` is a lower triangular Cholesky factor and `X` is a matrix whose rows are subject images.

of desired categories. These options may well prove useful to some users, especially those who are not concerned with the specifics of clustering. However, users interested in greater control over the size and shape of clusters will prefer the Boolean method.

Methodology

Thresholding a Gaussian Markov Random Field (GMRF) limits the ability to control the shape, size, and general qualities of the clustering. In addition, there is no such thing as a generalized multivariate Binomial distribution from which to generate dependent binary random draws, necessitating more creative approaches. The Boolean Model described in Section 2.2 is one such approach (Cressie and Wikle, 2011). The application of the Boolean Model here is as follows:

1. Define the dimensions of a space; in most cases the unit square suffices.
2. Use a Homogeneous Poisson Point Process (HPPP) with intensity/mean parameter λ to generate random spatial "events".
3. Randomly draw a number of radii equal to number of events; center circles with their respective radii at each event.
4. The union of the area within the circles is taken as the random set.
5. Map the random set onto a lattice/image space where locations within the random set/area to assigned values of 1, and locations outside the random set/area are assigned values of 0.

This approach can be tuned to produce greater flexibility in the observed qualities of random sets generated. The most obvious tuning approaches are to change the intensity/mean parameter λ and change the distribution of the random radii; tweaking only these parameters opens a world of clustering shapes, sizes, and densities exhibited by the random set. Choosing random sub-areas upon which to apply the HPPP allow even greater control; appropriately sized random sub-areas combined with careful selection of λ and radii distribution can result in mostly contiguous random sets. We implement this approach in the function `sim2D_binarymap()`; example images are shown in Figures 5 and 6.

Basic arguments

While its options are wide-ranging, `sim2D_binarymap()` can be run with a bare minimum of arguments. The default limits for the 2D space are `xlim = c(0, 1)` and `ylim = c(0, 1)`; unless there is an obvious reason to alter the default limits it is recommended to leave them as standardized. The sample size, i.e., number of images to be generated, is defined with `N`. As in other functions, image resolution is defined by `im.res`. The random radii are selected from a uniform distribution with lower and upper bounds defined by the first and second elements of `radius.bounds`, respectively. The intensity/mean parameter λ is defined by `lambda`, with default `lambda = 50`. A final option is `store.type = c("list", "matrix")`, which determines whether each lattice/image is stored as is in a list, or is vectorized and stored in a matrix such that each row is the vectorized image for a single subject; the default is `store.type = "list"`.

Random sub-areas

Random sub-areas upon which to apply the Boolean model greatly increase the ability to create desired clustering patterns, and can be combined with tuning `lambda` and `radius.bounds` to generate contiguous random sets. When `sub.area = TRUE`, random sub-areas are chosen over which to apply the Boolean Model. These sub-areas are rectangular in shape and can be tuned by altering `min.sa` and `max.sa`, which define the dimensions of the minimum and maximum possible sub-areas. The defaults, `min.sa = c(0.1, 0.1)` and `max.sa = c(0.3, 0.3)`, indicate the random sub-areas can take any dimensions between 0.10×0.10 and 0.30×0.30; note that these sub-area bounds should be within the image bounds specified with `xlim` and `ylim` and should always be positive; e.g., the default is the unit square, in which case sub-areas cannot exceed 1×1. When using random sub-areas it is also useful to specify radius bounds for both the minimum and maximum sub-areas, which is done with `radius.bounds.min.sa` and `radius.bounds.max.sa`. These values are used to scale the radius bounds for a given

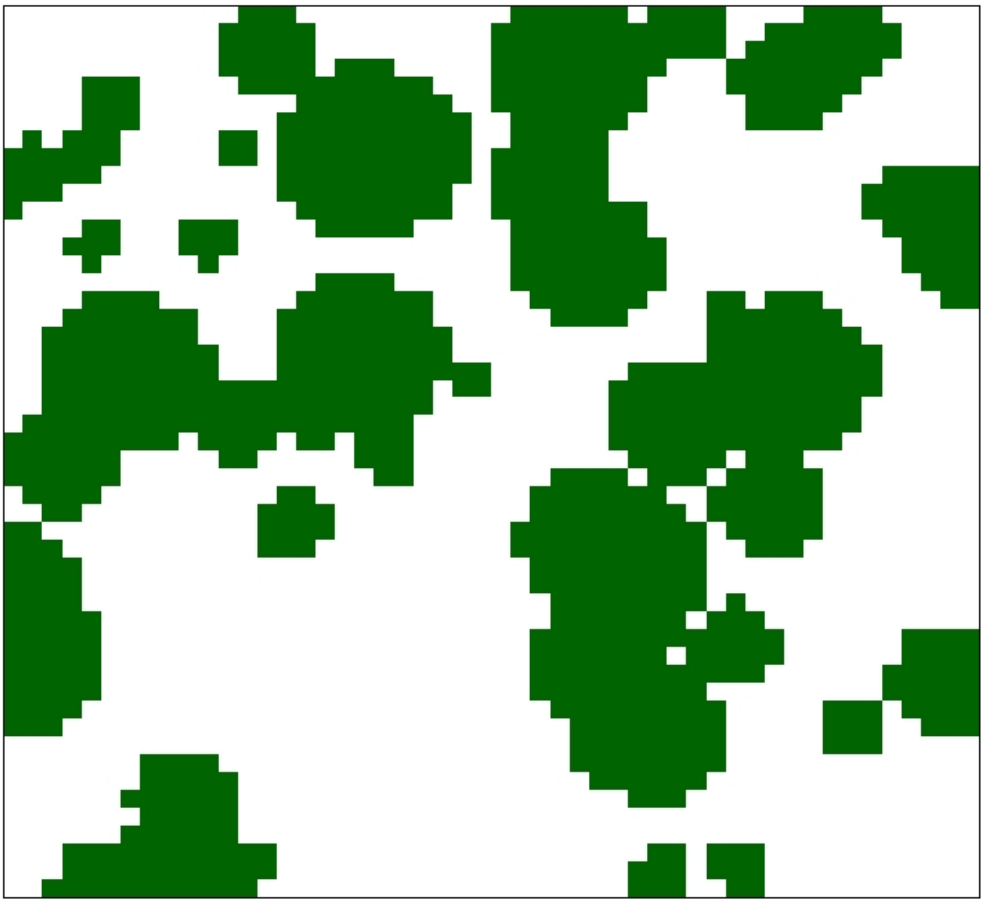

Figure 5: An example of a binary image generated by `sim2D_binarymap()` where the HPPP is applied to the entire area.

random sub-area based on the size of that sub-area. Tweaking these parameters allows for much greater flexibility in size and shape of the random sets.

Random lambda

A final option is to allow the intensity mean `lambda` (λ) to vary by subject by specifying `random.lambda = TRUE`, which places a prior distribution on `lambda` (λ). In this case `lambda` is the mean and `lambda.sd` is the standard deviation of the of the

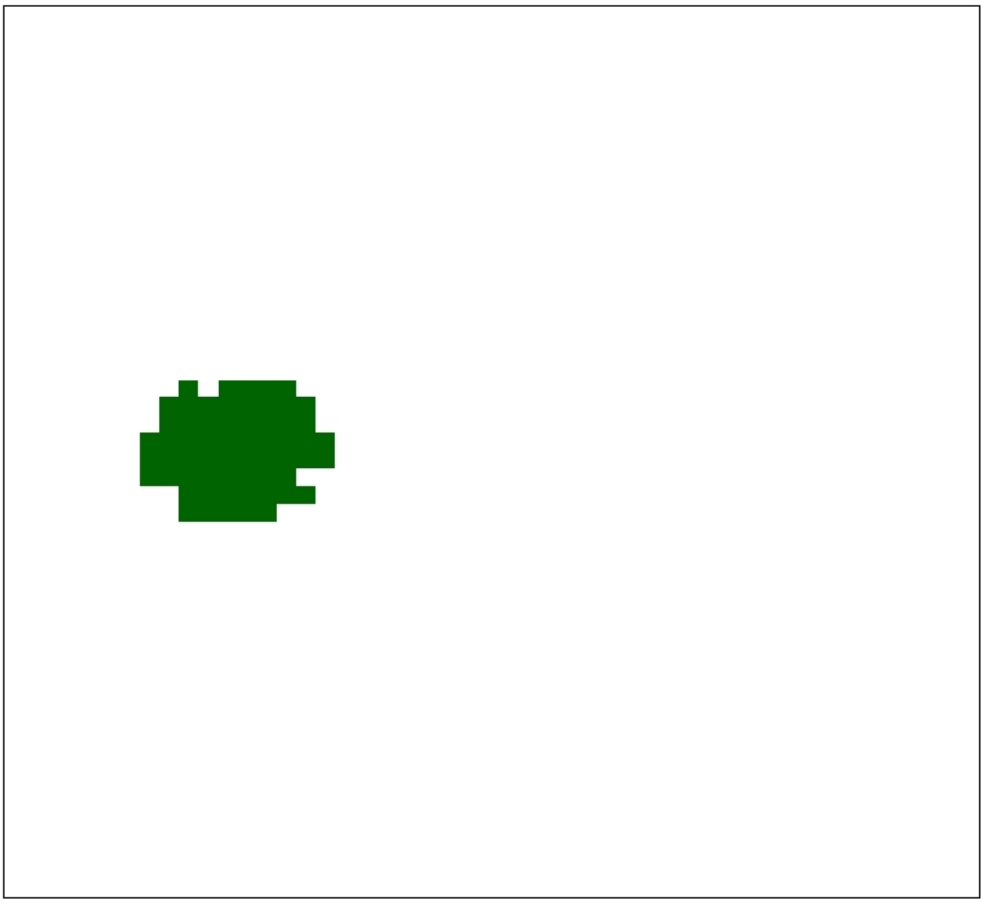

Figure 6: An example of a binary image generated by `sim2D_binarymap()` where the HPPP is applied to a randomly selected sub-area. In this particular case, the chosen parameter settings result in a single relatively small cluster.

prior distribution. Distributional options are specified by `prior = c("gaussian", "gamma")` to use either a Normal or Gamma distribution to draw subject-specific λ; the default is `prior = "gamma"`.

Additionally, one may desire to place accept/reject boundaries on the random λ, which can be done with `lambda.bound`, a 2-element vector whose first and second entries are the lower and upper bounds, respectively. If a drawn value is outside these bounds then

it is discarded and another draw taken; this is continued until a draw is taken that results in a value within the bounds.

3.4 Outcome Generation

Methodology

With subject images and parameter vectors in hand, scalar outcomes for each subject can now be generated. We incorporate `sim_MVN_X()` into `sim_Y_MVN_X()` and `sim2D_binarymap()` into `sim_Y_binary_X()`, respectively, in order to first generate the predictor images and from these images generate a scalar outcome for each subject; we create separate functions to minimize argument overload, but the basic structure of both functions are the same, with only design matrix generation arguments differing; i.e., both `sim_Y_MVN_X()` and `sim_Y_binary_X()` retain the arguments from their respective design matrix-generating functions, but the outcome-level arguments described below are the same.

Currently, there are three options for outcomes: Normal/Gaussian (continuous)), binary, and count. These options are specified using `dist = c("gaussian", "binomial", "poisson")`, respectively. All outcomes generated are independent of each other, i.e., subjects do not have repeated measures, and are assumed to have measurements independent of each other. Additionally, it is necessary to specify a parameter vector, β, which is easily obtained using `beta_builder()` as described in Section 3.1. If the desired outcome is Normally distributed, then the only remaining relevant argument is `rand.err`, which specifies the variance of the random error; i.e., for the data appropriate for the forms in Equations (2.1.1) and (2.1.2), the random error is assumed to follow a Normal distribution with zero mean and random error variance, σ_i^2; i.e., $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$.

However, as usual when departing from Normality, the options are more varied for generating binary or count data. The primary reason for this is that when the individual outcomes are Normally distributed, matrix algebra is available to generate all the outcomes

in a single operation using matrix algebra with Equation (2.1.2). However, binary and count data are often assumed to follow Binomial and Poisson distributions, respectively, and these distributions are most intuitively modeled in the framework of Generalized Linear Models of Equation (2.1.3). Unfortunately, the link functions $g(\cdot)$ appropriate for Binomial and Poisson distributions are not amenable to comparatively elegant matrix algebra approaches. Thus, if we insist upon using the GLM framework to draw directly from Binomial or Poisson distributions, then we must settle for subject-by-subject outcome generation. Each subject's expectation is calculated:

$$\mathrm{E}(Y_i) = g^{-1}(\beta_0 + X_{i1}\beta_1 + \cdots + X_{iJ}\beta_J) \tag{3.4.1}$$

Since the variances of the Binomial and Poisson distributions are functions of the expectation, once Equation (3.4.1) is used to find the subject-specific expectation, one can draw directly from either Binomial or Poisson random number generators (e.g., `rbinom()` and `rpois()`). This is the default approach implemented (i.e., `threshold.method = "none"`).

Alternatively, we can draw in the usual manner from a Normal distribution as in Section 3.2, then use various thresholding rules to obtain binary outcomes or counts. `sim_Y_MVN_X()` has two alternative approaches to binary outcome generation. Both approaches obtain a threshold, equal or below which, $Y_i = 0$, and above which $Y_i = 1$. The first approach is to specify `threshold.method = "manual"` and use `Y.thresh` to specify a raw threshold value. The second approach is the `threshold.method = "percentile"` and use `Y.thresh` to specify the percentile used to determine the threshold; e.g., if `Y.thresh = 0.80`, then the function calculates the 80^{th} percentile of the values drawn from the Normal distribution and sets all those equal or less than the 80^{th} percentile to 0, and those above to 1.

Similarly, count data can be obtained using draws from a Normal distribution. Here, we use a simple approach were values below or equal 0 are set to 0 and all positive values are rounding to the nearest whole number; this approach is implemented when `threshold.method = "round"`. Note that while this approach will result in count data, the distribution counts will often not resemble a Poisson distribution, so we recommend drawing directly from the Poisson distribution when one desires the draws to resemble the Poisson distribution.

Generating outcomes with spatially correlated predictors: example

As an example, consider simulating data for $N = 15$ subjects with 5×5 predictor images. After obtaining the Cholesky factor `Lex$L` and parameter vector `Bex$B`, we can easily use `simY_MVN_X()` to generate both the predictor images and scalar outcomes for all 15 subjects. Note that since the same seed was used to generate both data sets, we can verify that the 75^{th} percentile cutoff is binarizing the outcomes in the desired fashion.

```
Lex <- sim2Dpredictr::chol_s2Dp(
        corr.structure = "ar1",
        im.res = c(15, 15),
        matrix.type = "cov",
        corr.min = 0.01,
        return.cov = TRUE,
        rho = 0.5, sigma = 1,
        triangle = "lower",
        use.spam = TRUE)

Bex <- sim2Dpredictr::beta_builder(
        row.index = c(1, 1, 2),
        col.index = c(1, 2, 1),
```

```
        im.res = c(15, 15),
        B0 = 16, B.values = 3,
        index.type = "manual")

set.seed(26378)
gdat <- sim2Dpredictr::sim_Y_MVN_X(
        N = 15, mu = 0,
        L = Lex$L, S = Lex$S,
        use.spam = TRUE,
        B = Bex$B, dist = "gaussian")

set.seed(26378)
bdat <- sim2Dpredictr::sim_Y_MVN_X(
        N = 15, mu = 0,
        L = Lex$L, S = Lex$S,
        B = Bex$B, dist = "binomial",
        use.spam = TRUE,
        threshold.method = "percentile",
        Y.thresh = 0.75)
```

```
## The  75 th Percentile (threshold) is  20.10668 .
```

```
cbind(Y.gauss = gdat$Y, bdat)[, 1:6]
```

```
##     Y.gauss Y         X1         X2        X3        X4
## 1  20.88971 1  1.4377010  0.0222133  0.609156  0.564409
## 2  15.89933 0  0.1136221 -0.4097131 -0.872702 -0.484842
## 3   5.69438 0 -0.8288431 -0.8558688 -0.402442  0.237785
## 4  14.48040 0  0.0178685 -0.0395599  0.493261  0.393154
```

```
## 5  16.94924 0 -0.4917191  0.3245071 -0.589253 -1.889405
## 6  34.30102 1  1.9701192  1.0458329  0.566728 -0.473586
## 7  10.76760 0 -0.4065059 -0.6840246 -0.245032 -0.892260
## 8  20.10668 0  1.2675565  0.5662275  0.348832  1.639263
## 9  10.00988 0 -1.0704018 -0.1633097  0.816867  0.656865
## 10 19.63775 0  0.2881544  0.9854223  0.470944  2.312622
## 11 21.87931 1 -0.2644823  0.3734159  0.537972 -0.154805
## 12 17.14722 0  0.1477454 -0.6276245  0.857566 -0.100071
## 13  3.01856 0 -1.0164985 -2.3074985 -1.683845 -1.278653
## 14 15.70094 0  0.4390649 -0.4550332  0.889365  1.934803
## 15 21.81999 1  0.3978131 -0.0493428 -1.419200 -0.305921
```

3.5 Extracting Simulation Results

Simulations are generally intended to evaluate or explore the qualities of a statistical method; variable selection methods are judged according to whether they identify variables with non-zero parameters (i.e., Power) without carrying along variables with zero-valued parameters (i.e., false positives); the false discovery rate (FDR), family-wise error rate (FWER), and Power are some of the more useful metrics in such situations. Note that here we use "Power" to refer to a generalization: sample "Power" will refer to the proportion of non-zero parameters identified as non-zero. The average of these "Power proportions" will provide an estimate for the asymptotic "Power proportion". The function `sample_FP_Power()` provides sample summaries, i.e., sample false discovery proportion (FDP), family-wise error (FWE), and proportion of "true" non-zero parameters discovered (Power), given a vector of rejected hypothesis indicators (1 for rejection; 0 otherwise) and corresponding parameter vector, specified by the arguments `rejections` and `B`, respectively. If one has a vector of test statistics, then `rejections` can be calculated internally with the `test.statistic` argument, which is a vector of test statistics, e.g., F statistics or p values. The first ele-

ment of list argument `reject.threshold` is the threshold used to determine rejections, while the second element is one of `c("greater", "less", "2-tailed")`, which determines whether the reject for values above or below the threshold, or to perform a 2-tailed test, whose upper and lower thresholds are computed internally. While usually unnecessary, false positive and true positive vectors can be manually included with `FP` and `TP`, respectively. `B.incl.B0 = c(TRUE, FALSE)` and indicates whether `B` includes the intercept. Rarely does a vector of rejections contain an intercept, but quite often the parameter vector used to generate the data does; when `B.incl.B0 = TRUE`, the intercept is assumed to be the first element of `B`, and it is removed. Finally, `full.summary = TRUE` will additionally output total numbers of rejections, false positives, true positives, and non-zero parameters. Consider the following example, where `rejex` is a vector of rejections and `Bex` is a parameter vector:

```
rejex <- c(1, 0, 1, 1, 0, 0, 1, 1, 0, 0)
Bex <- c(15, 4.22, 3.16, 0, 0.98, 0, 0, 2.91, 0, 1.78, 0)
sim2Dpredictr::sample_FP_Power(
        rejections = rejex, B = Bex,
        B.incl.B0 = TRUE,
        full.summary = TRUE)
```

```
##   NumNonzeroB NumReject NumFP NumTP FDP FWE Power
## 1           5         5     2     3 0.4   1   0.6
```

We also demonstrate the a 2-tailed test using Z statistics and a cutoff of 1.96 and -1.96, and where the vector of parameters, not including the intercept, is $\boldsymbol{\beta} = \{0, 0, 4, 1, -2\}$:

```
zstat <- c(-0.5, 1.98, 2.01, 1.45, -1.99)


sim2Dpredictr::sample_FP_Power(
```

```
        test.statistic = zstat,
        reject.threshold = list(1.96, "2-tailed"),
        B = c(0, 0, 4, 1, -2), B.incl.B0 = FALSE,
        full.summary = TRUE)
```

```
##   NumNonzeroB NumReject NumFP NumTP     FDP FWE   Power
## 1           3         3     1     2 0.33333   1 0.66667
```

3.6 Visualizing Results

`inf_2Dimage()` is a simple function for visualizing inference results from a simulated data set and/or verifying that β corresponds to the desired spatial area; it was used to generate Figures 2, 3, and 7. It allows concurrently viewing the spatial extents of non-zero parameters and rejected parameters, thereby making it possible to see the spatial extents of true and false positives. As in `sample_FP_Power`, `rejections` is a vector of ordered rejected hypotheses corresponding to the same indices as the parameter vector, `B`; as before, the rejection vector can also be created internally by using `test.statistic` and `reject.threshold`. As in other functions, `im.res` is a vector whose first and second entries are the number of rows and columns of the image, respectively, while `B.incl.B0` determines whether the function expects the first entry of `B` to be an intercept. New arguments include `binarize.B`, which when set to `TRUE` will binarize the parameter vector `B` so that the non-zero parameters will all show up as the same color, and `grid.color`, which specifies the color of the grid on the plot. The legend labels denote true negatives (TN), false negatives (FN), true positives (TP), and false positives (FP).

4 Practical Application

4.1 Example: Lasso Analysis

A common variable selection method for large dimensional data is the lasso, a penalized regression introduced by Tibshirani (1996), which subjects parameter estimates

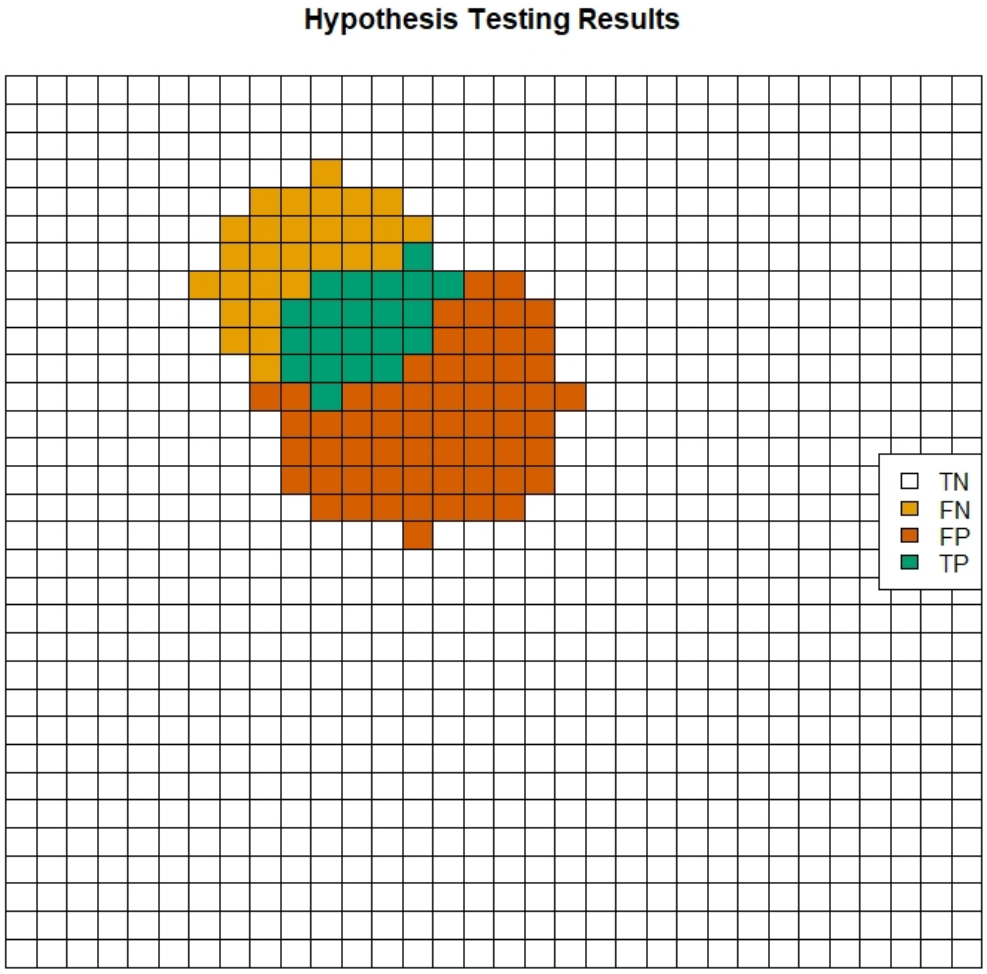

Figure 7: Here we see an example of model performance when the true non-zero parameters are known. Note that the area of "true" non-zero parameters is the same as in Figure 2.

to the constraint $\sum_{j=1}^{J} |\beta_j| \leq t$. Another formulation is to add the term $\lambda \sum_{j=1}^{J} |\beta_j|$ to the objective function, which allows us to interpret λ directly as a penalty parameter where larger values induce stronger shrinkage. Furthermore, the lasso has a Bayesian interpretation as placing a double exponential prior on the β_j, making it a useful tool in both Frequentist and Bayesian philosophical settings (Park and Casella, 2008). The lasso is particularly useful because it can fit models where the number of subjects is much smaller than the

number predictors, and can handle significant correlation between predictors. In addition, the lasso performs automatic variable selection by shrinking many parameters to exactly zero. Generalized linear models using the lasso penalty are easily fit using pathwise coordinate decent via the package `glmnet` (Friedman et al., 2007, 2010). Here we provide a short simulation using the lasso to show the capabilities of the functions introduced in previous sections.

4.2 Data Generation

We generate 15×15 (225 pixels) subject images with mean 0, standard deviation 1, and correlation function using Gaussian decay with $\phi = 0.75$ as in Equation (3.1.3); correlations less than 0.01 are set to 0, which allows for a relatively sparse matrix since only about 8% of correlations are non-zero (see Table 1). We also use a decay function to set the non-zero β_j resulting in a decay from $\beta_j = 0.5$ to $\beta_j = 0.0249$ (see Table 2 and Figure 8). This results in 29 (12.89%) non-zero parameters.

```
sim.res <- c(15, 15)


L <- sim2Dpredictr::chol_s2Dp(
        im.res = sim.res,
        corr.structure = "gaussian",
        use.spam = TRUE,
        phi = 0.75, corr.min = 0.01,
        triangle = "lower")


Sm <- as.matrix(L$S)


B <- sim2Dpredictr::beta_builder(
        im.res = sim.res,
```

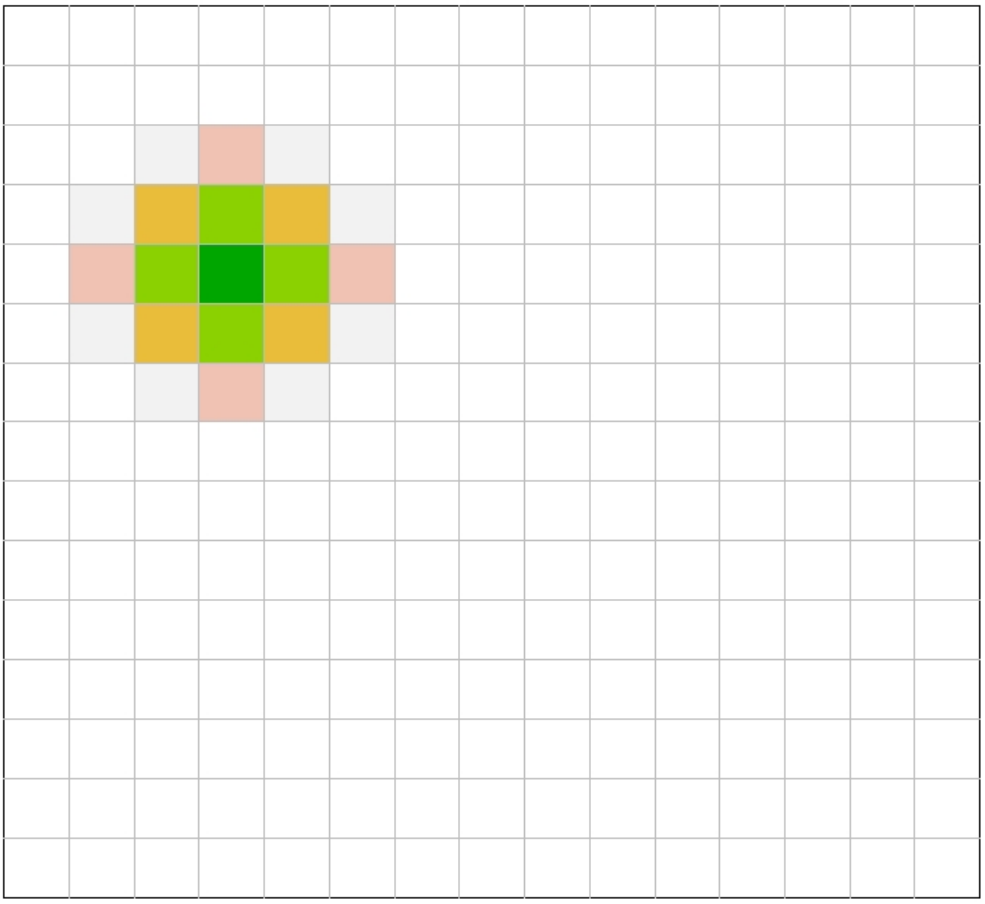

Figure 8: Darker colors represent larger magnitude parameters; the closer the color gets to white, the closer the magnitude is to zero, with white pixels indicating zero parameters.

```
        row.index = 5, col.index = 4,
        B.values = 0.5, index.type = "decay",
        decay.fn = "gaussian",
        phi = 1/3, max.d = 3,
        output.indices = FALSE)
```

We then generate $M = 1,000$ datasets of $N = 100$ subjects each with binary outcomes:

```r
set.seed(1383352)

sim.N <- 100

M <- 1000

dat <- list()

for (m in 1:M) {

  dat[[m]] <- sim2Dpredictr::sim_Y_MVN_X(

        N = sim.N, B = B,

        L = L$L, S = L$S,

        dist = "binomial",

        threshold.method = "none",

        incl.subjectID = FALSE)

}
```

Table 1: Summaries for the correlation imposed to generate images ($\boldsymbol{X}_i$) and parameter vector ($\boldsymbol{\beta}$).

Correlation		$\boldsymbol{\beta}$	
% Non-Zero	Maximum	% Non-Zero	# Non-Zero
0.0806914	0.4723666	0.1288889	29

Table 2: Frequencies for each non-zero parameter value in $\boldsymbol{\beta}$.

Value of β_j	0.0249	0.0347	0.0944	0.1318	0.2567	0.3583	0.5
Frequency	4	4	8	4	4	4	1

4.3 Results

We analyze the datasets with `cv.glmnet()`, which uses k-fold cross validation to choose the optimal penalty parameter λ according to some prediction error criteria. We use 5-fold cross validation and choose model whose value of λ maximizes cross validated AUC.

```
set.seed(28623)
for (m in 1:M) {
  dat.m <- dat[[m]]
  gm <- glmnet::cv.glmnet(
        x = as.matrix(dat.m[, -1 ]),
        y = dat.m$Y, family = "binomial",
        type.measure = "auc",
        nfolds = 5)
  gm.c <- glmnet::coef.glmnet(gm, s = gm$lambda.min)
  rej.m <- rep(0, prod(sim.res))
  rej.m[as.vector(gm.c)[-1] != 0] <- 1
  results.m <- cbind(
        auc = gm$cvm[gm$lambda == gm$lambda.min],
        lambda = gm$lambda.min,
        sim2Dpredictr::sample_FP_Power(
                B = B,
                rejections = rej.m))
  if (m > 1) {
    results <- rbind(results, results.m)
  } else {
    results <- results.m
  }
}
```

The full results are summarized in Table 3, where we find that for the generated data, on average the lasso identifies just under 20% of the non-zero parameters while about 31% of estimated non-zero parameters are spurious; futhermore, these observed values vary noticeably over the datasets, with standard deviations not too far off from the means. Nevertheless, the lasso performs well in this case with respect to cross validated AUC, averaging near 80% with standard deviation of about 6%. We can also see from Figure 9 that the distributions of the statistics are not generally symmetric; AUC is slightly left skewed and λ slightly right skewed, while FDP and Power are arguably mixture distributions; note that 37% of the time the lasso includes no false positives.

Table 3: Summaries of lasso performance.

Summary	AUC	lambda (λ)	FDP	FWE	Power
Mean	0.7953072	0.1030760	0.3109088	0.6300000	0.1996207
Median	0.0639995	0.0576828	0.2879007	0.4830459	0.1187947
SD	0.8006691	0.0996856	0.2857143	1.0000000	0.1724138
IQR	0.0794124	0.0732538	0.5714286	1.0000000	0.1724138

5 Summary and Discussion

The `R` package `sim2Dpredictr` is designed to make simulating scalar outcomes from spatially correlated predictors and spatially clustered parameters straightforward and user-friendly. The available functions allow users to flexibly and intuitively modify correlation structures in images and clustering in non-zero parameters. While other available packages, e.g., `RandomFields`, `gstat`, and `neuRosim`, offer simulation functions for spatial settings, we believe that `sim2Dpredictr` meets a different need in that it allows simulations that are tailored to using images as predictors rather than as outcomes, and in addition does not require significant background knowledge in specific practical applications or statistical theory in order to use effectively. The package is not intended to have a steep learning curve, and can often be used with a minimum of arguments; nevertheless, more

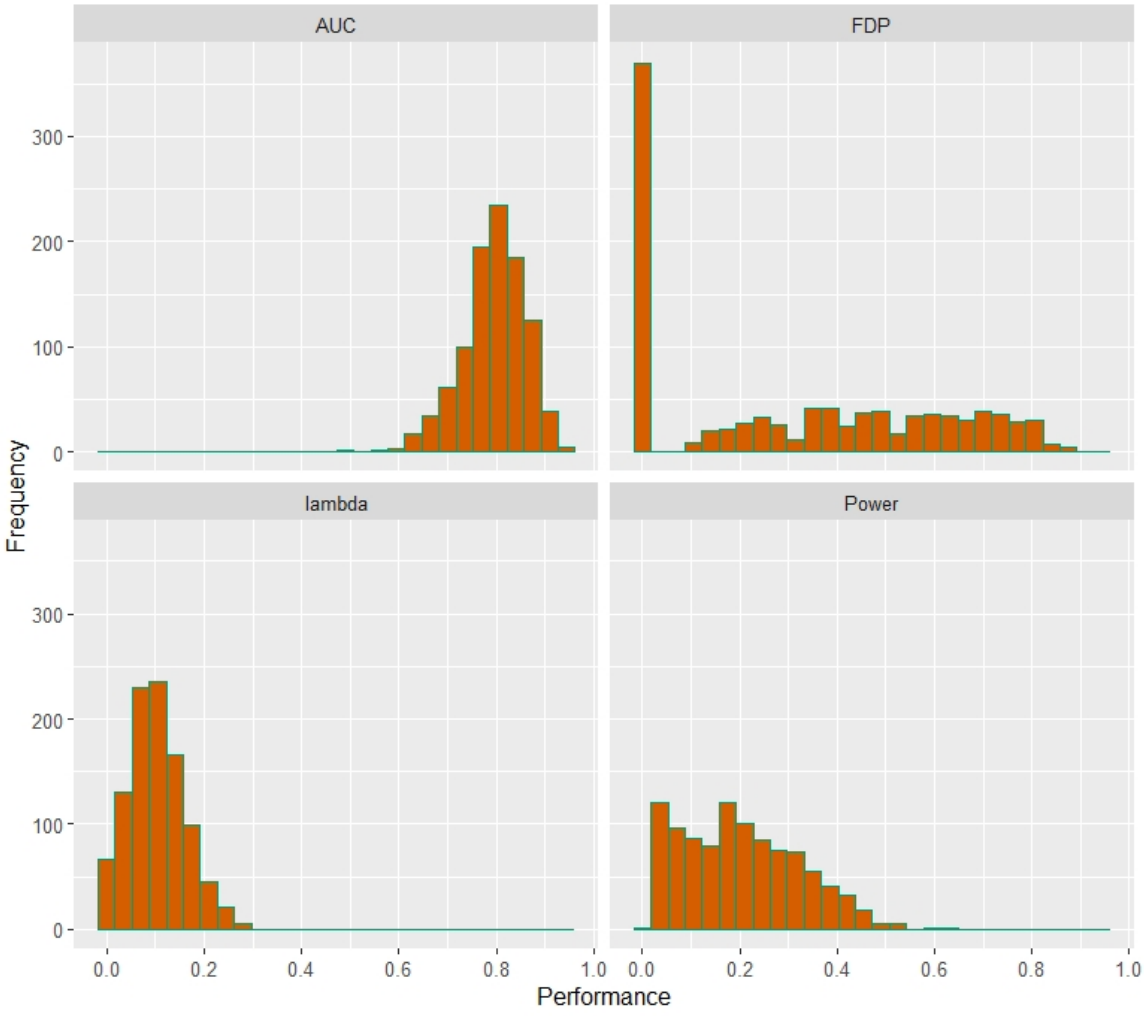

Figure 9: Distribution of performance statistics for the lasso over 1,000 datasets.

complex arguments are available for those who require their use, and the code is freely available so that it can be easily modified by others to suit different needs. Furthermore, the Boolean Method can offer far more flexibility than thresholding methods to generate binary images that resemble the practical application on hand, and `sim2Dpredictr` allows simulating either continuous or binary spatial predictors in addition to continuous, binary, or count outcomes within a single unified framework so that users can spend their time focusing on developing variable selection methods rather than fighting with code.

Computational Details

The results in this paper were obtained using `R` version 3.6.2.

REFERENCES

Andrews, D.L. and Mallows, C.L. =. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society B*, 36(1):99–102, 1974. doi: https://doi.org/10.1111/j.2517-6161.1974.tb00989.x.

Banerjee, Sudipto; Carlin, Bradley P., and Gelfand, Alan E. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, Florida, 2nd edition, 2015.

Bates, Elizabeth; Wilson, Stephen M.; Saygin, Ayse Pinar; Dick, Frederic; Sereno, Martin I.; Knight, Robert T., and Dronkers, Nina F. Voxel-based lesion-symptom mapping. *Nature Neuroscience*, 6(5):448–450, 2003. doi: 10.1038/nn1050.

Benjamini, Yoav and Hochberg, Yosef. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1): 289–300, 1995.

Benjamini, Yoav and Yekutieli, Daniel. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. doi: 10.1214/aos/1013699998.

Besag, Julian. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

Besag, Julian and Kooperberg, Charles. On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746, 1995. doi: 10.1093/biomet/82.4.733.

Cressie, Noel and Wikle, Christopher K. *Statistics for Spatio-Temporal Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, 2011.

Diggle, Peter J. *Time Series: A Biostatistical Introduction*. Oxford University Press, New York, New York, 1990.

Diggle, Peter J.; Heagerty, Patrick; Liang, Kung-Yee, and Zeger, Scott L. *Analysis of Longitudinal Data*. Oxford University Press, Oxford, UK, 2nd edition, 2002.

Farcomeni, Alessio. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 17: 347–388, 2008. doi: 10.1177/0962280206079046.

Friedman, J.; Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010. doi: 10.18637/jss.v033.i01.

Friedman, Jerome; Hastie, Trevor; Höfling, Holger, and Tibshirani, Robert. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007. doi: 10.1214/07-AOAS131.

Furrer, Reinhard and Sain, Stephan R. `spam`: A sparse matrix `R` package with emphasis on MCMC methods for Gaussian Markov random fields. *Journal of Statistical Software*, 36 (10):1–25, 2010. URL http://www.jstatsoft.org/v36/i10/.

Genovese, Christopher; Lazar, Nicole, and Nichols, Thomas. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15:870–878, 2002. doi: 10.1006/nimg.2001.1037.

Gräler, Benedikt; Pebesma, Edzer, and Heuvelink, Gerard. Spatio-temporal interpolation using `gstat`. *The R Journal*, 8:204–218, 2016. URL https://journal.r-project.org/archive/2016/RJ-2016-014/index.html.

Hastie, Trevor; Tibshirani, Robert, and Friedman, Jerome. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, 2nd edition, 2009. URL http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

Hedeker, Donald and Gibbons, Robert D. *Longitudinal Data Analysis*. John Wiley & Sons, Hoboken, New Jersey, 2006. doi: 10.1002/0470036486.

Jin, Xiaoping; Carlin, Bradley P., and Banerjee, Sudipto. Generalized hierarchical multivariate car models for areal data. *Biometrics*, 61(4):950–961, 2005. doi: 10.1111/j.1541-0420.2005.00359.x.

Mirman, Daniel; Landrigan, Jon-Frederick; Kokolis, Spiro; Verillo, Sean; Ferrara, Casey, and Pustina, Dorian. Corrections for multiple comparisons in voxel-based lesion-symptom mapping. *Neuropsychologia*, 115:112–123, 2018. doi: 10.1016/j.neuropsychologia.2017.08.025.

Park, Trevor and Casella, George. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008. doi: 10.1198/016214508000000337.

Pebesma, Edzer J. Multivariable geostatistics in `S`: the `gstat` package. *Computers & Geosciences*, 30:683–691, 2004. doi: 10.1016/j.cageo.2004.03.012.

`R` Core Team, . *`R`: A Language and Environment for Statistical Computing*. `R` Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

Ripley, Brian D. *Stochastic Simulation*. John Wiley & Sons, 1987. doi: 10.1002/9780470316726.

Rue, Håvard. Fast sampling of gaussian markov random fields. *Journal of the Royal Statistical Society B*, 63:325–338, 2001. doi: 10.1111/1467-9868.00288.

Rue, Håvard and Held, Leonhard. *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, Boca Raton, Florida, 2005.

Schlather, Martin; Malinowski, Alexander; Menck, Peter J.; Oesting, Marco, and Strokorb, Kirstin. Analysis, simulation and prediction of multivariate random fields with package

RandomFields. *Journal of Statistical Software*, 63(8):1–25, 2015. URL http://www.jstatsoft.org/v63/i08/.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *J.R. Statist. Soc.*, 58 (1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Welvaert, Marijke; Durnez, Joke; Moerkerke, Beatrijs; Verdoolaege, Geert, and Rosseel, Yves. neuRosim: An R package for generating fmri data. *Journal of Statistical Software*, 44(10):1–18, 2011. URL http://www.jstatsoft.org/v44/i10/.

INCORPORATING SPATIAL STRUCTURE INTO INCLUSION PROBABILITIES FOR BAYESIAN VARIABLE SELECTION

JUSTIN M. LEACH, INMACULADA ABAN, AND NENGJUN YI

In preparation for *Biostatistics*

Format adapted for dissertation

1 Introduction

Variable selection for linear models is a long-standing statistical problem tackled by both classical and Bayesian statisticians, and is useful for at least two reasons. The first reason is to select which variables are associated with an outcome or set of outcomes with the intention of providing an explicit scientific interpretation that relates the predictor(s) to the outcome. Variable selection is particularly useful when there are large numbers of potential predictors, since including all possible variables in a model may result in an excellent fit to the data on hand, but perform miserably when applied to an independent data set. It is thus especially important to apply some variable selection approach when the primary purpose of a model is to predict outcomes or generalize to new data.

Classical statistics often relies on hypothesis testing to select parameters that should remain in a (generalized) linear model (GLM), but due to the variability in the final model selected, prediction performance is often unacceptably poor, resulting in models that do not generalize well, despite removing many "noise" variables. Such issues helped to motivate the lasso model (Tibshirani, 1996). The lasso is a penalized approach to GLMs that decreases the variability of the selected model compared to, e.g., subset selection and stepwise approaches. Unlike other penalized approaches, e.g., ridge, the lasso can shrink parameter estimates to exactly zero and thus perform automatic variable selection. Another benefit of penalized approaches like the lasso is their ability to fit models when the number of predictors exceeds the number of the observations, whereas traditional GLMs are not identifiable in such cases.

Furthermore, while the lasso was initially developed in a classical framework, it has a naturally Bayesian interpretation as placing a double exponential prior on the parameters

of interest Park and Casella (2008). Bayesian variable selection has often relied on so-called spike-and-slab priors, which assume that the distribution of parameters can be modeled as a mixture of two distributions: a wide "slab" distribution to model "important" parameters that should be included in the model and a narrow "spike" distribution to model "unimportant" parameters that should be excluded from the model (Mitchell and Beauchamp, 1988; George and McCulloch, 1993). Ročková and George (2018) combine these approaches into the spike-and-slab lasso, which places double exponential priors on both the spike and slab distributions and thus allows for much greater flexibility in variable selection by trading a single penalty for an adaptive penalty based on the probabilities of inclusion. While the initial spike-and-slab lasso was proposed and described for regression, Tang et al. (2017) demonstrated a novel computational approach to using the spike-and-slab lasso prior to estimate parameters in GLMs, and showed the model improved prediction performance compared to the traditional lasso.

An additional benefit of penalized and/or Bayesian linear models is that they can provide estimates even when the predictors are highly correlated. However, in their most basic forms they do not model correlation among predictors, nor do they account for situations where clustering of non-zero parameters is expected. Specifically, when the predictors are vectorized images associated with subjects, it is reasonable to expect "important" and "unimportant" parameters will cluster near each other, respectively. In the language of variable selection, we should expect that the probability a parameter should remain in the model will be similar to the respective probabilities of spatially adjacent parameters.

However, the lasso has two primary downsides. The first is that when the number of predictors is much larger than the number of subjects, the number of non-zero parameters cannot exceed the number of subjects. The second is that when predictors are correlated it tends to select one predictor and discard the rest. In part, these issues inspired the elastic net, which proposed a compromise between ridge and lasso penalties (Zou and Hastie, 2005).

Unlike the ridge penalty, the elastic net solution is sparse, but unlike the lasso penalty it can include more non-zero parameters where appropriate. This flexibility may be desirable when images are used as predictors, where solutions should often be sparse, i.e., the number of "important" predictors is much smaller than entire set of predictors examined but still near to or larger than the number of subjects. In addition, images tend to exhibit strong correlation, and thus the lasso may select only a few out of the entire set of "important" predictors.

In what follows we extend the spike-and-slab lasso to a spike-and-slab elastic net that can explicitly incorporate spatial information into variable selection. Section 2 reviews the spike-and-slab lasso GLM and outlines the EM-algorithm used to fit the model. Section 3 introduces an extension of the spike-and-slab lasso GLM that generalizes the model to the elastic net and uses a variant of conditional autoregressions to model spatial structure. Section 4 details a small simulation study to demonstrate the potential of the proposed method and examine its properties. Finally, Section 5 summarizes the findings and discuss their implications.

2 Spike-and-Slab Lasso for GLM

2.1 Theory Overview

GLM's can model outcomes that are non-normal, and include the traditional linear model as a special case. The standard form of a GLM is given by:

$$g(y_i) = \boldsymbol{X}_i\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{J} x_{ij}\beta_j = \eta_i, \quad i = 1, \ldots, N \tag{2.1.1}$$

where $g(\cdot)$ is an appropriate link function, $\boldsymbol{X}_i$ is a $1 \times J$ subject-specific design vector, $\boldsymbol{\beta}$ is a $J \times 1$ parameter vector, β_0 is an intercept, J is the total number of predictors in the model, and N is the number of observations.

GLMs become Bayesian by assigning prior distributions to parameters. Recall that the classical lasso model introduced by Tibshirani (1996) is equivalent to placing double

exponential priors on the β_j (Park and Casella, 2008). Thus, the spike-and-slab lasso prior is incorporated as a mixture of double exponential distributions, where the narrow spike distribution shrinks "unimportant" parameters more severely than the wider slab distribution, which allows "important" parameters to remain nearer to their initial estimates. The explicit formulation of the spike-and-slab lasso is given by (Ročková and George, 2018):

$$\beta_j|\gamma_j, s_0, s_1 \sim (1-\gamma_j)DE(\beta_j|0, s_0) + \gamma_j DE(\beta_j|0, s_1) \tag{2.1.2}$$

where γ_j indicates whether the j^{th} variable is included in the model ($\gamma_j = 1$) or not ($\gamma_j = 0$), and $s_1 > s_0 > 0$; we can further simplify as follows:

$$\beta_j|\gamma_j, s_0, s_1 \sim DE(\beta_j|0, S_j) = \frac{1}{2S_j}\exp\left(-\frac{|\beta_j|}{S_j}\right) \tag{2.1.3}$$

where $S_j = (1-\gamma_j)s_0 + \gamma_j s_1$.

2.2 The EM-Coordinate Descent Algorithm

Bayesian and Classical models traditionally differ in that Bayesian models provide inference about posterior distributions for parameters that are inherently variable, whereas Classical models provide point estimates for theoretical fixed parameters. However, for very large numbers of predictors even relatively fast MCMC draws from the posterior distribution can impose prohibitive costs. Optimization approaches resulting solely in parameter estimates may be preferred in some practical settings, especially when the goal of the analysis is prediction and/or variable selection and not the posterior distribution. It is possible to fit GLM's with lasso, ridge, or elastic net penalties using the coordinate descent algorithm, which is much faster than MCMC for large dimensional data (Zou and Hastie, 2005; Friedman et al., 2007, 2010).

While spike-and-slab priors have an intuitive interpretation, various implementations and practical settings can require novel approaches to model fitting, and the Expectation

Maximization (EM) algorithm is a common model fitting paradigm; e.g., see Ročková and George (2014). Tang et al. (2017) use an EM algorithm to fit the spike-and-slab lasso by treating the model inclusion indicators γ_j as missing parameters. The log posterior density is given by:

$$\begin{aligned}\log p(\boldsymbol{\beta}, \phi, \boldsymbol{\gamma}, \theta | \boldsymbol{y}) &= \log p(y|\boldsymbol{\beta}, \phi) + \sum_{j=1}^{J} \log p(\beta_j | S_j) + \sum_{j=1}^{J} \log p(\gamma_j|\theta) + \log p(\theta) \\ &\propto \ell(\boldsymbol{\beta}, \phi) - \sum_{j=1}^{J} \frac{1}{S_j}|\beta_j| + \sum_{j=1}^{J} (\gamma_j \log\theta + (1-\gamma_j)\log(1-\theta)) \quad (2.2.1)\end{aligned}$$

where $\ell(\boldsymbol{\beta}, \phi) = \log p(y|\boldsymbol{\beta}, \phi)$. The algorithm proceeds by taking the expectation with respect to the γ_j conditional on the other parameters (E-step), inputs these conditional expecations into equation (2.2.1), and then maximizes over the remaining parameters (M-step), iterating over these steps until convergence. We now give explicit details regarding both the E- and M-step.

E-Step

It can be shown that the expectation of the log joint posterior density with respect to the conditional distributions of the γ_j is as follows (Tang et al., 2017):

$$\begin{aligned}p_j &= p(\gamma_j = 1 | \beta_j, \theta, \boldsymbol{y}) \\ &= \frac{p(\beta_j|\gamma_j = 1, s_1)p(\gamma_j = 1|\theta)}{p(\beta_j|\gamma_j = 0, s_0)p(\gamma_j = 0|\theta) + p(\beta_j|\gamma_j = 1, s_1)p(\gamma_j = 1|\theta)} \quad (2.2.2)\end{aligned}$$

It follows that the conditional posterior expectation of the j^{th} scale/penalty parameter S_j^{-1} is as follows:

$$\begin{aligned}E(S_j^{-1}|\beta_j) &= E\left(\frac{1}{(1-\gamma)s_0 + \gamma_j s_1}\middle|\beta_j\right) \\ &= \frac{1-p_j}{s_0} + \frac{p_j}{s_1} \quad (2.2.3)\end{aligned}$$

It is important to note that no matter the form of the p_j, once their values are obtained the conditional expectation of the S_j^{-1} immediately follows.

M-step

Equation (2.2.1) shows that $(\boldsymbol{\beta}, \phi)$ and θ are never within the same term simultaneously, and so can be updated separately by maximizing each the following expressions:

$$Q_1(\boldsymbol{\beta}, \phi) = \ell(\boldsymbol{\beta}, \phi) - \sum_{j=1}^{J} \frac{1}{S_j} |\beta_j| \tag{2.2.4}$$

$$Q_2(\theta) = \sum_{j=1}^{J} (\gamma_j \log \theta + (1 - \gamma_j) \log(1 - \theta)) \tag{2.2.5}$$

$Q_1(\boldsymbol{\beta})$ can be updated via the cyclic coordinate descent algorithm, because $Q_1(\boldsymbol{\beta})$ is equivalent to the lasso penalty with γ_j and S_j^{-1} traded for their conditional posterior distributions; e.g., by using the `R` package `glmnet`. By simple calculus, we update θ with $\theta = \frac{1}{J} \sum_{j=1}^{J} p_j$.

2.3 Extending the EMCD Algorithm to Incorporate Spatial Information

While the approach produced by Tang et al. (2017) can handle both ill-posed data and situations where predictors are highly correlated, it does not model the structure of correlation among parameter estimates and only allows for lasso penalties. While this approach may be appropriate when no assumptions can be made about the structure of correlation, spatial data are amenable to spatially structured priors because in such cases it is reasonable to expect that parameter values will exhibit spatially clustering. In addition, the expected correlation among predictors in spatial settings may make the lasso undesirable since it tends to pick one predictor and ignore the rest. In what follows we extend the spike-and-slab lasso GLM to address both of these issues.

3 The EMCD-IAR Model

3.1 The Spike-and-Slab Elastic Net

The elastic net penalty has a Bayesian interpretation as a mixture of normal and double exponential distributions (Zou and Hastie, 2005):

$$p(\beta_j|\lambda) \propto \exp\left[-\lambda\{(1-\xi)\beta_j^2 + \xi|\beta_j|\}\right] \tag{3.1.1}$$

The elastic net is easily extended to a spike-and-slab framework:

$$p(\beta_j|\gamma_j, s_0, s_1) = EN(\beta_j|0, S_j) \propto \exp\left[-\frac{1}{S_j}\{(1-\xi)\beta_j^2 + \xi|\beta_j|\}\right] \tag{3.1.2}$$

where $S_j = (1-\gamma_j)s_0 + \gamma_j s_1$ and $\xi \in [0,1]$. $\xi = 0$ would correspond to a "spike-and-slab ridge" penalty, while $\xi = 1$ produces the spike-and-slab lasso described by equation (2.1.3).

3.2 CAR and IAR Spatial Models for Inclusion Probabilities

The most direct approach to incorporating spatial structure into parameter estimates is to impose a correlation structure within the prior distributions for the β_j. However, this approach ultimately requires inverting a very large covariance matrix, which is computationally expensive and may be impractical for large sets of spatial predictors. In the context of Bayesian variable selection, a more subtle approach is to impose dependence upon the prior inclusion probabilities, which indirectly imposes spatial dependence on the estimates of β_j through the penalty $E(S_j^{-1}|\beta_j)$, since the conditional expectation in equation (2.2.3) is a function of $p_j = p(\gamma_j = 1|\beta_j, \theta_j, \boldsymbol{y})$, which is a function of the estimate for θ_j.

A common framework for modeling spatial structure is a special case of Gaussian Markov Random Fields (GMRF) known as conditional autoregressions (CAR); CAR models are multivariate Normal distributions specified by a conditional structure and have been used extensively in spatial modeling, and more relevant to our purposes, to impose spatial structure

via prior distributions (Banerjee et al., 2015; Brown et al., 2014; Cressie and Wikle, 2011; Rue and Held, 2005). The prior probabilities of inclusion, θ_j, must be transformed in order to use GMRF's to model their spatial dependence, since the multivariate Normal distribution has support on all real numbers, while probabilities are bounded between zero and one. The logit of the probabilities of inclusion, $\theta_j \in [0,1]$, $\psi_j = \text{logit}(\theta_j) = \log \frac{\theta_j}{1-\theta_j} \in (-\infty, \infty)$, has the same support as the (multivariate) Normal distribution. Thus, after modeling the ψ_j, we can obtain $\theta_j = \text{logit}^{-1}(\psi_j)$.

We impose conditional prior distributions of the ψ_j using the following CAR parameterization (Jin et al., 2005; Banerjee et al., 2015):

$$p(\psi_j|\psi_i, \tau) = \mathcal{N}\left(\alpha \sum_{j:j\sim i} \frac{\psi_i}{n_j}, (\tau^2 n_j)^{-1}\right) \tag{3.2.1}$$

where n_j is the number of neighbors for location j and τ is a common precision parameter so that the precision $\tau^2 n_j$ only varies based on the number of neighbors. The precision is the inverse of the variance, and is useful for two reasons. The first is that entries in the precision matrix describe conditional dependence structures, so that zero entries imply conditional independence between variables. The second reason follows in part from the first: the precision matrix is often sparse for CAR models because most common usages assume the vast majority variables are conditionally independent, which suggests computational benefits in many cases (Rue, 2001; Rue and Held, 2005). The parameter α is called a "propriety parameter", because it controls the degree of dependence in the ψ_j, where $\alpha = 0$ implies independence and $\alpha \to 1$ produces increasingly stronger dependence (Besag and Kooperberg, 1995; Banerjee et al., 2015). The joint distribution of this model then simplifies to the following (Jin et al., 2005; Banerjee et al., 2015):

$$\boldsymbol{\psi} = \mathcal{N}\left(\mathbf{0}, [\tau^2(\boldsymbol{D} - \alpha \boldsymbol{W})]^{-1}\right) \tag{3.2.2}$$

where $\boldsymbol{D} = \text{diag}(n_j)$ contains the number of neighbors for each location and $\boldsymbol{W} = \{w_{ij}\}$ is the adjacency matrix where $w_{ij} = \begin{cases} 1 & j \sim i \\ 0 & \text{otherwise} \end{cases}$.

Setting $\alpha = 1$ results in an Intrinsic Autoregressive model (IAR) (Banerjee et al., 2015; Besag and Kooperberg, 1995). The IAR does not have a proper joint distribution and one consequence is that the random variables do not vary about a global constant mean, but rather each is interpreted as varying about the mean of its neighbors.[71] This interpretive quality is one benefit of the IAR, but the IAR is often required for modeling stronger correlation among random variables, since values exceedingly close to $\alpha = 1$ are required to model correlations near 1 for spatially adjacent variables. While the IAR cannot be used as a data generating model, it has been especially useful as a prior distribution in spatial models (Besag and Kooperberg, 1995; Banerjee et al., 2015; Rue and Held, 2005). The IAR model is often formulated as a pairwise difference, which has both interpretive and computational benefits (Besag and Kooperberg, 1995; Morris et al., 2019a):

$$\log p(\boldsymbol{\psi}) \propto \frac{-\tau^2}{2}\left(\sum_{j:j<i}(\psi_j - \psi_i)^2\right) \tag{3.2.3}$$

3.3 Derivation of the Log Joint Posterior Distribution

The EM algorithm proceeds by taking the expectation of the "missing" data/parameters, replacing those parameters with their conditional expectation in the joint log likelihood, and then maximizing that likelihood to obtain estimates for the parameters of interest. Thus, we first obtain the joint posterior distribution:

$$p(\boldsymbol{\beta}, \phi, \boldsymbol{\gamma}, \boldsymbol{\psi}, \tau, \alpha | \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\beta}, \phi)\prod_{j=1}^{J} p(\beta_j|S_j)\prod_{j=1}^{J} p(\gamma_j|\theta_j)p(\boldsymbol{\psi}|\tau^{-1}, \alpha)p(\alpha)p(\tau^{-1}) \tag{3.3.1}$$

[71] Strictly speaking, the mean and covariance do not exist, but the precision matrix does exist. See Banerjee et al. (2015) for more details.

The log joint posterior will therefore look like equation (2.2.1), but with extra terms to account for the spatial dependence among inclusion probabilities. If we set $\alpha = 1$ to obtain the IAR model, then obviously the terms containing α drop out. In addition, while we might prefer to model τ directly, it is not always computationally feasible, and in practice it often suffices to set $\tau = 1$, e.g., as in Morris et al. (2019a). Adding spatial structure via the IAR prior on the prior inclusion probabilities and incorporating the elastic net prior for the β_j alters the joint posterior as follows:

$$\begin{aligned}\log p(\boldsymbol{\beta}, \phi, \boldsymbol{\gamma}, \boldsymbol{\psi}, \tau, \alpha) \propto & \underbrace{\ell(\boldsymbol{\beta}, \phi)}_{\text{log likelihood}} \underbrace{- \sum_{j=1}^{J} \frac{1}{S_j}\{(1-\xi)\beta_j^2 + \xi|\beta_j|\}}_{\text{log prior for } \boldsymbol{\beta}} \\ & \underbrace{+ \sum_{j=1}^{J} \gamma_j \log\theta_j + (1-\gamma_j)\log(1-\theta_j)}_{\text{log prior for } \boldsymbol{\gamma}} \\ & \underbrace{- \frac{1}{2}\left(\sum_{j:j<i} (\psi_j - \psi_i)^2\right)}_{\text{log prior for } \psi_j = \text{logit}(\theta_j)}\end{aligned} \tag{3.3.2}$$

We now show what changes must be made to the EM-algorithm to obtain parameter estimates when using an IAR prior on the probabilities of inclusion.

3.4 The Structure of the EM Algorithm

E-Step

Since we are treating the γ_j as missing parameters, we must take the conditional expectation of the γ_j given the other parameters in the model. Similar to Tang et al. (2017), by application of Bayes' Rule the conditional probability that a variable should be included the model is as follows:

$$\begin{aligned} p_j &= p(\gamma_j = 1 | \beta_j, \theta_j, \boldsymbol{y}) \\ &= \frac{p(\beta_j|\gamma_j = 1, s_1)p(\gamma_j = 1|\theta_j)}{p(\beta_j|\gamma_j = 1, s_1)p(\gamma_j = 1|\theta_j) + p(\beta_j|\gamma_j = 0, s_0)p(\gamma_j = 0|\theta_j)} \end{aligned} \tag{3.4.1}$$

where $p(\gamma_j = 1|\theta_j) = \theta_j$, $p(\gamma_j = 0|\theta_j) = 1 - \theta_j$, $p(\beta_j|\gamma_j = 1, s_1) = EN(\beta_j|0, s_1)$, and $p(\beta_j|\gamma_j = 1, s_1) = EN(\beta_j|0, s_0)$. Given p_j, the conditional posterior expectation of S_j^{-1} is the same as in Tang et al. (2017):

$$\begin{aligned} E(S_j^{-1}|\beta_j) &= E\left(\frac{1}{(1-\gamma_j)s_0 + \gamma_j s_1}\right) \\ &= \frac{1-p_j}{s_0} + \frac{p_j}{s_1} \end{aligned} \tag{3.4.2}$$

Therefore, the E-step differs from Tang et al. (2017) primarily in that rather than a single prior probability of model inclusion, θ, there are now J prior probabilities of inclusion, θ_j. The M-step then progresses by maximizing equation (3.3.2) with γ_j and S_j^{-1} exchanged for their conditional expectations.

M-Step

Once we have obtained the conditional expectation of the missing indicators for model inclusion, we can plug these into the joint log posterior distribution and maximize the expression. As before, we can divide the maximization into convenient pieces. Regardless of the spatial model, the first piece is still $Q_1(\boldsymbol{\beta}, \phi)$, similar to equation (2.2.4), except that now the full range of elastic net priors specified by $\xi \in [0, 1]$ is allowed; $Q_1(\boldsymbol{\beta}, \phi)$ can thus be maximized via cyclic coordinate descent as before, e.g., by using the `R` package `glmnet`. However, now that the θ_j have spatial dependence, Q_2 must now also update the logit of prior inclusion probabilities $\boldsymbol{\psi}$, and incorporate the spatial structure of the IAR model, respectively:

$$Q_{1,EN} = \underbrace{\ell(\boldsymbol{\beta}, \phi)}_{\text{log likelihood}} \underbrace{- \sum_{j=1}^{J} \frac{1}{S_j} \{(1-\xi)\beta_j^2 + \xi|\beta_j|\}}_{\text{log prior for } \boldsymbol{\beta}} \tag{3.4.3}$$

$$Q_{2,IAR} = \underbrace{\sum_{j=1}^{J} \gamma_j \log \theta_j + (1-\gamma_j)\log(1-\theta_j)}_{\text{log prior for } \boldsymbol{\gamma}} - \underbrace{\frac{1}{2}\left(\sum_{j:j<i} (\psi_j - \psi_i)^2\right)}_{\text{log prior for } \psi_j = \text{logit}(\theta_j)} \tag{3.4.4}$$

We replace the γ_j with their conditional expectations and maximize to find parameter estimates for $(\boldsymbol{\psi}, \tau)$, and adapt the `stan` model from Morris (2017) and Morris et al. (2019a) to perform the maximization of $Q_{2,IAR}$ using the `R` package `rstan`. We then iterate the E- and M-steps until convergence, and follow Tang et al. (2017) to assess convergence by the criterion:

$$\frac{|d^{(t)} - d^{(t-1)}|}{(0.1 + |d^{(t)}|)} < \epsilon \tag{3.4.5}$$

where $d^{(t)} = -2 \log \ell(\boldsymbol{\beta}^{(t)}, \phi^{(t)})$ is the estimated deviance at iteration t. A development version of an `R` package to fit these models is available on github.

4 Simulations

4.1 Simulation Framework

We performed simulations in a limited setting to demonstrate the methodology and show its potential for improved prediction. The simulations consist of 5,000 data sets containing $N = \{25, 50, 100\}$ subjects.[72] Each subject design vector arises from a 32×32

[72] These sample sizes were chosen due to our interest in prediction performance in smaller samples. The monte carlo standard errors (MCSE) for AUC, MSE, MAE, misclassification, and proportion of true parameters selected were under 0.001 in nearly all cases. The MCSE for FDR were all under 0.01. We did not choose the simulation sizes to optimize MCSE below a set value, but comment here on their estimates to give interested readers an idea of the variability of the simulation study results inherent to taking simulation samples. See,

two dimensional image generated from a multivariate Normal distribution with zero mean and unit variance; a correlation structure was imposed such that the correlation between two variables was given by:

$$\text{corr}(X_{ij}, X_{ik}) = 0.90^{d(X_{ij}, X_{ik})}, \quad j, k = 1, \ldots, J;\ i = 1, \ldots, N$$

where $d(X_{ij}, X_{ik})$ is the Euclidean distance between X_{ij} and X_{ik}. An example of a subject image is shown in Figure 1. When predictors exhibit spatial correlation it is often reasonable to expect "important" parameters to exhibit spatial clustering. We therefore generated parameter vectors clustered in two-dimensional space before vectorization, as in Figure 2.[73] We simulated binary outcomes for each of $i = 1, \ldots, N$ subjects from a $Bernoulli(\boldsymbol{X}_i\boldsymbol{\beta})$ distribution for two scenarios: $\beta_j = 0.5$ and $\beta_j = 0.1$. Thus, the $\beta_j = 0.5$ and $\beta_j = 0.1$ data sets correspond to a 64.87% and 10.51% increase in odds of an "event" occurring, respectively. The data were generated using the `R` package `sim2Dpredictr`, which is available on CRAN.

We analyze each dataset with both elastic net ($\xi = 0.5$; a halfway compromise between ridge and lasso) and lasso ($\xi = 1$) priors under the traditional framework, the spike-and-slab framework without spatial structure, and the spike-and-slab framework with spatial structure, using a combination of the `R` packages `glmnet`, `BhGLM`, and `ssnet`. We employ 10-fold cross validation for $N = \{50, 100\}$ and 5-fold cross validation for $N = 25$ to estimate measures of model fit/variable selection criteria. For the traditional elastic net models we allow `cv.glmnet()` to internally select the optimal parameter, and for the spike-and-slab lasso models we set the slab scale parameter to $s_1 = 1$ and manually choose the sequence $s_0 = \{0.01, 0.02, 0.03, \ldots, 0.3\}$ over which to choose the value of spike parameter that minimizes cross-validated prediction error.

e.g., Koehler et al. (2009) or Morris et al. (2019b).

[73] Note that in principle Figure 2 does not follow from Figure 1. For example, we can conceive of various sizes and shapes for the parameter cluster(s). It also is possible to conceive of theoretical situations where measurements were uncorrelated in space, but parameter values were spatially clustered, or vice versa.

We report several metrics to evaluate two aspects of model performance, prediction and variable selection. Prediction accuracy is assessed with cross-validated measures of deviance, mean square error (MSE), mean absolute error (MAE), area under the ROC curve (AUC), and misclassification (MC).[74] False discovery rate (FDR) and Power are used to evaluate variable selection performance. FDR is the average proportion of parameters true value was zero out of all those whose estimates were non-zero. We define Power here as the average proportion of true non-zero parameters discovered.[75] Well performing models will have lower values for deviance, MSE, MAE, MC, and FDR, and will have higher values for AUC and Power. In all cases we choose the model whose penalty minimizes the cross-validated Deviance. Note that Deviance in this case is defined as -2 times the log likelihood, but with respect to the held-out data, not the training data; i.e., it is an estimate of model fitness to independent data, rather than observed data, which can help prevent overfitting models that generalize poorly.

4.2 Simulation Results

As seen in Tables 1 and 2, over the 5,000 simulated data sets for each sample size the models with spatially structured priors have the best cross-validated prediction error compared to models without spatially structured priors, and that this is the case for both effect sizes whether we use elastic net with $\xi = 0.5$ or the lasso. While the spike-and-slab elastic net (SSEN) with IAR priors tends to outperform the spike-and-slab lasso (SSL) with IAR priors, this difference is less noticeable than that exhibited by including the spatial structure versus not; i.e., most of the improvement in prediction error appears to be the result of the IAR priors as opposed to compromising between ridge and lasso priors/penalties. The SSEN with IAR priors consistently has the lowest deviance, mean squared error (MSE), and

[74] Misclassification is defined as $\frac{1}{n}\sum_{i=1}^{n} I(|y_i - \hat{y}_i| > 0.5)$ where $I(\cdot)$ is an indicator function whose value is 1 when the argument is true, and zero otherwise.

[75] Usually Power is the average proportion of the time that a model would identify a parameter as non-zero, given that it really was non-zero. Here we simply generalize to a set of non-zero parameters we want to discover. We use "Power" for ease of interpretation.

misclassification, as well as the highest AUC; when $\beta_j = 0.1$, it also has the lowest mean absolute error (MAE), but the SSL with IAR priors has the lowest MAE when $\beta_j = 0.5$.

With respect to variable selection, the traditional elastic net captures the highest proportion of true non-zero parameters for all scenarios, and for any given scenario the elastic net captures a higher proportion of true non-zero parameters compared to the lasso. However, in most scenarios examined, the traditional elastic net had the highest estimated FDR. In general, it appears to be the case as well that including the IAR priors compromises between the traditional elastic net/lasso and the spike-and-slab elastic net/lasso in that it tends to have FDR and proportion of non-zero parameters discovered in between the other two frameworks. However, in general it was the case for all models considered that FDR and proportion of true non-zero parameters is far from what most researchers would consider optimal.

Table 1: Model Performance for $\beta_j = 0.5$

N	Model	s_0	Dev.†	AUC	MSE	MAE	MC‡	FDR	Power
25	Lasso	0.0487	18.94	0.8926	0.1225	0.2401	0.1805	0.6786	0.0990
	EN	0.0646	17.68	0.9049	0.1143	0.2290	0.1671	0.7162	0.4293
	SSL	0.1575	11.03	0.9751	0.0609	0.1701	0.0700	0.2422	0.0293
	SSEN	0.2032	15.33	0.9596	0.0883	0.2370	0.1012	0.5122	0.2192
	SSL (IAR)	0.1968	9.37	0.9882	0.0483	0.1488	0.0485	0.3033	0.0301
	SSEN (IAR)	0.1375	9.31	0.9939	0.0448	0.1538	0.0347	0.4109	0.0405
50	Lasso	0.0232	24.93	0.9563	0.0788	0.1667	0.1144	0.6532	0.1756
	EN	0.0357	25.58	0.9564	0.0804	0.1709	0.1146	0.6926	0.5744
	SSL	0.1464	18.99	0.9791	0.0547	0.1379	0.0698	0.2265	0.0452
	SSEN	0.1310	21.24	0.9824	0.0585	0.1649	0.0663	0.3875	0.1512
	SSL (IAR)	0.2004	13.50	0.9929	0.0351	0.1054	0.0385	0.3059	0.0535
	SSEN (IAR)	0.1387	13.41	0.9959	0.0324	0.1096	0.0290	0.3733	0.0729
100	Lasso	0.0149	39.24	0.9750	0.0611	0.1316	0.0869	0.6402	0.2657
	EN	0.0230	40.11	0.9749	0.0621	0.1356	0.0876	0.6863	0.6877
	SSL	0.1267	29.52	0.9874	0.0429	0.1051	0.0564	0.2053	0.0718
	SSEN	0.1004	31.57	0.9888	0.0440	0.1195	0.0539	0.3282	0.1458
	SSL (IAR)	0.1927	21.94	0.9945	0.0297	0.0832	0.0353	0.3235	0.0842
	SSEN (IAR)	0.1286	21.34	0.9963	0.0271	0.0847	0.0286	0.3773	0.1135

†Deviance
‡Misclassification

Table 2: Model Performance for $\beta_j = 0.1$

N	Model	s_0	Dev.†	AUC	MSE	MAE	MC‡	FDR	Power
25	Lasso	0.1248	28.11	0.7602	0.1918	0.3639	0.3038	0.7039	0.0543
	EN	0.1956	27.20	0.7698	0.1854	0.3554	0.2923	0.7314	0.2161
	SSL	0.1858	20.50	0.8560	0.1329	0.2932	0.1973	0.5125	0.0292
	SSEN	0.2116	22.80	0.8452	0.1484	0.3260	0.2162	0.7041	0.1653
	SSL (IAR)	0.1966	14.55	0.9473	0.0863	0.2117	0.1089	0.5612	0.0206
	SSEN (IAR)	0.1283	13.73	0.9618	0.0786	0.2076	0.0928	0.6480	0.0242
50	Lasso	0.0799	48.35	0.8318	0.1607	0.3253	0.2407	0.6951	0.1003
	EN	0.1394	49.35	0.8270	0.1641	0.3306	0.2452	0.7023	0.3048
	SSL	0.1189	38.26	0.8999	0.1216	0.2561	0.1716	0.3338	0.0313
	SSEN	0.1039	38.80	0.9005	0.1231	0.2698	0.1731	0.5569	0.0834
	SSL (IAR)	0.1796	28.72	0.9453	0.0878	0.1945	0.1191	0.6071	0.0337
	SSEN (IAR)	0.1198	25.63	0.9608	0.0762	0.1820	0.0993	0.6791	0.0422
100	Lasso	0.0592	91.22	0.8611	0.1493	0.3087	0.2180	0.6490	0.1557
	EN	0.1072	92.17	0.8593	0.1509	0.3129	0.2203	0.6522	0.3890
	SSL	0.0912	77.79	0.9003	0.1243	0.2542	0.1777	0.3285	0.0497
	SSEN	0.0722	76.97	0.9029	0.1229	0.2553	0.1748	0.5147	0.0756
	SSL (IAR)	0.1501	62.06	0.9368	0.0969	0.1982	0.1361	0.6499	0.0504
	SSEN (IAR)	0.1050	52.56	0.9563	0.0805	0.1743	0.1107	0.7509	0.0617

†Deviance
‡Misclassification

It is reasonable to wonder why the cross validated prediction error is in such contrast to the traditional variable selection measures. Some insight is gained by focusing on the parameter estimates themselves. Consider figure 3, which displays average parameter estimates by sample size and model when true non-zero parameters are equal to 0.5. In general, all the models tend to prioritize one parameter over the others, as evidenced by the peaks in each plot, and often there are several clusters beneath this peak. As advertised, the spike-and-slab framework shrinks parameters deemed "important", i.e., true non-zero parameters in the simulation setting, are noticeably less than the the traditional frameworks, and the peaks for the traditional methods are all obviously lower compared to the spike-and-slab models, which peak closer to the true value for non-zero parameters.

The spike-and-slab models without the IAR prior on inclusion probabilities tend to have the highest peaks, so why do the models with spatial structure tend to have better fits?

One explanation is that the models with spatially structured priors tend to keep more true non-zero parameters closer to the peak, even if that peak is slightly lower as a result. Notice also that for each analysis there are several clusters of non-zero variables. The models with spatially structured priors tend to have more concentrated clustering, and aside from the peak, the center, i.e. the mean value, of these clusters are slightly above the center of the clusters for the models without spatially structured priors, which indicates that on average those models are providing estimates closer to the true values for non-zero parameters. While the results are somewhat more complicated for the $\beta_j = 0.1$, the general trends are similar. For $N = 25$, the SSL, SSL (IAR), and SSEN (IAR) models are peaking near 0.1. At $N = 50$, these 3 models plus SSEN peak near 0.15, while the traditional lasso and EN are closer to 0.075. By $N = 100$, the traditional models are peaking near 0.1, while the spike-and-slab models peak closer to 0.2. Thus, for the smaller effect size, the models are over-shooting the effect size considerably, at least for the largest parameters.

However, more insight is gained by considering summaries for estimates of zero and non-zero parameters, as shown in tables 3 and 4 and figures 5 and 6, which show the mean and standard deviations for estimates of non-zero and zero parameters, respectively. Not surprisingly, the average parameter estimates increase as the sample size increases. With respect to non-zero parameters, we can see that in fact adding spatially structured priors to the model increases the average estimate for every sample size and for both the elastic net and lasso priors. This is an interesting contrast to the proportion of true non-zero parameters captured, where the traditional methods appeared to have the best performance. With respect to the true zeroes, as was often the case with FDR, the models with spatial structure were not as low as the spike-and-slab models without spatial structure, but they were lower than the traditional models. This sheds light on how the spatial structure is leading to improved prediction error - presumably the spike-and-slab models are providing closer to the "true" non-zero parameter values. The average parameter values for the true zero parameters are also fairly close to zero, and so even when they are included, i.e., and leading to higher FDR,

they are small enough compared to the estimates for true non-zero parameters that they do not introduce as much noise to the prediction; this is another benefit of the spike-and-slab framework, where again as advertised the spike prior shrinks "unimportant" parameters more than "important" parameters, at least on average.

However, the mean parameter estimates do not necessarily show how the elastic with spatial structure is outperforming the spike-and-slab lasso with spatial structure, especially when the true non-zero parameters are equal to 0.5. This might be explained by considering the variation in these estimates. Compared to the average estimates, their standard deviations are relatively large, often larger than the average estimate. Narrowing in on the comparison betwen the SSL (IAR) and SSEN (IAR) models, we see that the variation about the former is quite a bit larger than the latter. It appears that a bias-variance tradeoff may explain the improved performance for larger effect size, especially for the larger effect sizes, and that the lower variance in general contributes to better prediction.

5 Discussion

We have presented a novel approach to using the spike-and-slab prior on the elastic net when predictors exhibit spatial structure. The elastic net can be preferred to the lasso in circumstances where the number of predictors far exceeds the sample size and when the predictors exhibit strong correlations and since both the lasso and elastic net are expressible in a Bayesian framework they are reasonably amenable to a spike-and-slab prior framework, e.g., as explored in Ročková and George (2014) and Ročková and George (2018). Our primary contribution has been to show how to further incorporate spatial information into the model fitting process by placing intrinsic autoregressive priors on the logit of the probabilities of inclusion and to fit this model for GLMs by extending the EM algorithm presented in Tang et al. (2017).

Furthermore, through a simulation study, we have explored some of the properties of this model. While this simulation study was limited to only a few effect sizes and binary

Table 3: Average Parameter Estimates

$\beta_j = 0.5$

True Non-Zero Parameters						
N	Lasso	SSL	SSL (IAR)	EN	SSEN	SSEN (IAR)
25	0.07101	0.08100	0.08538	0.06513	0.04920	0.06789
50	0.11791	0.13124	0.14007	0.10597	0.10002	0.11935
100	0.16498	0.19382	0.20738	0.14959	0.16303	0.18240
True Zero Parameters						
25	0.00196	0.00062	0.00091	0.00244	0.00090	0.00102
50	0.00211	0.00049	0.00112	0.00254	0.00067	0.00120
100	0.00209	0.00063	0.00144	0.00243	0.00067	0.00147

$\beta_j = 0.1$

True Non-Zero Parameters						
N	Lasso	SSL	SSL (IAR)	EN	SSEN	SSEN (IAR)
25	0.02617	0.03868	0.04847	0.02483	0.02609	0.04027
50	0.03834	0.06476	0.07274	0.03407	0.05494	0.06820
100	0.04713	0.07711	0.09168	0.04253	0.07501	0.09410
True Zero Parameters						
25	0.00116	0.00084	0.00132	0.00138	0.00103	0.00129
50	0.00108	0.00059	0.00143	0.00117	0.00069	0.00144
100	0.00091	0.00055	0.00130	0.00098	0.00059	0.00142

outcomes, several important lessons can be extracted. First, we have demonstrated the potential for spike-and-slab models with spatially structured priors to improve upon their spatially unstructured counterparts; it is also noteworthy that the spike-and-slab models in general outperformed the traditional models with respect to cross validated prediction error, showing also that this prior framework may be a fruitful one for spatial data. While at least in the settings examined, the FDR and proportion of true non-zero parameters captured is not impressive for any model, and larger sample sizes would be necessary to achieve reasonable results on these metrics, the summaries for the parameter estimates themselves lend insight into why the spike-and-slab models with spatial structure are better fit. That is, in general the spike-and-slab models with spatially structured priors tend to produce estimates closer to the "true" values for "important" parameters and correspondingly shrink "unimportant" parameter estimates more strongly than do the traditional methods, which

Table 4: Standard Deviations for Parameter Estimates

$\beta_j = 0.5$						
True Non-Zero Parameters						
N	Lasso	SSL	SSL (IAR)	EN	SSEN	SSEN (IAR)
25	0.07402	0.15189	0.15015	0.01573	0.08740	0.08898
50	0.10151	0.26354	0.20648	0.02059	0.17889	0.12286
100	0.10313	0.26443	0.21328	0.02181	0.18733	0.13153
True Zero Parameters						
25	0.02578	0.03858	0.03925	0.01058	0.01775	0.03029
50	0.02441	0.03013	0.03847	0.01166	0.02065	0.03151
100	0.02274	0.03275	0.04040	0.01204	0.02432	0.03316

$\beta_j = 0.1$						
True Non-Zero Parameters						
N	Lasso	SSL	SSL (IAR)	EN	SSEN	SSEN (IAR)
25	0.03369	0.07671	0.07713	0.00951	0.04135	0.05482
50	0.03661	0.11648	0.10305	0.01185	0.09836	0.07802
100	0.03840	0.12757	0.11375	0.01430	0.11191	0.09366
True Zero Parameters						
25	0.01665	0.03381	0.03558	0.00669	0.01472	0.02596
50	0.01345	0.02974	0.03126	0.00643	0.02258	0.02478
100	0.01025	0.02204	0.02530	0.00582	0.01995	0.02196

demonstrates the power of combining spike-and-slab models with shrinkage penalties.

While larger sample sizes may be desirable, many if not most real-world scenarios using images do not have very large sample sizes. Thus, it is relevant to probe how methods perform with smaller sample sizes and to understand what we can expect to learn in such circumstances. This approach to modeling thus has many possible future directions and possible applications. Since the model is fit for GLM, it is amenable to the full range of non-linear outcomes ordinarily analyzed with GLM's, and thus applications beyond Gaussian outcomes, or as shown here binary outcomes, is immediately possible without revision to the model. Given the strong performance with respect to cross validated prediction error, the algorithm may also be useful in classification problems and may be extended to handle more complex spatial relationships to better address difficult classification problems.

Appendix: Figures

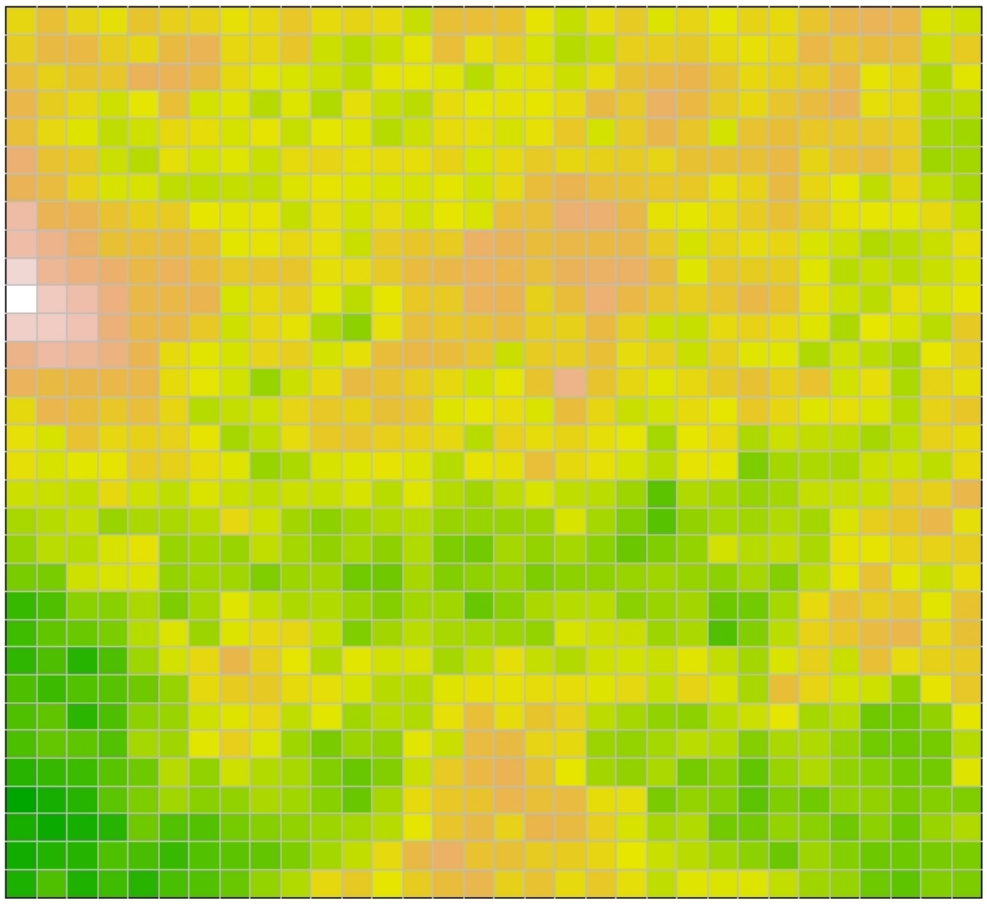

Figure 1: Example of a subject image before being vectorized into a row of the design matrix.

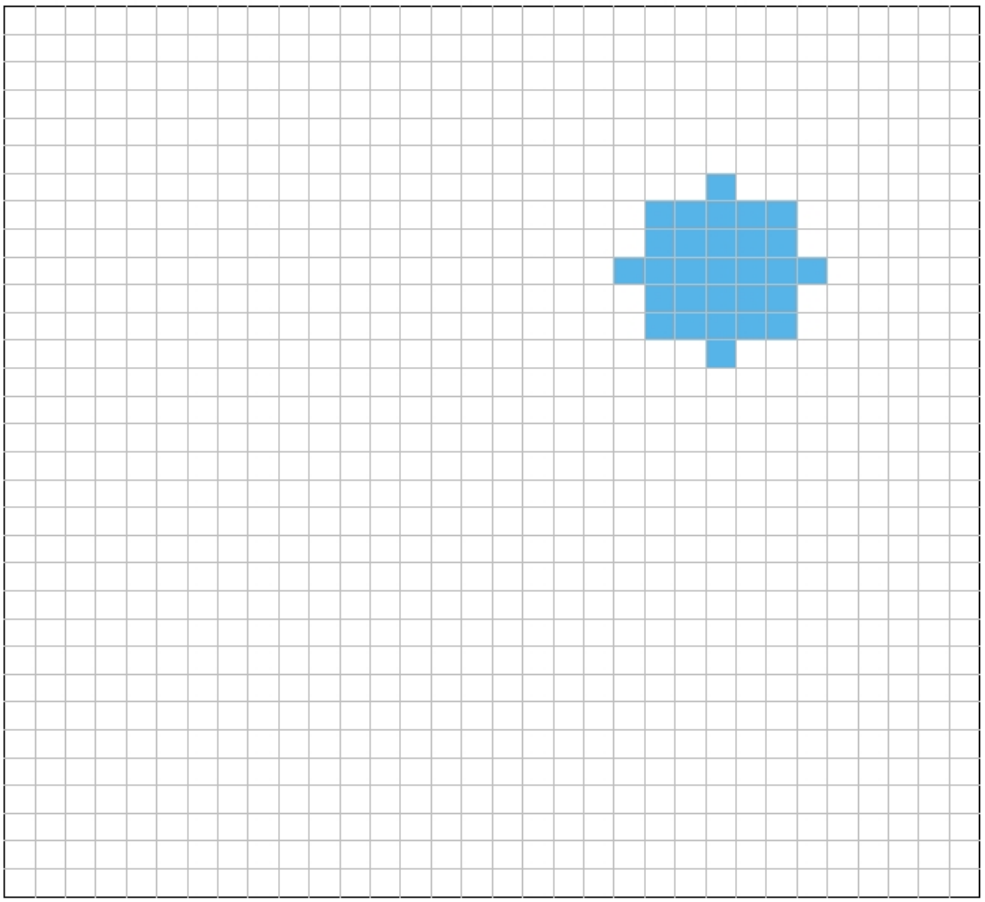

Figure 2: Example of clustered non-zero parameters. The blue pixels represent $\beta_j \neq 0$.

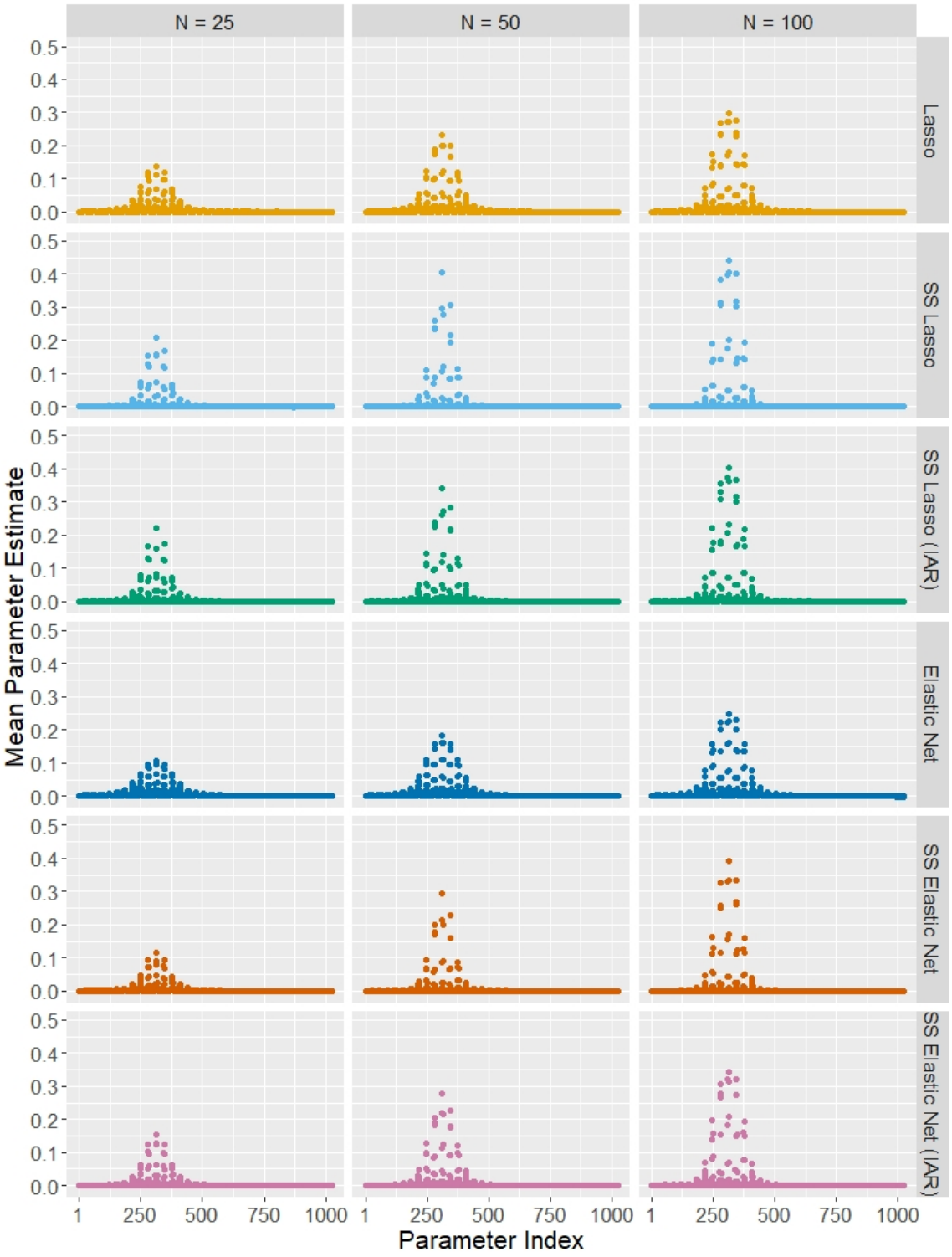

Figure 3: Average parameter estimates for each model when true non-zero $\beta_j = 0.5$. Each dot is an average parameter estimate for a specific parameter.

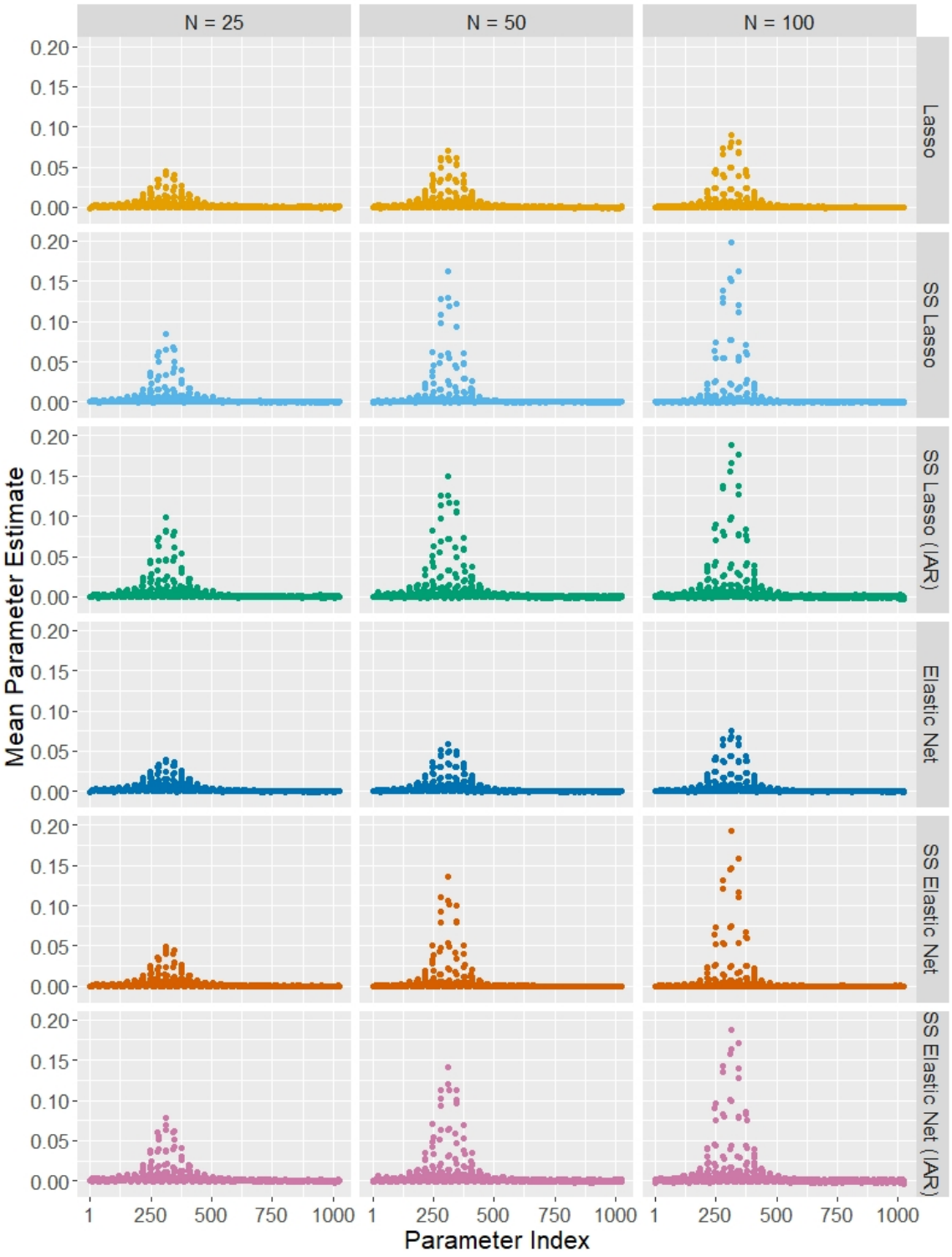

Figure 4: Average parameter estimates for each model when true non-zero $\beta_j = 0.1$. Each dot is an average parameter estimate for a specific parameter.

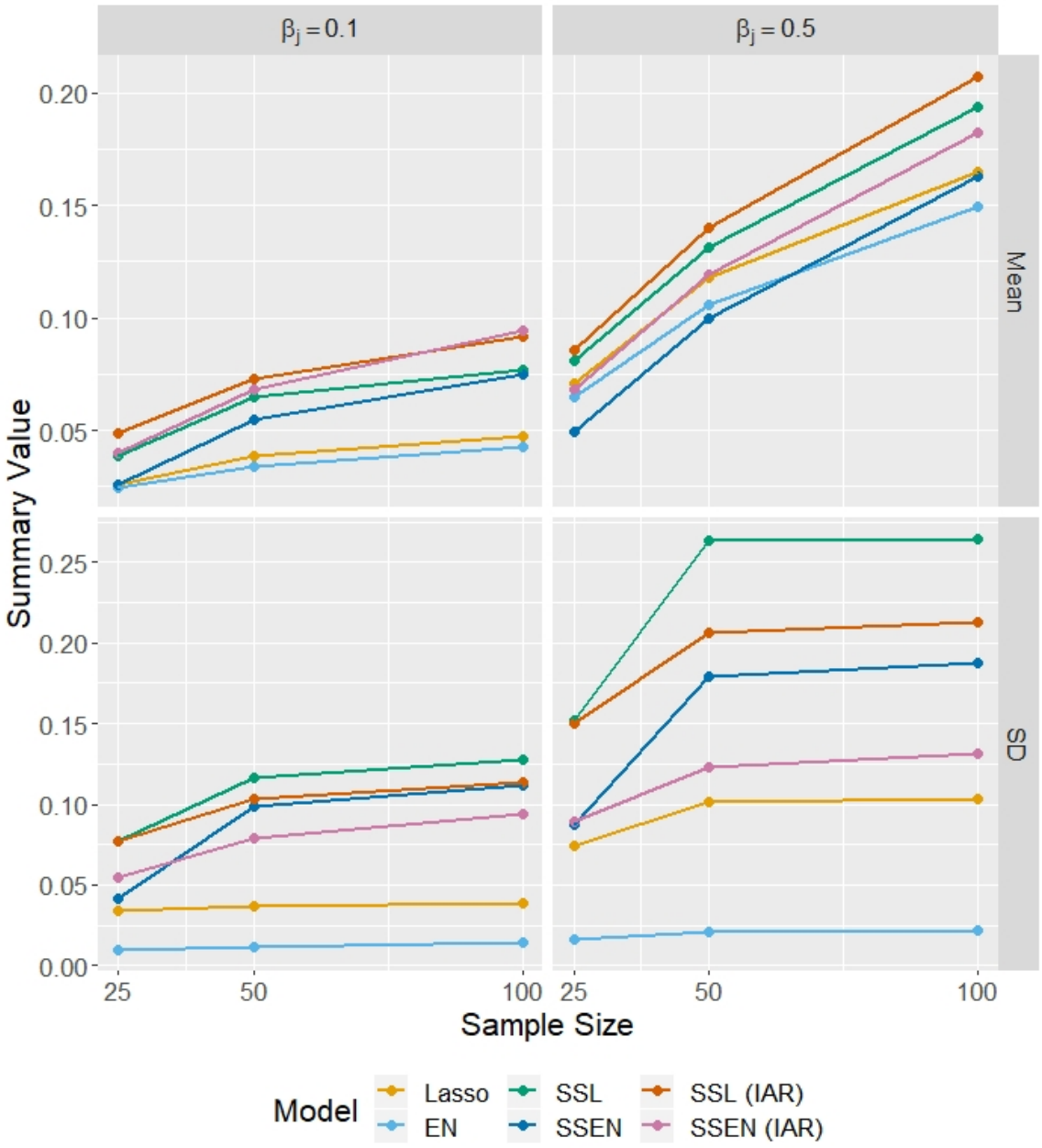

Figure 5: Collective average estimates for true non-zero parameters. Each dot is the summary measure of all estimates for true non-zero parameters.

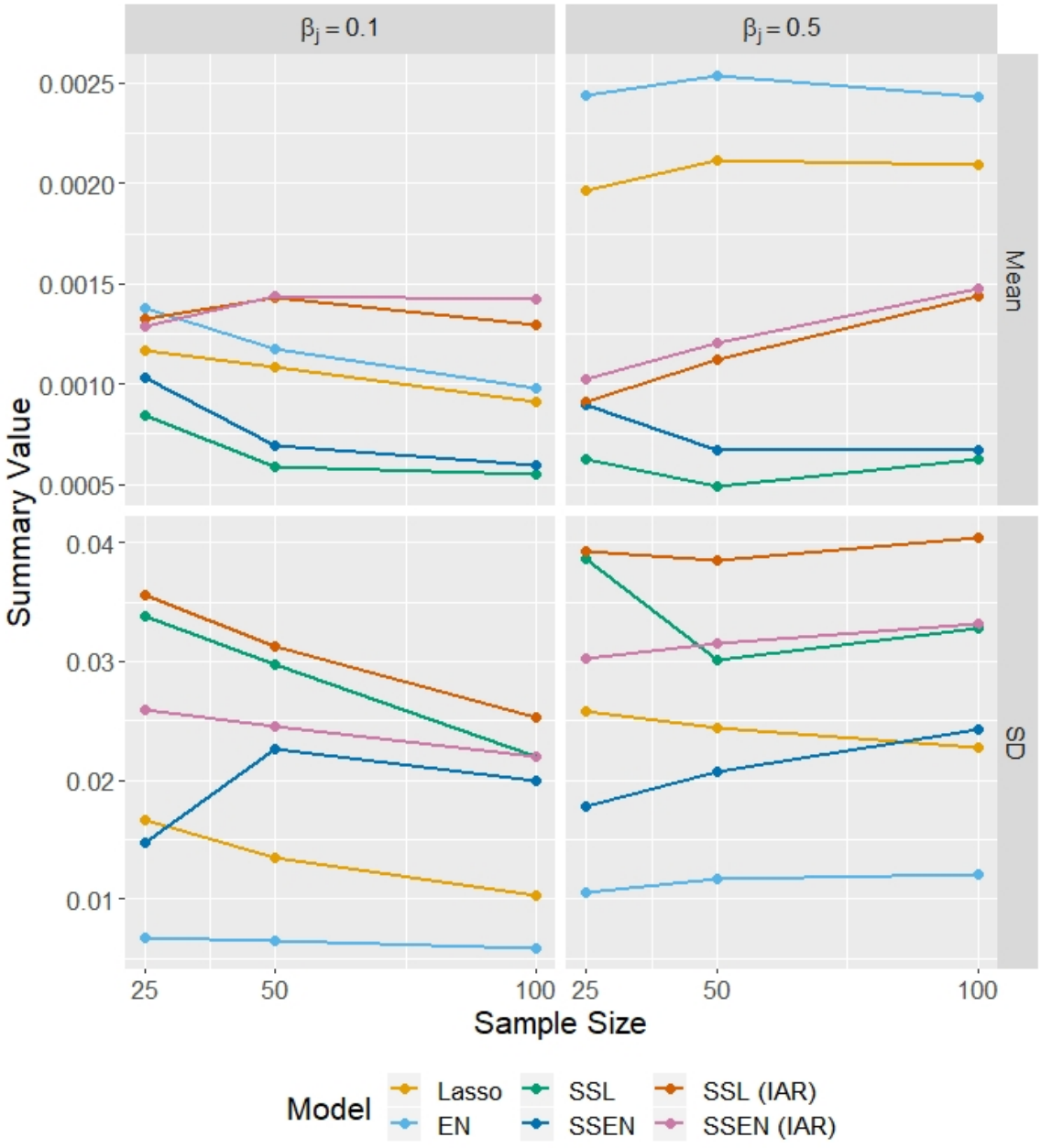

Figure 6: Collective average estimates for true zero parameters. Each dot is the summary measure of all estimates for true zero parameters. The β_j labels in this case correspond to the simulation scenario, but the true value for the parameters estimated here is zero.

REFERENCES

Banerjee, Sudipto; Carlin, Bradley P., and Gelfand, Alan E. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, Florida, 2nd edition, 2015.

Besag, Julian. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

Besag, Julian and Kooperberg, Charles. On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746, 1995. doi: 10.1093/biomet/82.4.733.

Brown, D. Andrew; Lazar, Nicole A.; Datta, Gauri S.; Jang, Woncheol, and McDowell, Jennifer. Incorporating spatial dependence into bayesian multiple testing of statistical parametric maps in functional neuroimaging. *NeuroImage*, 84:97–112, 2014. doi: 10.1016/j.neuroimage.2013.08.024.

Cressie, Noel and Wikle, Christopher K. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, Hoboken, New Jersey, 2011.

Dempster, A.P.; Laird, N.M., and Rubin, D.B. Maximum likelihood from incomplete data via the em algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B*, 39:1–38, 1977.

Friedman, J.; Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010. doi: 10.18637/jss.v033.i01.

Friedman, Jerome; Hastie, Trevor; Höfling, Holger, and Tibshirani, Robert. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007. doi: 10.1214/07-AOAS131.

George, Edward I. and McCulloch, Robert E. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993. doi: 10.1080/01621459.1993.10476353.

George, Edward I. and McCulloch, Robert E. Approaches for bayesian variable selection. *Statistica Sinica*, 7:339–373, 1997.

Hastie, Trevor; Tibshirani, Robert, and Friedman, Jerome. *The Elements of Statistical Learning*. Springer, 2009.

Jin, Xiaoping; Carlin, Bradley P., and Banerjee, Sudipto. Generalized hierarchical multivariate car models for areal data. *Biometrics*, 61(4):950–961, 2005. doi: 10.1111/j.1541-0420.2005.00359.x.

Koehler, Elizabeth; Brown, Elizabeth, and Haneuse, Sebastien. On the assessment of monte carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2): 155–162, 2009. doi: 10.1198/tast.2009.0030.

Künsch, Hans R. Intrinsic autoregressions and related models on the two-dimensional lattice. *Biometrika*, 74(3):517–524, 1987. doi: 10.2307/2337341.

Leach, Justin M. and Aban, Inmaculada. sim2dpredictr: An r package for simulating scalar outcomes with spatially dependent predictors. *Unpublished*, 2020.

Mitchell, T.J. and Beauchamp, J.J. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988. doi: 10.1080/01621459.1988.10478694.

Morris, Mitzi. Spatial models in stan: Intrinsic auto-regressive models for areal data. *Stan Case Studies*, 4, 2017. URL https://mc-stan.org/users/documentation/case-studies/icar_stan.html.

Morris, Mitzi; Wheeler-Martin, Katherine; Simpson, Dan; Mooney, Stephen J.; Gelman, Andrew, and DiMaggio, Charles. Bayesian hierarchical spatial models: Implementing the baseg york mollié model in stan. *Spatial and Spatio-temporal Epidemiology*, 31, 2019a. doi: 10.1016/j.sste.2019.100301.

Morris, Tim P.; White, Ian R., and Crowther, Michael J. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38:2074–2102, 2019b. doi: 10.1002/sim/8086.

Park, Trevor and Casella, George. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008. doi: 10.1198/016214508000000337.

R Core Team, . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

Ročková, Veronica and George, Edward. Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846, 2014. doi: 10.1080/01621459.2013.869223.

Ročková, Veronica and George, Edward. The spike and slab lasso. *Journal of the American Statistical Association*, 113:431–444, 2018. doi: 10.1080/01621459.2016.1260469.

Rue, Håvard. Fast sampling of gaussian markov random fields. *J.R. Statist. Soc. B*, 63: 325–338, 2001.

Rue, Håvard and Held, Leonhard. *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, Boca Raton, Florida, 2005.

Tang, Zaixiang; Shen, Yueping; Zhang, Xinyan, and Yi, Nengjun. The spike and slab lasso generalized linear models for prediction and associated genes detection. *Genetics*, 205: 77–88, 2017. doi: 10.1534/genetics.116.192195.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *J.R. Statist. Soc.*, 58 (1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Zou, Hui and Hastie, Trevor. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Methodological)*, 67:301–320, 2005. doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x.

THE SPIKE-AND-SLAB ELASTIC NET AS A CLASSIFICATION TOOL IN ALZHEIMER'S DISEASE

JUSTIN M. LEACH, LLOYD EDWARDS, RAJESH KANA, KRISTINA VISSCHER,

NENGJUN YI, AND INMACULADA ABAN

In preparation for *Statistics in Medicine*

Format adapted for dissertation

1 Introduction

1.1 Alzheimer's Disease Overview

Dementia has a long history as a scourge on the quality of life for aging persons, and in recent decades has been a leading cause of death. The medical research community has thus devoted considerable time and effort to understanding dementia's leading cause, Alzheimer's disease (AD). In consequence, the understanding of AD, and dementia in general, has developed significantly in the last century, and is evolving still (Bondi et al., 2017).

Early stages of dementia have received increasing research attention in the most recent several decades. These prodromal stages of neurodegenerative disease, more commonly known as mild cognitive impairment (MCI), are characterized by memory loss that is worse than would be expected given an individual's age, but not severe enough to meet the criteria of dementia. There are five general criteria for classification as MCI: the subject has subjective memory complaints, objective memory impairment that is advanced given the subject's age, relatively unaffected general cognition, unaffected daily activities, and is not diagnosed with dementia.

However, MCI also has many subtypes, and diagnosing MCI is not so straightforward as these criteria suggest, especially given that there appear to be substantial issues with false positives in diagnosing MCI; Bondi et al. (2017) provides a more general overview of these issues. Nevertheless, studies focusing on, or including, MCI can be useful for understanding the development of dementia in its early stages, as well as for identifying individuals who are at greater risk developing dementia.

Biomarker research has progressed alongside an increased understanding of AD's etiology. AD pathology is characterized primarily by amyloid plaques and neurofibrillary tangles (Lane et al., 2018). One influential theory, the amyloid cascade hypothesis, proposes that amyloid deposits underlie the generation of abnormal tau protein aggregation, which then leads to the neurofibrillary tangles that damage neurons and thus results in the observed deficits in cognitive and functional ability (Jack et al., 2010, 2013). However, subsequent research casts significant doubt on the viability of this hypothesis, specifically related to its temporal assumptions, and alternative theories have arisen, but a stable consensus is not obtained (Braak et al., 2011; Braak and Del Tredici, 2014; Crary et al., 2014). Given this uncertainty, temporal agnosticism is the current norm with respect to biomarker-related risks, e.g., see Jack et al. (2016a) and Bondi et al. (2017).

The popularity of the amyloid cascade hypothesis drove significant focus on amyloid-pathology-related imaging approaches, but doubt surrounding the hypothesis' validity, and evidence that tau-pathology more closely aligns with disease severity, created an interest in imaging approaches that capture tau protein pathology (Brosch et al., 2017). To date, Positron Emisson Tomography (PET) imaging with the tracer [^{18}F]AV-1451, or flortaucipir, is the most widely image modality for tau protein imaging. Another important biomarker, brain atrophy, predates the amyloid cascade hypothesis, indicates the extent of neurodegeneration, and is correlated with both tau deposition and neuropsychological deficits (Frisoni et al., 2010). Brain atrophy is not confined to AD, and should generally not be used in isolation to diagnose AD, but nevertheless patterns of brain atrophy have been useful in AD research.

1.2 Classification and Model Selection

The availability of neuroimaging-based biomarkers resulted in research on algorithmic approaches to classification of, and prediction of progression to, MCI and AD (Arbabshirani et al., 2017; Bondi et al., 2017; Rathore et al., 2017). Often the task is binary

classification, i.e., imaging data is input to the algorithm/model, which classifies the subject as either having or not having the disease. For a given classification paradigm, we need to both select the "optimal" model, and provide an objective description of how well the model distinguishes between subjects with or without the disease. These are related issues since the criteria for the optimal model/classifier is how well the classifier classifies subjects it has not yet seen. The general process, as shown in figure 1, is to divide data into "training" and "test" sets. The subjects in the training set are used to fit the model, and then the model is used to predict the classes for subjects in the test set, after which we evaluate the classifier's performance.

Many classifiers have free/tuning parameters that must be chosen by the investigator.[76] Classification problems require wariness towards overfitting models, because to be useful a model must generalize to new data. Thus, free/tuning parameter value choices should be based on corresponding estimates of expected prediction error. Prediction error estimates can be obtained via k-fold cross validation (Hastie et al., 2009). K-fold cross validation divides the subjects into k independent sets, and fits the model k times, each time holding out a different set as the test set. The average prediction error across the sets provides an estimate for the expected prediction error, and parameter values that minimize the estimate of expected prediction error are used to fit the classifier/model using all available data. Since many datasets have too small a sample size to hold out a separate test set, and since k-fold cross validation provides estimates of expected prediction error, it also allows for evaluating model performance while making use of all available data to build the model.

Improvement in the accuracy of these algorithms has the potential to allow for earlier detection of MCI and/or AD, and if sufficiently accurate in predicting progression, may allow researchers to identify patients who are at high risk for progression to AD. While many algorithms used thus far have shown promise, statistical learning approaches continue

[76] In Bayesian contexts, free/tuning parameters usually correspond to the hyperparameters in prior distributions.

to advance, and it is beneficial to apply, extend, and evaluate these algorithms with respect to MCI/AD classification. In this work we add to the classification literature by applying and evaluating the performance of the spike-and-slab elastic net with spatially informative priors as a classifier. First, we review several relevant statistical methods that build up to the methods used in this paper.

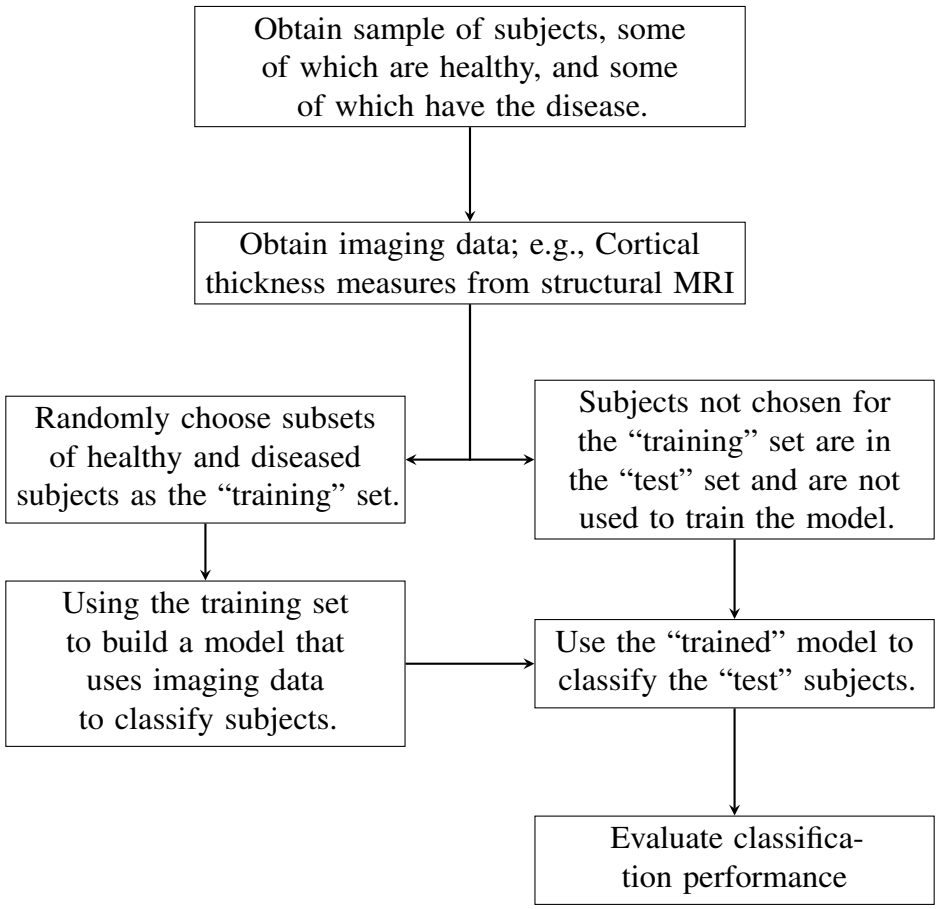

Figure 1: This flow chart describes the basic process of building and evaluating a classifier.

1.3 Penalized and Bayesian Generalized Linear Models

Penalized linear models are a useful approach to statistical learning, the most well-known of which are the ridge, lasso, and elastic net; the latter compromises between the first two (Tibshirani, 1996; Zou and Hastie, 2005). All three of these models have Bayesian interpretations, and in particular the latter two are useful for variable selection, because they tend to produce sparse solutions, i.e., applying the penalty results in many, if not most, of

the parameter estimates being zero, which allows for variable selection without resorting to null hypothesis significance testing (NHST).

The primary benefit of the elastic net is that it avoids the pitfalls of both the ridge and lasso. The ridge cannot provide automatic variables selection because all parameter estimates are non-zero, and when the number of predictors far outnumbers the number of subjects, the number of non-zero estimates allowed from the lasso is capped at the number of subjects; a related issues is when predictors are highly correlated, the lasso tends to choose one and discard the rest. In contrast, the elastic net can provide sparse solutions while leaving more parameter estimates non-zero, which makes it an attractive approach when using images as predictors, since there are often more predictors than subjects, and "relevant" predictors may often spatially cluster and be highly correlated. While penalty parameters, or prior distributions, must be chosen by the researcher, cross validation provides a principled approach to penalty parameter selection. These models can also produce estimates under ill-posed data, i.e., situations where there are more predictors than subjects or observations.

While penalized models produce biased estimates, the trade-off is typically substantial reduction in variance around the parameter estimates and predicted outcomes, which can improve the generalizeability of the models. The elastic net framework is thus attractive for both variable selection and prediction, and can be adapted to classification problems, e.g., by applying classification rules to a penalized/Bayesian logistic regression. An interesting consequence of the elastic net is that the sparse solutions produced by the initial model mean that the classifier will remove "unnecessary" predictors/features during the model fitting process by shrinking their estimates to zero, so that only a subset of the initial set of possible predictors/features is used in the classifier; the automatic variable selection of elastic net models equates to automatic feature selection in a statistical learning context.

The Bayesian interpretation of these models allows for useful extensions with respect to variable selection. In particular, the spike-and-slab lasso of Ročková and George

(2018) combines the lasso penalty with a common Bayesian approach to variable selection: the spike-and-slab prior, which models parameters are arising from a mixture of two distributions, one each for parameters that are and are not relevant to modeling the outcome of interest, respectively (Mitchell and Beauchamp, 1988; George and McCulloch, 1993). This approach extends the lasso such that there is stronger shrinkage imposed on parameters that are irrelevant and weaker shrinkage applied to relevant parameters, which leaves the final estimates of relevant parameters closer to their initial estimates, and drives estimates of irrelevant parameters to zero.

1.4 Outline

The elastic net is applicable to a wide range of problems, including imaging data, especially since it can perform variable/feature selection in cases where there are many more possible predictors than subjects. The elastic net has been used in other AD classification studies (de Vos et al., 2018; Schouten et al., 2016; Teipel et al., 2017; Trzepacz et al., 2016); however, the combination of spike-and-slab priors with the elastic net framework is a relatively new methodology, and to our knowledge it has not yet been explored in AD classification. The primary aim of this work is to demonstrate the potential of the methodology to AD classification, and by extension, to neuroimaging-based classification in general.

In previous work, we extended the spike-and-slab lasso to accommodate the elastic net penalty and explicitly model dependence among predictors (Leach et al., 2020). This class of models contains (Bayesian) logistic regression as a special case, and by using thresholding rules we can create a classifier. That is, the parameter estimates from a logistic regression can be used to obtain estimated/predicted probabilities of subjects having the disease; we can then use a threshold, say 0.5, above which subjects are classified as having the disease. In what follows, we show how this class of models can classify subjects by disease status; specifically, cognitive normal (CN) versus MCI, CN versus AD, and MCI

versus AD. The outline is as follows: In section 2, we briefly describe the ADNI data, discuss specific methods and outcomes used for classification, review the statistical details relevant to understanding the classifiers, and discuss several metrics used to evaluate classification performance. In section 3, we classify subjects' disease status using cortical thickness and tau PET images as predictors, and compare the results across several classifiers built from the methods presented in Section 2. Finally, in section 4, we discuss the implications of the results, outline future research directions, and address limitations in the present study.

2 Methods

2.1 ADNI Methods

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).[77]

As mentioned in section 1.1, brain atrophy is a useful metric for studying AD. A related measure is cortical thickness, which can be assessed in a cross sectional setting, whereas true atrophy measures would require longitudinal measures. Cortical thickness measures have been used as classification features throughout the literature (Rathore et al., 2017). However, cortical thickness is not trivial to measure because the topology of the cortex is that of a densely folded 2D sheet, and manual measurement is both laborious and error-prone. Nevertheless, reasonably accurate assessments of the properties of the cortical surface, whether surface area, volume, or thickness, can be obtained by "unfolding" the cortex. In companion papers, Dale et al. (1999) and Fischl et al. (1999) present an

[77]This is directly lifted from ADNI's manuscript citations document - they require this language in manuscripts/publications that use ADNI data.

approach to do just this by first segmenting white and grey matter voxels, then estimating the white/grey boundary at the subvoxel level using a triangular tessellation, and finally "inflating" that boundary out towards the pial surface to obtain the outer boundary, all the while minimizing metric distortions. Accurate cortical thickness measures may then be obtained by finding the distance between these two surfaces at a given point (Fischl and Dale, 2000). This entire process can be performed with FreeSurfer software, which results in a wide array of data, including surface area, volume, and cortical thickness measures (Fischl, 2012). FreeSurfer-processed ADNI data is available as summaries for brain regions specified by Desikan-Killiany atlas (Desikan et al., 2006).[78]

In addition to MRI, ADNI-3, the most recent renewal of the ongoing ADNI study, also has [^{18}F]AV-1451 Positron Emisson Tomography (PET) imaging; in fact, one of the motivations of ADNI-3 was to incorporate "innovative technologies", which includes a focus on tau PET imaging (Weiner et al., 2017). In ADNI tau PET images have been processed with Freesurfer, and SUVR summaries are available by regions of the Desikan-Killiany atlas.

Since both cortical thickness and tau PET images can be reasonably expected to provide information regarding AD status, in the present work we pursue two separate paths of analysis: one using cortical thickness summaries as features, and the other using tau PET SUVR summaries as features.

2.2 Statistical Methods

Logistic regression arises from a generalized linear model (GLM) where outcomes are binary; subject-specific probabilities of being in one class or the other can be extracted from such models, and we can then apply thresholding rules to build a classifier.[79]

[78] This paper contains a "high level" summary of the most relevant FreeSurfer methods. A more technical summary is found on the FreeSurferWiki, as a more detailed description is beyond the scope of this document; nevertheless, the several papers most relevant to this process are contained in the references.

[79] A technical detail: strictly speaking, GLM's model $\mathrm{E}(y_i)$, and so when we assume y_i arises from a binomial distribution, the observed y_i will be 0 or 1, but the *expectation* $\mathrm{E}(y_i)$ will be a probability, i.e., take

The mathematical form of a GLM is as follows:

$$g(y_i) = \boldsymbol{X}_i\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{J} x_{ij}\beta_j = \eta_i, \quad i = 1, \ldots, N \tag{2.2.1}$$

where for the i^{th} subject, y_i is an observed outcome, $\boldsymbol{X}_i$ is $1 \times (J+1)$ vector of measured predictors, $\boldsymbol{\beta}$ is a $(J+1) \times 1$ vector of unknown parameters, and $g(\cdot)$ is an appropriate link function; for logistic regression, $g(\cdot)$ is the logit function. In this work, the outcome is disease class, e.g., whether the subject is CN or has dementia, and the predictors are the subject's average cortical thickness or tau disposition measures for each region in Desikan-Killiany atlas. Thus, each subject will have a single outcome indicating disease status, and sixty-eight predictors, one for atlas region.

Ročková and George (2018) introduced the spike-and-slab lasso, which combines spike-and-slab priors with the lasso penalty. This model penalizes estimates for "irrelevant" parameters more strongly than those of "relevant" parameters, and so shrinks more "irrelevant" parameter estimates to zero, while allowing "relevant" parameter estimates to remain larger. While Ročková and George (2018) focus on linear models, Tang et al. (2017) describe how to fit the spike-and-slab lasso for GLM's by using an Expectation-Maximization Coordinate Descent algorithm. As discussed in Section 1.3, the lasso may have downsides when predictors arise from images, and so we construct a spike-and-slab elastic net. As shown in Leach et al. (2020), the spike-and-slab lasso can be generalized to a spike-and-slab elastic net, which can still be fit by the Expectation-Maximization Coordinate Descent algorithm. The spike-and-slab elastic net prior is as follows:

$$p(\beta_j|\gamma_j, s_0, s_1) \propto \exp\left[-\frac{1}{S_j}\{(1-\xi)\beta_j^2 + \xi|\beta_j|\}\right] \tag{2.2.2}$$

where $S_j = (1-\gamma_j)s_0 + \gamma_j s_1$, s_1 is the slab scale, s_0 is the spike scale, $s_1 > s_0 > 0$, and

on some value between 0 and 1.

$\xi \in [0, 1]$.[80] A spike-and-slab ridge is obtained when $\xi = 0$ and spike-and-slab lasso when $\xi = 1$. The $\gamma_j \in \{0, 1\}$ are indicator variables for model inclusion, and are assigned a Bernoulli distribution with unknown probability of inclusion given by θ_j.[81]

Spatial information is often relevant to determining which parameters affect an outcome; i.e., we expect relevant parameters to cluster spatially.[82] Unlike classical GLM's, Bayesian GLM's will provide solutions when predictors are highly correlated, but will not incorporate spatial information into model selection unless explicitly included in the prior distributions. It is conceivable to explicitly model correlation among the parameters, β_j, in their prior distributions, but a more computationally viable approach is to model correlation among the prior probabilities of inclusion, θ_j, whose conditional estimates affect the degree of shrinkage applied to parameters; correlation among these estimates mean that a given estimate for β_j depends in part upon the shrinkage applied to its neighbors' estimates. A variant of Conditional Autoregressions (CAR) known as Intrinsic Autoregressions (IAR) have been used to incorporate spatial information into a wide range of practical applications, and can be used as a prior distribution on the logit probabilities of inclusion (Banerjee et al., 2015; Besag, 1974; Besag and Kooperberg, 1995; Rue and Held, 2005). Below is the prior distribution for the logit of probabilities of inclusion:

$$\log p(\psi_j | \psi_i, \tau) \propto \frac{-\tau^2}{2} \left(\sum_{j:j<i} (\psi_j - \psi_i)^2 \right) \tag{2.2.3}$$

where $\psi_j = \text{logit}(\theta_j)$.[83] Note that in practice, we often set $\tau = 1$, which is the convention applied in this work (Morris et al., 2019).

[80] The slab distribution is wide, to allow parameters to take larger values, while the spike distribution is narrow, to shrink estimates severely toward zero. This is why $s_1 > s_0$.

[81] The γ_j are treated as missing data by the EM algorithm; their conditional expectations are estimated and plugged into the joint posterior distribution, which is the maximized over the remaining parameters. This iterative approach continues unto convergence.

[82] For example, many multiple testing procedures in neuroimaging explicitly make this assumption by taking into account the size, extent, and general properties of clusters of statistically significant voxels.

[83] Since $\theta_j \in [0, 1]$, the multivariate Normal distribution, of which the CAR model is a special case, is not a valid prior for θ_j; however, $\text{logit}(\theta_j) \in \mathbb{R}$, which is the same support as the multivariate Normal distribution, making the CAR a valid prior for $\text{logit}(\theta_j)$.

2.3 *Classification and Model Evaluation*

The goal of this work is to examine the classification ability of the spike-and-slab elastic net class of models. There are therefore two levels of statistical concern. The first is that the spike-and-slab elastic net has several "free" parameters, which must be selected by the user.[84] K-fold cross validation techniques provide a principled way to make such choices, by providing reasonable estimates of prediction error, as assessed by cross-validated estimates of metrics like model deviance, mean-squared error, or area under the ROC curve (AUC) (Hastie et al., 2009). We fit the model for several values of each free parameter and compare their prediction error estimates.[85] We can then select the model with the lowest prediction error estimates.[86] This is often a useful approach to model selection, even as the goals of various models differ by circumstance. Figure 2 shows how we apply k-fold cross validation in this work.

In many applications logistic regression is employed to obtain estimates for effects of predictors on an outcome, typically odds ratios with corresponding interval estimates, and evaluation of the model would revolve around those estimates. However, in this work we are not explicitly focused on such "effect" estimates; rather, we need to apply a classification rule to the estimated probabilities arising from the logistic regression. Since the predicted outcomes obtained from a logistic regression are probabilities, one may apply a threshold to classify subjects as either having or not having the disease. While in theory any value between zero and one can be used as a threshold, in practice subjects are typically placed in the class for which they have the highest predicted probability, which in binary classification means if the predicted probability is great than one half, they are classified as having the

[84] In this case, we must choose ξ, which determines the compromise between ridge and lasso penalities, as well as the spike (s_0) and slab (s_1) scales. However, this process applies in general for other situations with different "free" parameters.

[85] At this point, which measure is used is not important; what matters for the sake understanding this process is that we estimate prediction error, which provides an idea of how well the model would generalize. See Hastie et al. (2009), Chapter 7 for more details on the relationship between prediction error and cross validation.

[86] In contrast to most prediction error estimates, when using AUC we choose the model with highest value; however, generally speaking, higher cross-validated AUC estimates correspond to smaller prediction errors.

disease. Thus, the usefulness of the model with respect to classification is not in the accuracy or interpretation of odds ratios, but whether after applying a threshold to the subjects' predicted outcomes, the result is accurate classification, both of the current subjects and of subjects the model has not yet seen, i.e., independent data.

Assessing classifiers requires evaluation with respect to several metrics in order to avoid being deceived. An obvious concern is the classifier's accuracy, that is, what is the probability that the algorithm correctly classifies a subject. However, this metric can be misleading, especially in situations where the sample sizes for each class are unbalanced.[87] Thus, in addition to accuracy, it is important to estimate and consider several other metrics in evaluating classification performance:

1. *Sensitivity*: the probability that the subject is classified as having the disease, given that the subject has the disease.

2. *Specificity*: the probability that the subject is not classified as having the disease, given that the subject does not have the disease.

3. *Positive Predictive Value (PPV)*: the probability that a subject has the disease, given that the subject was classified as having the disease.

4. *Negative Predictive Value (NPV)*: the probability that the subject does not have the disease, given that the subject is classified as not having the disease.

In the authors' experience, PPV and NPV are often neglected, but neglecting these metrics makes the same mistake as only reporting accuracy, because these metrics can also be deceiving, especially under unbalanced data.[88] Ideally, all five metrics should be near

[87] For example, consider a sample of 1,000 subjects, 950 of which are healthy, and 50 of which have the disease. A classifier that assigns "healthy" to all subjects no matter their characteristics would have 95% accuracy, but miss every disease case. This is clearly an undesirable classifier, despite its high accuracy.

[88] Let us return to the case where 950 subjects do not have a disease and 50 do. Suppose the sensitivity and specificity are both 90%, meaning 45/50 subjects with the disease and 855/950 subjects without the disease are so classified. Now, 855/860 (99.42%) of those who tested negative were actually disease free, but only 45/140 (32.14%) of subjects who tested positive actually had the disease, which means roughly 2/3 of positive tests impose undue anxiety in their recipients.

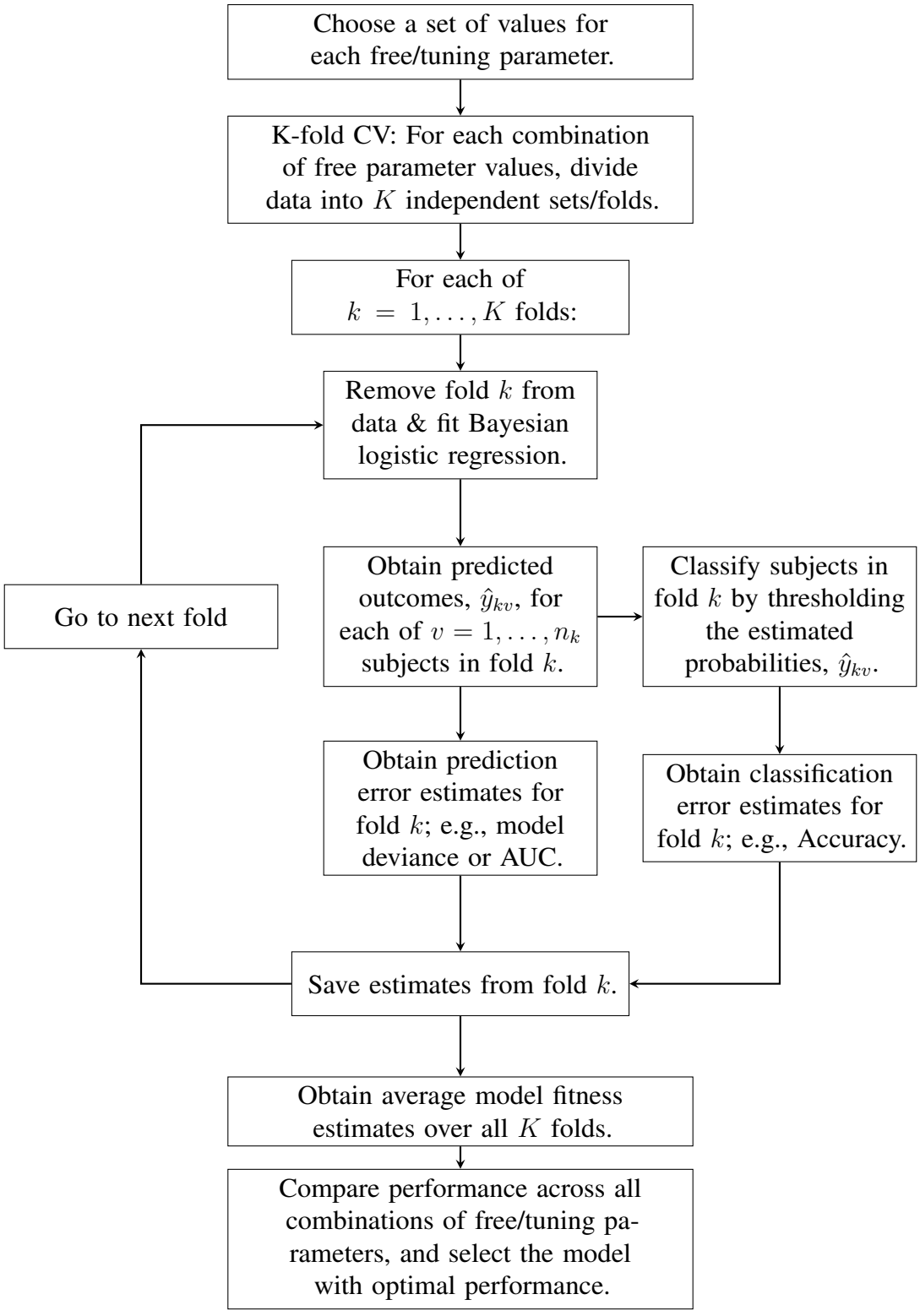

Figure 2: This flowchart describes application of k-fold cross validation employed in this paper.

one if the classifier is performing well, and these metrics should be estimated within a cross-validation process, just like measures of prediction error before applying the classifier, because we are most concerned with estimating what these metrics would be if we applied the classifier to independent data. Therefore, as we move on to the analyses, it is important the bear in mind that we will be assessing the models' performance on two levels: the first is in basic model selection, that is, regardless of the classification rule, how well do we predict the initial model would generalize. Secondly, after applying the classification rule to create a classifier, how well do we predict the classifier would classify new subjects? These two levels are the heart of the present examination, with a particular focus on the latter, classification ability.

3 Analysis

3.1 Analysis Framework

Two cross sectional datasets were employed for classification, one contained cortical thickness measures and one contained standardized uptake value ratios (SUVR) from tau PET images.[89] The cortical thickness data set included 234 (60.15%) cognitively normal (CN) subjects, 116 (29.82%) with MCI, and 39 (10.03%) with dementia. The tau PET dataset included 262 (60.93%) cognitively normal (CN) subjects, 127 (29.53%) with MCI, and 41 (9.53%) with dementia. For each data set, we evaluate the algorithm's ability to classify CN vs. MCI, CN vs. dementia, and MCI vs. dementia.

We fit two sets of Bayesian logistic regression models. The first set of models is fit with the traditional lasso ($\xi = 1$), spike-and-slab lasso (SSL), and spike-and-slab lasso with IAR priors on the inclusion probabilities (SSL-IAR). The second set of models is a halfway compromise between the ridge and lasso ($\xi = 0.5$), resulting in what we refer to as the traditional elastic net (EN), spike-and-slab elastic net (SSEN), and spike-and-slab elastic net with IAR priors on the inclusion priors (SSEN-IAR). Recall that including IAR priors

[89] See, e.g., Vemuri et al. (2017) for details regarding SUVR in AD research.

on inclusion probabilities is how we explicitly model spatial structure; thus, the three levels of models can be seen as gradually extending the elastic net from its traditional form to spike-and-slab form to a spike-and-slab form with an explicit modeling of spatial structure. For a given model and set values of s_1 and s_0 we obtain model fits and prediction error statistics via 5-fold cross validation; since the resulting held-out sets were relatively small, we performed 5-fold cross validation 10 times each case in order to obtain more stable estimates.

Prior scale values, i.e., s_1 and s_0, must be selected in a principled way. The traditional models have that $s_1 = s_0$, in which case a single parameter value must be chosen. We use the `R` package `glmnet` to fit a grid of models and select the model whose scale parameter minimizes the cross-validated deviance.[90] When using spike-and-slab priors we fit models on a grid of spike priors, $s_0 = \{0.01, 0.02, \ldots, 0.5\}$, and slab priors, $s_1 = \{1, 2, 3, 4, 5, 7.5, 10, 12.5, 15, 17.5, 20, 22.5, 25, 27.5, 30\}$ and select the model that minimizes the cross-validated deviance. That is, we perform 5-fold cross validation to obtain parameter and prediction error estimates at each combination of values for s_0 and s_1, and then choose the values of s_0 and s_1 that minimize the prediction error as measured by cross-validated deviance. While models are selected using deviance, for the final six models in each case we report the mean squared error (MSE), mean absolute error (MAE), area under the ROC curve (AUC), and misclassification (MC)[91] to enable comparison across models.

Classification is performed in each case by placing an observation in the class that has the highest probability; e.g., when comparing CN and MCI, if the estimated probability is > 0.5, then the subject is classified as MCI. Classification performance estimates are obtained

[90] Without going into great detail, the key distinction between cross validated deviance and model deviance is that the typical model deviance measures the fit of a model with the data used to fit the model, i.e., the training error, whereas in the context of k-fold cross-validation we can see how well a model fits the held out set in each of the k models, by using deviance, MSE, etc. This process provides us with a better estimate of prediction error. See Hastie et al. (2009), Chapter 7, for details.

[91] Misclassification is defined as $\frac{1}{n}\sum_{i=1}^{n} I(|y_i - \hat{y}_i| > 0.5)$ where $I(\cdot)$ is an indicator function whose value is 1 when the argument is true, and zero otherwise.

within the cross validation process. Following Section 2.3, we evaluate the classification performances of each model using estimates for accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

3.2 Analysis Results

Cortical thickness

Table 1 shows the estimated prediction error statistics when using cortical thickness as features. For each classification scenario, SSEN-IAR had the lowest model deviance, but in each case was closely followed by SSL-IAR, and occasionally SSL-IAR slightly outperformed SSEN-IAR on other metrics, e.g., MSE for CN vs. dementia, AUC for CN vs. MCI, and misclassification for CN vs. dementia. In general, model performance varied widely across classification scenario. The most noticeable variation was with AUC, which was above 0.94 for CN vs. dementia, between 0.62 and 0.68 for CN vs. MCI, and between 0.79 and 0.86 for MCI vs. dementia; see also figure 3. Similar differences by classification scenario were present for MSE, MAE, and misclassification (MC).

Classification performance, as shown in table 2 and the left-hand side of figures 5, 6, and 7, presents a more complicated story. Accuracy, specificity, PPV, and NPV were generally very similar across models within a given classification scenario. Nevertheless, the spike-and-slab models tended to have slightly higher accuracies and predictive values by about 1% to 3%. With respect to sensitivity, the spike-and-slab models noticeably outperformed the traditional models between roughly 7% and 19%.

There were noticeable differences in model performance across classification scenarios. For CN vs. dementia, all models had accuracy above 0.93 (max: SSEN, 0.952), specificity above 0.98 (max: SSEN, 0.985), PPV above 0.86 (max: SSEN, 0.898), and NPV above 0.94 (max: SSEN, 0.966). Sensitivity had the most obvious differences across models, with SSEN having the highest (0.79), the other spike-and-slab models above 0.75, and the traditional models below 0.70. See figure 5, left-hand side.

For CN vs. MCI, accuracy (max: SSEN, 0.733) and NPV (max: SSL-IAR, 0.736) were both above 0.71 for all models. Specificity was above 0.91 (max: EN, 0.974). The best sensitivity was SSL (IAR) at 0.335, while the other spike-and-slab models were above 0.29, and the traditional models around 0.23. PPV was higher for the traditional models (near 0.80), while the spike-and-slab models were between 0.64 and 0.75. See also figure 6, left-hand side.

For MCI vs. dementia, accuracy was above 0.81 for the spike-and-slab models (max: SSEN-IAR, 0.832) and below 0.80 for the traditional models. Specificity was above 0.91 for all models (max: EN, 0.939). SSEN-IAR had the highest NPV (0.862), while the other spike-and-slab models were above 0.84, and the traditional models below 0.82). SSEN-IAR had the highest sensitivity (0.559), while the other spike-and-slab models were above 0.50, and the traditional models below 0.40. SSEN-IAR also had the highest PPV (0.712), with SSEN and SSL-IAR near 0.69, SSL and EN near 0.67, and traditional lasso at 0.64. See also figure 7, left-hand side.

Tau PET imaging

Table 3 shows the estimated prediction error statistics when using tau PET imaging as features. Unlike with cortical thickness, no specific model consistently outperformed all others, but the spike-and-slab models always outperformed the traditional models; by deviance the best fits were SSL for CN vs. dementia, SSL-IAR for CN vs. MCI and SSEN-IAR for MCI vs. dementia.

Similar to cortical thickness, model performance differed considerably by classification scenario, and again the most noticeable metric was AUC, which was greater than 0.89 for all models under CN vs. dementia (max: SSL, 0.933), above 0.68 for CN vs. MCI (max: SSEN-IAR, 0.728), and above 0.68 for MCI vs. dementia (max: SSEN-IAR, 0.785); it is also noticeable that for MCI vs. Dementia, the models with IAR priors were around 0.78, the spike-and-slab models without IAR priors just under 0.74, and the traditional models

under 0.70. ROC curves for tau PET are shown in figure 4, which highlights the differing performances of the models by classification scenario.

Table 4 displays the classification performance when using tau PET measures as features. As with cortical thickness, the accuracy, specificity, and NPV were fairly similar across models within a classification scenario, but the spike-and-slab models outperformed the traditional models with respect to accuracy and NPV in every case by approximately 1% to 4%. While sensitivity was lower for tau PET compared to cortical thickness, the spike-and-slab models noticeably outperformed the traditional models by approximately 6% to 15%. PPV was higher for the spike-and-slab models for both CN vs. dementia ($4\% - 7\%$) and MCI vs. dementia ($10\% - 17\%$), but fairly similar with mixed results for CN vs. MCI.

As with cortical thickness, there were noticeable differences across classification scenarios. For CN vs. dementia, accuracy was above 0.91 (max: SSL, 0.936), specificity above 0.97 (max: SSEN-IAR, 0.985), and NPV above 0.92 (max: SSL, 0.947). SSL had the highest sensitivity at 0.651, with the other spike-and-slab models near 0.60 and the traditional models a little above 0.5. SSEN (IAR) had the highest PPV (0.861), with the other spike-and-slab models above 0.83, and the traditional models at or below 0.80. Accuracy was around 0.93 for the spike-and-slab models and 0.91 for the traditional models. See figure 5, right hand side.

For CN vs. MCI, specificity was above 0.92 for all models. SSL (IAR) had the highest accuracy and NPV (both 0.765), while the other spike-and-slab models were above 0.74, and the traditional models about 0.73. SSL (IAR) had the highest sensitivity (0.406), while the other spike-and-slab models were between 0.33 and 0.37, and the traditional models below 0.30. PPV was highest for SSL (0.766), and all other models were above 0.71. See figure 6, right hand side.

For MCI vs. dementia specificity was above 0.94 for all models (max: SSEN 0.967). The spike-and-slab models had NPV of approximately 0.82 (max: SSEN-IAR, 0.822), while

the traditional models were approximately 0.79. Similarly, accuracy was approximately 0.81 for all spike-and-slab models (max: SSL and SSEN, 0.811) and approximately 0.77 for the traditional models. PPV was highest for the spike-and-slab models without spatial structure with SSEN being the highest (0.761), while the models with IAR priors were approximately 0.69, and the traditional models less than 0.60. SSEN (IAR) had the highest sensitivity (0.363), while the other spike-and-slab models were above 0.32 and the traditional models below 0.20. See figure 7, right hand side.

3.3 Conclusion

When classifying CN vs. demented subjects, all the algorithms tend to perform well with respect to accuracy, specificity, and NPV (all > 0.90), regardless of the imaging modality, and while the spike-and-slab models tended to outperform the traditional models on these measures, the differences were generally not extreme. The similarity across models with respect to accuracy, specificity, and NPV continues with CN vs. MCI and MCI vs. dementia, but performance is considerably weaker for accuracy and NPV. Specificity remains above 0.90 for all models, but accuracy and NPV drop to between $0.70 - 0.76$ for CN vs. MCI and $0.75 - 0.86$ for MCI vs. dementia. It is worth noticing that accuracy appears to be driven by NPV, since these metrics are nearly indistinguishable in every case, as visualized in figures 5, 6, and 7.

However, accuracy can be misleading, especially with unbalanced data. When we turn attention to sensitivity and PPV, clear differences emerge between both models and imaging modality. Sensitivity is consistently lower, and has a larger range, than the other metrics ($0.18 <$ sensitivity < 0.79) in all cases), and for CN vs. dementia and MCI vs. dementia, cortical thickness has higher sensitivity compared to tau PET. In all cases except CN vs. dementia using cortical thickness, a model using the IAR prior had the highest sensitivity; in this exception the SSEN model had the highest sensitivity. In every scenario the spike-and-slab models noticeably outperformed the traditional models with respect the

sensitivity. PPV was consistently higher than sensitivity, in most cases above 0.60. When using tau PET images and for CN vs. dementia and MCI vs. dementia, PPV was noticeably higher for the spike-and-slab models; this was true to a lesser extent with cortical thickness. There was not a clear trend for CN vs. MCI with respect to PPV.

In conclusion, all the models tend not to classify subjects as being at a later stage than they are, i.e., avoiding false positives even when classifying subjects who are more similar to each other. When classifying dementia vs. either CN or MCI, a positive result from the spike-and-slab models is more likely to be correct compared to the traditional models, and usually with acceptably high probability, especially for CN vs. dementia. However, especially for comparisons other than CN vs. dementia, the models tend to miss many subjects who have progressed further. While the spike-and-slab and/or IAR priors helped to improve sensitivity, the improvement did not always result in sensitivity values that would be considered desirable.

4 Discussion

In this work we have applied the spike-and-slab elastic net as a classifier in ADNI data. While this is not the first paper to apply the elastic net to neuroimaging data related to AD, it is to our knowledge the first attempt to apply the spike-and-slab prior framework to classification in AD data; an additional novelty is the explicit modeling of spatial information within the spike-and-slab elastic net framework. In the analyses presented, the classifiers had greatest success in classifying CN vs. dementia, followed by MCI vs. dementia, and lastly CN vs. MCI, regardless of the imaging modality. Classification ability drops off noticeably when MCI enters the picture, but this is perhaps not too surprising since presumably the difference between CN and MCI subjects and MCI and dementia subjects will tend to be less than that between CN and dementia subjects, and in addition MCI diagnoses are known to be subject to false positive issues (Bondi et al., 2017).

The classification accuracy estimates presented here are often comparable to that of other classification methods in the literature, e.g., see table 2 in Rathore et al. (2017), which suggests that the presented algorithms may prove useful in wider contexts. While the models using spike-and-slab priors tended to outperform traditional models with respect to accuracy, specificity, PPV, and NPV, the difference across models within a classification scenario was often relatively small. The clearest benefit of the spike-and-slab elastic net framework was with respect to sensitivity. While in many cases, especially CN vs. MCI, the sensitivities still low enough to be undesirable, the models using spike-and-slab models always outperformed the traditional models, and most cases including the IAR prior on inclusion probabilities yielded an additional benefit to sensitivity.

There are several limitations related to the data used in the current study, several of which point towards ways in which model performance may be improved. As described in Rathore et al. (2017), many classification approaches use atlases to reduce the dimension of the predictors/features. However, given the ability of the presented class of models to handle high dimensional spatial data, restricting analyses to data averaged within each region of the Desikan-Killiany atlas, which process reduces thousands of measurements to 68 per subject, may reduce the effectiveness of the models used in this work. While dimension reduction is necessary to fit many models in the first place, variants of the elastic net can handle situations where there are many more predictors/features than subjects, and the addition of the IAR prior on inclusion probabilities can also incorporate spatial information into variable/feature selection. Another concern with atlas use is that important areas that overlap several atlas regions may not be detectable after reduction to the atlas, which may harm model performance.

Despite our intuitions regarding dimension reduction, algorithms that use reduced features tend to perform better than those based on voxel or vertex level data (Park et al., 2012). However, even if vertex or voxel level data is too noisy to improve performance,

there are a wide range of dimension reduction methods available, and it is possible that an approach to dimension reduction yielding many more than 68 features would improve model performance, and take greater advantage these methods' strengths. Given that the spike-and-slab elastic net framework is specifically designed to handle "noisy" situations, we may reasonably expect the difference in performance between the spike-and-slab and traditional elastic net to be larger as the number of variables/features increases. More flexible approaches to dimension reduction may also yield different, and potentially more relevant, feature sets for tau PET imaging and cortical thickness data sets, which may also improve model performance. We also restricted analysis to cross sectional data, but it would useful to perform a study to determine whether the algorithm could predict whether subjects would progress from CN to MCI, MCI to dementia, or even CN to dementia.

While there are several limitations to the current study, the spike-and-slab elastic net models tended to outperform the traditional elastic net models across several metrics, noticeably improved sensitivity estimates, and showed comparable classification accuracy to other algorithms in the literature. In addition, there are many future directions for extending the model to incorporate flexible dimension reduction as a step before applying the spike-and-slab elastic net, which may allow one to create a feature set that better exploits the strengths of the model.

Appendix: Tables

Table 1: Cortical Thickness: Prediction Error Estimates

				Cross-Validated Average				
	Model	s_0	s_1	Dev.	AUC	MSE	MAE	MC
CN vs. Dem.	Lasso	0.003	0.003	96.199	0.946	0.049	0.096	0.061
	SSL	0.250	7.500	78.079	0.966	0.038	0.068	0.052
	SSL-IAR	0.250	10.000	71.314	0.972	0.034	0.062	0.048
	EN	0.001	0.001	87.383	0.955	0.044	0.088	0.060
	SSEN	0.150	12.500	84.156	0.956	0.040	0.082	0.043
	SSEN-IAR	0.260	20.000	70.646	0.973	0.035	0.068	0.049
CN vs. MCI	Lasso	0.007	0.007	421.441	0.625	0.204	0.409	0.275
	SSL	0.140	7.500	414.751	0.657	0.200	0.393	0.285
	SSL-IAR	0.140	10.000	404.155	0.681	0.194	0.383	0.276
	EN	0.009	0.009	420.883	0.627	0.204	0.409	0.273
	SSEN	0.070	4.000	413.268	0.658	0.199	0.397	0.267
	SSEN-IAR	0.140	7.500	404.120	0.679	0.194	0.385	0.273
MCI vs. Dem.	Lasso	0.013	0.013	140.356	0.796	0.148	0.290	0.217
	SSL	0.140	5.000	124.898	0.841	0.129	0.248	0.186
	SSL-IAR	0.150	4.000	124.109	0.845	0.128	0.243	0.175
	EN	0.012	0.012	135.041	0.812	0.142	0.280	0.203
	SSEN	0.140	5.000	124.198	0.843	0.128	0.254	0.180
	SSEN-IAR	0.140	5.000	121.553	0.851	0.125	0.244	0.168

Table 2: Cortical Thickness: Classification Performance

				Cross-Validated Average				
	Model	s_0	s_1	Accu.	Sens.	Spec.	PPV	NPV
CN vs. Dem.	Lasso	0.003	0.003	0.939	0.672	0.984	0.874	0.947
	SSL	0.250	7.500	0.948	0.751	0.981	0.870	0.960
	SSL-IAR	0.250	10.000	0.952	0.769	0.982	0.880	0.962
	EN	0.001	0.001	0.940	0.685	0.982	0.864	0.949
	SSEN	0.150	12.500	0.957	0.790	0.985	0.898	0.966
	SSEN-IAR	0.260	20.000	0.951	0.751	0.984	0.885	0.960
CN vs. MCI	Lasso	0.007	0.007	0.725	0.233	0.969	0.788	0.718
	SSL	0.140	7.500	0.715	0.309	0.916	0.645	0.728
	SSL-IAR	0.140	10.000	0.724	0.335	0.917	0.667	0.736
	EN	0.009	0.009	0.727	0.228	0.974	0.812	0.718
	SSEN	0.070	4.000	0.733	0.290	0.952	0.750	0.730
	SSEN-IAR	0.140	7.500	0.727	0.328	0.925	0.684	0.735
MCI vs. Dem.	Lasso	0.013	0.013	0.783	0.315	0.941	0.642	0.803
	SSL	0.140	5.000	0.814	0.508	0.916	0.673	0.847
	SSL-IAR	0.150	4.000	0.825	0.554	0.916	0.690	0.859
	EN	0.012	0.012	0.797	0.374	0.939	0.672	0.817
	SSEN	0.140	5.000	0.820	0.505	0.926	0.697	0.848
	SSEN-IAR	0.140	5.000	0.832	0.559	0.923	0.712	0.862

Table 3: Tau PET Imaging: Prediction Error Estimates

				Cross-Validated Average				
	Model	s_0	s_1	Dev.	AUC	MSE	MAE	MC
CN vs. Dem.	Lasso	0.002	0.002	142.133	0.895	0.065	0.125	0.086
	SSL	0.440	17.500	110.925	0.933	0.051	0.097	0.064
	SSL-IAR	0.160	7.500	118.995	0.919	0.054	0.105	0.070
	EN	0.001	0.001	136.890	0.904	0.062	0.123	0.083
	SSEN	0.270	30.000	117.081	0.916	0.053	0.107	0.070
	SSEN-IAR	0.470	30.000	113.763	0.926	0.052	0.106	0.067
CN vs. MCI	Lasso	0.003	0.003	442.126	0.685	0.191	0.383	0.263
	SSL	0.270	20.000	425.962	0.726	0.181	0.359	0.255
	SSL-IAR	0.250	15.000	423.400	0.727	0.178	0.352	0.235
	EN	0.002	0.002	443.969	0.682	0.192	0.383	0.269
	SSEN	0.110	20.000	427.318	0.719	0.182	0.365	0.256
	SSEN-IAR	0.270	20.000	424.092	0.728	0.181	0.361	0.253
MCI vs. Dem.	Lasso	0.010	0.010	172.852	0.683	0.168	0.324	0.229
	SSL	0.200	12.500	158.931	0.739	0.147	0.288	0.189
	SSL-IAR	0.170	7.500	154.885	0.778	0.145	0.279	0.195
	EN	0.022	0.022	170.016	0.690	0.165	0.326	0.230
	SSEN	0.190	12.500	158.446	0.738	0.148	0.294	0.189
	SSEN-IAR	0.300	22.500	153.490	0.785	0.145	0.280	0.195

Table 4: Tau PET Imaging: Classification Performance

				Cross-Validated Average				
	Model	s_0	s_1	Accu.	Sens.	Spec.	PPV	NPV
CN vs. Dem.	Lasso	0.002	0.002	0.914	0.522	0.975	0.770	0.929
	SSL	0.440	17.500	0.936	0.651	0.980	0.840	0.947
	SSL-IAR	0.160	7.500	0.930	0.600	0.982	0.837	0.940
	EN	0.001	0.001	0.917	0.507	0.981	0.809	0.927
	SSEN	0.270	30.000	0.930	0.588	0.984	0.852	0.938
	SSEN-IAR	0.470	30.000	0.933	0.602	0.985	0.861	0.940
CN vs. MCI	Lasso	0.003	0.003	0.737	0.294	0.952	0.750	0.735
	SSL	0.270	20.000	0.745	0.365	0.929	0.715	0.751
	SSL-IAR	0.250	15.000	0.765	0.406	0.940	0.766	0.765
	EN	0.002	0.002	0.731	0.277	0.952	0.737	0.730
	SSEN	0.110	20.000	0.744	0.330	0.945	0.746	0.744
	SSEN-IAR	0.270	20.000	0.747	0.354	0.939	0.737	0.749
MCI vs. Dem.	Lasso	0.010	0.010	0.771	0.198	0.957	0.597	0.787
	SSL	0.200	12.500	0.811	0.346	0.961	0.739	0.820
	SSL-IAR	0.170	7.500	0.805	0.359	0.950	0.698	0.821
	EN	0.022	0.022	0.770	0.188	0.957	0.589	0.785
	SSEN	0.190	12.500	0.811	0.327	0.967	0.761	0.817
	SSEN-IAR	0.300	22.500	0.805	0.363	0.948	0.696	0.822

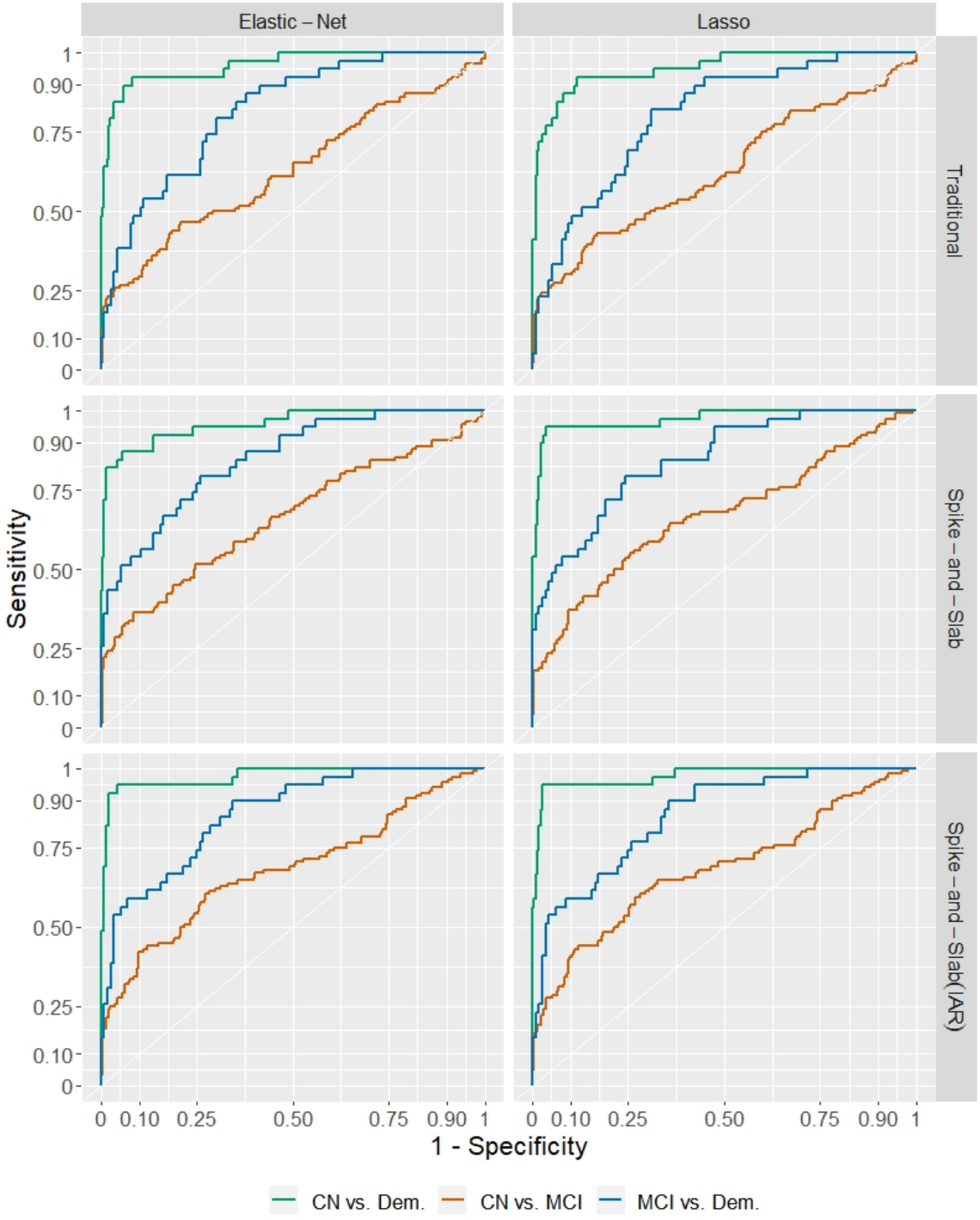

Figure 3: Estimated ROC curves for cortical thickness as features/predictors.

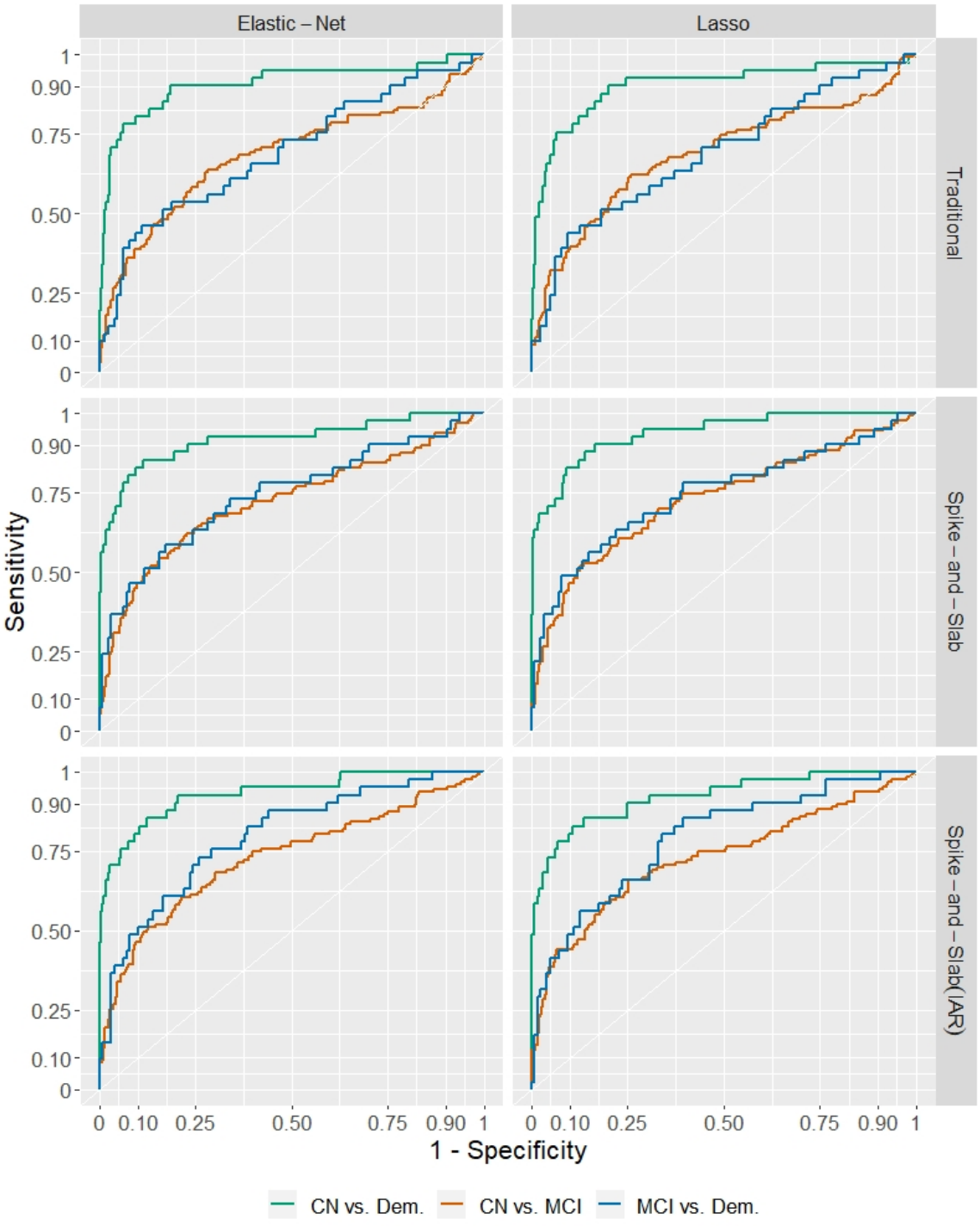

Figure 4: Estimated ROC curves for tau PET as features/predictors.

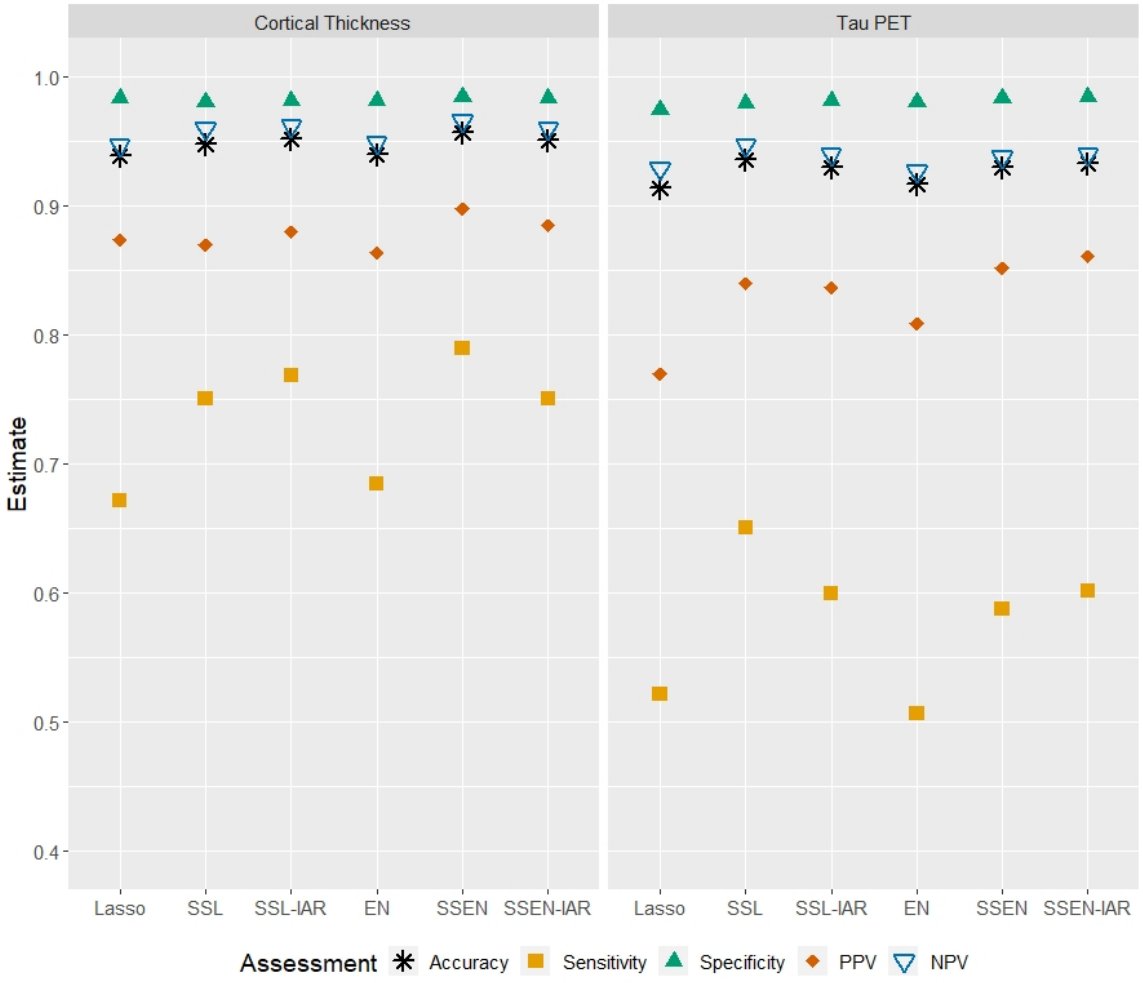

Figure 5: Model classification performance for cognitive normal vs. demented subjects.

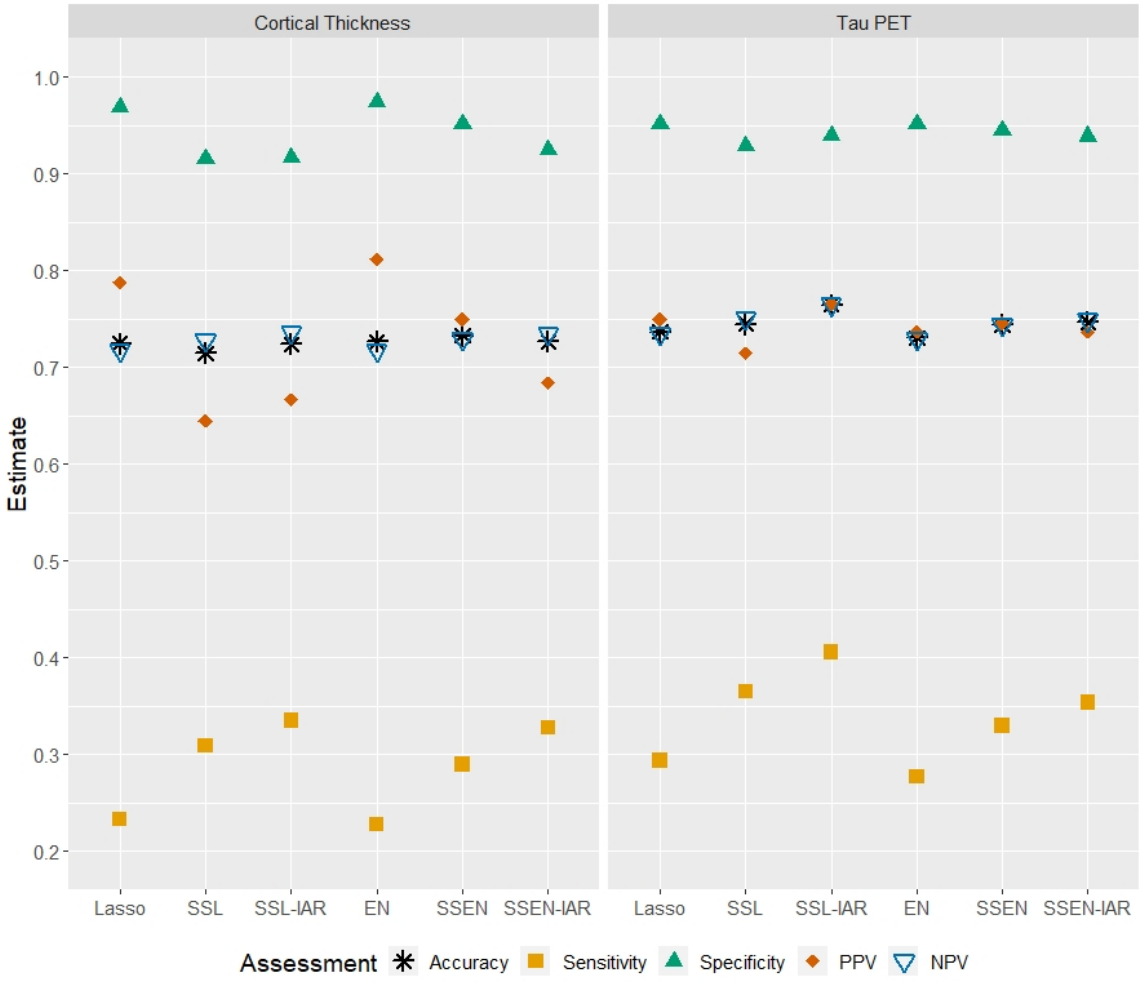

Figure 6: Model classification performance for cognitive normal vs. MCI subjects.

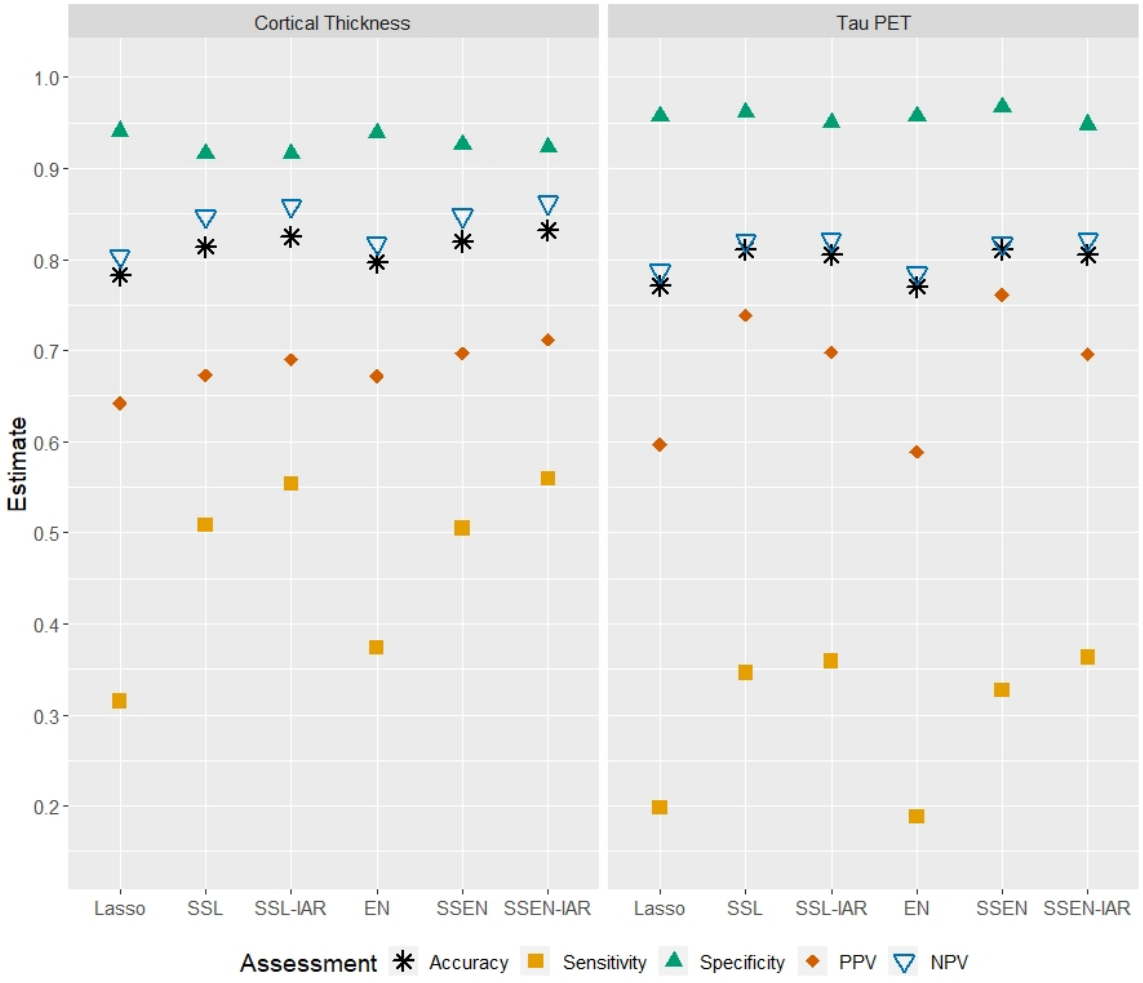

Figure 7: Model classification performance for MCI vs. demented subjects.

5 Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.;Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.;Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

REFERENCES

Arbabshirani, Mohammad R.; Plis, Sergey; Sui, Jing, and Calhoun, Vince D. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 145: 137–165, 2017. ISSN 10959572. doi: 10.1016/j.neuroimage.2016.02.079.

Banerjee, Sudipto; Carlin, Bradley P., and Gelfand, Alan E. *Hierarchical Modeling and*

Analysis for Spatial Data. Chapman & Hall/CRC, Boca Raton, Florida, second edition, 2015.

Besag, Julian. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

Besag, Julian and Kooperberg, Charles. On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746, 1995. doi: 10.1093/biomet/82.4.733.

Bondi, Mark W.; Edmonds, Emily C., and Salmon, David P. Alzheimer's disease: Past, present, and future. *Journal of the International Neuropsychological Society*, 23:818–831, 2017. ISSN 14697661. doi: 10.1017/S135561771700100X.

Braak, Heiko and Del Tredici, Kelly. Are cases with tau pathology occurring in the absence of Aβ deposits part of the AD-related pathological process? *Acta Neuropathologica*, 128 (6):767–772, 2014. ISSN 14320533. doi: 10.1007/s00401-014-1356-1.

Braak, Heiko; Thal, Dietmar R.; Ghebremedhin, Estifanos, and Del Tredici, Kelly. Stages of the pathologic process in alzheimer disease: Age categories from 1 to 100 years. *Journal of Neuropathology and Experimental Neurology*, 70(11), 2011. ISSN 00223069. doi: 10.1097/NEN.0b013e318232a379.

Brosch, Jared R.; Farlow, Martin R.; Risacher, Shannon L., and Apostolova, Liana G. Tau Imaging in Alzheimer's Disease Diagnosis and Clinical Trials. *Neurotherapeutics*, 14: 62–68, 2017. ISSN 18787479. doi: 10.1007/s13311-016-0490-y.

Crary, John F.; Trojanowski, John Q.; Schneider, Julie A.; Abisambra, Jose F.; Abner, Erin L.; Alafuzoff, Irina; Arnold, Steven E.; Attems, Johannes; Beach, Thomas G.; Bigio, Eileen H.; Cairns, Nigel J.; Dickson, Dennis W.; Gearing, Marla; Grinberg, Lea T.; Hof, Patrick R.; Hyman, Bradley T.; Jellinger, Kurt; Jicha, Gregory A.; Kovacs, Gabor G.; Knopman, David S.; Kofler, Julia; Kukull, Walter A.; Mackenzie, Ian R.; Masliah, Eliezer; McKee, Ann; Montine, Thomas J.; Murray, Melissa E.; Neltner, Janna H.;

Santa-Maria, Ismael; Seeley, William W.; Serrano-Pozo, Alberto; Shelanski, Michael L.; Stein, Thor; Takao, Masaki; Thal, Dietmar R.; Toledo, Jonathan B.; Troncoso, Juan C.; Vonsattel, Jean Paul; White, Charles L.; Wisniewski, Thomas; Woltjer, Randall L.; Yamada, Masahito, and Nelson, Peter T. Primary age-related tauopathy (PART): a common pathology associated with human aging. *Acta Neuropathologica*, 128(6):755–766, 2014. ISSN 14320533. doi: 10.1007/s00401-014-1349-0.

Dale, Anders M. and Serano, Martin I. Improved localization of cortical activity by combining eeg and meg with mri cortical surface reconstruction: A linear approach. *Journal of Cognitive Neuroscience*, 5(2):162–176, 1993. doi: 10.1162/jocn.1993.5.2.162.

Dale, Anders M.; Fischl, Bruce, and Sereno, Martin I. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999. ISSN 10538119. doi: 10.1006/nimg.1998.0395.

de Vos, Frank; Koini, Marisa; Schouten, Tijn M.; Seiler, Stephan; va der Grond, Jeroen; Lechner, Anita; Schmidt, Reinhold; de Rooij, Mark, and Rombouts, Serge A.R.B. A comprehensive analysis of resting state fmri measures to classify individual patients with alzheimer's disease. *NeuroImage*, 167:62–72, 2018. doi: 10.1016/j.neuroimage.2017.11.025.

Desikan, Rahul S.; Ségonne, Florent; Fischl, Bruce; Quinn, Brian T.; Dickerson, Bradford C.; Blacker, Deborah; Buckner, Randy L.; Dale, Anders M.; Maguire, R. Paul; Hyman, Bradley T.; Albert, Marilyn S., and Killiany, Ronald J. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, 31(3):968 – 980, 2006. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2006.01.021.

Edmonds, Emily C.; Eppig, Joel; Bondi, Mark W.; Leyden, Kelly M.; Goodwin, Bailey; Delano-Wood, Lisa, and McDonald, Carrie R. Heterogeneous cortical atrophy patterns in

MCI not captured by conventional diagnostic criteria. *Neurology*, 87(20), 2016. ISSN 1526632X. doi: 10.1212/WNL.0000000000003326.

Fischl, Bruce. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2012.01.021.

Fischl, Bruce and Dale, Anders M. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20):11050–11055, 2000. ISSN 00278424. doi: 10.1073/pnas.200033797.

Fischl, Bruce; Sereno, Martin I., and Dale, Anders M. Cortical surface-based analysis: II. Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195–207, 1999. ISSN 10538119. doi: 10.1006/nimg.1998.0396.

Fischl, Bruce; Liu, Arthur, and Dale, Anders M. Automated manifold surgery: Constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Transactions on Medical Imaging*, 20(1):70–80, 2001. ISSN 02780062. doi: 10.1109/42.906426.

Fischl, Bruce; Salat, David H.; Busa, Evelina; Albert, Marilyn; Dieterich, Megan; Haselgrove, Christian; van der Kouwe, Andre; Killiany, Ron; Kennedy, David; Klaveness, Shuna; Montillo, Albert; Makris, Nikos; Rosen, Bruce, and Dale, Anders M. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33:341–355, 2002. doi: 10.1016/s0896-6273(02)00569-x.

Friedman, J.; Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010. doi: 10.18637/jss.v033.i01.

Friedman, Jerome; Hastie, Trevor; Höfling, Holger, and Tibshirani, Robert. Pathwise

coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007. doi: 10.1214/07-AOAS131.

Frisoni, Giovanni B.; Fox, Nick C.; Jack, Clifford R.; Scheltens, Philip, and Thompson, Paul M. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6(2), 2010. ISSN 17594758. doi: 10.1038/nrneurol.2009.215.

George, Edward I. and McCulloch, Robert E. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993. doi: 10.1080/01621459.1993.10476353.

George, Edward I. and McCulloch, Robert E. Approaches for bayesian variable selection. *Statistica Sinica*, 7:339–373, 1997.

Hastie, Trevor; Tibshirani, Robert, and Friedman, Jerome. *The Elements of Statistical Learning*. Springer, 2009.

Jack, Clifford R.; Knopman, David S.; Jagust, William J.; Shaw, Leslie M.; Aisen, Paul S.; Weiner, Michael W.; Petersen, Ronald C., and Trojanowski, John Q. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, 12(2):207–216, 2010. ISSN 14744422. doi: 10.1016/S1474-4422(09)70299-6.

Jack, Clifford R.; Knopman, David S.; Jagust, William J.; Petersen, Ronald C.; Weiner, Michael W.; Aisen, Paul S.; Shaw, Leslie M.; Vemuri, Prashanthi; Wiste, Heather J.; Weigand, Stephen D.; Lesnick, Timothy G.; Pankratz, Vernon S.; Donohue, Michael C., and Trojanowski, John Q. Tracking pathophysiological processes in Alzheimer's disease: An updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, 12(2): 207–216, 2013. ISSN 14744422. doi: 10.1016/S1474-4422(12)70291-0.

Jack, Clifford R.; Bennett, David A.; Blennow, Kaj; Carrillo, Maria C.; Feldman, Howard H.; Frisoni, Giovanni B.; Hampel, Harald; Jagust, William J.; Johnson, Keith A.; Knopman, David S.; Petersen, Ronald C.; Scheltens, Philip; Sperling, Reisa A., and Dubois, Bruno.

A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers. *Neurology*, 87(5):539–547, 2016a. ISSN 1526632X. doi: 10.1212/WNL.0000000000002923.

Jack, Clifford R.; Knopman, David S.; Chételat, Gaël; Dickson, Dennis; Fagan, Anne M.; Frisoni, Giovanni B.; Jagust, William; Mormino, Elizabeth C.; Petersen, Ronald C.; Sperling, Reisa A.; van der Flier, Wiesje M.; Villemagne, Victor L.; Visser, Pieter J., and Vos, Stephanie J. B. Suspected non-Alzheimer disease pathophysiology — concept and controversy. *Nature Reviews Neurology*, 12(2):117–124, 2016b. ISSN 1759-4758. doi: 10.1038/nrneurol.2015.251. URL http://www.nature.com/articles/nrneurol.2015.251.

Lane, C. A.; Hardy, J., and Schott, J. M. Alzheimer's disease. *European Journal of Neurology*, 2018. ISSN 14681331. doi: 10.1111/ene.13439.

Leach, Justin M.; Aban, Inmaculada, and Yi, Nengjun. Incorporating spatial structure into inclusion probabilities for bayesian variable selection. *Unpublished*, 2020.

Mitchell, T.J. and Beauchamp, J.J. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988. doi: 10.1080/01621459.1988.10478694.

Morris, Mitzi; Wheeler-Martin, Katherine; Simpson, Dan; Mooney, Stephen J.; Gelman, Andrew, and DiMaggio, Charles. Bayesian hierarchical spatial models: Implementing the baseg york mollié model in stan. *Spatial and Spatio-temporal Epidemiology*, 31, 2019. doi: 10.1016/j.sste.2019.100301.

Park, Hyunjin; Yang, Jin Ju; Seo, Jongbum, and Lee, Jong Min. Dimensionality reduced cortical features and their use in the classification of Alzheimer's disease and mild cognitive impairment. *Neuroscience Letters*, 529(2):123–127, 2012. ISSN 03043940. doi: 10.1016/j.neulet.2012.09.011.

Park, Trevor and Casella, George. The bayesian lasso. *Journal of the American Statistical*

Association, 103(482):681–686, 2008. doi: 10.1198/016214508000000337.

Petersen, R. C.; Aisen, P. S.; Beckett, L. A.; Donohue, M. C.; Gamst, A. C.; Harvey, D. J.; Jack, C. R.; Jagust, W. J.; Shaw, L. M.; Toga, A. W.; Trojanowski, J. Q., and Weiner, M. W. Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology*, 74:201–209, 2010. ISSN 1526632X. doi: 10.1212/WNL.0b013e3181cb3e25.

Plant, Claudia; Teipel, Stefan J.; Oswald, Annahita; Böhm, Christian; Meindl, Thomas; Mourao-Miranda, Janaina; Bokde, Arun W.; Hampel, Harald, and Ewers, Michael. Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. *NeuroImage*, 50:162–174, 2010. ISSN 10538119. doi: 10.1016/j.neuroimage.2009.11.046.

Rathore, Saima; Habes, Mohamad; Iftikhar, Muhammad Aksam; Shacklett, Amanda, and Davatzikos, Christos. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage*, 155:530–548, 2017. ISSN 10959572. doi: 10.1016/j.neuroimage.2017.03.057.

R Core Team, . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

Ročková, Veronica and George, Edward. The spike and slab lasso. *Journal of the American Statistical Association*, 113:431–444, 2018. doi: 10.1080/01621459.2016.1260469.

Rue, Håvard and Held, Leonhard. *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, Boca Raton, Florida, 2005.

Schouten, Tijn M.; Koini, Marisa; De Vos, Frank; Seiler, Stephan; Van Der Grond, Jeroen; Lechner, Anita; Hafkemeijer, Anne; Möller, Christiane; Schmidt, Reinhold; De Rooij, Mark, and Rombouts, Serge A.R.B. Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate

Alzheimer's disease. *NeuroImage: Clinical*, 11:46–51, 2016. ISSN 22131582. doi: 10.1016/j.nicl.2016.01.002.

Segonne, Florent; Dale, Anders M.; Busa, Evelina; Glessner, Maureen; Salat, David H.; Hahn, Horst K., and Fischl, Bruce. A hybrid approach to the skull stripping problem in mri. *NeuroImage*, 22(3):1060–1075, 2004. doi: 10.1016/j.neuroimage.2004.03.032.

Segonne, Florent; Pacheco, Jenni, and Fischl, Bruce. Geometrically accurate topology-correction of cortical surfacces using nonseparating loops. *IEEE Transactions on Medical Imaging*, 26(4):518–529, 2007. doi: 10.1109/TMI.2006.887364.

Sled, John G.; Zijdenbos, Alex P., and Evans, Alan C. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Transactions on Medical Imaging*, 17(1):87–97, 1998. doi: 10.1109/42.668698.

Tang, Zaixiang; Shen, Yueping; Zhang, Xinyan, and Yi, Nengjun. The spike and slab lasso generalized linear models for prediction and associated genes detection. *Genetics*, 205: 77–88, 2017. doi: 10.1534/genetics.116.192195.

Teipel, Stefan J.; Grothe, Michel J.; Metzger, Coraline D.; Grimmer, Timo; Sorg, Christian; Ewers, Michael; Franzmeier, Nicolai; Meisenzahl, Eva; Kloppel, Stephan; Borchardt, Viola; Walter, Martin, and Dyrba, Martin. Robust detection of impaired resting state functional connectivity networks in alzheimer's disease using elastic net regularized regression. *Frontiers in Aging Neuroscience*, 8(318), 2017. doi: 10.3389/fnagi.2016.00318.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *J.R. Statist. Soc.*, 58 (1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Trzepacz, Paula T.; Hochstetler, Helen; Yu, Peng; Castelluccio, Peter; Witte, Michael M., and Degenhardt, Grazia Dell'Agnello Elisabeth K. Relationship of hippocampal volume

to amyloid burden across diagnostic stages of alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, 41:68–79, 2016. doi: 10.1159/000441351.

Vemuri, Prashanthi; Lowe, Val J.; Knopman, David S.; Senjem, Matthew L.; Kemp, Bradley J.; Schwarz, Christopher G.; Przybelski, Scott A.; Machulda, Mary M.; Petersen, Ronald C., and Jr., Clifford R. Jack. Tau-pet uptake: Regional variation in average suvr and impact of amyloid deposition. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, 6:21–30, 2017. doi: 10.1016/j.dadm.2016.12.010.

Weiner, Michael W.; Veitch, Dallas P.; Aisen, Paul S.; Beckett, Laurel A.; Cairns, Nigel J.; Green, Robert C.; Harvey, Danielle; Jack, Clifford R.; Jagust, William; Morris, John C.; Petersen, Ronald C.; Salazar, Jennifer; Saykin, Andrew J.; Shaw, Leslie M.; Toga, Arthur W., and Trojanowski, John Q. The Alzheimer's Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement, 2017. ISSN 15525279.

Wyman, Bradley T.; Harvey, Danielle J.; Crawford, Karen; Bernstein, Matt A.; Carmichael, Owen; Cole, Patricia E.; Crane, Paul K.; Decarli, Charles; Fox, Nick C.; Gunter, Jeffrey L.; Hill, Derek; Killiany, Ronald J.; Pachai, Chahin; Schwarz, Adam J.; Schuff, Norbert; Senjem, Matthew L.; Suhy, Joyce; Thompson, Paul M.; Weiner, Michael, and Jack, Clifford R. Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimer's and Dementia*, 9(3):332–337, 2013. ISSN 15525279. doi: 10.1016/j.jalz.2012.06.004.

Zou, Hui and Hastie, Trevor. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Methodological)*, 67:301–320, 2005. doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x.

CONCLUSIONS AND FUTURE DIRECTIONS

1 General Summary and Conclusions

This work has covered a wide range of statistical topics, but the central theme throughout is that of modeling scalar outcomes using spatial data. Specifically, we have focused on the role of Bayesian variable selection when the predictors/variables are spatial measurements. Bayesian variable selection has often been employed using so-called spike-and-slab priors, which models parameter values as arising from a mixture distribution of "relevant" or "irrelevant" parameters, whose indicators are assumed drawn from a binomial distribution with some probability of inclusion in the model. Alternatively, penalized approaches, such as the lasso can perform automatic variable selection by shrinking many estimates to exactly zero (Tibshirani, 1996). Ročková and George (2018) combined the spike-and-slab prior framework with the lasso penalty, which allows one to adaptively shrink parameter estimates such that parameters with relatively lower probabilities of inclusion have their estimates shrunk more aggressively than parameters with comparatively higher probabilities of inclusion.

Our primary contribution was to extend the spike-and-slab lasso to better handle situations where parameter values can be expected to cluster spatially, which is generally the case when imaging data are used as predictors. The primary extension explicitly incorporated spatial structure into variable selection by placing an Intrinsic Autoregressive (IAR) prior on the logit of probabilities of inclusion. The practical effect of the IAR prior on the logit of inclusion probabilities is that the shrinkage applied to a given estimate will in part depend on the shrinkage applied to spatially proximate parameters. We fit this model by extending the

EM algorithm in Tang et al. (2017), which was proposed to fit spike-and-slab lasso GLM's. In addition, we tweaked this algorithm to handle the full range of the elastic net, of which the lasso is a special case. The models can be fit using the `R` package `ssnet`, which builds upon the `R` packages `BhGLM` and `glmnet`.

We also built an `R` package, `sim2Dpredictr`, which can flexibly simulate a wide range of spatial data using variations on the multivariate Normal distribution (MVN) to generate continuous-valued images, and an adaptation of the Boolean method to generate a wide range of binary images. The images are then used to generate data within a GLM framework, and there are additionally functions that allow one to easily obtain estimates of FDR, FWER, and Power from a set of simulations, and functions for creating spatially clustered "true" parameter values. This package allowed us to conduct a simulation study tailored to our exact research questions, and we believe it will be useful to other researchers interested in similar problems.

We used `sim2Dpredictr` to run a simulation study where "relevant" parameters were clustered spatially, and found that compared to traditional elastic net and spike-and-slab models without IAR priors, the models that included IAR priors had consistently lower average prediction error estimates, as estimated by 10-fold cross validation. AUC estimates were consistently higher for the models with spatial structure, and more noticeably so when the "true" magnitude of parameters was relatively small; this is important because many practical situations will have relatively small effects for changes in measures at each spatial location. Raw estimates for false discovery rate (FDR) and the proportion of "relevant" parameters (Power) kept in the model were generally poor, but on average estimates for "irrelevant" parameters were very close to zero (< 0.003 in all cases), and the majority of parameter were shrunk to zero in any given case, which helped limit noise, and allowed the "relevant" parameters that were included in a model to dominate the predicted values for any given set of predictors. The spike-and-slab elastic net models also tended to shrink

"irrelevant" parameter's estimates more severely than traditional methods, while on average producing estimates for "relevant" parameters that were closer to the "true" values. Finally, while on average the spike-and-slab elastic net models were more biased, the estimates were less variable than the spike-and-slab lasso estimates, which likely accounts for the comparatively superior prediction errors and shows the utility of generalizing the spike-and-slab lasso to the spike-and-slab elastic net when predictors are spatial measurements.

The results of the simulation study led us to focus less on traditional measures of variable selection, e.g., controlling FDR and maximizing Power, and more on the ability of proposed methods to generalize well, as evidenced by the high estimates for AUC and low estimates for prediction error, e.g., model deviance or mean squared error (MSE). A statistical problem that prioritizes generalizeability of models is classification. GLM's contain logistic regression as special case, and for any given set of predictors and parameter estimates, one can obtain predicted probabilities, e.g., the probability that a subject has a disease. These probabilities can then be thresholded to classify subjects as either having or not having the disease.

We used neuroimaging data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to compare the classification ability of the traditional elastic net, the spike-and-slab elastic net without spatial structure, and the spike-and-slab elastic net with IAR priors. Across several classification scenarios, cognitive normal (CN) vs. dementia, CN vs. mild cognitive impairment (MCI), and MCI vs. dementia, all models performed best for CN vs. dementia. While the spike-and-slab models tended to outperform the traditional models with respect to accuracy, specificity, positive predictive value (PPV), and negative predictive value (NPV), the performance was not drastically different across models. However, the spike-and-slab models showed noticeably higher sensitivity in each case examined, and the models that additionally had IAR priors had the highest sensitivity for a given scenario in most cases, although the difference between the spike-and-slab and traditional models was

generally larger than that between the spike-and-slab models with and without IAR priors on the inclusion probabilities.

These analyses show that compared to traditional elastic net classifiers, the spike-and-slab elastic net can improve classification performance, especially with respect to sensitivity. In addition, the data used consisted of measurements on the Desikan-Killiany atlas, which contains only 68 regions. Since the spike-and-slab elastic net with IAR priors is specifically designed to handle high dimensional data and given the results of the simulation study, it is reasonable to predict that the models with IAR priors would have greater separation from the other models for data with higher dimensionality.

2 Future Directions

There are several future research directions available, and we detail a few of them here. In terms of explicitly statistical directions, we only applied our extensions to GLM's, but these priors could be applied to survival analysis or multinomial regression, both of which would require adapting related algorithms. The latter situation, multinomial regression, may have specific potential in classification problems. When limited to a Bayesian logistic regression, we have to model each pair separately, but a multinomial regression would allow to simultaneous classification of, say CN, MCI, and demented subjects.

In addition, this dissertation has only scratched the surface of potential practical applications of the developed methods. As mentioned above, we used average measurements within regions of the Desikan-Killiany atlas, which reduces several thousand measurements to 68 per subject. We learned useful lessons analyzing this data, but the methods are designed to be able to handle higher dimensional data. Thus, it would be beneficial to explore the methods' performance using data that has not been reduced to an atlas, or at the very least has been reduced such that there are still several thousand, or even several hundred, measurements. It is possible, if not likely, that the spike-and-slab elastic net with IAR priors on logit inclusion probabilities would reveal their full potential in such a setting.

Another practical application we did not explore was predicting progression to a more severe disease state based on current neuroimaging biomarkers. It is possible that the current methods would prove useful in this setting. Especially in this setting, it is possible that adapting the methods to handle longitudinal data would be useful in predicting disease progression.

3 Final Comment

In summary, the overall contribution of this dissertation is to show that the spike-and-slab elastic net, and its extensions with IAR priors on the logit probabilities of inclusion, have the potential to improve the generalizeability of models applied to problems in imaging statistics. In addition, there are many practical future directions of research to improve and apply the methods.

GENERAL LIST OF REFERENCES

Aguirre, G.; Zarahn, E., and D'Esposito, M. The variability of human, bold hemodynamic responses. *NeuroImage*, 8(4):360–369, 1998. doi: 10.1006/nimg.1998.0369.

Amaro Jr., Edson and Barker, Gareth J. Study design in fmri: Basic principles. *Brain and Cognition*, 60(3):220–232, 2006. doi: 10.1006/nimg.1998.0369.

Andrews, D.L. and Mallows, C.L. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society B*, 36(1):99–102, 1974. doi: https://doi.org/10.1111/j.2517-6161.1974.tb00989.x.

Arbabshirani, Mohammad R.; Plis, Sergey; Sui, Jing, and Calhoun, Vince D. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 145: 137–165, 2017. ISSN 10959572. doi: 10.1016/j.neuroimage.2016.02.079.

Banerjee, Sudipto; Carlin, Bradley P., and Gelfand, Alan E. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, Florida, 2nd edition, 2015.

Bates, Elizabeth; Wilson, Stephen M.; Saygin, Ayse Pinar; Dick, Frederic; Sereno, Martin I.; Knight, Robert T., and Dronkers, Nina F. Voxel-based lesion-symptom mapping. *Nature Neuroscience*, 6:448–450, 2003. doi: 10.1038/nn1050.

Benjamini, Yoav and Hochberg, Yosef. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Benjamini, Yoav and Yekutieli, Daniel. The control of the false discovery rate in multiple

testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. doi: 10.1214/aos/1013699998.

Bennett, Craig M.; Wolford, George L., and Miller, Michael B. The principled control of false postives in neuroimaging. *SCAN*, 4:417–422, 2009. doi: 10.1093/scan/nsp053.

Berger, James O. Could fisher, jeffreys, and neyman have agreed on testing? *Statistical Science*, 18(1):1–32, 2003. doi: 10.1214/ss/1056397485.

Berger, James O. and Wolpert, Robert L. *The Likelihood Principle*, volume 6 of *Lecture Notes-Monograph Series*. Institute of Mathematical Statistics, 2 edition, 1988.

Berry, Donald and Hochberg, Yosef. Bayesian perspectives on multiple comparisons. *Journal Of Statistical Planning and Inference*, 82:215–227, 1999. doi: 10.1016/S0378-3758(99)00044-0.

Berry, Kenneth; Mielke, Paul, and Howard, Mielke. The fisher-pitman permutation test: An attractive alternative to the f test. *Psychological Reports*, 90:495–502, 2002. doi: 10.2466/pr0.2002.90.2.495.

Besag, Julian. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

Besag, Julian and Kooperberg, Charles. On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746, 1995. doi: 10.1093/biomet/82.4.733.

Best, A.; Greenberg, B., and Glick, M. From tea tasting to t test: a p value ain't what you think it is. *JADA*, 147(7), 2016. doi: 10.1016/j.adaj.2016.05.004.

Bilmes, Jeff. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, International Computer Science Institute, 1998.

Bondi, Mark W.; Edmonds, Emily C., and Salmon, David P. Alzheimer's disease: Past,

present, and future. *Journal of the International Neuropsychological Society*, 23:818–831, 2017. ISSN 14697661. doi: 10.1017/S135561771700100X.

Bowman, F. DuBois; Caffo, Brian; Spear, Bassett Susan, and Kilts, Clinton. A bayesian hierarchical framework for spatial modeling of fmri data. *NeuroImage*, 39:146–156, 2008. doi: 10.1016/j.neuroimage.2007.08.012.

Braak, Heiko and Del Tredici, Kelly. Are cases with tau pathology occurring in the absence of Aβ deposits part of the AD-related pathological process? *Acta Neuropathologica*, 128 (6):767–772, 2014. ISSN 14320533. doi: 10.1007/s00401-014-1356-1.

Braak, Heiko; Thal, Dietmar R.; Ghebremedhin, Estifanos, and Del Tredici, Kelly. Stages of the pathologic process in alzheimer disease: Age categories from 1 to 100 years. *Journal of Neuropathology and Experimental Neurology*, 70(11), 2011. ISSN 00223069. doi: 10.1097/NEN.0b013e318232a379.

Bradley, James. *Distribution-Free Statistical Tests.* Prentice-Hall, Inc., Edgewood Cliffs, N.J., 1968.

Brosch, Jared R.; Farlow, Martin R.; Risacher, Shannon L., and Apostolova, Liana G. Tau Imaging in Alzheimer's Disease Diagnosis and Clinical Trials. *Neurotherapeutics*, 14: 62–68, 2017. ISSN 18787479. doi: 10.1007/s13311-016-0490-y.

Brown, D. Andrew; Lazar, Nicole A.; Datta, Gauri S.; Jang, Woncheol, and McDowell, Jennifer. Incorporating spatial dependence into bayesian multiple testing of statistical parametric maps in functional neuroimaging. *NeuroImage*, 84:97–112, 2014. doi: 10.1016/j.neuroimage.2013.08.024.

Casella, George and Berger, Roger L. *Statistical Inference, 2nd Ed.* The Wadsworth Group, 2002.

Crary, John F.; Trojanowski, John Q.; Schneider, Julie A.; Abisambra, Jose F.; Abner, Erin L.; Alafuzoff, Irina; Arnold, Steven E.; Attems, Johannes; Beach, Thomas G.; Bigio,

Eileen H.; Cairns, Nigel J.; Dickson, Dennis W.; Gearing, Marla; Grinberg, Lea T.; Hof, Patrick R.; Hyman, Bradley T.; Jellinger, Kurt; Jicha, Gregory A.; Kovacs, Gabor G.; Knopman, David S.; Kofler, Julia; Kukull, Walter A.; Mackenzie, Ian R.; Masliah, Eliezer; McKee, Ann; Montine, Thomas J.; Murray, Melissa E.; Neltner, Janna H.; Santa-Maria, Ismael; Seeley, William W.; Serrano-Pozo, Alberto; Shelanski, Michael L.; Stein, Thor; Takao, Masaki; Thal, Dietmar R.; Toledo, Jonathan B.; Troncoso, Juan C.; Vonsattel, Jean Paul; White, Charles L.; Wisniewski, Thomas; Woltjer, Randall L.; Yamada, Masahito, and Nelson, Peter T. Primary age-related tauopathy (PART): a common pathology associated with human aging. *Acta Neuropathologica*, 128(6):755–766, 2014. ISSN 14320533. doi: 10.1007/s00401-014-1349-0.

Craven, Peter and Wahba, Grace. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979. doi: 10.1007/BF01404567.

Cressie, Noel and Wikle, Christopher K. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, Hoboken, New Jersey, 2011.

Dale, Anders M. Optimal experimental design for event-related fmri. *Human Brain Mapping*, 8:109–114, 1999. doi: 10.1002/(SICI)1097-0193(1999)8:2/3<109::AID-HBM7>3.0.CO;2-W.

Dale, Anders M. and Serano, Martin I. Improved localization of cortical activity by combining eeg and meg with mri cortical surface reconstruction: A linear approach. *Journal of Cognitive Neuroscience*, 5(2):162–176, 1993. doi: 10.1162/jocn.1993.5.2.162.

Dale, Anders M.; Fischl, Bruce, and Sereno, Martin I. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999. ISSN 10538119. doi: 10.1006/nimg.1998.0395.

Davison, A.C. and Hinkley, D.V. *Bootstrap Methods and their Application*. Cambridge University Press, 1997.

de Vos, Frank; Koini, Marisa; Schouten, Tijn M.; Seiler, Stephan; va der Grond, Jeroen; Lechner, Anita; Schmidt, Reinhold; de Rooij, Mark, and Rombouts, Serge A.R.B. A comprehensive analysis of resting state fmri measures to classify individual patients with alzheimer's disease. *NeuroImage*, 167:62–72, 2018. doi: 10.1016/j.neuroimage.2017.11.025.

Dempster, A.P.; Laird, N.M., and Rubin, D.B. Maximum likelihood from incomplete data via the em algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B*, 39:1–38, 1977.

Desikan, Rahul S.; Ségonne, Florent; Fischl, Bruce; Quinn, Brian T.; Dickerson, Bradford C.; Blacker, Deborah; Buckner, Randy L.; Dale, Anders M.; Maguire, R. Paul; Hyman, Bradley T.; Albert, Marilyn S., and Killiany, Ronald J. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, 31(3):968 – 980, 2006. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2006.01.021.

Diggle, Peter J. *Time Series: A Biostatistical Introduction*. Oxford University Press, New York, New York, 1990.

Diggle, Peter J.; Heagerty, Patrick; Liang, Kung-Yee, and Zeger, Scott L. *Analysis of Longitudinal Data*. Oxford University Press, Oxford, UK, 2nd edition, 2002.

Dinov, Ivo D.; Boscardin, John W.; Mega, Michael S.; Sowell, Elizabeth L., and Toga, Arthur W. A wavelet-based statistical analysis of fmri data. *Neuroinformatics*, 3, 2005. doi: 10.1385/NI:03:04:1.

Edmonds, Emily C.; Eppig, Joel; Bondi, Mark W.; Leyden, Kelly M.; Goodwin, Bailey; Delano-Wood, Lisa, and McDonald, Carrie R. Heterogeneous cortical atrophy patterns in MCI not captured by conventional diagnostic criteria. *Neurology*, 87(20), 2016. ISSN 1526632X. doi: 10.1212/WNL.0000000000003326.

Efron, Bradley and Tibshirani, Robert J. *An Introduction to the Bootstrap*. Chapman and

Hall, New York, 1993.

Farcomeni, Alessio. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 17: 347–388, 2008. doi: 10.1177/0962280206079046.

Fischl, Bruce. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2012.01.021.

Fischl, Bruce and Dale, Anders M. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20):11050–11055, 2000. ISSN 00278424. doi: 10.1073/pnas.200033797.

Fischl, Bruce; Sereno, Martin I., and Dale, Anders M. Cortical surface-based analysis: II. Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195–207, 1999. ISSN 10538119. doi: 10.1006/nimg.1998.0396.

Fischl, Bruce; Liu, Arthur, and Dale, Anders M. Automated manifold surgery: Constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Transactions on Medical Imaging*, 20(1):70–80, 2001. ISSN 02780062. doi: 10.1109/42.906426.

Fischl, Bruce; Salat, David H.; Busa, Evelina; Albert, Marilyn; Dieterich, Megan; Haselgrove, Christian; van der Kouwe, Andre; Killiany, Ron; Kennedy, David; Klaveness, Shuna; Montillo, Albert; Makris, Nikos; Rosen, Bruce, and Dale, Anders M. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33:341–355, 2002. doi: 10.1016/s0896-6273(02)00569-x.

Friedman, J.; Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010. doi: 10.18637/jss.v033.i01.

Friedman, Jerome; Hastie, Trevor; Höfling, Holger, and Tibshirani, Robert. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007. doi: 10.1214/07-AOAS131.

Frisoni, Giovanni B.; Fox, Nick C.; Jack, Clifford R.; Scheltens, Philip, and Thompson, Paul M. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6(2), 2010. ISSN 17594758. doi: 10.1038/nrneurol.2009.215.

Friston, Karl; Jezzard, Peter, and Turner, Robert. Analysis of functional mri time-series. *Human Brain Mapping*, 1:153–171, 1994. doi: 10.1002/hbm.460010207.

Friston, Karl and others, . Comparing function (pet) images: The assessment of significant change. *Journal of Cerebral Blood Flow and Metabolism*, 11:690–699, 1991. doi: 10.1038/jcbfm.1991.122.

Friston, K.J.; Holmes, A.P.; Poline, J.B.; Grasby, P.J.; Williams, S.C.; Frackowiak, R.S., and Turner, R. Analysis of fmri time-series revisited. *Neuroimage*, 2(1):45–53, 1995. doi: 10.1006/nimg.1995.1007.

Furrer, Reinhard and Sain, Stephan R. spam: A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields. *Journal of Statistical Software*, 36 (10):1–25, 2010. URL http://www.jstatsoft.org/v36/i10/.

Gelman, Andrew. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534, 2006. doi: 10.1214/06-BA117A.

Gelman, Andrew and Shalizi, Cosma R. Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66:8–38, 2013. doi: 10.1111/j.2044-8317.2011.02037.x.

Gelman, Andrew; Jakulin, Aleks; Pittau, Maria Grazia, and Su, Yu-Sung. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008. doi: 10.1214/08-AOAS191.

Gelman, Andrew; Hill, Jennifer, and Yajima, Masanao. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5: 189–211, 2012. doi: 10.1080/19345747.2011.618213.

Gelman, Andrew; Carlin, John B.; Stern, Hal S.; Dunson, David B.; Vehtari, Aki, and Rubin, Donald B. *Bayesian Data Analysis, 3rd E.* Taylor and Francis Group, LLC., 2013.

Genovese, Christopher and Wasserman, Larry. Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101(476):1408–1417, 2006. doi: 10.1198/016214506000000339.

Genovese, Christopher; Lazar, Nicole, and Nichols, Thomas. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15:870–878, 2002. doi: 10.1006/nimg.2001.1037.

George, Edward I. and McCulloch, Robert E. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993. doi: 10.1080/01621459.1993.10476353.

George, Edward I. and McCulloch, Robert E. Approaches for bayesian variable selection. *Statistica Sinica*, 7:339–373, 1997.

Glover, Gary H. Deconvolution of impulse response in event-related bold fmri. *NeuroImage*, 42(2):416–429, 1999. doi: 10.1006/nimg.1998.0419.

Golub, Gene H.; Heath, Michael, and Wahba, Grace. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979. doi: 10.1080/00401706.1979.10489751.

Goutte, Cyril; Nielsen, Finn Årup, and Hansen, Lars Kai. Modeling the haemodynamic response in fmri using smooth fir filters. *IEEE Transactions on Medical Imaging*, 19(12): 1188–1201, 2000. doi: 10.1109/42.897811.

Greenland, Sander. Bayesian perspectives for epidemiological research: I. foundations and basic methods. *International Journal of Epidemiology*, 35:765–775, 2006. doi: 10.1093/ije/dyi312.

Greenland, Sander; Senn, Stephen J.; Rothman, Kenneth J.; Carlin, John B.; Poole, Charles; Goodman, Steven N., and Altman, Douglas G. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31:337–350, 2016. doi: 10.1007/s10654-016-0149-3.

Gräler, Benedikt; Pebesma, Edzer, and Heuvelink, Gerard. Spatio-temporal interpolation using `gstat`. *The R Journal*, 8:204–218, 2016. URL https://journal.r-project.org/archive/2016/RJ-2016-014/index.html.

Hastie, Trevor; Tibshirani, Robert, and Friedman, Jerome. *The Elements of Statistical Learning*. Springer, 2009.

Häusser, Michael and Mel, Bartlett. Dendrites: bug or feature? *Current Opinion in Neurobiology*, 13(3):372–383, 2003. doi: 10.1016/S0959-4388(03)00075-8.

Hedeker, Donald and Gibbons, Robert D. *Longitudinal Data Analysis*. John Wiley & Sons, Hoboken, New Jersey, 2006. doi: 10.1002/0470036486.

Hochberg, Yosef. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–802, 1988. doi: 10.1093/biomet/75.4.800.

Hochberg, Yosef and Tamhane, Ajit. *Multiple Comparisons Procedures*. John Wiley and Sons, Inc., 1987.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

Holmes, A.; Blair, R.; Watson, D., and Ford, I. Nonparametric analysis of statistic images

from functional mapping experiments. *Journal of Cerebral Blood Flow and Metabolism*, 16:7–22, 1996. doi: 10.1097/00004647-199601000-00002.

Hubbard, Raymond and Lindsay, R. Murray. Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, 18(1):69–88, 2008. doi: 10.1177/0959354307086923.

Jack, Clifford R.; Knopman, David S.; Jagust, William J.; Shaw, Leslie M.; Aisen, Paul S.; Weiner, Michael W.; Petersen, Ronald C., and Trojanowski, John Q. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, 12(2):207–216, 2010. ISSN 14744422. doi: 10.1016/S1474-4422(09)70299-6.

Jack, Clifford R.; Knopman, David S.; Jagust, William J.; Petersen, Ronald C.; Weiner, Michael W.; Aisen, Paul S.; Shaw, Leslie M.; Vemuri, Prashanthi; Wiste, Heather J.; Weigand, Stephen D.; Lesnick, Timothy G.; Pankratz, Vernon S.; Donohue, Michael C., and Trojanowski, John Q. Tracking pathophysiological processes in Alzheimer's disease: An updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, 12(2): 207–216, 2013. ISSN 14744422. doi: 10.1016/S1474-4422(12)70291-0.

Jack, Clifford R.; Bennett, David A.; Blennow, Kaj; Carrillo, Maria C.; Feldman, Howard H.; Frisoni, Giovanni B.; Hampel, Harald; Jagust, William J.; Johnson, Keith A.; Knopman, David S.; Petersen, Ronald C.; Scheltens, Philip; Sperling, Reisa A., and Dubois, Bruno. A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers. *Neurology*, 87(5):539–547, 2016a. ISSN 1526632X. doi: 10.1212/WNL.0000000000002923.

Jack, Clifford R.; Knopman, David S.; Chételat, Gaël; Dickson, Dennis; Fagan, Anne M.; Frisoni, Giovanni B.; Jagust, William; Mormino, Elizabeth C.; Petersen, Ronald C.; Sperling, Reisa A.; van der Flier, Wiesje M.; Villemagne, Victor L.; Visser, Pieter J., and Vos, Stephanie J. B. Suspected non-Alzheimer disease pathophysiology — concept and

controversy. *Nature Reviews Neurology*, 12(2):117–124, 2016b. ISSN 1759-4758. doi: 10.1038/nrneurol.2015.251. URL http://www.nature.com/articles/nrneurol.2015.251.

Jeffreys, William H. and Berger, James O. Ockham's razor and bayesian analysis. *American Scientist*, 80(1):64–72, 1992.

Jin, Xiaoping; Carlin, Bradley P., and Banerjee, Sudipto. Generalized hierarchical multivariate car models for areal data. *Biometrics*, 61(4):950–961, 2005. doi: 10.1111/j.1541-0420.2005.00359.x.

Joober, Ridha; Schmitz, Norbert; Annable, Lawrence, and Boksa, Patricia. Publication bias: What are the challenges and can they be overcome? *Journal of Psychiartry Neuroscience*, 37(3), 2012. doi: 10.1503/jpn.120065.

Joseph, Max. Exact sparse car models in stan. *Stan Case Studies*, 3, 2016. URL https://mc-stan.org/users/documentation/case-studies/mbjoseph-CARStan.html.

Koehler, Elizabeth; Brown, Elizabeth, and Haneuse, Sebastien. On the assessment of monte carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2): 155–162, 2009. doi: 10.1198/tast.2009.0030.

Künsch, Hans R. Intrinsic autoregressions and related models on the two-dimensional lattice. *Biometrika*, 74(3):517–524, 1987. doi: 10.2307/2337341.

Lane, C. A.; Hardy, J., and Schott, J. M. Alzheimer's disease. *European Journal of Neurology*, 2018. ISSN 14681331. doi: 10.1111/ene.13439.

Liang, Kung-Yee and Zeger, Scott L. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986. doi: 10.1093/biomet/73.1.13.

Lieberman, Matthew D. and Cunningham, William A. Type i and type ii error concerns in fmri research: re-balancing the scale. *SCAN*, 4:423–428, 2009. doi: 10.1093/scan/nsp052.

Lindley, Dennis V. The philosophy of statistics. *The Statistician*, 49(3):293–337, 2000. doi: 10.1111/1467-9884.00238.

Lindquist, Martin. The statistical analysis of fmri data. *Statistical Science*, 23(4):439–464, 2008. doi: 10.1214/09-STS282.

Lindquist, Martin and Meija, Amanda. Zen and the art of multiple comparisons. *Psychosom Med.*, 77(2):114–125, 2015. doi: 10.1097/PSY.0000000000000148.

Liu, Thomas T. The development of event-related fmri designs. *Neuroimage*, 62(2): 1157–1162, 2012. doi: 10.1097/00004647-199601000-00002.

Liu, Thomas T.; Frank, Lawrence R.; Wong, Eric C., and Buxton, Richard B. Detection power, estimation efficiency, and predictability in event-related fmri. *NeuroImage*, 13(4): 759–773, 2001. doi: 10.1006/nimg.2000.0728.

Loh, Ji Meng; Lindquist, Martin A., and Wager, Tor D. Residual analysis for detecting mis-modeling in fmri. *Statistica Sinica*, 18:1421–1448, 2008.

Long, Michael; Berry, Kenneth, and Mielke, Paul. A note on permutation tests of significance for multiple regression coefficients. *Psychological Reports*, 100:339–345, 2007. doi: 10.2466/pr0.100.2.339-345.

Marquardt, Donald W. and See, Ronald D. Ridge regression in practice. *The American Statistician*, 29(1):3–20, 1975. doi: 10.2307/2683673.

Mayo, Deborah. *Error and the Growth of Experimental Knowledge*. The University of Chicago Press, 1996.

McCulloch, Charles E.; Searle, Shayle R., and Neuhaus, John M. *Generalized, Linear, and Mixed Models, Second Edition*. John Wiley and Sons, Hoboken, New Jersey, 2008.

Mezrich, Reuben. A perspective on k-space. *Radiology*, 195(2):297–315, 1995. doi: 10.1148/radiology.195.2.7724743.

Mirman, Daniel; Landrigan, Jon-Frederick; Kokolis, Spiro; Verillo, Sean; Ferrara, Casey, and Pustina, Dorian. Corrections for multiple comparisons in voxel-based lesion-symptom mapping. *Neuropsychologia*, 115:112–123, 2018. doi: 10.1016/j.neuropsychologia.2017.08.025.

Mitchell, T.J. and Beauchamp, J.J. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988. doi: 10.1080/01621459.1988.10478694.

Morris, Mitzi. Spatial models in stan: Intrinsic auto-regressive models for areal data. *Stan Case Studies*, 4, 2017. URL https://mc-stan.org/users/documentation/case-studies/icar_stan.html.

Morris, Mitzi; Wheeler-Martin, Katherine; Simpson, Dan; Mooney, Stephen J.; Gelman, Andrew, and DiMaggio, Charles. Bayesian hierarchical spatial models: Implementing the baseg york mollié model in stan. *Spatial and Spatio-temporal Epidemiology*, 31, 2019a. doi: 10.1016/j.sste.2019.100301.

Morris, Tim P.; White, Ian R., and Crowther, Michael J. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38:2074–2102, 2019b. doi: 10.1002/sim/8086.

Myers, Raymond H. and Milton, Janet S. *A First Course in the Theory of Linear Statistical Models*. The McGraw-Hill Companies, Inc., 1998.

Nichols, Thomas and Hayasaka, Satoru. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, 12:419–446, 2003. doi: 10.1191/0962280203sm341ra.

Nichols, Thomas and Holmes, Andrew. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15:1–25, 2001. doi: 10.1002/hbm.1058.

Ombao, Hernando; Lindquist, Martin; Thompson, Wesley, and Aston, John, editors. *Handbook of Neuroimaging Data Analysis*. Chapman & Hall, 2017.

Park, Hyunjin; Yang, Jin Ju; Seo, Jongbum, and Lee, Jong Min. Dimensionality reduced cortical features and their use in the classification of Alzheimer's disease and mild cognitive impairment. *Neuroscience Letters*, 529(2):123–127, 2012. ISSN 03043940. doi: 10.1016/j.neulet.2012.09.011.

Park, Trevor and Casella, George. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008. doi: 10.1198/016214508000000337.

Pebesma, Edzer J. Multivariable geostatistics in `S`: the `gstat` package. *Computers & Geosciences*, 30:683–691, 2004. doi: 10.1016/j.cageo.2004.03.012.

Petersen, R. C.; Aisen, P. S.; Beckett, L. A.; Donohue, M. C.; Gamst, A. C.; Harvey, D. J.; Jack, C. R.; Jagust, W. J.; Shaw, L. M.; Toga, A. W.; Trojanowski, J. Q., and Weiner, M. W. Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology*, 74:201–209, 2010. ISSN 1526632X. doi: 10.1212/WNL.0b013e3181cb3e25.

Plant, Claudia; Teipel, Stefan J.; Oswald, Annahita; Böhm, Christian; Meindl, Thomas; Mourao-Miranda, Janaina; Bokde, Arun W.; Hampel, Harald, and Ewers, Michael. Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. *NeuroImage*, 50:162–174, 2010. ISSN 10538119. doi: 10.1016/j.neuroimage.2009.11.046.

Rathore, Saima; Habes, Mohamad; Iftikhar, Muhammad Aksam; Shacklett, Amanda, and Davatzikos, Christos. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage*, 155:530–548, 2017. ISSN 10959572. doi: 10.1016/j.neuroimage.2017.03.057.

`R` Core Team, . *`R`: A Language and Environment for Statistical Computing*. `R` Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

Ripley, Brian D. *Stochastic Simulation*. John Wiley & Sons, 1987. doi: 10.1002/9780470316726.

Roiland, Justin. Rick potion no. 9, Jan 2014.

Romano, J. and Wolf, M. Control of generalized error rates in multiple testing. *The Annals of Statistics*, 35(4):1378–1408, 2007. doi: 10.1214/009053606000001622.

Rosenow, Felix; Klein, Karl Martin, and Hamer, Hajo M. Non-invasive eeg evaluation in epilepsy diagnosis. *Expert Review of Neurotherapeutics*, 15(4):425–444, 2015. doi: 10.1586/14737175.2015.1025382.

Rothman, Kennth. No adjustments are needed for multiple comparisons. *Epidemiology*, 1 (1):43–46, 1990. doi: 10.1097/00001648-199001000-00010.

Ročková, Veronica and George, Edward. Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846, 2014. doi: 10.1080/01621459.2013.869223.

Ročková, Veronica and George, Edward. The spike and slab lasso. *Journal of the American Statistical Association*, 113:431–444, 2018. doi: 10.1080/01621459.2016.1260469.

Rue, Håvard. Fast sampling of gaussian markov random fields. *J.R. Statist. Soc. B*, 63: 325–338, 2001.

Rue, Håvard and Held, Leonhard. *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, Boca Raton, Florida, 2005.

Saini, Jagriti and Dutta, Maitreyee. An extensive review on development of eeg-based computer-aided diagnosis systems for epilepsy detection. *Network: Computation in Neural Systems*, 28(1):1–27, 2017. doi: 10.1080/0954898X.2017.1325527.

Schacter, Daniel L.; Buckner, Randy L.; Koutstaal, Wilma; Dale, Anders M., and Rosen,

Bruce M. Late onset of anterior prefrontal activity during true and false recognition: An event-related fmri study. *NeuroImage*, 6(4):259–269, 1997. doi: 10.1006/nimg.1997.0305.

Schlather, Martin; Malinowski, Alexander; Menck, Peter J.; Oesting, Marco, and Strokorb, Kirstin. Analysis, simulation and prediction of multivariate random fields with package `RandomFields`. *Journal of Statistical Software*, 63(8):1–25, 2015. URL http://www.jstatsoft.org/v63/i08/.

Schouten, Tijn M.; Koini, Marisa; De Vos, Frank; Seiler, Stephan; Van Der Grond, Jeroen; Lechner, Anita; Hafkemeijer, Anne; Möller, Christiane; Schmidt, Reinhold; De Rooij, Mark, and Rombouts, Serge A.R.B. Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease. *NeuroImage: Clinical*, 11:46–51, 2016. ISSN 22131582. doi: 10.1016/j.nicl.2016.01.002.

Scott, James G. and Berger, James O. An exploration of aspects of bayesian multiple testing. *Journal of Statistical Planning and Inference*, 126:2144–2162, 2006. doi: 10.1016/j.jspi.2005.08.031.

Scott, James G. and Berger, James O. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619, 2010.

Segonne, Florent; Dale, Anders M.; Busa, Evelina; Glessner, Maureen; Salat, David H.; Hahn, Horst K., and Fischl, Bruce. A hybrid approach to the skull stripping problem in mri. *NeuroImage*, 22(3):1060–1075, 2004. doi: 10.1016/j.neuroimage.2004.03.032.

Segonne, Florent; Pacheco, Jenni, and Fischl, Bruce. Geometrically accurate topology-correction of cortical surfacces using nonseparating loops. *IEEE Transactions on Medical Imaging*, 26(4):518–529, 2007. doi: 10.1109/TMI.2006.887364.

Sled, John G.; Zijdenbos, Alex P., and Evans, Alan C. A nonparametric method for automatic

correction of intensity nonuniformity in mri data. *IEEE Transactions on Medical Imaging*, 17(1):87–97, 1998. doi: 10.1109/42.668698.

Soric, Branko. Statistical "discoveries" and effect size estimation. *Journal of the American Statistical Association*, 84:608–610, 1989. doi: 10.1214/10-AOS792.

Storey, John D. The positive false discovery rate: A bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003. doi: 10.1214/aos/1074290335.

Tang, Zaixiang; Shen, Yueping; Zhang, Xinyan, and Yi, Nengjun. The spike and slab lasso generalized linear models for prediction and associated genes detection. *Genetics*, 205: 77–88, 2017. doi: 10.1534/genetics.116.192195.

Tang, Zaixiang; Shen, Yueping; Li, Yan; Zhang, Xinyan; Wen, Jia; Qian, Chen'ao; Zhuang, Wenzhou; Shi, Xinghua, and Yi, Nengjun. Group spike and slab lasso generalized linear models for disease prediction and associated genes detection by incorporating pathway information. *Bioinformatics*, 34(6):901–910, 2018. doi: 10.1093/bioinformatics/btx684.

Teh, Y. Dirichlet process. *Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning.*, 2011. doi: 10.1007/978-1-4899-7687-1_219.

Teipel, Stefan J.; Grothe, Michel J.; Metzger, Coraline D.; Grimmer, Timo; Sorg, Christian; Ewers, Michael; Franzmeier, Nicolai; Meisenzahl, Eva; Kloppel, Stephan; Borchardt, Viola; Walter, Martin, and Dyrba, Martin. Robust detection of impaired resting state functional connectivity networks in alzheimer's disease using elastic net regularized regression. *Frontiers in Aging Neuroscience*, 8(318), 2017. doi: 10.3389/fnagi.2016.00318.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *J.R. Statist. Soc.*, 58 (1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Trafimow, D. Editorial. *Basic and Applied Social Psychology*, 36(1):1–2, 2014.

Trafimow, D. and Marks, M. Editorial. *Basic and Applied Social Psychology*, 37:1–2, 2015.

Trappenberg, Thomas. *Fundamentals of Computational Neuroscience*. Oxford University Press Inc., New York, 2 edition, 2010.

Trzepacz, Paula T.; Hochstetler, Helen; Yu, Peng; Castelluccio, Peter; Witte, Michael M., and Degenhardt, Grazia Dell'Agnello Elisabeth K. Relationship of hippocampal volume to amyloid burden across diagnostic stages of alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, 41:68–79, 2016. doi: 10.1159/000441351.

Tuerlinckx, Francis; Rijmen, Frank; Verbeke, Geert, and De Boeck, Paul. Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, 59:225–255, 2006. doi: 10.1348/000711005X79857.

Vemuri, Prashanthi; Lowe, Val J.; Knopman, David S.; Senjem, Matthew L.; Kemp, Bradley J.; Schwarz, Christopher G.; Przybelski, Scott A.; Machulda, Mary M.; Petersen, Ronald C., and Jr., Clifford R. Jack. Tau-pet uptake: Regional variation in average suvr and impact of amyloid deposition. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, 6:21–30, 2017. doi: 10.1016/j.dadm.2016.12.010.

Wasserstein, Ronald and Lazar, Nicole. The asa's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016. doi: 10.1080/00031305.2016.1154108.

Weiner, Michael W.; Veitch, Dallas P.; Aisen, Paul S.; Beckett, Laurel A.; Cairns, Nigel J.; Green, Robert C.; Harvey, Danielle; Jack, Clifford R.; Jagust, William; Morris, John C.; Petersen, Ronald C.; Salazar, Jennifer; Saykin, Andrew J.; Shaw, Leslie M.; Toga, Arthur W., and Trojanowski, John Q. The Alzheimer's Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement, 2017. ISSN 15525279.

Welvaert, Marijke; Durnez, Joke; Moerkerke, Beatrijs; Verdoolaege, Geert, and Rosseel,

Yves. neuRosim: An R package for generating fmri data. *Journal of Statistical Software*, 44(10):1–18, 2011. URL http://www.jstatsoft.org/v44/i10/.

Worsley, K.; Evans, A.; Marrett, S., and Neelin, P. A three-dimensional statistical analysis for cbf activation studies in human brain. *Journal of Cerebral Blood Flow and Metabolism*, 12:900–918, 1992. doi: 10.1038/jcbfm.1992.127.

Worsley, Keith and Friston, Karl. Analysis of fmri time-series revisited - again. *Neuroimage*, 2(3):173–81, 1995. doi: 10.1006/nimg.1995.1023.

K.Worsley, Marrett; NeelinS., P.; Vandal, A.; Friston, K., and Evans, A. A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4:58–73, 1996. doi: 10.1002/(SICI)1097-0193(1996)4:1%3C58::AID-HBM4%3E3.0.CO;2-O.

Wyman, Bradley T.; Harvey, Danielle J.; Crawford, Karen; Bernstein, Matt A.; Carmichael, Owen; Cole, Patricia E.; Crane, Paul K.; Decarli, Charles; Fox, Nick C.; Gunter, Jeffrey L.; Hill, Derek; Killiany, Ronald J.; Pachai, Chahin; Schwarz, Adam J.; Schuff, Norbert; Senjem, Matthew L.; Suhy, Joyce; Thompson, Paul M.; Weiner, Michael, and Jack, Clifford R. Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimer's and Dementia*, 9(3):332–337, 2013. ISSN 15525279. doi: 10.1016/j.jalz.2012.06.004.

Yi, Nengjun and Ma, Shuangge. Hierarchical shrinkage priors and model fitting for high-dimensional generalized linear models. *Stat. Appl. Genet. Mol. Biol.*, 11(6), 2012. doi: 10.1515/1544-6115.1803.

Yi, Nengjun; Zhi, Degui, and Li, Jun. Hierarchical generalized linear models for multiple groups of rare and common variants: Jointly estimating group and individual-variant effects. *PLOS Genetics*, 7(12), 2011. doi: 10.1371/journal.pgen.1002382.

Yi, Nengjun; Xu, Shizhong; Lou, Xiang-Yang, and Mallick, Himel. Multiple comparisons

in genetic association studies: A hierarchical modeling approach. *Stat. Appl. Genet. Mol. Biol.*, 13(1):35–48, 2014. doi: 10.1515/sagmb-2012-0040.

Zou, Hui and Hastie, Trevor. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Methodological)*, 67:301–320, 2005. doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x.

APPENDIX A

NEUROIMAGING OVERVIEW

1 Introduction

The study of the brain, and in particular the human brain, incorporates a wide range of disciplines and has become increasingly complex as research increasingly focuses on the analysis of images taken from the brains of live subjects. Applications of physics and engineering resulted in Magnetic Resonance Imaging (MRI) machines, which can obtain snapshots of the the brain's structure or even follow brain activity over time. Like all images, these brain images are the result of discrete intensity values than can be thought of as the average intensity within a particular area or volume; in two dimensions we call these locations pixels and in three dimensions we call them voxels. The biological correlate to the intensities depends upon the image medium, but these intensity values can be used to learn about brain structure, development, function, and so forth.[1] Psychology, the burgeoning discipline of Neuroscience, and medical practitioners spur biological research questions related to brain function as it relates to cognition, healthy and diseased brain activity and structure, as well as general brain development. Intersections of mathematics, statistics, and engineering aim to improve the analysis of the data arising from brain images and understand the functions and interactions of various brain regions. Navigating the literature requires some dedication to an understanding of these intersections; it is neither wise nor particularly useful to attempt an application of statistical or mathematical methodology without at least a rudimentary understanding of how brain images are obtained; henceforth, one may begin to understand the methodology that has arisen to handle analysis of these images. There are computational issues in analyzing massive datasets along with statistical

[1] The most common imaging mediums are Magnetic Resonance Images (MRI), Electroencephalograms (EEG), Positron Emission Tomography (PET), and Magnetoencephalography (MEG). A detailed explanation of each is beyond the scope of this document.

issues in accounting for non-independence of data. The literature on neuroimaging data analysis is vast and often difficult to compile into a coherent whole, but textbooks are now arriving on the scene. In particular, *The Handbook of Neuroimaging Data Analysis* provides an excellent overview of current techniques (Ombao et al., 2017); we cite further works as needed, but much of this document is distilled from this resource. In this overview, we aim to provide some basic facts about the human brain, the basic ideas inherent to obtaining images, and some relevant history of at least a few prominent analysis techniques.

2 The Human Brain

2.1 Basic Facts

The human brain weighs on average 1.5 kg, but varies by gender, age, body size, etc. (Ombao et al., 2017). Brain cells are broadly divided into neurons and glia, the first of which process information, and the second of which play supporting roles. The brain also divides into 3 rather distinct types of tissue: Grey matter (GM), consisting of cell bodies and dendrites, white matter (WM), consisting of myelinated axons connecting the cell bodies, and cerebrospinal fluid (CSF), contained within the ventricles and subarachnoid spaces. The distinction between these 3 types of tissue is often visibly obvious in images, although some voxels will not strictly contain 1 tissue type; this latter problem is called the partial volume effect. More generally, the brain is parcellated into the cerebrum, cerebellum, and brain stem. The cerebrum, consisting of grey matter, is split into hemispheres, which are connected by the corpus callosum, consisting of white matter. Many studies focus on aspects of either grey or white matter. The question of whether a brain area is implicated in some task generally refers to cell behavior in grey matter, and in particular the outer, folded layer of the cerebrum known as the the cerebral cortex. The folds of the cerebral cortex are described by cortical ridges and depressions, called gyri and sulci, respectively. Studies of white matter commonly consists of mapping white matter that connects the various regions of the cerebral cortex. Other studies may focus on volume changes or lesions in various

brain areas. This is, however, only a sample of the many creative ways neuroscience is using to learn more about the brain and human cognition and disease.

2.2 Levels of Abstraction

While the vast portion of this document shall address analyses intended primarily to apply to variations on Magnetic Resonance Imaging (MRI), it is helpful to differentiate between various levels of neuroscience research because the kind of data extracted, and therefore the kinds of appropriate analyses, vary by level of abstraction. The levels of abstraction can be seen as a kind of continuum describing behavior at the molecular level and extending all the way to organism behavior. Disciplinary expertise may vary by the scale or question of interest. While not exhaustive, disciplines intersecting in brain research include physics, biology, mathematics, statistics, psychology, medicine, and of course neuroscience.

We do not intend to provide an overview of all neuroscience research here, but rather distinguish the kinds of data one might in practice. In the most basic sense, one may study either individual neurons or populations of neurons. Due to ethics conerns, capturing data from neural firings in response to some stimuli is usually performed on animals rather than humans and often the data consists of times that the neuron fires. More detailed studies may investigate more complicated biological processes and the data may then consist of types of action potentials (sodium- versus calcium-based), the number of action potentials generated (i.e. spike trains), or perhaps changes in synaptic efficiency (Häusser and Mel, 2003; Trappenberg, 2010). Detailing the types of data possible in neuroscience would probably necessitate a book of near infinite length, but we highlight these levels of abstraction in order clarify the importance of understanding the form of data to be analyzed.

Generally speaking, most human research is not at the level of individual neurons, or even networks of neurons at the resolution of individual neurons. Many studies use electrical signals (EEG, MEG); for example, EEG is commonly used to study and diagnose epilepsy

(Rosenow et al., 2015; Saini and Dutta, 2017). However, we shall focus on MRI, whose methodology is summarized below. MRI takes advantage of the fact that after neurons fire, blood flow increases to the region where firing transpired. MRI is therefore a method of using blood oxygen levels to either distinguish between tissue types or infer activity. The data obtained consists of an array of voxels, a generalization of pixels to three dimensions, within each of which an intensity value is measured. These arrays create a fine resolution structural image or a time series of intensities describing blood oxygen levels within a voxel. For functional MRI, higher intensity shall correspond to neural firing and for structural MRI, intensity differences shall distinguish between tissue classes. Therefore, what we are measuring is blood flow and not the firing of individual neurons.

3 MRI

3.1 General Types of MRI Techniques

There are three primary MRI techniques:

1. **Structural MRI (sMRI)** results in a single three dimensional volume for each subject.
2. **Diffusion MRI (dMRI)** (usually) maps white matter tracts.
3. **Functional MRI (fMRI)** consists of a series of three dimensional volumes in short succession and are used to infer brain activity by analyzing time series of blood flow within sampled voxels.

For each of these techniques there are four general steps to analyzing the data:

1. *MRI Scan:* The subject is scanned in the MRI machine and data is obtained and transformed into magnitude and phase images; usually magnitude images are of primary research interest.
2. *Preprocessing:* Regardless of technique, raw data is not appropriate for analysis. Adjustments must be made for confounding factors such as subject movement and

inhomogeneities in the magnetic fields; adjustments for field inhomogeneities is commonly referred to as “inhomogeneity correction". Structural images typically require removing the skull from images and distinguishing between white matter, grey matter, and cerebrospinal fluid; removing the skull from the image is often called “masking" or “skull stripping". Time series acquired from fMRI typically require adjustments for breathing and heart rate. Both sMRI and fMRI require correcting for subject movement and often must be warped into a template to enable comparisons among subjects. The literature on preprocessing is vast, but largely beyond the scope of the current work. We note, however, that in the course of data analysis, the analyst should ensure appropriate preprocessing takes place.

3. *Subject Analysis:* Statistical analysis can be conducted at the subject level. At the most basic level, we may ask whether a particular voxel is “activated", or we may map an individual subject’s white matter tracts via Diffusion Tensor Imaging.

4. *Group Analysis:* Statistical analysis can be conducted on a sample of subjects. For instance, we may ask whether activation in some brain region of interest varies by treatment or disease status. At the most basic level, we can look at a particular voxel and ask whether activation differs between some groups.

3.2 MRI Physics

3.2.1 General Description

Complete description of MRI physics is beyond the scope of this work, but it is useful and perhaps indispensable to have some intuition about the process. We give a short description here based on Lindquist (2008) and Ombao et al. (2017). MRI is a re-naming of Nuclear Magnetic Resonance (NMR) and the 3-D volume obtained is not obtained at once, but arises from interpolating a sequence of 2-D slices.

We now describe the relevant physics. Water is ubiquitous in the body and is made

up of 2 hydrogen atoms and one oxygen atom. Atomic nuclei with odd numbers of protons or neutrons have angular momentum. Since hydrogen atoms have only 1 proton, the angular momentum of hydrogen atoms can be employed to measure the density of those hydrogen atoms. Protons may be represented as positively charged spheres, spinning about their axes and producing a net magnetic moment along the direction of the spins. The net magnetization is defined as the average angular momenta from all nuclear spins and may be described as a vector with a longitudinal component parallel to the magnetic field and a transverse component perpendicular to the magnetic field. The net magnetization is the source of MR signal.

Absent a magnetic field, the nuclei have random orientation and there is no net magnetization. The MR machine introduces a powerful external magnet, usually between 1.5-7 Tesla, which coerces the spin axes into orientations either parallel or anti-parallel to the magnet. The spin axes then precess around the magnet's axis at the Larmor frequency with a net magnetization in the direction of the magnetic field. A radio frequency electromagnetic field (RF) pulse is then employed to change the spin orientations; the angle of this change is determined by amplitude of the RF pulse. Afterwards, the system of nuclei briefly precess in phase and the net magnetization is detected by receiver coils. Essentially, the RF pulse aligns the phase of the nuclei and rotates them into the transverse plane, after which the spin axes begin to recede back into alignment with the scanner magnetic field; this is called *longitudinal relaxation*, and the rate of relaxation is described by the time constant *longitudinal relaxation time,* T_1. After the RF pulse ends, the spins also fall out of phase, called *transverse relaxation*. The time constant *transverse relaxation time,* T_2, describes how quickly the spins fall out of phase. Both T_1 and T_2 vary by tissue, which allows for well defined structural images. Lastly, we note that local inhomogeneity in the magnetic field may result in some nuclei de-phasing quicker than others; the inhomogeneity is often due to blood flow changes and oxygenation. A variation on T_2 relaxation, T_2^* is often used to counteract these ill effects. It is useful to summarize the key parts of the MRI scanner and

their respective purposes:

1. *Strong Super-Conducting Electromagnet:* This magnet is present throughout the scanning process and aligns spin axes parallel to its direction along the longitudinal plane; in mathematical formulations it is often denoted β_0.

2. *Radio Frequency Coils:* Located near the head of the subject, these are turned on and off to alter the orientation of the nuclei's spins. When turned on, the spin axis orientations are disrupted out of alignment with the super-conducting magnet. When turned off, the spin axis orientations gradually return to alignment with the super-conducting magnet.

3. *Electromagnetic Gradient Coils:* Theses create specific spatial variation in magnetic field strength and is necessary for obtaining spatial information in addition to signal.

3.2.2 Image Contrasts

Image contrast results from the frequency with which we disrupt the nuclei's spins out of alignment with the super-conducting magnet. *Time to repetition (TR)* is the frequenct of disruption or excitation of the nuclei and *Time to echo (TE)* is when we measure net magnetization. The measured signal is approximately represented as follows:

$$S(T_1, T_2) = M_0(1 - e^{-TR/T_1})e^{-TE/T_2} \tag{3.2.1}$$

where M_0 is the initial net magnetization prior to exciting the nuclei. The following are common image contrasts:

1. **Proton Density**: *Long TR, Short TE*. This yields a signal which is approximately M_0 and thus proportional to the density of protons in the tissue. Grey matter will appear white, while white matter is dark and is generally not well distinguished from CSF.

2. $\mathbf{T_1}$**-Weighted**: *Moderate TR, Short TE* (≈ 20 ms). T_1-weighted images are typically

employed in structural MRI (sMRI) to highlight anatomical structure. White matter is bright, grey matter is dark grey, and CSF black.

3. $\mathbf{T_2}$**-Weighted**: *Long TR, Moderate/Long TE.* Also typically employed in structural MRI. CSF and grey matter are bright while air is dark.

4. $\mathbf{T_2^*}$**-Weighted**: As mentioned above, adjustments for magnetic field inhomogeneities are possible by employing particular pulse sequences with magnetic gradients. T_2^*-weighted images are key in functional MRI (fMRI) due to sensitivity to blood flow and oxygenation.

Notice how the definition of each corresponds to the definition of signal above; e.g., T_1-weighted images make TR larger and therefore the TR term, e^{-TR/T_1}, will be further from 1, increasing the influence of the time constant T_1, while the shrinking TE will make $e^{-TE/T_2} \to 1$, allowing less influence for the time constant T_2.

3.2.3 Image Acquisition

As mentioned above, gradient coils allow for spatial partitioning of net magnetization signal. We shall not go into great detail, but it is important to understand the origin of the data to be analyzed. We follow Ombao et al. (2017), Chapters 6 and 8. We have briefly mentioned that MRI data are acquired as 2-D slices which are the interpolated to construct a 3-D image. The gradient coils sequentially control spatial homogeneity in each slice and each measurement is represented as a Fourier transformation of the spin density at a point in the frequency domain, or k-space. Therefore, the MR signal at time t_j is described as

$$S(t_j) = \int_x \int_y M(x,y) e^{-2\pi i (k_x(t_j)x + k_y(t_j)y} \, dx \, dy \tag{3.2.2}$$

where $M(x,y)$ is the spin density at point (x,y) and $k_x(t_j), k_y(t_j)$ is a discrete location in k-space. Note that the "i" in equation 3.2.2 is not an index, but rather denotes the complex number such that $i^2 = -1$. Also, while equation 3.2.2 is a continuous Fourier transform, in

practice one generally uses a discrete Fourier transform, since the data are sampled discretely. Examples of points in k-space include:

$$k_x(t_j) = \frac{\gamma}{2\pi} \int_0^{t_j} G_x(\tau)\, d\tau$$
$$k_y(t_j) = \frac{\gamma}{2\pi} \int_0^{t_j} G_y(\tau)\, d\tau$$

where $xG_x(t_j) + yG_y(t_j) = B_z(r, t_j)$ is the gradient coil change, for $r^2 = x^2 + y^2$ and some constant γ. Note that points in k-space have no one-to-one correspondence with points in the image space; rather, points near the edges of k-space contribute more to image resolution while points near the center contribute more to image contrast, and thus choices about where and how to sample from k-space affect the resulting image in terms of both resolution and contrast. The locations of sampled data in k-space are determined by the strength of the frequency and phase-encoding gradients, with stronger gradients resulting in sampling further from the center of k-space; a third gradient determines the relevant characteristics of the sampled slice (e.g. location, orientation, and thickness), but this gradient is not affected by k-space considerations; for a detailed discussion on k-space and how its manipulation affects the result image see Mezrich (1995). The most common sampling technique is echo-planar imaging (EPI), which employs uniform sampling about its origin. EPI ensures that we may use the Fast Fourier Transform (FFT) in image reconstruction. The raw k-space data is complex valued and measurement error for both the real and imaginary components is assumed to be Gaussian. Once the data is transformed into image space via the inverse Fourier transform, the measurements are separated into magnitude and phase images, after which the magnitude images follow a Rice distribution rather than a normal distribution.[2] Most analyses use only the magnitude images.

[2]The Rice distribution has the following probability density function: $P(Z) = \frac{Z}{\sigma^2} \exp\left(-\frac{Z^2+|V|^2}{2\sigma^2}\right) I_0\left(\frac{Z|V|}{\sigma^2}\right)$ where $I_0(z)$ is a modified Bessel function of the first kind, $z, \sigma > 0$, and when $|V| = 0$ the distribution is known as the Rayleigh distribution.

4 Modeling fMRI Time Series

4.1 The Basic Idea: BOLD Contrast

Functional MRI consists of taking multiple 3-D MRI volumes of a subject in one session, so that each subject has a time series of 3-D images; more specifically, each voxel has a corresponding time series of intensity values. The most common contrast for fMRI is Blood-Oxygen-Level-Dependent (BOLD) contrast, which is based on differences in magnetic properties between oxygenated and deoxygenated hemoglobin. Increasing neural demands correspond to increasing metabolic demands for oxygen. Neural firing extracts oxygen from hemoglobin in the blood, from which follows that hemoglobin will become paramagnetic. The magnetic field distortions result in decreased T_2^* signal, i.e. faster signal decay, which implies a local decrease in BOLD signal. An over-compensation in blood flow follows, increasing the density of oxygenated hemoglobin, which then manifests as increased BOLD signal in the region. BOLD fMRI essentially uses hemodynamic response to neural firing to infer neuronal activity.

4.2 The Hemodynamic Response Function

The hemodynamic response function (HRF) describes the increase and subsequent decrease in blood flow and is characterized by a signal increase approximately 1-2 seconds post-neural activity, peaking approximately 5-8 seconds post-neural activity, followed by the signal falling to below baseline from which it takes approximately 10 seconds to recover. This last is called the *post-stimulus undershoot*, which occurs because blood flow decreases quicker than blood volume so that deoxygenated hemoglobin has a higher concentration. Explicit modeling of the HRF is primarily necessary under experimental paradigms; that is, some stimulus is presented and the response modeled. The HRF is commonly modeled in one of the following ways:

1. *Balloon Models:* This approach is a non-linear physiologically based model described by ordinary differential equations. Changes in blood volume, blood inflow, deoxyhe-

moglobin, and flow inducing signals are explicitly modeled to describe changes in the observed BOLD signal. While the Balloon model is perhaps the most biologically plausible model, it requires the estimation on many parameters, has difficulty in estimating parameters in the presence of noisy data, and lacks a direct framework for inference.

2. *Linear Time Invariant (LTI) Models:* This approach simplifies the biology so that the neuronal activity is the input (impulse) and the HRF is the response function. The signal at time t, $x(t)$ is modeled by the convolution of the stimulus function, $v(t)$, and the HRF, $h(t)$:

$$x(t) = (v * h)(t) \tag{4.2.1}$$

where the form and complexity of $h(t)$ is determined by the researcher. The LTI system has the following qualities/assumptions:[3]

   (a) *Scaling:* Scaling the input by a factor b results in scaling the BOLD signal by b; this assumption allows us to assume that amplitude differences reflect differences in neuronal activity.

   (b) *Superposition:* The sum of individual responses to two stimuli shall describe the response when the stimuli are applied simultaneously.

   (c) *Time-invariance:* Shifting a stimulus in time by amount Δt results in a response shifted in time by Δt.

   We must also assume the following:

   (a) A linear BOLD response assumption is necessary for the convolution framework in equation 4.2.1 to be a valid approach (Lindquist, 2008).

[3] We say assumptions because while this modeling framework has these qualities, the underlying phenomena must reasonably be approximated by the modeling framework.

(b) The neural activity function is correctly modeled; i.e., the experimental paradigm makes sense.

(c) The HRF is correctly modeled, or at least approximated to an acceptable degree of accuracy.

The HRF under LTI assumptions usually takes one of the following forms:

(a) *Canonical HRF:* This form is a linear combination of two gamma functions:

$$h(t) = c\left(\frac{t^{\alpha_1-1}\beta_1^{\alpha_1}e^{-\beta_1 t}}{\Gamma(\alpha_1)}\right)\left(\frac{t^{\alpha_2-1}\beta_2^{\alpha_2}e^{-\beta_2 t}}{\Gamma(\alpha_2)}\right) \tag{4.2.2}$$

where $\alpha_1 = 6, \alpha_2 = 16, \beta_1 = \beta_2 = 1, c = 1/6$ and Γ denotes the gamma function. These values are based on empirical research in the visual cortex, but the HRF is known to vary in its parameters by brain region (Aguirre et al., 1998; Schacter et al., 1997). The canonical HRF is a popular choice in fMRI data analysis, but it lacks flexibility, and using empirical results from only one region of the brain makes it susceptible to mis-modeling signal (Loh et al., 2008). Flexibility can be improved by modeling the HRF along with its temporal and dispersion derivatives, which allows differences in onset and width.

(b) *Temporal Basis Function Models:* Here the HRF is modeled as a linear combination of basis functions:

$$h(t) = \sum_i^m \beta_i f_i(t) \tag{4.2.3}$$

where $i = 1, \ldots, m$ is the number of basis functions. Then the BOLD response model is still the convolution of the stimulus function and the HRF:

$$x(t) = \sum_i^m \beta_i f(s * f_i)(t) \tag{4.2.4}$$

where β_i is the weight of the i^{th} component. This approach is far more flexible than the canonical HRF and accommodates models as complex as the *Finite Impulse Response (FIR)* model, which consists of a basis set with 1 free parameter for every time point following stimulation and allows for arbitrary HRF shapes in each voxel (Glover, 1999; Goutte et al., 2000).

4.3 Statistical Parametric Maps (SPM)

The size of datasets, i.e., number of voxels to be analyzed often frustrates the application of multivariate methods. Early work from Karl Friston developed and popularized the use of "mass-univariate" approaches in neuroimaging, whereby voxel-specific statistics are obtained (Friston et al., 1991, 1994, 1995; Worsley and Friston, 1995; Worsley et al., 1996). The voxel level statistics can be based on t-, F-, or Z-statistics depending on the context and complexity of modeling and analysis. These are generally based on time-series within-voxel. As mentioned above, these may correspond to within-subject or group-level inferences. SPM approaches are known to have multiple testing issues. This is a difficult issue, but many approaches have been developed in an attempt to control for false positives while allowing a reasonable ability to detect true signal (Lindquist and Meija, 2015; Nichols and Hayasaka, 2003).

4.4 Conclusion

This is summary has by no means been exhaustive. As mentioned EEG and MEG are also important in many studies, but in addition alternative approaches to modeling fMRI time series can also be found in the literature. For example, Dinov et al. (2005) proposed using wavelets in statistical analyses of fMRI data. Correspondingly, statistical approaches are vast and vary by both across and with research questions. However, this summary provides a reasonable overview of the literature with the caveat that a complete review would be textbook length.

APPENDIX B

THE EM ALGORITHM

1 Motivation

The EM algorithm and much of its relevant theory was established by Dempster et al. (1977), although some of the groundwork had already been laid. The approach was developed specifically to handle incomplete data, although what counts as "incomplete data" is quite broad. One application is when data is missing, but one still wants to estimate parameters without excluding subjects. However, the approach has found application in a wide range of analytically difficult situations where it makes sense to assume some parameters exist and are missing or hidden (Bilmes, 1998). The iterative process consists of two steps: expectation (E-step) and maximization (M-step). Generally, one defines a likelihood function that is a function of the parameters of interest, the observed data, and the missing data/parameters. The E-step takes the expectation of the (log of) this function in terms of the missing data/parameters, given an estimate for the parameters of interest and the observed data. The M-step then maximizes the expectation from the E-step in terms of the parameters of interest. The process continues iteratively until convergence.

Generally, consider sample spaces, $\mathcal{X}, \mathcal{Y}$ and a many to one mapping from $\mathcal{X} \to \mathcal{Y}$. We observe data $\boldsymbol{y}$, a realization from $\mathcal{Y}$ and must infer about $\boldsymbol{x} \in \mathcal{X}$, which is not observed. There are then two associated density functions: the complete-data, $f(\boldsymbol{x}|\boldsymbol{\theta})$ and incomplete, $g(\boldsymbol{y}|\boldsymbol{\theta})$, where importantly both distributions depend upon the same $r \times 1$ vector of parameters of interest, $\boldsymbol{\theta}$. These distributions are related to each other by:

$$g(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{\mathcal{X}(\boldsymbol{y})} f(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x} \tag{1.1}$$

This is to say that we must use the observed $\boldsymbol{y}$ to find $\boldsymbol{\theta}$, but while we do not have access to

$\boldsymbol{x}$ we can use information contained in the structure of the distribution function $f(\boldsymbol{x}|\boldsymbol{\theta})$.

2 Exponential Families

We begin with the simpler case of exponential families. If $f(\boldsymbol{x}|\boldsymbol{\theta})$ is a regular exponential family and $\boldsymbol{\theta}$ are the natural parameters, then

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = b(\boldsymbol{x}) \exp\{\boldsymbol{t}(\boldsymbol{x})\boldsymbol{\theta}\}/a(\boldsymbol{\theta}) \tag{2.1}$$

where $\boldsymbol{t}(\boldsymbol{x})$ is the $1 \times r$ vector of sufficient statistics and $\boldsymbol{\theta}$ is restricted to an r-dimensional convex set Θ such that equation 2.1 is a probability density function for all $\boldsymbol{\theta} \in \Theta$. The EM algorithm in this framework is then:

1. **E-Step:** Estimate the complete-data sufficient statistic, $\boldsymbol{t}(\boldsymbol{x})$, for the p^{th} iteration via

$$\boldsymbol{t}^{(p)} = E(\boldsymbol{t}(\boldsymbol{x})|\boldsymbol{y}, \boldsymbol{\theta}^{(p)}) \tag{2.2}$$

2. **M-Step:** Estimate $\boldsymbol{\theta}^{(p+1)}$ by solving for $\boldsymbol{\theta}$ in

$$E(\boldsymbol{t}(\boldsymbol{x})|\boldsymbol{\theta}) = \boldsymbol{t}^{(p)} \tag{2.3}$$

Notice that equation 2.3 has the form of the likelihood equations for maximum likelihood when data arises from an exponential family; i.e., the solution $\boldsymbol{\theta}^{(p)}$ is a maximum likelihood estimator for $\boldsymbol{\theta}$. Repeating these steps results in a value $\boldsymbol{\theta}^*$ that maximizes $L(\boldsymbol{\theta}) = \log g(\boldsymbol{y}|\boldsymbol{\theta})$. After some relevant manipulation, one may differentiate $L(\boldsymbol{\theta})$ to obtain $\boldsymbol{D}L(\boldsymbol{\theta}) = -E(\boldsymbol{t}|\boldsymbol{\theta}) + E(\boldsymbol{t}|\boldsymbol{y}, \boldsymbol{\theta})$. In the limit as $p \to \infty$, if $\boldsymbol{\theta}^{(p)} = \boldsymbol{\theta}^{(p+1)} = \boldsymbol{\theta}^*$, then $E(\boldsymbol{t}|\boldsymbol{\theta}^*) = E(\boldsymbol{t}|\boldsymbol{y}, \boldsymbol{\theta}^*) \to \boldsymbol{D}L(\boldsymbol{\theta}) = \mathbf{0}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

3 Generalized EM Algorithm

The EM algorithm is generalizable beyond exponential families. Define

$$Q(\boldsymbol{\theta}'|\boldsymbol{\theta}) = E(\log f(\boldsymbol{x}|\boldsymbol{\theta}')|\boldsymbol{y}, \boldsymbol{\theta}) \tag{3.1}$$

which is assumed to exist for all pairs $(\boldsymbol{\theta}', \boldsymbol{\theta})$ and $f(\boldsymbol{x}|\boldsymbol{\theta}) > 0$ almost everywhere in $\mathcal{X}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Then we have

1. **E-Step:** Compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)})$.
2. **M-Step:** Find $\boldsymbol{\theta}^{(p+1)} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max} \left\{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)})\right\}$.

That is, we compute an expectation and then maximize that expectation. Dempster et al. (1977) show that this process increases the log-likelihood with respect to $\boldsymbol{\theta}$, $L(\boldsymbol{\theta})$, at each step and converges to (at least) a local maximum; further details are found in Dempster et al. (1977), but these are beyond the scope of this work.

APPENDIX C

PROBABILITY DISTRIBUTIONS

Beta Distribution

1. *Probability Density Function:* $f_Y(y) = \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}$
2. *Support/Parameters:* $y \in [0,1]$; $\alpha > 0$, $\beta > 0$
3. $\mathrm{E}(Y) = \frac{\alpha}{\alpha+\beta}$
4. $\mathrm{var}(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Binomial Distribution

1. *Probability Mass Function:* $f_Y(y) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$
2. *Support/Parameters:* $y = 1, \ldots, n$; $\theta \in [0,1]$
3. $\mathrm{E}(Y) = n\theta$
4. $\mathrm{var}(Y) = n\theta(1-\theta)$

Double Exponential (Laplace) Distribution

1. *Probability Density Function:* $f_Y(y) = \frac{\lambda}{2} e^{\lambda|y-\mu|}$
2. *Support/Parameters:* $y \in \mathbb{R}$; $\mu \in \mathbb{R}$, $\sigma > 0$
3. $\mathrm{E}(Y) = \mu$
4. $\mathrm{var}(Y) = \frac{2}{\lambda^2}$

Gamma Distribution

1. *Probability Density Function*: $f_Y(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta}$
2. *Support/Parameters:* $y \in \mathbb{R};\ \alpha > 0; \beta > 0$
3. $\mathrm{E}(Y) = \alpha\beta$
4. $\mathrm{var}(Y) = \alpha\beta^2$

Normal Distribution

1. *Probability Density Function*: $f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2}$
2. *Support/Parameters:* $y \in \mathbb{R};\ \mu \in \mathbb{R},\ \sigma^2 > 0$
3. $\mathrm{E}(Y) = \mu$
4. $\mathrm{var}(Y) = \sigma^2$

Poisson Distribution

1. *Probability Mass Function:* $f_Y(y) = \frac{e^{-\lambda}\lambda^y}{y!}$
2. *Support/Parameters:* $y = 1, 2, \ldots;\ \lambda \geq 0$
3. $\mathrm{E}(Y) = \lambda$
4. $\mathrm{var}(Y) = \lambda$