Bayesian Multiple Testing Under Dependence with Application to Functional Magnetic Resonance Imaging

by

Derek Andrew Brown

(Under the direction of Nicole A. Lazar and Gauri S. Datta)

Abstract

The analysis of functional neuroimaging data often involves the simultaneous testing for activation at thousands of voxels, leading to a massive multiple testing problem. This is true whether the data analyzed are time courses observed at each voxel or a collection of summary statistics such as statistical parametric maps (SPMs). It is known that classical multiplicity corrections become strongly conservative in the presence of a massive number of tests. Some more popular approaches for thresholding imaging data, such as the Benjamini-Hochberg procedure for false discovery rate control, tend to lose precision or power when the assumption of independence of the data does not hold. Bayesian approaches to large scale simultaneous inference also often rely on the assumption of independence. This dissertation introduces a spatial dependence structure into a Bayesian testing model for the analysis of SPMs. By using SPMs rather than the voxel time courses, much of the computational burden of Bayesian analysis is mitigated. Increased power is demonstrated by using the dependence model to draw inference on a real dataset collected in a fMRI study of cognitive control. The model also is shown to lead to improved identification of neural activation patterns known to be associated with eye movement tasks.

INDEX WORDS: Multiple Testing Problem, Bayesian Statistics, Conditional Autoregressive Model, False Discovery Rate, fMRI, Saccades

Bayesian Multiple Testing Under Dependence with Application to Functional Magnetic Resonance Imaging

by

Derek Andrew Brown

B.S., Georgia Institute of Technology, 2006

M.S., The University of Georgia, 2010

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2013

Bayesian Multiple Testing Under Dependence with Application to Functional Magnetic Resonance Imaging

by

Derek Andrew Brown

Approved:

Major Professors:	Nicole A. Lazar Gauri S. Datta
Committee:	Jennifer McDowell Jaxk Reeves Paul Schliekelman Lynne Seymour

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2013

Dedication

This is for Dad.

Acknowledgements

To acknowledge everyone and everything who has played a part in the accomplishment of this work would be an almost endless task. There are so many people that deserve recognition that some will not receive due credit, regardless of what I write. A few short sentences here are but a meager attempt to tip my proverbial hat to some of the more prominent figures that have played a role in shaping my professional and academic career.

First, I must give my gratitude to God, the Divine Architect without whose intervention I would no doubt be in a very different place, if any at all. I can only hope to have gained a small bit of the wisdom necessary to separate what I can control from what I cannot and to focus my thoughts appropriately. Of course, my loving family has been nothing but a gift, undeserved but not taken for granted. Mom, Dad, and Russell have put up with me, yet loved me, throughout my life. They refused to give up on me, even in times where I nearly gave up on myself. My Grandma has always been and continues to be an inspiration to me. She is the strongest woman I have ever known. I think of my cousins Chad, Chip, and Krista as not just kin, but friends. My Grandaddy will always be in my thoughts. I count myself lucky for having known him. To my entire family, I want you to know that I will never forget who I am or where I come from; without you I would be nothing.

My wife Christy is the single greatest thing to happen to me in life. She knows better than anyone what it took for me to survive graduate school. There is not a doubt in my mind that I would have failed if it were not for her encouragement and care. There is nothing I can write here that will do you justice, Christy. Our life together is only beginning. I love you.

Nobody can get a PhD without guidance. I am fortunate to have found two great advisors. I will forever be indebted to Nicole and Gauri for investing their thoughts and time into turning me into a researcher. Thanks go to my doctoral committee, as well. Woncheol has always been helpful, even from the other side of the world. Jennifer continues to be extraordinarily patient with me while I fumble my way through learning about neuroimaging. Jaxk seems to have a penchant for asking questions that nobody else thinks of. Both Paul and Lynne were good to step in on short notice.

Of course, many others deserve praise for making my life better, even if they don't realize it. James and I somehow manage to still encourage each other despite only seeing each other once every couple of years. Brandon, Josh, Jeremy, and Travis have proven to be lifelong friends. Lynne, Adam, Lindsay and the Blind Pig helped me to keep my sanity. (What's the point of grad school without ninja wings and cheese fries?) Tim and Jimmy were always good for talking about something besides statistics. I thank all of you for helping me to escape the stress.

Much of the work in this dissertation has been submitted for publication in *NeuroImage*. I would be remiss if I did not mention the three reviewers' comments concerning that manuscript. Their suggestions were instrumental in improving this work. As of the date of this writing, I do not know whether or not the manuscript will be accepted.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The analysis of high throughput data presents many challenges to researchers across a variety disciplines. Many of the problems that must be dealt with are ubiquitous in the field of statistics but are magnified or exacerbated when the data sets are on an extremely large scale. Often, the goal is to perform hypothesis tests or otherwise infer the presence or absence of a signal at an extremely large number of points. Observing thousands (or millions, in some cases) of points at the same time greatly increases the chances of some data spuriously exhibiting a signal just by random variation. The situation is further complicated when the data are dependent so that there is redundant information between observations being tested. Failure to account for this dependence structure can have an adverse effect on the ability to detect interesting signals. The need to balance a high risk for false positives with sensitivity for detection then becomes a foremost concern, as does the development of appropriate techniques for dealing with dependence.

The multiple testing problem has a rich enough history that there is a consensus that it has been satisfactorily addressed in small- to moderate-sized settings. The fact that performing multiple hypothesis tests increases the probability of declaring a false positive was first addressed by Fisher (1935). More ideas followed, including methods for following up an analysis of variance with multiple contrasts (Scheffé, 1953) and pairwise comparisons (e.g. Tukey, 1952; Duncan, 1955). Around that same time a procedure was introduced for comparing multiple treatments to a control by Dunnett (1955). The commonly used Bonferroni procedure for constructing simultaneous confidence intervals may be attributed to Dunn (1961). Such procedures have

long been known to have desirable theoretical properties such as control of the strong family-wise error rate (SFWER), the probability of falsely rejecting one or more null hypotheses. Methods that control the SFWER but do not facilitate constructing simultaneous confidence intervals include the so-called REGWR procedure (Ryan, 1960; Einot and Gabriel, 1975; Welsch, 1977) and Holm's method (Holm, 1979), a modification to the Bonferroni correction based on ordering the p-values. Other improvements to the Bonferroni procedure were introduced in Simes (1986) and Hochberg (1988), among others. See Oehlert (2000, Chapter 5) for a broad survey of multiple testing procedures.

Prior to the last decade, most multiple testing procedures were constructed with the intent of controlling the overall error rate for a relatively small number of simultaneous tests. The advent of high throughput technology revealed, however, that classical procedures can be inappropriate in the presence of thousands of simultaneous tests. As an example, consider the Bonferroni correction. This procedure takes the nominal significance level α and divides it by the number of tests being performed, say K. The threshold for p-values at which significance is declared then becomes α/K. The correction is a simple and effective method of controlling the SFWER. If one is performing, say, $60,000$ one-sided tests at the $\alpha = .05$ significance level, though, the correction only identifies as signals those points with p-values less than $.05/60,000 \approx .00000083$, or equivalently a Gaussian test statistic of at least $z \approx 4.79$. This extremely conservative threshold would result in only the most extreme points in the data being selected, or none at all. There is no reason to think, though, that many points in the data are not true signals just because there are thousands of others being tested at the same time.

Any procedure controlling the overall family-wise error rate must necessarily make the individual rejection criterion more stringent as the number of tests increases. This approach has proven to be overly conservative in many cases, leading to much new

research on multiple testing. Using the analysis of gene microarrays as a motivating example, Dudoit et al. (2003) discussed several procedures for testing in the high throughput setting. A more recent overview of techniques for large-scale testing is Efron (2010), including effect size estimation and combining results to draw inference on sets of observations with enrichment analysis.

A particularly important breakthrough in the statistical analysis of massive data sets came when Benjamini and Hochberg (1995) introduced the false discovery rate (FDR) along with a simple procedure for its control. The Benjamini-Hochberg "step-up" procedure was ahead of its time when it was first introduced and hence not widely accepted by the statistical community (Benjamini, 2010). Its value was more fully appreciated with its application to high throughput data such as gene microarrays and neuroimaging. The procedure controls the expected proportion of discoveries (hypothesis rejections) that are false, as opposed to controlling the overall probability of committing *any* Type I error via the SFWER. This makes it much easier to scale up to larger data sets without becoming overly conservative. A conditional version of FDR, the positive false discovery rate, has been explored by Storey (2002, 2003).

Regardless of the thresholding rule chosen, the null distribution used to calculate p-values can itself be problematic. Efron (2004) showed that in the large scale scenario, the distribution of test statistics under the null hypothesis can be different from what theory predicts because of unobserved covariates or some other kind of dependence in the data. The difference in distributions may seem slight. For instance, when using standard Gaussian test statistics, the appropriate null mean may be -.36 rather than a theoretical standard Gaussian with a mean of zero. Using the theoretical distribution, a test statistic of $z = 1.4$ has a one-sided p-value of .08, failing to be significant at the $\alpha = .05$ level. Under the correct null distribution, this same statistic has a p-value of .04 and hence would be declared significant. All locations with such marginally significant test statistics would fail to be detected under the theoretical distribution,

despite a small (one-third of a standard deviation) difference between its mean and the truth. Even small differences, then, can change the long-run operating characteristics of any significance testing procedure, leading to considerable inconsistencies in results when the procedure is applied thousands of times. Even the Benjamini-Hochberg procedure will fail if the wrong null distribution is used to calculate p-values. This can mean the difference between identifying hundreds of interesting cases and none at all (Efron, 2007). Empirical estimation of the null distribution of transformed p-values was discussed in Efron (2004).

In addition to the frequentist methods described above, Bayesian statistics also provides means of dealing with high throughput data. In the case of complete independence (e.g. a non-hierarchical structure), Berry and Hochberg (1999) showed that the resulting probabilities evaluated at each point depend only on the data and parameters at that point; they are unaffected by any other data that may have been simultaneously observed so that no adjustment is made. Berry and Hochberg argued, however, that such independence is rarely the case. In variable selection problems where a subset of points are to be selected as interesting cases among many possible candidates, there is a need for multiplicity correction in the sense that the estimated probability of any particular point (variable) being non-null should adjust accordingly as the total number of points being observed increases. While most Bayesian analyses contain intrinsic penalties for model complexity (Jefferys and Berger, 1992), the *multiplicity* adjustment is only induced through the modeling of the prior probability of a case being non-null (Scott and Berger, 2010), in which case the model automatically accounts for the number of tests in *a posteriori* probability statements (see Section 2.2). This inherent self-calibration is appealing since it relieves researchers of explicitly correcting the threshold for discoveries. For discussions of Bayesian approaches to the multiple comparisons problem, see Waller and Duncan (1969), Westfall et al. (1997), Berry and Hochberg (1999), and Scott and Berger (2006, 2010), among others.

Close parallels between Bayesian posterior calculations and estimated false discovery rates have been demonstrated in the literature. Under independent mixture distributions of the test statistics, Storey (2003) showed how the positive false discovery rate may be expressed as the Bayesian posterior probability of the null hypothesis being true, given that a test statistic lies in a specified rejection region. Efron (2010) defined a *local false discovery rate* for some test quantity, z, as $\text{fdr}(z) = \pi_0 f_0(z)/f(z)$, where $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$ is the mixture density of the test statistics and $\pi_0 = 1 - \pi_1$ is the probability that a test statistic arises from its null density, f_0.

Much of the work thus far developed in the literature is based on the assumption of independence of the data. This assumption, which does not hold in many applications, can yield drastically different results from what would otherwise be obtained by accounting for the true dependence in the data. Consider a simple t-test. A one- or two-sample t-test is based on an effect size, or the scaled distance between two values. Inappropriately assuming independence causes the scale (standard deviation) to be poorly estimated so that the true difference between observed and null values is obscured. Even when the "within-group" correlation is accounted for so that each individual statistic is calculated appropriately, there can still be considerable "between-group" association so that the test statistics themselves are correlated. This can result in the null distribution of the observed test statistics being over- or under-dispersed relative to the theoretical null, even though the null distribution of each individual test statistic is correct (Efron, 2007). Consequently, either too few or too many test statistics may be declared significant. The original Benjamini-Hochberg procedure controls the FDR in the case of a dependence structure known as positive regression dependence (PRD), although Efron (2010, page 51) commented, "even PRD is unlikely to hold in practice." Benjamini and Yekutieli (2001) introduced a modified procedure for controlling the FDR under arbitrary dependence structures. In spite of this, FDR control still tends to lose precision under dependence (Efron, 2007).

The deleterious impact correlation can have on empirical Bayes methods and FDR control was investigated in Qiu et al. (2005). Genovese et al. (2006) gave a method for weighting p-values whereby known spatial structure can be effectively exploited. Efron (2007) expanded upon previous work to focus on the effects dependence can have on the distribution of test statistics. Further work on large-scale testing under dependence within the frequentist paradigm may be found in Sun and Cai (2009), who used a dependence structure induced by a hidden Markov model to construct a "local index of significance" in place of a p-value.

In this work, we consider the analysis of functional magnetic resonance imaging (fMRI) data. Functional neuroimaging provides a set of tools that record state-related brain signals that are subsequently used to generate maps of the neural circuitry activation associated with that state. An imaging study may require a researcher to perform hypothesis tests on a large number of parameters simultaneously to infer the presence or absence of signal changes throughout the brain. See Chapter 3 for more details about fMRI data.

Much research has been done on effectively identifying regions of activation in brain images from fMRI studies. A widely-used approach based on thresholding contiguous clusters of voxels was introduced in Forman et al. (1995). The use of random fields toward this end was pioneered in Worsley et al. (1992) and Worsley (1994). See Worsley (2003) for further development of this work. FDR control was introduced to the neuroimaging community by Genovese et al. (2002). More recently, Benjamini and Heller (2007) used fMRI to demonstrate clustering of points together in spatial data sets, then controlling the FDR on the clusters. Chumbley and Friston (2009) compared the FDR procedure to random field analysis, demonstrating some dramatic differences in results that the two methods can produce. Computationally-intense permutation tests were proposed to remedy the null distribution problem in Holmes et al. (1996) and Nichols and Holmes (2001). Nichols and Hayasaka (2003) compared

permutation tests, Bonferroni, and random field theory methods for controlling the family-wise error rate. That Bayesian and FDR procedures share the common characteristic of adapting to features of the data was remarked upon in Marchini and Presanis (2004) in the context of statistical parametric maps (SPMs; Friston et al., 1995). They concluded that, in general, the use of Bayesian modeling is the most powerful approach to identifying regions of activation when compared to thresholding via FDR control or random field theory.

In neuroimaging, many collected data lend themselves to hierarchical modeling, which fits naturally into the Bayesian framework. Images are constructed by partitioning the brain into voxels. Since multiple brains may be analyzed simultaneously, voxels are nested in participants who may in turn be nested in group-level structures (Lazar, 2008, page 173). Early work in the Bayesian analysis of brain images appeared in Genovese (2000), Gössl et al. (2001), and Friston and Penny (2003). More recent work includes Smith and Fahrmeir (2007), who used a binary Markov random field to model spatial correlation among the voxel-specific indicators of activation, and Bowman et al. (2008), who took advantage of hierarchical structures to separate local dependence and inter-regional correlation of predetermined regions of interest. SPMs were treated as arising from cluster point processes with both population-level and individual-level centers in Xu et al. (2009), allowing for the estimation of the proportion of individuals exhibiting activation at certain locations. Point processes were also used in Kang et al. (2011) for Bayesian meta-analysis about reported foci from imaging studies and the variability among participants. A review of Bayesian procedures in fMRI may be found in Woolrich (2012).

One of the biggest obstacles to Bayesian inference being more widely accepted in the neuroimaging community is the prohibitive computational burden it often imposes. This computing problem has become one of the primary concerns for researchers who work with massive data sets but still need reasonable computation time

to get results. Early attempts to deal with this in fMRI analysis were Penny et al. (2003) and Penny et al. (2005), who used variational Bayes to obtain approximations to posterior distributions. Friston and Penny (2003) suggested using empirical Bayes methods as opposed to fully hierarchical Bayes to reduce computational loads. Some empirical Bayes approaches to the more general multiple testing problem may be found in, e.g., Bogdan et al. (2008), Muralidharan (2010), and Martin and Tokdar (2012). See Scott and Berger (2010) for a discussion of conditions under which a multiplicity adjustment can be induced with both empirical Bayes and hierarchical Bayes.

Many of the Bayesian models currently found in the imaging literature rely on modeling the entire time series collected at each voxel (e.g. Genovese, 2000; Gössl et al., 2001; Fahrmeir and Gössl, 2002; Penny et al., 2003, 2005; Smith and Fahrmeir, 2007; Harrison et al., 2008). We present in this dissertation a Bayesian model that works directly with reduced imaging data. In particular, each observation is a test statistic quantifying the change in blood-oxygenation-level-dependent (BOLD) signal over the course of an fMRI experiment, averaging over the temporal dimension and thus vastly reducing the size of the data sets to be analyzed. By applying Bayesian thresholding to SPMs, it will be easier for researchers to enjoy the added flexibility of modeling complex spatial and hierarchical structures while maintaining reasonable computation times to get results.

FMRI analysis is complicated by the fact that imaging data tend to be spatially correlated. There is some literature on modeling the spatial structure of fMRI data (e.g. Hartvig and Jensen, 2000; Gössl et al., 2001; Smith and Fahrmeir, 2007; Bowman et al., 2008). Much of the work thus far developed is based on the assumption of independence of the data, though. In this work we extend the Bayesian multiple testing model considered in Scott and Berger (2006) to account for the correlation present in imaging data. Spatial dependence is introduced through a Gaussian autoregressive

model on the underlying continuous signals. This facilitates a sharing of information between voxels, allowing test statistics to not simply be evaluated on their own, but to be viewed in the context of the behavior around them. We demonstrate our model's ability to improve upon detection of task-related activation. In particular, we show how the spatial correlation induces the identification of larger clusters of activated regions, which carries more physiological meaning to neuroscientists than individually-selected voxels.

The remainder of this dissertation is organized as follows: In Chapter 2, we review the relevant background and introduce the necessary methods for constructing our spatial Bayesian testing model. This includes a brief overview of Bayesian statistics, multiple testing, and spatial dependence. We present an fMRI data set in Chapter 3 to motivate our problem and propose our testing model that incorporates spatial dependence for analyzing the data. The results of simulation studies are in Chapter 4, where we analyze the performance of both the Scott-Berger model and our own on simulated spatial data. In Chapter 5, we show results from applying the model to a real fMRI data set. We compare these to the Bayesian model under an independence assumption as well as results obtained from false discovery rate control under arbitrary dependence structure. We conclude in Chapter 6 with a discussion of these results and commentary regarding future research directions.

Chapter 2

Foundations

The work in this dissertation approaches the large-scale multiple testing problem from a Bayesian perspective. Bayesian statistics is generally less well-known among researchers (or less understood) than its classical counterpart. Indeed, computational limitations significantly hindered the Bayesian approach until the advent of reliable computing technology within the last thirty years. We briefly discuss in this chapter the rationale of Bayesian analysis and specifically its role in high throughput significance testing.

Of particular interest to us are the types of correlation induced by the spatial proximity of observations. Our focus in this dissertation is on a particular type of spatial correlation based on conditional specification of the joint distribution of the data, introduced in this chapter. For an in-depth treatment of spatial statistics, we refer interested readers to Cressie (1993). A more recent text on spatial data analysis is Schabenberger and Gotway (2005).

2.1 The Bayesian Paradigm

In general, the field of statistics may be described as a union of three ways of thinking, or what Efron (1998) termed "the statistical triangle". The Bayesian and frequentist philosophies are the two usually at loggerheads. The third approach, the likelihood or Fisherian philosophy, serves as a compromise between the two. An underlying assumption in frequentist statistics is that of a fixed, but unknown, parameter value. Knowledge of the parameter values allows probability distributions to be exactly specified or approximated so that the long-run relative frequency of events may be

calculated. It is these long-run frequencies that determine critical values or thresholds quantifying the limits of plausibility. The Bayesian paradigm is fundamentally different from the frequentist way of thinking in that it usually does not assume any single value as being the true parameter generating the observed data. Rather, a distribution of plausible parameter values is specified before any data are observed. In this context, probability distributions are treated as tools with which a researcher's uncertainty about the values that a parameter may take can be quanitfied (Berger, 1985, Chapter 3). Broad overviews of Bayesian data analysis in practice may be found in Gelman et al. (2004) and Carlin and Louis (2009). For theoretical justifications and deeper philosophical foundations, see Berger (1985), Bernardo and Smith (1994), or Robert (2007).

The aim of any Bayesian analysis is to obtain the *posterior distribution* of a (possibly vector-valued) parameter. The posterior is simply the prior distribution of a parameter, $\pi(\theta)$, after it is updated using the likelihood of the observed data, $f(y \mid \theta)$. This update is done using Bayes' Rule,

$$\pi(\theta \mid y) = \frac{f(y \mid \theta)\pi(\theta)}{g(y)},$$

where $g(y) = \int f(y, \theta) d\theta = \int f(y \mid \theta)\pi(\theta) d\theta$. The posterior is used to draw inference about θ. For certain classes of models, the posterior distribution can be worked out exactly so that moments, quantiles, etc. are precisely known.

A simple example is illustrated in Figure 2.1, in which both the prior and posterior densities are Gaussian. In this illustration, a researcher may feel *a priori* that the parameter θ is unlikely to be larger than five in magnitude, based on previous information. This is reflected in a prior density that places most of its probability mass between -5 and 5. After observing the data, the distribution of plausible values is updated so that θ is determined to be between 0 and 5 with high probability. The

Bayesian approach thus provides a formal way of justifying actual probability statements, as opposed to more awkward confidence levels that must be adopted under classical statistics.

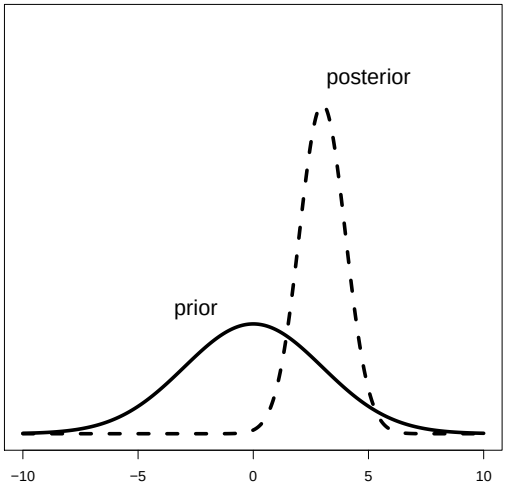


Figure 2.1: An example of prior and posterior distributions when both the likelihood and prior are Gaussian.

The previous example is an illustration of a simple case. In practice, many Bayesian models result in posterior distributions that are not known or obtainable in closed form. In such instances, indirect approaches are necessary to approximate the posterior distribution. Thus it was not until the growth of computing power and its use in Monte Carlo simulations that Bayesian analysis really burgeoned. Popular non-iterative Monte Carlo techniques include importance sampling (Hammersley and Handscomb, 1964; Berger, 1985; Geweke, 1989) and rejection sampling (Ripley, 1987).

Markov chain Monte Carlo (MCMC) is a very general class of methods that, as its name implies, uses the theory of Markov chains to approximate probability distributions. Rather than determining a limiting distribution from a known set of transition probabilities, MCMC turns the problem around and specifies a desired stationary distribution (e.g. the posterior) from which a set of transition probabilities are de-

rived. Tools for doing this may be found as early as Metropolis et al. (1953) and Hastings (1970). The literature on Markov chains is too vast to provide extensive references, but introductions may be found in, e.g., Hoel et al. (1972), Resnick (1992), or Ross (2007). The seminal work by Geman and Geman (1984) laid the foundation for MCMC methods by introducing the Gibbs sampling algorithm. Gelfand and Smith (1990) engendered interest within the Bayesian community by demonstrating how the Gibbs sampler can be used for simulating draws from complicated distributions. A gentle overview of the Metropolis-Hastings algorithm was given in Chib and Greenberg (1995). A general reference for MCMC, with numerous illustrations and case studies, is Gilks et al. (1996).

2.2 Bayesian Multiple Testing

Consider the problem of analyzing a large array of test statistics. At each point, the calculated test statistic contains information about the change in a signal observed in response to some treatment. The statistics may be values quantifying changes in light emission from astronomical observations of a cluster of stars, the amount of differential gene expression in a microarray, or the change in BOLD signal at each voxel in a brain over the course of an fMRI experiment. For J observations, we assume that each data point, y_j, $j = 1, \ldots, J$, is a realization from a Gaussian distribution with mean θ_j and common variance σ^2. That is, we suppose each statistic has its own location-specific mean. The problem is to determine whether or not the mean generating the statistic is non-zero, indicating a shift from baseline conditions.

In the context of gene microarray analysis, Scott and Berger (2006) supposed that, *a priori*, the probability of y_j being uninteresting noise is given by some unknown value, p, so that the underlying parameter is a non-null case with probability $1 - p$. The uninteresting cases are modeled by placing a prior degenerate at zero on the signal means. For the interesting, nonzero mean cases, a Gaussian distribution is

used to model plausible values. Hence the means generating the data points may be modeled using a "spike and slab" mixture prior,

$$\pi(\theta_j \mid p, \tau^2) = p\delta_0(\theta_j) + (1-p)\phi_{0,\tau^2}(\theta_j), \tag{2.1}$$

where $\delta_0(\cdot)$ is the Dirac delta spike taking θ_j to be zero with probability one, and $\phi_{0,\tau^2}(\cdot)$ is the Gaussian density with mean zero and variance τ^2.

Priors of this form are standard in the Bayesian variable selection framework. They were introduced by Mitchell and Beauchamp (1988) for variable selection in linear regression, who coined the phrase "spike and slab mixture". The mixture model was used in Geweke (1996), who provided a procedure for selecting models subject to order constraints among the variables included in each model. A similar approach was taken in George and McCulloch (1993), who treated each regression coefficient as arising from a mixture of two normal distributions with different variances for stochastic search variable selection. The model of George and McCulloch was modified in Chipman (1996) to facilitate restricted variable selection in which, for instance, interaction variables cannot be included without the lower-order main effects terms. Smith and Kohn (1996) used the spike and slab prior to select appropriate knots for nonparametric regression. This class of priors has further been used to induce adaptive regularization in genomics and proteomics by modeling sparsity of significant features (West, 2003; Lucas et al., 2006; Carvalho et al., 2008; Morris et al., 2011). Literature on Bayesian variable selection was reviewed in Clyde and George (2004). Scott and Berger (2010) studied Bayesian variable selection priors and discussed the conditions under which multiplicity correction can be induced.

For the prior specified in (2.1), it is convenient to introduce a binary indicator of activity, γ_j, so that we can use the expanded parameterization $\theta_j = \gamma_j\mu_j$, where $\mu_j \sim N(0, \tau^2)$ and $P(\gamma_j = 0) = p = 1 - P(\gamma_j = 1)$ is the probability of any y_j being

a null case. Here, $N(0, \tau^2)$ represents the Gaussian distribution with mean 0 and variance τ^2. Equation (2.1) is then a consequence of writing $\pi(\theta_j) = \pi(\theta_j \mid \gamma_j = 0)P(\gamma_j = 0) + \pi(\theta_j \mid \gamma_j = 1)P(\gamma_j = 1)$. Suitable priors may be specified for both the noise variance σ^2 and the hypervariance in the prior, τ^2.

The critical element in the Bayesian multiplicity adjustment is in the prior specified for p, the common probability of any observation being a null case (Scott and Berger, 2010). For modeling the probability of an observation being a non-null case (the *inclusion probability*; Barbieri and Berger, 2004), Scott and Berger (2006) used $\pi_p(p) = \alpha p^{\alpha-1}$, a Beta density with the second shape parameter set to one. This density is determined by specifying the parameter α, which in turn fixes the mean, mode, and variance. In Figure 2.2, we see that the density shifts closer to one as α increases. The density has mean $\alpha/(\alpha + 1)$, indicating that the *a priori* probability of being an interesting case decreases with increasing α. Taking this view, α can be used as a tuning parameter to model a researcher's initial beliefs about the likely prevalence of non-null cases in a data set.

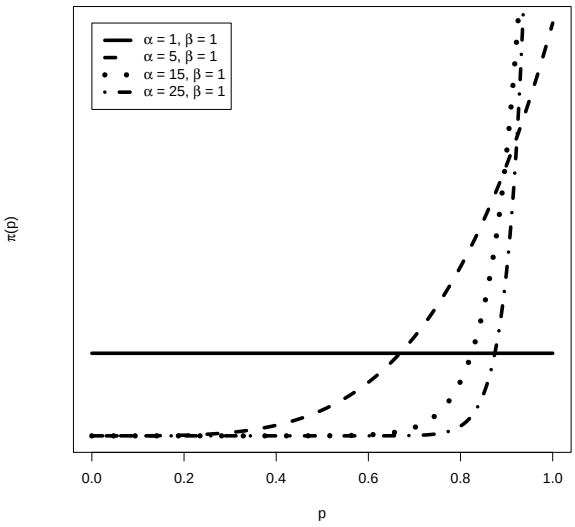


Figure 2.2: Beta densities of the form $\pi_p(p) = \alpha p^{\alpha-1}$, for selected values of α

The *a posteriori* effect of α is seen by noting the updated shape parameter in the posterior distribution. Conditional on the indicators of activity, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)^T$, the α parameter is updated to $\alpha' := J - \sum \gamma_i + \alpha$, where J is the number of tests being simultaneously performed, i.e. the number of locations at which data have been observed. The second shape parameter changes from 1 to $\beta' := \sum \gamma_i + 1$ in the posterior, so that the posterior mean becomes $\alpha'/(\alpha'+\beta') = (J-\sum \gamma_i+\alpha)/(J+\alpha+1)$. The consequence is that α and $\boldsymbol{\gamma}$ have opposite effects on the posterior probability of being non-null. Greater values of α are required as $\sum \gamma_i$ increases to offset an effect that would otherwise select a larger number of data points. A larger value of α is, in effect, a stronger statement about the likelihood of observing a non-null signal.

In general, strongly informative priors have a greater influence on the posterior so that more information is required to reduce their effect. Such priors are often necessary in the presence of a large number of data points. An informative prior on p can be a way of modeling the expected sparsity of interesting cases to be observed (West, 2003; Lucas et al., 2006; Carvalho et al., 2008). It can be used to reflect a researcher's belief that interesting cases are few relative to the total number of observations by specifying an *a priori* lower probability that any particular point is non-null.

Defining $p_j := P(\gamma_j = 0 \mid \mathbf{y})$ to be the marginal posterior probability of a null case at location j, Scott and Berger (2006, Lemma 3) showed that

$$p_j = \int_{[0,\infty)\times[0,\infty)\times[0,1]} \left[1 + \frac{1-p}{p}\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left(\frac{y_j^2\tau^2}{2\sigma^2(\sigma^2+\tau^2)}\right)\right]^{-1} d\Pi(\tau^2, \sigma^2, p \mid \mathbf{y}),$$

where $\Pi(\tau^2, \sigma^2, p \mid \mathbf{y})$ is the joint posterior distribution of the variance parameters and p. With increasing J and the other parameters held fixed, p converges in probability to one in the posterior so that $(1-p)/p \approx 0$ with high probability under a large number of tests. When this term is zero in the integral, the rest of the integrand is

that of a probability density and thus integrates to one over its support. In other words, p_j approaches one as J increases, forcing the probability of y_j being null to increase. This provides an intrinsic adjustment for the number of tests.

The inherent calibration of the Bayesian procedure is certainly appealing, as is its simple interpretation. This model is limited, however, in that it assumes no dependence structure between the data points. The signal means that we are ultimately interested in are modeled to be independent of one another, as are the observations themselves.

2.3 Conditional Autoregressive Models

Suppose $\{X_i = X(\mathbf{s}_i) : \mathbf{s}_i \in \mathcal{D} \subset \mathbb{R}^p, i = 1, \ldots, J\}$ is a set of random variables constituting a spatial process. Each X_i is a realized value from some distribution at location $\mathbf{s}_i$. A *conditional autoregressive*, or CAR, model (Besag, 1974) gives the joint distribution of $\mathbf{X} = (X_1, \ldots, X_J)^T$ by specifying the conditional distribution at each X_i, given all the other values in the field. It is assumed that each random variable depends on the others only through its immediate neighbors. Spatial processes such as these are known as *Markov random fields*. They can be viewed as spatial analogs to the more commonly known Markov chains. It should be noted that specifying an arbitrary set of conditional distributions does not guarantee the existence of a valid joint distribution. Besag (1974), using Brook's Lemma (Brook, 1964) and the Hammersley-Clifford Theorem (Hammersley and Clifford, 1971), provided conditions under which the joint distribution is guaranteed to exist. The Gaussian CAR model is one such example (Besag and Kooperberg, 1995). This model assumes that

$$X_i \mid \mathbf{x}_{(-i)} \sim N\left(\eta_i + \sum_{j=1}^{J} c_{ij}(x_j - \eta_j), \quad \sigma_i^2\right), \tag{2.2}$$

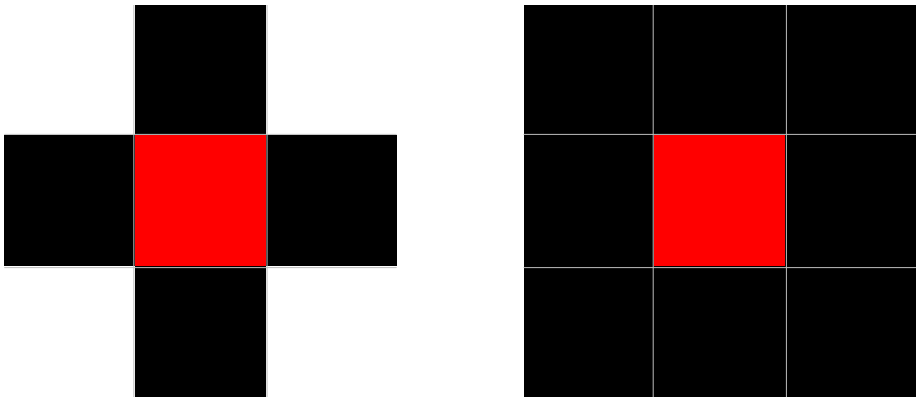
where $c_{ii} = 0$, $c_{ij} = 0$ except when $\mathbf{s}_i$ and $\mathbf{s}_j$ are neighbors, and $c_{ij} = c_{ji}$. Here, $\mathbf{x}_{(-i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_J)^T$ is the vector of all observations except x_i. This conditional distribution is determined in part by the set $\{X_j : c_{ij} \neq 0\}$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_J)^T$, a vector of location parameters providing a center or baseline for each of the conditional distributions of X_i. It is seen, then, that $X_i - \eta_i$ is a random variable centered around some linear combination of its neighbors. This model also has location-specific variances, σ_i^2, thus allowing for the neighbors to influence the variability of X_i as well. The resulting joint density, when it exists, is given by

$$f(\mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\eta})^T \mathbf{D}^{-1}(\mathbf{I} - \mathbf{C})(\mathbf{x} - \boldsymbol{\eta})\right),$$

where $\mathbf{C} = \{c_{ij}\}_{i,j=1}^J$ and $\mathbf{D} = \text{diag}\{\sigma_i^2, i = 1, \ldots, J\}$. The precision matrix $\mathbf{D}^{-1}(\mathbf{I} - \mathbf{C})$ is a measure of the reliability of information contained in a sample from the distribution of $\mathbf{X}$. Since the variability of X_i with respect to X_j is the same as the variability of X_j with respect to X_i, this matrix must be symmetric. The condition that $c_{ij}/\sigma_i^2 = c_{ji}/\sigma_j^2$ is thus required for all pairs of locations i and j.

There are any number of ways that a neighbor may be defined in a CAR model. For example, we could take a "rook" structure, where $\mathbf{s}_i$ and $\mathbf{s}_j$ are neighbors if and only if they share a side, or we could borrow more neighboring information by using a "queen" structure, taking those locations with shared sides or shared corners as a neighborhood (Figure 2.3). The c_{ij} parameters can also be used to assign neighbors differing weights in a neighborhood. For example, inverse distance weighting can be used to downweight the information contained in points further away from the center without ignoring them altogether.

A special case of the Gaussian CAR model can be developed by defining adjacency indicators $w_{ij} := I(i \sim j)$, where $I(\cdot)$ is the indicator function and $i \sim j$ if and only if sites $\mathbf{s}_i$ and $\mathbf{s}_j$ are neighbors. Let $w_{i\cdot} = \sum_{j=1}^J w_{ij}$ be the number of neighbors for



(a) Rook (b) Queen

Figure 2.3: Two examples of neighborhood structures for a CAR model

location i and define the the adjacency matrix to be $\mathbf{W} = \{w_{ij}\}_{i,j=1}^{J}$. By taking $\sigma_i^2 = \sigma^2/w_{i\cdot}$ and $c_{ij} = w_{ij}/w_{i\cdot}$ in (2.2), the CAR model becomes

$$X_i \mid \mathbf{x}_{(-i)} \sim N\left(\eta_i + \frac{1}{w_{i\cdot}}\sum_{j=1}^{J} w_{ij}(x_j - \eta_j), \quad \frac{\sigma^2}{w_{i\cdot}}\right), \tag{2.3}$$

with precision matrix

$$\mathbf{D}^{-1}(\mathbf{I} - \mathbf{C}) \equiv \sigma^{-2}(\mathbf{D}_w - \mathbf{W}), \tag{2.4}$$

where $\mathbf{D}_w = \text{diag}\{w_{i\cdot}, i = 1, \ldots, J\}$. This is known as an *intrinsic autoregressive (IAR) model* (Besag et al., 1991).

The IAR is appealing because of its intuitive interpretation. At each location i, the average of the neighbors is used as the center for the distribution of $X_i - \eta_i$. The IAR is particularly attractive from a Bayesian point of view because of the ease with which it can be incorporated into Gibbs sampling for simulating posterior distributions. See Banerjee et al. (2004) for a discussion of CAR models in Bayesian inference.

The interpretability of IAR models makes them desirable as priors on the parameters governing the distributions of observed values. The potential difficulty in working with IAR models, though, is seen by noting that $(\mathbf{D}_w - \mathbf{W})(1, \ldots, 1)^T = (0, \ldots, 0)^T$,

implying that the precision matrix is singular. The inverse of this matrix is needed in the normalizing constant of the joint density. The absence of a normalizing constant implies that $\int f(\mathbf{x})d\mathbf{x} = \infty$ so that the distribution is *improper*. When using improper priors, care needs to be taken to ensure that the posterior distribution is proper (integrates to one). Otherwise, inference would attempt to find moments, percentiles, etc. that do not exist. In many applications the impropriety of a prior distribution is of no consequence because the posterior will still be proper. Nevertheless, our proposed model requires the IAR model to be proper. We discuss why this is true in Chapter 3.

We now introduce a parameter ρ so that the precision matrix is redefined to be $\Sigma^{-1} := \sigma^{-2}(\mathbf{D}_w - \rho\mathbf{W})$. By placing appropriate bounds on ρ, we can guarantee that Σ^{-1} is positive definite and thus a valid precision matrix. This result was given in Banerjee et al. (2004). We state it as a Lemma, though, and provide a proof since the notation and ideas contained therein will be used in Section 3.2. The proof uses a simple but useful result that we state as a preliminary Lemma. We borrow the terminology of Banerjee et al. and call ρ a "propriety parameter".

Lemma 1. *For any $n \times n$ matrix $\mathbf{A}$ with eigenvalues δ_i, $i = 1, \ldots, n$, and for any constants a and $b \neq 0$, the eigenvalues of $a\mathbf{I} + b\mathbf{A}$, ψ_i, can be put in one-to-one correspondence with the eigenvalues of $\mathbf{A}$ as $\psi_i = a + b\delta_i$, $i = 1, \ldots, n$.*

Proof. Let δ be an eigenvalue of $\mathbf{A}$ with eigenvector $\mathbf{e}$. Then

$$\begin{aligned}(a\mathbf{I} + b\mathbf{A})\mathbf{e} &= a\mathbf{e} + b\mathbf{A}\mathbf{e} \\ &= a\mathbf{e} + b\delta\mathbf{e} \\ &= (a + b\delta)\mathbf{e},\end{aligned}$$

so $a + b\delta \equiv \psi$ is an eigenvalue of $a\mathbf{I} + b\mathbf{A}$ with eigenvector $\mathbf{e}$. Since $b \neq 0$, this transformation is bijective, establishing the result. □

Lemma 2. *A sufficient condition for the matrix* $(\mathbf{D}_w - \rho\mathbf{W})$ *to be positive definite is* $\lambda_1^{-1} < \rho < \lambda_J^{-1}$*, where* $\lambda_1 < 0$ *and* $\lambda_J > 0$ *are the smallest and largest eigenvalues of* $\mathbf{D}_w^{-1/2}\mathbf{W}\mathbf{D}_w^{-1/2}$*, respectively.*

Proof. We must choose ρ such that $\mathbf{D}_w - \rho\mathbf{W} > 0$, where the notation $\mathbf{A} > 0$ means the matrix $\mathbf{A}$ is positive definite. Since

$$\mathbf{D}_w - \rho\mathbf{W} = \mathbf{D}_w^{1/2}(\mathbf{I} - \rho\mathbf{W}^*)\mathbf{D}_w^{1/2},$$

$\mathbf{W}^* = \mathbf{D}_w^{-1/2}\mathbf{W}\mathbf{D}_w^{-1/2}$, and $\mathbf{D}_w > 0$, it suffices to restrict ρ such that $\mathbf{I} - \rho\mathbf{W}^* > 0$.

Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_J$ be the ordered eigenvalues of $\mathbf{W}^*$. Then we have by Lemma 1 that the eigenvalues of $\mathbf{I} - \rho\mathbf{W}^*$ are $1 - \rho\lambda_j$, $j = 1, \ldots, J$. Since $\mathbf{D}_w$ is diagonal and the diagonal elements of $\mathbf{W}$ are zero, the diagonal elements of $\mathbf{W}^* = \mathbf{D}_w^{-1/2}\mathbf{W}\mathbf{D}_w^{-1/2}$ are also zero and thus $\text{tr}(\mathbf{W}^*) = 0 = \sum_{j=1}^{J} \lambda_J$. It must then be true that there are $r_1 > 0$ negative eigenvalues and $r_2 > 0$ positive eigenvalues of $\mathbf{W}^*$, since $r := r_1 + r_2 = \text{rank}(\mathbf{W}^*) > 0$.

Finally, $\mathbf{I} - \rho\mathbf{W}^* > 0$ if and only if $1 - \rho\lambda_j > 0 \Rightarrow 1 > \rho\lambda_j$, for all j. It follows that $\lambda_j^{-1} < \rho$, $j = 1, \ldots, r_1$ and $\lambda_j^{-1} > \rho$, $j = r - r_2 + 1, \ldots, J$. In other words, it must be the case that $\lambda_1^{-1} < \rho < \lambda_J^{-1}$. This completes the proof. □

For massive spatial arrays such as those encountered in fMRI, $\mathbf{D}_w^{-1/2}\mathbf{W}\mathbf{D}_w^{-1/2}$ will be of large dimension, making eigenvalue computation difficult. The adjacency matrix depends only on the spatial structure of the data and no unknown parameters, though, so the eigenvalues only need to be calculated once and not updated in an MCMC routine. In cases where eigenvalue calculation is prohibitive, Carlin and Banerjee (2003) showed that it is unnecessary, using the scaled adjacency matrix to rewrite the precision matrix with a parameter that only has the restriction $|\rho| < 1$. The intrinsic autoregression can thus be viewed as a limiting case of the conditional autoregression where the propriety parameter approaches one (Besag and Kooperberg, 1995).

Another problem with either the IAR or its proper counterpart is that the location-specific variances are scaled by the number of neighbors adjacent to the location in question. Since the locations along the edges of a spatial array have differing numbers of neighbors, the random variables at these locations will have different marginal and conditional variances, and hence different correlation structure, than the rest of the data. Besag and Kooperberg (1995) addressed this problem by providing an algorithm attributed to Dempster (1972) which constructs a covariance structure with uniform conditional variances among all data points. They also discussed another approach of simply replacing the missing neighbors on the edges with a typical value of the data such as the median. Such remediation is more of a concern when dealing with smaller data sets so that a greater proportion of the observations lie on the edge. For fMRI, the large number of observations in the interior of the array tends to mitigate the impact of the heterogenous correlation structures on the edge. For this reason, we ignore edge effects here.

Chapter 3

Motivation and Methods

With the background established and the necessary tools in place, we are ready to narrow our focus to the specific application underlying this dissertation. This chapter introduces an experimental data set collected in an imaging study. In presenting it, we describe a common technique for the analysis of such data, the voxelwise general linear model, and how it leads to the construction of statistical parametric maps. The features of this data set and statistical issues associated with them serve as our motivation for proposing a Bayesian method for its analysis. We propose two models. The first is tailored for implementation in `WinBUGS` (Lunn et al., 2000), a free program for modeling and drawing posterior inference via MCMC for a wide range of hierarchical models. Such software is useful in the embryonic stages of model development when sensitivity and operating characteristics are the dominating interests over computational issues. A second model also is proposed as a more objective alternative to the first, but its implementation requires a researcher to write customized sampling algorithms.

3.1 Motivation

The Analysis of fMRI Data with the General Linear Model

The study of functional magnetic resonance imaging has seen a remarkable growth in the last decade. It is the study of brain activation associated with various stimuli that are presented under experimental conditions. Neuronal activity is observed indirectly by changes in a magnetic field that result from greater oxygenation in the blood, of which an increase in the blood-oxygenation-level-dependent (BOLD) signal is a

direct consequence. Magnetic resonance (MR) imaging provides data that are images of the brain, each consisting of multiple slices partitioned into three-dimensional pixels called voxels. Voxels are collected in a series of matrices, each representing one slice of many that make up the full three-dimensional brain volume. The resolution of the matrices is set by the experimenter and varies across studies depending on the type of data collected and experimental protocol invoked. Regardless of whether higher or lower resolution is acquired, the data to be analyzed contain a large number of voxels, on the order of thousands per slice. For functional MR images, each of these voxels contains information about the signal at a particular location over time. This is what makes it *functional* imaging as opposed to MRI, which only maps the structural anatomy without observing time-dependent signals.

It is common practice to analyze BOLD fMRI data using a voxelwise general linear model (GLM; Friston et al., 1995). This approach models the MR signal at voxel j using a relationship of the form

$$\mathbf{y}_j = X\boldsymbol{\beta}_j + \mathbf{e}_j,$$

where $\mathbf{y}_j = (y_{j,1}, \ldots, y_{j,T})^T$ is the time course of MR signals at voxel j, X is the $T \times p$ design matrix, $\boldsymbol{\beta}_j$ is the $p \times 1$ vector of regression coefficients, and $\mathbf{e}_j$ is the (often serially correlated) error in the measurements. This equation is sometimes written as

$$\mathbf{y}_j = Z\boldsymbol{\eta}_j + H\boldsymbol{\delta}_j + \mathbf{e}_j$$

to separate the effects of interest, $\boldsymbol{\eta}_j$, from nuisance covariates such as motion correction and low-frequency trends represented by $\boldsymbol{\delta}_j$. Inference then focuses on the coefficients in $\boldsymbol{\eta}_j$ such as the stimulus time course to determine which voxels are significantly associated with the experimental stimulus. Analyzing the full time course at each voxel in a three-dimensional image can involve processing tens of millions of

observations, creating huge computational demands (Friston and Penny, 2003; Smith and Fahrmeir, 2007).

Alternatively, researchers may choose to analyze fMRI data through the use of statistical parametric maps (SPMs). With this approach, each voxel is assigned a summary statistic quantifying the effect of the factor of interest. Inference then focuses on the observed statistic at each voxel, which has a known distribution under a null hypothesis. One of the advantages of the SPM approach is that the data to be processed are collapsed over the temporal dimension. The most computationally intensive step of the analysis (e.g. Bayesian MCMC) can thus be simplified by working with the reduced data.

Experimental Data

The data we consider in this work are from a study by Camchong et al. (2008) investigating the differences in neural activation patterns associated with cognitive control tasks that require generation of volitional saccades. The task involved alternating between blocks of fixation (baseline) and the volitional saccade task known as an antisaccade. Antisaccades require that participants inhibit a glance towards a prepotent cue and generate one to the mirror image location (opposite side of the screen, same distance from center). During fixation blocks, participants fixed the gaze on a point for a duration of 22.5 seconds. For the antisaccade blocks, a single central point was presented for 1.7 seconds followed by a dot presented 8 degrees to the left or right for 1.25 seconds. The participants were asked to move their eyes to the mirror image of the target as quickly and as accurately as possible. Each run consisted of nine fixation periods alternated with eight antisaccade trials. One blocked run was recorded for each participant.

Data were obtained using a GE Signa Horizon LX 1.5T MRI scanner. Each functional run collected T_2^*-weighted images with in-plane resolution 3.75×3.75 millime-

ters, TE = 40 milliseconds, TR = 1912 milliseconds, 3800 milliseconds acquisition time, flip angle 77 degrees. The images were processed with AFNI software (Cox, 1996). Volumes were registered to a middle volume to correct for minor head motion. A Gaussian filter with full width, half-maximum of 4 millimeters was applied to smooth the data. The data were transformed to Talaraich space (Talaraich and Tournoux, 1988), resulting in 4×4×4 millimeter resolution. Each run was modeled with a linear regression including covariates for linear drift and head motion as well as the stimulus time course. For details about the data collection and preprocessing, see Dyckman et al. (2007) and Camchong et al. (2008).

The data analyzed in this work are the resulting SPMs from fourteen healthy participants in the study, with slices of dimension 40×48. Voxels are masked so that only those inside the brain are used in the analysis, about 700-900 voxels per subject. The statistics in each slice are treated as the observed data, y_j, $j = 1, \ldots, J_m$, where J_m is the total number of non-masked voxels specific to participant m, $m = 1, \ldots, 14$. For each of the participants, we analyze the slice located at $Z = +40$ in Talaraich coordinates.

For each voxel j, $j = 1, \ldots, J_m$, the test statistic is calculated as

$$y_j = \frac{\bar{Z}_j^{\text{task}} - \bar{Z}_j^{\text{baseline}}}{se},$$

where $\bar{Z}_j^{\text{task}}$ is the average BOLD signal observed at voxel j during all task periods, $\bar{Z}_j^{\text{baseline}}$ is the average signal observed during all baseline periods, and se is the standard error of $\bar{Z}_j^{\text{task}} - \bar{Z}_j^{\text{baseline}}$. These statistics have a large number of degrees of freedom so that we may assume each data point, y_j, is a realization from a Gaussian distribution with mean θ_j and common variance σ^2. The objective is to determine whether or not the mean generating each statistic is non-zero, indicating a shift from baseline conditions.

In a single slice from one individual, there are approximately 800 tests conducted simultaneously; more than 11,000 if single slices from each subject are considered together in a groupwise analysis. This creates a multiple testing problem exacerbated by dependence in the data. On a large scale, the correlation among voxels is not necessarily a function of Euclidean distance since distinct regions of the brain may support the same behaviors. Locally, though, it is sensible to allow information to be shared across small neighborhoods of voxels. This is reasonable since the voxels are artificial partitions of the brain. Neither the BOLD signal changes nor the neural correlates underlying that signal are constrained by voxel boundaries. It is likely that activation is spread across regions when it occurs (Forman et al., 1995). By allowing the inferences made at one voxel to be influenced by adjacent voxels, there is potential for more power to detect task-related activation. Conversely, the absence of anything interesting in the neighborhood of what would otherwise be viewed as an activated voxel would make it more likely for this point to be dismissed as spurious. Our model is proposed to incorporate this intuition by allowing voxels that are close to each other to share information.

3.2 Proposed Model

Choice of Prior Distributions

The statistic calculated at each voxel in an SPM is an areal summary of a small part of the brain, usually on the order of two to five cubic millimeters. The resulting lattice structure lends itself to CAR models. We can account for local spatial dependence in the data by using a modified IAR model for the joint distribution of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{J_m})^T$, defining a neighborhood of a voxel to be the set of all voxels that share a side or a corner with it so that a voxel's neighbors are the eight voxels surrounding it in a two-dimensional slice (as in Figure 2.3b). The IAR model must be adjusted to ensure that it is proper.

To see why propriety is required in the prior for $\boldsymbol{\mu}$, consider the case when there is no activation anywhere. Then $\gamma_j = 0$ for all j and each test statistic comes from a Gaussian distribution with mean 0. The vector $\boldsymbol{\mu}$ does not appear in the resulting likelihood, so its marginal distribution is not updated in the posterior. The posterior of $\boldsymbol{\mu}$ is exactly equal to the prior. For this reason, the priors on parameters that only appear in certain components of mixture distributions must be proper. The argument is the same for using a proper prior on the hypervariance in the prior for $\boldsymbol{\mu}$. See McLachlan and Peel (2000, Chapter 4) or Gelman et al. (2004, Chapter 18) for discussions of Bayesian mixture modeling.

To force the joint prior on the signal means to be proper, we introduce a propriety parameter ρ into the precision matrix so that it becomes $\sigma^{-2}(\mathbf{D}_w - \rho\mathbf{W})$. By Lemma 2, a sufficient condition to ensure that the prior is proper is for ρ to be between λ_1^{-1} and $\lambda_{J_m}^{-1}$, where $\lambda_1 < 0$ and $\lambda_{J_m} > 0$ are the smallest and largest eigenvalues, respectively, of $\mathbf{D}_w^{-1/2}\mathbf{W}\mathbf{D}_w^{-1/2}$. We thus specify the conditional distributions as

$$\mu_j \mid \boldsymbol{\mu}_{(-j)}, \tau^2 \sim N\left(\frac{\rho}{w_j.}\sum_{j\sim i}\mu_i, \frac{\tau^2}{w_j.}\right), \quad j = 1, \ldots, J_m, \tag{3.1}$$

where $w_j.$ and $j \sim i$ are as defined in Section 2.3.

There is no obvious value to use for ρ. It can be estimated or, alternatively, assigned a prior distribution. We only assume that there is a positive association between a voxel and its neighbors. We thus assign a uniform prior on the interval $(0, \lambda_{J_m}^{-1})$. An advantage of modeling ρ in this manner is that only the largest eigenvalue associated with the adjacency matrix needs to be calculated. Posterior simulation results usually yield estimates of ρ in the range (.995, .999) for the slices we consider in this work. This parameter then is only a minor adjustment to the otherwise more desirable IAR, which is not unusual. Appreciable interaction in this type of model seems to require the propriety parameter to be close to one (Banerjee et al., 2004).

We also wish to avoid strong information about either the noise variance, σ^2, or the hypervariance, τ^2. We specify the prior distribution of the data standard deviation σ to be uniform on the interval $(0, 1000)$. This vague prior avoids strong *a priori* influence on the possible variation, allowing the data themselves more flexibility in determining reasonable values. The proper uniform is used to approximate an improper prior in `WinBUGS`, an approach suggested by Gelman (2006) and Carlin and Louis (2009) when using the software. Gelman (2006) notes that supposedly noninformative hypervariance priors may have disproportionate effects on inference. In Section 5.3 below, we perform a sensitivity analysis comparing the results with different variance priors, including inverse Gamma, uniform, and folded-t, following those considered in Gelman (2006). The results are found to be quite insensitive to the prior chosen. We decide to use a Gamma distribution on the precision parameter, $1/\tau^2 \sim \text{Ga}(.001, .001)$. The distribution of $1/\tau^2$ then has a mean of $.001/.001 = 1$ with standard deviation $\sqrt{.001(1/.001)^2} \approx 32$, allowing a wide range of plausible values. We discuss later in this chapter an alternative model that uses the variance priors suggested by Scott and Berger (2006).

Lastly, we note that the alternative hypothesis reflected in the continuous component of the mixture prior allows for $\theta_j < 0$ so that the prior is actually performing a two-sided test. In fMRI, some neural circuits actually decrease their activity as tasks are performed (Murphy et al., 2009; Cole et al., 2010; Margulies et al., 2010; Power et al., 2012; Saad et al., 2012). Task-related activation, though, is characterized by an increase in the BOLD signal during stimulus. Thus, while the mixture prior is modeling non-null voxels that either increase or decrease in BOLD signal, we restrict our attention to those exhibiting positive change in order to identify task-related activation. We are, of course, implicitly assuming positive and negative changes are symmetric in distribution. This is a shortcoming that will be addressed in future work.

Implementation

Our proposed models are modifications of the Scott-Berger model presented in Section 2.2. That is, we treat the observed test statistics as arising from a two-component mixture, $y_j \mid \mu_j, \sigma, p \sim pN(0, \sigma^2) + (1-p)N(\mu_j, \sigma^2)$, $j = 1, \ldots, J_m$. Equivalently, we may write this as $y_j \mid \gamma_j, \mu_j, \sigma \sim N(\gamma_j \mu_j, \sigma^2)$ with $\gamma_j \sim \text{Bernoulli}(1-p)$. We follow Scott and Berger and assign p a prior of the form $\pi_p(p) = \alpha p^{\alpha-1}$. Our main modification to the Scott-Berger model is in the joint prior distribution of the continuous, underlying signals, $\boldsymbol{\mu}$. We summarize Model 1 as follows:

Model 1:

- $y_j \mid \gamma_j, \mu_j, \sigma \overset{ind}{\sim} N(\gamma_j \mu_j, \sigma^2), \quad j = 1, \ldots, J_m$
- $\gamma_j \mid p \overset{ind}{\sim} \text{Bernoulli}(1-p), \quad j = 1, \ldots, J_m$
- $\mu_j \mid \boldsymbol{\mu}_{(-j)}, \tau, \rho \sim N\left(\frac{\rho}{w_{j\cdot}} \sum_{j \sim i} \mu_i, \frac{\tau^2}{w_{j\cdot}}\right), \quad j = 1, \ldots, J_m$
- $\sigma \sim \text{U}(0, 1000)$
- $p \sim \text{Beta}(\alpha, 1), \quad \alpha \geq 1$
- $\rho \sim \text{U}(0, \lambda_{J_m}^{-1})$
- $\tau^{-2} \sim \text{Ga}(.001, .001)$

We write the distributions of μ_j, $j = 1, \ldots, J_m$, conditionally to underscore the fact that working with the conditional distributions is easier than the full joint distribution. The mixture distribution of the data makes the full posterior distribution difficult to specify analytically. Using MCMC, it can be approximated through iterative sampling. This is fairly straightforward, but with a couple of difficulties. We discuss these in the following remarks.

Remark 1. *In each iteration of an MCMC algorithm, the conditional distribution of μ_j is only updated when $\gamma_j = 1$. In exploring the posterior parameter space, an*

algorithm risks reaching areas where $\gamma_j = 0$ for all j, or nearly so. In this case, the mean parameters μ_j are not updated, potentially causing the procedure to take a long time to escape this part of the space, leading to very slow convergence. If this event has sufficiently low probability, the issue can be ignored (Gelman et al., 2004, Chapter 18). Otherwise, a delicate MCMC procedure may be necessary for this step.

Remark 2. *Related to Remark 1 is the fact that parameterizing the signal means as $\theta_j = \gamma_j \mu_j$ is an overparameterization. With only one observation at each point, the data contain no information about μ_j when $\gamma_j = 0$. Our interest here, though, is only in the θ_j and, in particular, which of them are positive. If one is interested in the amplitudes of the continuous signals throughout an image, rather than just thresholding to locate potentially interesting voxels, additional constraints are necessary to ensure identifiability, specifying $\mu_j = 0$ when $\gamma_j = 0$, for instance.*

Remark 3. *In fitting our model to fMRI data, we find an extreme sensitivity to α in the prior for p. While this parameter can be viewed as a tuning parameter to reflect prior belief about the prevalence of activation (see Section 2.2), the best way to determine it in practice is not clear.*

In this work we only consider one slice from each participant to keep the computations relatively simple while still elucidating the behavior of our model. This simplification makes it feasible to implement our model using `WinBUGS`. Functions exist for both `R` (R Development Core Team, 2012) and `MATLAB` (The MathWorks, Inc., Natick, MA) with which users can call `WinBUGS` as part of a larger program. We use the `arm` library (Gelman et al., 2012) in `R` and the `matbugs` function (Murphy and Mahdaviani, 2005) in `MATLAB`.

Our experience has been that `WinBUGS` can work for relatively small-scale cases such as single-slice analysis. In practice, a three-dimensional or whole-brain analysis is usually desired, in which case more efficient Monte Carlo algorithms are required.

We do not recommend using `WinBUGS` for multi-slice or whole brain analysis, as the program becomes unstable with larger data sets. The computational burden increases nonlinearly; required computing time can increase ten-fold for even a doubling of the size of the data set. Scott and Berger (2006) describe an importance sampling algorithm in implementing their independence model which may also be feasible in our case. The conditional specification of the CAR structure, though, makes Gibbs sampling better suited for our model, with Metropolis-Hastings steps incorporated for the non-standard distributions.

Alternative Specification with an Improper Prior

In situations where `WinBUGS` is no longer desirable, we are free to use improper priors (provided the posterior is proper). We no longer need to approximate an improper prior on σ with a uniform prior over an interval of finite Lebesgue measure. This leads us to propose an alternative model. We again follow Scott and Berger (2006) and suggest specifying σ^2 and τ^2 with a joint prior given by

$$\pi_{(\tau^2,\sigma^2)}(\tau^2,\sigma^2) = (\tau^2+\sigma^2)^{-2}; \quad \sigma^2,\tau^2 > 0.$$

This prior can be written as $\{(1/\sigma^2)(1+\tau^2/\sigma^2)^{-2}\}(1/\sigma^2) \equiv \pi_{\tau^2|\sigma^2}(\tau^2 \mid \sigma^2)\pi_{\sigma^2}(\sigma^2)$, preserving the required propriety on τ^2 while incorporating a noninformative prior for σ^2. To accommodate additional objectivity in the alternative specification, we consider a prior for ρ in which the uniform interval is extended from $(0, \lambda_{J_m}^{-1})$ to $(\lambda_1^{-1}, \lambda_{J_m}^{-1})$. We call this alternative Model 2, summarized as follows:

Model 2:

- $y_j \mid \gamma_j, \mu_j, \sigma \overset{ind}{\sim} N(\gamma_j\mu_j, \sigma^2), \quad j = 1,\ldots,J_m$
- $\gamma_j \mid p \overset{ind}{\sim} \text{Bernoulli}(1-p), \quad j = 1,\ldots,J_m$
- $\mu_j \mid \boldsymbol{\mu}_{(-j)}, \tau, \rho \sim N\left(\frac{\rho}{w_{j\cdot}}\sum_{j\sim i}\mu_i, \frac{\tau^2}{w_{j\cdot}}\right), \quad j = 1,\ldots,J_m$

- $p \sim \text{Beta}(\alpha, 1), \quad \alpha \geq 1$
- $\rho \sim \text{U}(\lambda_1^{-1}, \lambda_{J_m}^{-1})$
- $\pi_{\tau^2|\sigma^2}(\tau^2 \mid \sigma^2) = \left(\frac{1}{\sigma^2}\right)\left(1 + \frac{\tau^2}{\sigma^2}\right)^{-2}, \quad \tau^2 > 0$
- $\pi_{\sigma^2}(\sigma^2) = \frac{1}{\sigma^2}, \quad \sigma^2 > 0$

In this dissertation, we implement Model 1 to take advantage of the computational convenience offered by `WinBUGS`. This allows us to focus our attention on the modeling aspects of the Bayesian analysis, investigating the effects of hyperparameter choices and combined versus individual analyses. Model 2 is proposed as a more flexible alternative to Model 1 that can be used when a more large-scale analysis is desired. We establish the propriety of Model 2 in the next subsection to ensure that any inferences based on the second model would be valid. Specialized algorithms beyond the scope of `WinBUGS` are needed to sample from the posterior in Model 2, although samples from the posterior of Model 1 via `WinBUGS` could possibly be used as candidate draws in an importance sampling algorithm for Model 2. We keep discussion of computing time or efficiency to a minimum in this dissertation, though, as they are not our intended foci. We wish only to raise awareness of computational issues, not to be prescriptive.

Propriety of the Posterior Distributions

For Model 1, we have only proper priors, so the posterior distribution is necessarily proper since, for any proper prior $\pi(\theta)$, $f(y, \theta) = f(y \mid \theta)\pi(\theta)$ is a valid joint distribution on (y, θ) and thus the posterior normalizing constant is $g(y) = \int f(y \mid \theta)\pi(\theta)d\theta < \infty$. When using improper priors, as we suggest in Model 2, it is important to check that the posterior distribution is proper. We claim this as a Proposition and provide a proof. In the proof, we make use of a simple inequality. We state it as a preliminary Lemma before establishing the main result.

Lemma 3. *For any positive constants a, b, and c,*

$$\frac{b}{ba+c} < \frac{\max\{b,1\}}{a+c}.$$

Proof. Let $a, c \in \mathbb{R}^+$. Then, for $0 < b < 1$,

$$\begin{aligned} \frac{b}{ba+c} - \frac{1}{a+c} &= \frac{b(a+c) - ba - c}{(ba+c)(a+c)} \\ &= \frac{c(b-1)}{(ba+c)(a+c)} \\ &< 0, \end{aligned}$$

and for $b > 1$,

$$\begin{aligned} \frac{b}{ba+c} - \frac{b}{a+c} &= \frac{b(a+c) - b(ba+c)}{(ba+c)(a+c)} \\ &= \frac{ba(1-b)}{(ba+c)(a+c)} \\ &< 0. \end{aligned}$$

□

Proposition. *The posterior distribution of Model 2 is proper.*

Proof. First, note that for a general parameter θ and data y, the posterior is given by

$$\pi(\theta \mid y) = \frac{f(y \mid \theta)\pi(\theta)}{g(y)},$$

where $g(y) = \int f(y \mid \theta)\pi(\theta)d\theta = \int f(y, \theta)d\theta$. Thus, to show that the posterior is proper, it suffices to show that $\int f(y, \theta)d\theta < \infty$, for all y (*a.e.*).

Let $\boldsymbol{\Theta} = (\boldsymbol{\mu}^T, \sigma^2, \tau^2, \rho)^T$ and define $J := J_m$ to suppress the dependence of the number of tests on m. Then for Model 2, we need to show that

$$\int_{(\boldsymbol{\Theta},\boldsymbol{\gamma},p)} dF(\mathbf{y},\boldsymbol{\Theta},\boldsymbol{\gamma},p) = \sum_{\boldsymbol{\gamma}\in\{0,1\}^J} \int_p \int_{\boldsymbol{\Theta}} f(\mathbf{y},\boldsymbol{\Theta},\boldsymbol{\gamma},p) d\boldsymbol{\Theta} dp < \infty. \tag{3.2}$$

First, note that for any $\boldsymbol{\gamma} \in \{0,1\}^J$,

$$\begin{aligned} f(\mathbf{y},\boldsymbol{\Theta},\boldsymbol{\gamma},p) &= (2\pi)^{-J/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{J}(y_j-\gamma_j\mu_j)^2\right) \\ &\quad \times (2\pi)^{-J/2}|\tau^2(\mathbf{D}_w-\rho\mathbf{W})^{-1}|^{-1/2}\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\mu}^T(\mathbf{D}_w-\rho\mathbf{W})\boldsymbol{\mu}\right) \\ &\quad \times (\sigma^2+\tau^2)^{-2}\pi_\rho(\rho)\alpha p^{J-\sum_j\gamma_j+\alpha-1}(1-p)^{\sum_j\gamma_j} \\ &\equiv f(\mathbf{y},\boldsymbol{\Theta}\mid\boldsymbol{\gamma})\pi_{(\boldsymbol{\gamma},p)}(\boldsymbol{\gamma},p), \end{aligned}$$

where $\pi_{(\boldsymbol{\gamma},p)}(\boldsymbol{\gamma},p) = \alpha p^{J-\sum_j\gamma_j+\alpha-1}(1-p)^{\sum_j\gamma_j}$. Hence, the integral inside the summation in (3.2) is

$$\begin{aligned} \int_p\int_{\boldsymbol{\Theta}} f(\mathbf{y},\boldsymbol{\Theta},\boldsymbol{\gamma},p)d\boldsymbol{\Theta}dp &= \int_p\int_{\boldsymbol{\Theta}} f(\mathbf{y},\boldsymbol{\Theta}\mid\boldsymbol{\gamma})\pi_{(\boldsymbol{\gamma},p)}(\boldsymbol{\gamma},p)d\boldsymbol{\Theta}dp \\ &\propto \left(\int_0^1 p^{J-\sum_j\gamma_j+\alpha-1}(1-p)^{\sum_j\gamma_j}dp\right)\int_{\boldsymbol{\Theta}} f(\mathbf{y},\boldsymbol{\Theta}\mid\boldsymbol{\gamma})d\boldsymbol{\Theta}. \end{aligned}$$

But $\int_0^1 p^{J-\sum_j\gamma_j+\alpha-1}(1-p)^{\sum_j\gamma_j}dp < \infty$ since it is the integral of the kernel of a Beta density over its sample space. Also, the summation over $\boldsymbol{\gamma}$ has $2^J < \infty$ terms, so it suffices to establish that

$$\int_\rho\int_{\tau^2}\int_{\sigma^2}\int_{\boldsymbol{\mu}} f(\mathbf{y},\boldsymbol{\mu},\sigma^2,\tau^2,\rho\mid\boldsymbol{\gamma})d\boldsymbol{\mu}d\sigma^2d\tau^2d\rho < \infty, \quad \forall\boldsymbol{\gamma}\in\{0,1\}^J.$$

We begin with the case when $\gamma_j = 1$, for all $j \in \{1,\ldots,J\}$ and define $f_{\mathbf{1}}(\mathbf{y}\mid\boldsymbol{\mu}) := f(\mathbf{y}\mid\boldsymbol{\mu},\boldsymbol{\gamma}=\mathbf{1})$ to simplify notation. We have that

$$
\begin{aligned}
f_{\mathbf{1}}(\mathbf{y} \mid \boldsymbol{\mu}) &= \prod_{j=1}^{J} f(y_j \mid \mu_j, \gamma_j = 1) \\
&= (2\pi)^{-J/2}(\sigma^2)^{-J/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{J}(y_j - \mu_j)^2\right).
\end{aligned}
$$

The joint prior density for $\boldsymbol{\mu}$ is

$$
\pi_{\boldsymbol{\mu}}(\boldsymbol{\mu} \mid \tau^2, \rho) = (2\pi)^{-J/2}|\tau^2(\mathbf{D}_w - \rho\mathbf{W})^{-1}|^{-1/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\mu}^T(\mathbf{D}_w - \rho\mathbf{W})\boldsymbol{\mu}\right).
$$

However, $\rho \in (\lambda_1^{-1}, \lambda_J^{-1})$ implies that $(\mathbf{D}_w - \rho\mathbf{W}) > 0$ by Lemma 2, so the prior on $\boldsymbol{\mu}$ has the density of a $N_J(\mathbf{0}, \tau^2(\mathbf{D}_w - \rho\mathbf{W})^{-1})$ distribution. We can thus integrate $f_{\mathbf{1}}(\mathbf{y} \mid \boldsymbol{\mu})\pi_{\boldsymbol{\mu}}(\boldsymbol{\mu} \mid \tau^2, \rho)$ with respect to $\boldsymbol{\mu}$ as the convolution of two normal densities. That is, if $\mathbf{y} \mid \boldsymbol{\mu}$ has a $N_J(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ density and $\boldsymbol{\mu}$ has a $N_J(\mathbf{0}, \tau^2(\mathbf{D}_w - \rho\mathbf{W})^{-1})$ density, the marginal density of $\mathbf{y}$ is that of a $N_J(\mathbf{0}, \sigma^2\mathbf{I} + \tau^2(\mathbf{D}_w - \rho\mathbf{W})^{-1})$ distribution (Lindley and Smith, 1972). We thus know that the integration yields

$$
\begin{aligned}
f_{\mathbf{1}}(\mathbf{y}, \sigma^2, \tau^2, \rho) &= \pi_{(\tau^2,\sigma^2)}(\tau^2, \sigma^2)\pi_\rho(\rho) \int_{\mathbb{R}^J} f_{\mathbf{1}}(\mathbf{y} \mid \boldsymbol{\mu}, \sigma^2)\pi_{\boldsymbol{\mu}}(\boldsymbol{\mu} \mid \tau^2, \rho) d\boldsymbol{\mu} \\
&\propto |\sigma^2\mathbf{I} + \tau^2(\mathbf{D}_w - \rho\mathbf{W})^{-1}|^{-1/2} \\
&\quad \times \exp\left(-\frac{1}{2}\mathbf{y}^T(\sigma^2\mathbf{I} + \tau^2(\mathbf{D}_w - \rho\mathbf{W})^{-1})^{-1}\mathbf{y}\right) \\
&\quad \times (\sigma^2 + \tau^2)^{-2}\pi_\rho(\rho).
\end{aligned}
$$

Now, reparameterize the variance components by defining $\eta := \tau^2/\sigma^2$ so that

$$
\begin{aligned}
f_{\mathbf{1}}(\mathbf{y}, \sigma^2, \eta, \rho) &\propto (\sigma^2)^{-J/2}|\mathbf{I} + \eta(\mathbf{D}_w - \rho\mathbf{W})^{-1}|^{-1/2} \\
&\quad \times \exp\left(-\frac{1}{2\sigma^2}\mathbf{y}^T(\mathbf{I} + \eta(\mathbf{D}_w - \rho\mathbf{W})^{-1})^{-1}\mathbf{y}\right) \\
&\quad \times (\sigma^2)^{-1}(1 + \eta)^{-2}\pi_\rho(\rho).
\end{aligned}
$$

We then integrate over σ^2 to obtain

$$\begin{aligned} f_{\mathbf{1}}(\mathbf{y},\eta,\rho) &\propto \pi_\rho(\rho)(1+\eta)^{-2}|\mathbf{I}+\eta(\mathbf{D}_w-\rho\mathbf{W})^{-1}|^{-1/2} \\ &\quad \times \int_0^\infty (\sigma^2)^{-(J/2)-1}\exp\left(-\frac{1}{2\sigma^2}\mathbf{y}^T(\mathbf{I}+\eta(\mathbf{D}_w-\rho\mathbf{W})^{-1})^{-1}\mathbf{y}\right)d\sigma^2 \\ &= \pi_\rho(\rho)(1+\eta)^{-2}|\mathbf{I}+\eta(\mathbf{D}_w-\rho\mathbf{W})^{-1}|^{-1/2}\Gamma(J/2) \\ &\quad \times \left(\frac{\mathbf{y}^T(\mathbf{I}+\eta(\mathbf{D}_w-\rho\mathbf{W})^{-1})^{-1}\mathbf{y}}{2}\right)^{-J/2} \\ &\quad \times \int_0^\infty \left(\frac{\mathbf{y}^T(\mathbf{I}+\eta(\mathbf{D}_w-\rho\mathbf{W})^{-1})^{-1}\mathbf{y}}{2}\right)^{J/2}(\Gamma(J/2))^{-1}(\sigma^2)^{-\frac{J}{2}-1} \\ &\quad \times \exp\left(-\frac{1}{2\sigma^2}\mathbf{y}^T(\mathbf{I}+\eta(\mathbf{D}_w-\rho\mathbf{W})^{-1})^{-1}\mathbf{y}\right)d\sigma^2 \\ &\propto \pi_\rho(\rho)(1+\eta)^{-2}\frac{|\mathbf{I}+\eta(\mathbf{D}_w-\rho\mathbf{W})^{-1}|^{-1/2}}{(\mathbf{y}^T(\mathbf{I}+\eta(\mathbf{D}_w-\rho\mathbf{W})^{-1})^{-1}\mathbf{y})^{J/2}}, \end{aligned}$$

since the integral is that of an inverse gamma density over its sample space. We can rewrite the quadratic form in the denominator of the last expression as

$$\begin{aligned} \mathbf{y}^T(\mathbf{I}+\eta(\mathbf{D}_w-\rho\mathbf{W})^{-1})^{-1}\mathbf{y} &= \mathbf{y}^T\mathbf{D}_w^{1/2}(\mathbf{D}_w+\eta(\mathbf{I}-\rho\mathbf{W}^*)^{-1})^{-1}\mathbf{D}_w^{1/2}\mathbf{y} \\ &= \mathbf{x}^T(\mathbf{D}_w+\eta(\mathbf{I}-\rho\mathbf{W}^*)^{-1})^{-1}\mathbf{x}, \end{aligned}$$

where $\mathbf{W}^*=\mathbf{D}_w^{-1/2}\mathbf{W}\mathbf{D}_w^{-1/2}$ and $\mathbf{x}=\mathbf{D}_w^{1/2}\mathbf{y}$. If we let $w_{(J)}=\max_{1\le j\le J} w_{j\cdot}$, then the matrix $w_{(J)}\mathbf{I}-\mathbf{D}_w$ is diagonal with nonnegative entries and thus positive semidefinite, denoted $w_{(J)}\mathbf{I}-\mathbf{D}_w\ge 0$. It is established in the proof of Lemma 2 that $\rho\in(\lambda_1^{-1},\lambda_J^{-1}) \Rightarrow (\mathbf{I}-\rho\mathbf{W}^*)>0 \Rightarrow \eta(\mathbf{I}-\rho\mathbf{W}^*)^{-1}>0$. By adding and subtracting $\eta(\mathbf{I}-\rho\mathbf{W}^*)^{-1}$, we obtain

$$w_{(J)}\mathbf{I}-\mathbf{D}_w = w_{(J)}\mathbf{I}+\eta(\mathbf{I}-\rho\mathbf{W}^*)^{-1}-(\mathbf{D}_w+\eta(\mathbf{I}-\rho\mathbf{W}^*)^{-1})\ge 0.$$

Making use of the fact that $\mathbf{B} > 0, \mathbf{A} - \mathbf{B} \geq 0 \Rightarrow \mathbf{B}^{-1} - \mathbf{A}^{-1} \geq 0$ (Rao, 1973), it follows that

$$(\mathbf{D}_w + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1})^{-1} - (w_{(J)}\mathbf{I} + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1})^{-1} \geq 0$$

$$\Rightarrow \mathbf{x}^T(\mathbf{D}_w + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1})^{-1}\mathbf{x} \geq \mathbf{x}^T(w_{(J)}\mathbf{I} + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1})^{-1}\mathbf{x}$$

$$\Rightarrow (\mathbf{x}^T(\mathbf{D}_w + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1})^{-1}\mathbf{x})^{-J/2} \leq (\mathbf{x}^T(w_{(J)}\mathbf{I} + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1})^{-1}\mathbf{x})^{-J/2}.$$

Now, $\mathbf{W}^*$ is symmetric, so it has a spectral decomposition of the form $\mathbf{W}^* = \mathbf{PMP}^T$, where $\mathbf{P}$ is the orthogonal matrix of eigenvectors of $\mathbf{W}^*$ and $\mathbf{M}$ is the diagonal matrix of eigenvalues (see, e.g., Strang, 1988). Let $\mathbf{u} = \mathbf{P}^T\mathbf{x} \Rightarrow \mathbf{x} = \mathbf{Pu}$ so that

$$\begin{aligned}
\mathbf{x}^T(w_{(J)}\mathbf{I} + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1})^{-1}\mathbf{x} &= \mathbf{u}^T\mathbf{P}^T(w_{(J)}\mathbf{I} + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1})^{-1}\mathbf{Pu} \\
&= \mathbf{u}^T(w_{(J)}\mathbf{P}^T\mathbf{P} + \eta\mathbf{P}^T(\mathbf{I} - \rho\mathbf{W}^*)^{-1}\mathbf{P})^{-1}\mathbf{u} \\
&= \mathbf{u}^T(w_{(J)}\mathbf{I} + \eta(\mathbf{I} - \rho\mathbf{M})^{-1})^{-1}\mathbf{u} \\
&= \sum_{j=1}^{J} \frac{(1 - \rho\lambda_j)u_j^2}{w_{(J)}(1 - \rho\lambda_j) + \eta}.
\end{aligned}$$

Again referring to the proof of Lemma 2, we see that $\mathbf{W}^*$ has r_1 negative eigenvalues, r_2 positive eigenvalues, and $J - r_1 - r_2$ zero eigenvalues, where $r_1 + r_2 = r = \text{rank}(\mathbf{W}^*)$. The summation in the last line can then be separated according to the sign of the eigenvalue in each term as

$$\underbrace{\sum_{j=1}^{r_1} \frac{(1 - \rho\lambda_j)u_j^2}{w_{(J)}(1 - \rho\lambda_j) + \eta}}_{\lambda_j < 0} + \underbrace{\sum_{j=J-r_2+1}^{J} \frac{(1 - \rho\lambda_j)u_j^2}{w_{(J)}(1 - \rho\lambda_j) + \eta}}_{\lambda_j > 0} + \underbrace{\sum_{j=r_1+1}^{J-r_2} \frac{u_j^2}{w_{(J)} + \eta}}_{\lambda_j = 0}. \tag{3.3}$$

If $\lambda_1^{-1} < \rho < 0$, then $0 < 1 - \rho\lambda_j < 1$ for $j = 1, \ldots, r_1$ and $1 - \rho\lambda_j > 1$ for $j = J - r_2 + 1, \ldots, J$, so

$$
\begin{aligned}
(3.3) \quad &\geq \sum_{j=J-r_2+1}^{J} \frac{(1-\rho\lambda_j)u_j^2}{w_{(J)}(1-\rho\lambda_j)+\eta} + \sum_{j=r_1+1}^{J-r_2} \frac{u_j^2}{w_{(J)}+\eta} \\
&\geq \sum_{j=J-r_1+1}^{J} \frac{u_j^2}{w_{(J)}+\eta} + \sum_{j=r_1+1}^{J-r_2} \frac{u_j^2}{w_{(J)}+\eta} \\
&= (w_{(J)}+\eta)^{-1} \sum_{j=r_1+1}^{J} u_j^2,
\end{aligned}
$$

where the second line follows from noticing that

$$
\begin{aligned}
\frac{1}{w_{(J)}+\eta} - \frac{1-\rho\lambda_j}{w_{(J)}(1-\rho\lambda_j)+\eta} &= \frac{w_{(J)}(1-\rho\lambda_j)+\eta-(w_{(J)}+\eta)(1-\rho\lambda_j)}{(w_{(J)}+\eta)(w_{(J)}(1-\rho\lambda_j)+\eta)} \\
&= \frac{\rho\lambda_j w_{(J)}}{(w_{(J)}+\eta)(w_{(J)}(1-\rho\lambda_j)+\eta)} \\
&< 0,
\end{aligned}
$$

for $\lambda_j > 0$. Similarly, if $0 < \rho < \lambda_J^{-1}$, then

$$
\begin{aligned}
(3.3) \quad &\geq \sum_{j=1}^{r_1} \frac{(1-\rho\lambda_j)u_j^2}{w_{(J)}(1-\rho\lambda_j)+\eta} + \sum_{j=r_1+1}^{J-r_2} \frac{u_j^2}{w_{(J)}+\eta} \\
&\geq \sum_{j=1}^{r_1} \frac{u_j^2}{w_{(J)}+\eta} + \sum_{j=r_1+1}^{J-r_2} \frac{u_j^2}{w_{(J)}+\eta} \\
&= (w_{(J)}+\eta)^{-1} \sum_{j=1}^{J-r_2} u_j^2.
\end{aligned}
$$

Thus, for all $\rho \in (\lambda_1^{-1}, \lambda_J^{-1})$,

$$
\mathbf{x}^T(w_{(J)}\mathbf{I} + \eta(\mathbf{I}-\rho\mathbf{W}^*)^{-1})^{-1}\mathbf{x} \geq k(w_{(J)}+\eta)^{-1}
$$

$$
\Rightarrow (\mathbf{x}^T(w_{(J)}\mathbf{I} + \eta(\mathbf{I}-\rho\mathbf{W}^*)^{-1})^{-1}\mathbf{x})^{-J/2} \leq k'(w_{(J)}+\eta)^{J/2}
$$

where $0 < k' < \infty$ is constant.

Next, we see that

$$\begin{aligned} |\mathbf{I} + \eta(\mathbf{D}_w - \rho\mathbf{W})^{-1}| &= |\mathbf{D}_w^{-1/2}||\mathbf{D}_w + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1}||\mathbf{D}_w^{-1/2}| \\ &= |\mathbf{D}_w|^{-1}|\mathbf{D}_w + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1}| \\ &\equiv K|\mathbf{D}_w + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1}|. \end{aligned}$$

Letting $w_{(1)} = \min_{1 \le j \le J} w_j.$, and again adding and subtracting $\eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1}$, we get

$$\mathbf{D}_w - w_{(1)}\mathbf{I} \ge 0 \Rightarrow \mathbf{D}_w + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1} - (w_{(1)}\mathbf{I} + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1}) \ge 0.$$

But $\mathbf{B} > 0, \mathbf{A} - \mathbf{B} \ge 0$ implies $|\mathbf{A}| \ge |\mathbf{B}|$ (Rao, 1973), so we find that

$$|\mathbf{D}_w + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1}| \ge |w_{(1)}\mathbf{I} + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1}|$$

$$\Rightarrow K|\mathbf{D}_w + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1}| \ge K|w_{(1)}\mathbf{I} + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1}|.$$

The proof of Lemma 2 implies that $(1 - \rho\lambda_j)^{-1}$, $j = 1, \ldots, J$, are the eigenvalues of $(\mathbf{I} - \rho\mathbf{W}^*)^{-1}$, so it follows from Lemma 1 that the eigenvalues of $w_{(1)}\mathbf{I} + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1}$ are $w_{(1)} + \eta(1 - \rho\lambda_j)^{-1}$, $j = 1, \ldots, J$. Therefore,

$$\begin{aligned} K|w_{(1)}\mathbf{I} + \eta(\mathbf{I} - \rho\mathbf{W}^*)^{-1}| &= K\prod_{j=1}^{J}(w_{(1)} + \eta(1 - \rho\lambda_j)^{-1}) \\ &= \frac{K\prod_{j=1}^{J}((1 - \rho\lambda_j)w_{(1)} + \eta)}{\prod_{j=1}^{J}(1 - \rho\lambda_j)}, \end{aligned}$$

and hence

$$|\mathbf{I} + \eta(\mathbf{D}_w - \rho\mathbf{W})^{-1}|^{-1/2} \le K'\left(\frac{\prod_{j=1}^{J}((1 - \rho\lambda_j)w_{(1)} + \eta)}{\prod_{j=1}^{J}(1 - \rho\lambda_j)}\right)^{-1/2},$$

where $0 < K' < \infty$ is constant. But $1 - \rho\lambda_j > 0,$ for all j, so by Lemma 3

$$\frac{1-\rho\lambda_j}{(1-\rho\lambda_j)w_{(1)}+\eta} \leq \frac{\max\{1-\rho\lambda_j, 1\}}{w_{(1)}+\eta}, \quad \forall j \in \{1,\ldots,J\}$$

$$\Rightarrow \frac{\prod_{j=1}^{J}(1-\rho\lambda_j)}{\prod_{j=1}^{J}((1-\rho\lambda_j)w_{(1)}+\eta)} \leq \frac{\prod_{j=1}^{J}\max\{1-\rho\lambda_j,1\}}{(w_{(1)}+\eta)^J}$$

$$\Rightarrow |\mathbf{I}+\eta(\mathbf{D}_w-\rho\mathbf{W}^*)^{-1}|^{-1/2} \leq \frac{K'\prod_{j=1}^{J}\max\{(1-\rho\lambda_j)^{1/2},1\}}{(w_{(1)}+\eta)^{J/2}}.$$

We have now established that the density function satisfies

$$\begin{aligned} f(\mathbf{y},\eta,\rho) &\leq C(w_{(J)}+\eta)^{J/2}\left(\frac{\prod_{j=1}^{J}\max\{(1-\rho\lambda_j)^{1/2},1\}}{(w_{(1)}+\eta)^{J/2}}\right)(1+\eta)^{-2}\pi_\rho(\rho) \\ &= \begin{cases} C\left(\frac{w_{(J)}+\eta}{w_{(1)}+\eta}\right)^{J/2}\left(\frac{\prod_{j=r_1+1}^{J}(1-\rho\lambda_j)^{1/2}}{(1+\eta)^2}\right)\pi_\rho(\rho), & \rho<0 \\ C\left(\frac{w_{(J)}+\eta}{w_{(1)}+\eta}\right)^{J/2}\left(\frac{\prod_{j=1}^{J-r_2}(1-\rho\lambda_j)^{1/2}}{(1+\eta)^2}\right)\pi_\rho(\rho), & \rho>0 \end{cases} \end{aligned}$$

where $C < \infty$ is constant.

Now, $\pi_\rho(\rho) \propto I(\lambda_1^{-1} < \rho < \lambda_J^{-1})$,

$$\int_{\lambda_1^{-1}}^{0}\prod_{j=r_1+1}^{J}(1-\rho\lambda_j)^{1/2}d\rho < \int_{\lambda_1^{-1}}^{0}(1-\rho\lambda_J)^{\frac{J-r_1}{2}}d\rho < \infty,$$

and

$$\int_{0}^{\lambda_J^{-1}}\prod_{j=1}^{J-r_2}(1-\rho\lambda_j)^{1/2}d\rho < \int_{0}^{\lambda_J^{-1}}(1-\rho\lambda_1)^{\frac{J-r_2}{2}}d\rho < \infty.$$

Also, we see that

$$\left(\frac{w_{(J)}+\eta}{w_{(1)}+\eta}\right)^{J/2}(1+\eta)^{-2} \sim (1+\eta)^{-2}$$

as $\eta \to \infty$ and $\int_0^\infty (1+\eta)^{-2} d\eta = 1 < \infty$, implying that (e.g. Bartle, 1976)

$$\int_0^\infty \left(\frac{w_{(J)} + \eta}{w_{(1)} + \eta} \right)^{J/2} (1+\eta)^{-2} d\eta < \infty.$$

From this it follows that

$$\begin{aligned} \int_{\lambda_1^{-1}}^{\lambda_J^{-1}} \int_0^\infty f_{\mathbf{1}}(\mathbf{y}, \eta, \rho) d\eta d\rho &\leq C \left(\int_{\lambda_1^{-1}}^{\lambda_J^{-1}} \pi_\rho(\rho) \prod_{j=1}^{J} \max\{(1-\rho\lambda_j)^{1/2}, 1\} d\rho \right) \\ &\quad \times \left(\int_0^\infty \left(\frac{w_{(J)} + \eta}{w_{(1)} + \eta} \right)^{J/2} (1+\eta)^{-2} d\eta \right) \\ &< \infty, \end{aligned}$$

showing that $\int_{\boldsymbol{\Theta}} f(\mathbf{y}, \boldsymbol{\Theta} \mid \boldsymbol{\gamma}) d\boldsymbol{\Theta} < \infty$ when $\gamma_1 = \gamma_2 = \cdots = \gamma_J = 1$.

We now turn our attention to a non-degenerate case, $|\{\gamma_j : \gamma_j = 1\}| = J_1 < J$, where $|\cdot|$ denotes the cardinality of a set. Let $\boldsymbol{\gamma}^*$ denote such an arbitrary configuration of $\boldsymbol{\gamma}$. Further, let $S_1 = \{j : \gamma_j = 1\} \subsetneq \{1, \ldots, J\}$, $\mathbf{y}_1 = \{y_j : j \in S_1\}$, and $\mathbf{y}_0 = \{y_j : j \in S_1^c\}$ so that we may partition $\mathbf{y}$ as $\mathbf{y} = (\mathbf{y}_0^T, \mathbf{y}_1^T)^T$. Our strategy here is to show $h(\mathbf{y}_0, \mathbf{y}_1) \equiv h(\mathbf{y}) := \int_{\boldsymbol{\Theta}} f(\mathbf{y}, \boldsymbol{\Theta} \mid \boldsymbol{\gamma}^*) d\boldsymbol{\Theta} < \infty$ by showing that

$$\int_{\mathbf{y}_1} h(\mathbf{y}_0, \mathbf{y}_1) d\mathbf{y}_1 < \infty$$

for all S_1 and $\mathbf{y}_1$. The key elements in the argument are that $J_1 < J \Rightarrow J - J_1 > 0$ and that we can factor a double integral over, e.g., $\mathbf{y}$ and $\boldsymbol{\mu}$ as

$$\int_{\mathbb{R}^J} \int_{\mathbb{R}^J} f(\mathbf{y} \mid \boldsymbol{\mu}) \pi(\boldsymbol{\mu}) d\boldsymbol{\mu} d\mathbf{y} = \left(\int_{\mathbb{R}^J} f(\mathbf{y} \mid \boldsymbol{\mu}) d\mathbf{y} \right) \left(\int_{\mathbb{R}^J} \pi(\boldsymbol{\mu}) d\boldsymbol{\mu} \right),$$

since the integral of a (proper) probability density over its sample space is one, thus free from other parameters.

Now, we have

$$\begin{aligned}
\int_{\mathbf{y}_1} h(\mathbf{y}_0, \mathbf{y}_1) d\mathbf{y}_1 &= \int_{\sigma^2}\int_{\tau^2}\int_{\rho}\int_{\boldsymbol{\mu}}\int_{\mathbf{y}_1} f(\mathbf{y}_0, \mathbf{y}_1, \tau^2, \sigma^2, \rho, \boldsymbol{\mu} \mid \boldsymbol{\gamma}^*) d\mathbf{y}_1 d\boldsymbol{\mu} d\rho d\tau^2 d\sigma^2 \\
&\propto \int_0^\infty (\sigma^2)^{-\frac{J-J_1}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{j\in S_1^c} y_j^2\right) \\
&\quad \times \underbrace{\int_{\mathbb{R}^{J_1}} (\sigma^2)^{-\frac{J_1}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{j\in S_1}(y_j-\mu_j)^2\right) d\mathbf{y}_1}_{=(2\pi)^{J_1/2}} \\
&\quad \times \underbrace{\int_{\mathbb{R}^J} |\tau^2(\mathbf{D}_w - \rho\mathbf{W})^{-1}|^{-1/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\mu}^T(\mathbf{D}_w-\rho\mathbf{W})\boldsymbol{\mu}\right) d\boldsymbol{\mu}}_{=(2\pi)^{J/2}} \\
&\quad \times \underbrace{\int_{\lambda_1^{-1}}^{\lambda_J^{-1}} (\lambda_J^{-1}-\lambda_1^{-1})^{-1} d\rho}_{=1} \int_0^\infty (\sigma^2+\tau^2)^{-2} d\tau^2 d\sigma^2 \\
&\propto \int_0^\infty (\sigma^2)^{-\frac{J-J_1}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{j\in S_1^c} y_j^2\right) \int_0^\infty (\sigma^2+\tau^2)^{-2} d\tau^2 d\sigma^2 \\
&= \int_0^\infty (\sigma^2)^{-\frac{J-J_1}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{j\in S_1^c} y_j^2\right) \left[-\frac{1}{\sigma^2+\tau^2}\right]_{\tau^2=0}^\infty d\sigma^2 \\
&= \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{J-J_1}{2}+1} \exp\left[-\left(\frac{1}{\sigma^2}\right)\left(\frac{\sum_{j\in S_1^c} y_j^2}{2}\right)\right] d\sigma^2 < \infty,
\end{aligned}$$

since the integral is that of the kernel of an inverse gamma density over its sample space.

The proof is completed by considering the other degenerate case, $\gamma_j = 0,$ for all j. The preceding argument still applies to this case, though, with $S_1 = \emptyset$ and $J_1 = 0$. The result is therefore established. □

Chapter 4

Simulation Study

This chapter investigates and compares the performance of our proposed Model 1 and the Scott-Berger (SB) model. The intrinsic multiplicity adjustment is demonstrated for both models with simulated signals among an increasing number of noise observations. Since our model relies on spatial structure in the data, we simulate spatial arrays with non-null observations within them. We explore selected error rates of both models, measured as the proportion of missed signals, proportion of false positives, and misclassification proportion. We do this while varying both the non-null signal strength relative to the noise and the strength of spatial association among the non-null locations.

4.1 The Multiplicity Adjustment

To demonstrate the multiplicity adjustment under the SB model, we simulate data $y_i \sim N(\theta_i, 1)$, $i = 1, \ldots, J$, for different values of J. For each J, five points are generated with $\theta_1 = 10, \theta_2 = 4, \theta_3 = 1, \theta_4 = -.2$, and $\theta_5 = -15$. These represent signals hidden among uninteresting noise. The remaining noise observations are simulated by drawing from a $N(0, 1)$ distribution. The posterior distribution is simulated via MCMC and $1 - p_i = P(\gamma_i = 1 \mid \mathbf{y})$ is estimated with $(1/N) \sum_{k=1}^{N} I(\gamma_i^{(k)} = 1)$, where $\gamma_i^{(k)}$ is the k^{th} draw from the marginal posterior of γ_i and $N = 5,000$ is the number of samples drawn from the posterior distribution. Table 4.1 shows the results when the number of tests being performed is 50, 500, and 5,000. While the most extreme values generated from $\theta_1 = 10$ and $\theta_5 = -15$ persist as extremely strong evidence of interesting signals, the estimated probabilities that the other values come from

non-zero means decrease as J increases. This is the multiplicity adjustment. Points with an appreciable probability of having a signal in a small number of simultaneous tests become more likely to be just noise as the number of tests increases.

Table 4.1: Estimates of non-null probabilities under the Scott-Berger model for simulated signals among increasing noise. For each θ and for each J, a single value is drawn from a $N(\theta, 1)$ distribution, with the remaining $J-5$ observations drawn from $N(0, 1)$.

	θ				
	10	4	1	-.2	-15
$J = 50$	1.000	.958	.058	.019	1.000
$J = 500$	1.000	.933	.004	.001	1.000
$J = 5,000$	1.000	.797	0.000	0.000	1.000

We also illustrate the posterior calibration by simulating a 20×20 spatial array of data and fitting both the SB model and our CAR testing model. The binary non-null patterns are simulated by drawing from an Ising model, $p(\mathbf{x}) \propto \exp\{\beta \sum_{i \sim j} I(x_i = x_j)\}, \mathbf{x} \in \{0, 1\}^{400}$. We refer interested readers to Higdon (1994) for discussion of this model. Here it is only important to note that β is an interaction parameter reflecting the strength of association between neighbors in the spatial array. We use $\beta = .45$, an arbitrary value chosen to induce strong clustering of "activated" regions. Using the generated activation pattern, the observed data are then simulated as $y_i \sim N(\theta_i, 1)$, where $\theta_i = 0$ at the null cases. For the non-null cases, we use the universal threshold of Donoho and Johnstone (1994), $\sqrt{2 \log(20 \times 20)} \approx 3.46$. This threshold may be regarded as a minimum detectable distance when the number of true signals is small compared to the noise observations (Bogdan et al., 2008, Remark 1). The binary activation map and the resulting simulated data are shown in Figure 4.1.

The posterior probabilities of activation for these data are estimated using both the SB independence model and our CAR testing model. First, we fit the models using only the 20×20 grid so that there are $J = 400$ tests being performed. We then take the same data array and put it in the middle of a 20×100 array, with the rest

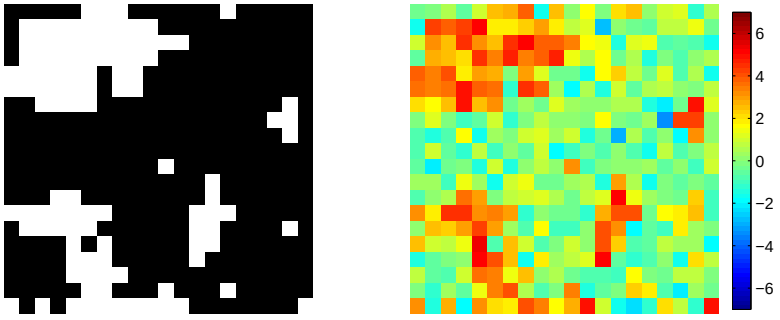


Figure 4.1: Simulated binary map of the non-null pattern (left) and corresponding simulated observations (right). Data are drawn from $N(3.46, 1)$ in the non-null cases, and from $N(0, 1)$ in the null cases.

of the observations being random noise generated from a $N(0, 1)$ distribution. In this case, there are $J = 2,000$ locations being tested, the original 20×20 data array and an additional 1,600. Figures 4.2 and 4.3 depict the log-scale posterior probabilities for $J = 400$ and $J = 2,000$, as well as the thresholded activation maps, using .95 as the threshold. In both the SB and CAR models, we take p to have a uniform prior, $\alpha = 1$. For the maps generated with $J = 2,000$ observations, only the subset corresponding to the original 20×20 grid is displayed for easier comparison.

We can see that both models correct for multiplicity. All of the posterior probabilities are being penalized to reflect the greater number of tests performed. The SB model makes a relatively strong adjustment. The estimated posterior probabilities are also lower in the CAR model. The CAR model imposes a less conservative penalty on the estimated probabilities, preserving power in the presence of more tests. The adjustment can also be seen in the threshold maps for both models, where fewer points exceed the .95 threshold when $J = 2,000$ tests are considered versus $J = 400$.

4.2 Error Rates

We compare the performance of the two models by examining three error measures. Let $\delta_j = 1$ if the testing procedure selects location j as a non-null case, 0 otherwise,

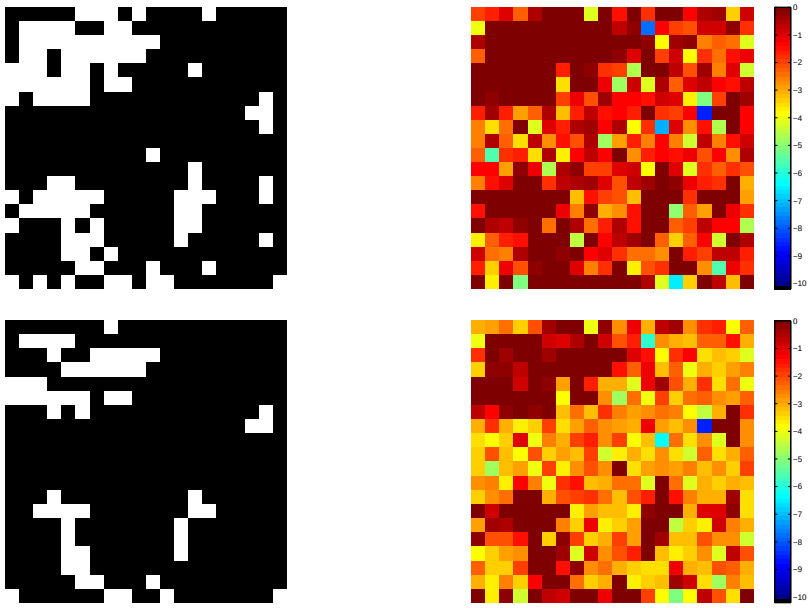


Figure 4.2: Thresholded activation maps and log-scale posterior probabilities from the Scott-Berger model with $J = 400$ (top row) and $J = 2,000$ tests (bottom row). In the thresholded maps, points with estimated non-null probability of at least .95 are selected.

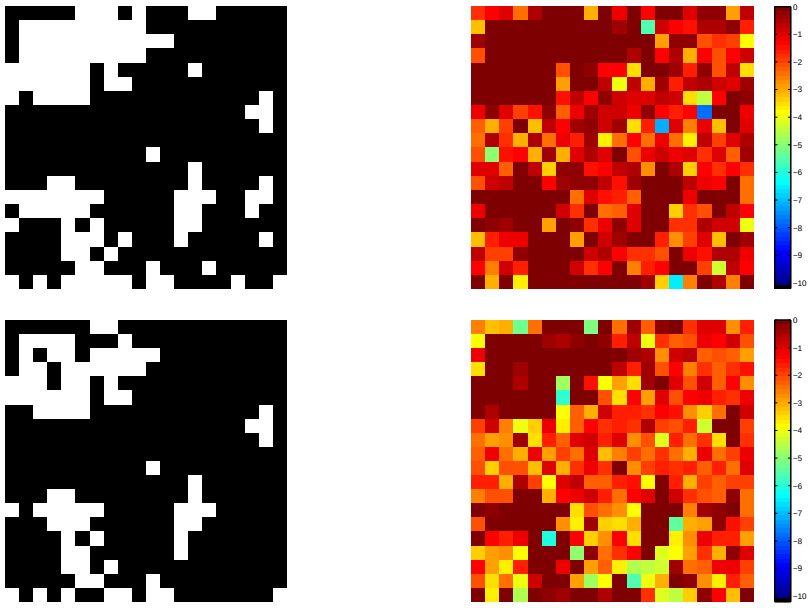


Figure 4.3: Thresholded activation maps and log-scale posterior probabilities from the CAR testing model with $J = 400$ (top row) and $J = 2,000$ tests (bottom row). In the thresholded maps, points with estimated non-null probability of at least .95 are selected.

and let γ_j indicate whether or not location j is truly "active". For both $J = 400$ and $J = 2,000$, we calculate the false discovery proportion, FDP $= I(\sum \delta_j > 0) \sum \delta_j(1 - \gamma_j)/\sum \delta_j$, the false non-discovery proportion, FNP $= I(\sum(1 - \delta_j) > 0) \sum \gamma_j(1 - \delta_j)/\sum(1 - \delta_j)$, and the overall misclassification proportion, MP $= (\sum \delta_j(1 - \gamma_j) + \sum \gamma_j(1 - \delta_j))/J$. Table 4.2 summarizes these error measures for both the SB and CAR models. The CAR model has a lower false non-discovery proportion in both cases, and an overall lower misclassification proportion.

Table 4.2: False discovery proportion (FDP), false non-discovery proportion (FNP), and overall misclassification proportion (MP) for both models with $J = 400$ and $J = 2,000$ tests. These are calculated using the results displayed in Figures 4.2 and 4.3.

	SB			CAR		
	FDP	FNP	MP	FDP	FNP	MP
$J = 400$	.0693	.0435	.0500	.1043	.0140	.0400
$J = 2,000$	0.0000	.1433	.1225	0.0000	.0728	.0575

We further compare the performance of the SB and CAR models using different non-null signal strengths and different values of β in the Ising model. We take $\beta = .2$ and $\beta = .45$. The lower value of β induces weaker clustering of like values in the map so that there is a less pronounced spatial association between the null and non-null cases. For both weak ($\beta = .2$) and strong ($\beta = .45$) clustering, we simulate the non-null cases from two different alternative distributions. First, we generate the non-null cases from $N(4, 1)$, representing a strong signal. For the weaker signal, we use a $N(2.1, 1)$ distribution. The second non-null mean follows Lange et al. (1999) and Hartvig and Jensen (2000) by supposing that, in fMRI data, the distribution of signals is approximated as Gaussian with mean $\theta = m\sqrt{SSD_x}$, where SSD_x is the corrected sum of squares of the binary experimental stimulus pattern and m is the ratio of the activation magnitude to its standard deviation. Hartvig and Jensen estimate m to be $\hat{m} = .43$ and take $SSD_x = 24$ for a typical stimulus pattern. The activation patterns and subsequent data arrays are displayed in Figure 4.4.

Tables 4.3 and 4.4 report the FDP, FNP, and MP across non-null means for both values of β. With $\theta = 4$, it appears that the SB model performs slightly better than our model. This could be because the CAR model is lowering the estimated probability of activation at some locations by using the information in nearby weaker observations, leading to a higher FNP. With the weak signal ($\theta = 2.1$), there is a strong contrast between the SB and CAR models. The SB model fails to detect any non-null cases at all whereas the CAR model detects at least some of the true signals. Both procedures have an undesirable FNP and MP, but this is unavoidable with weak signal-to-noise ratios.

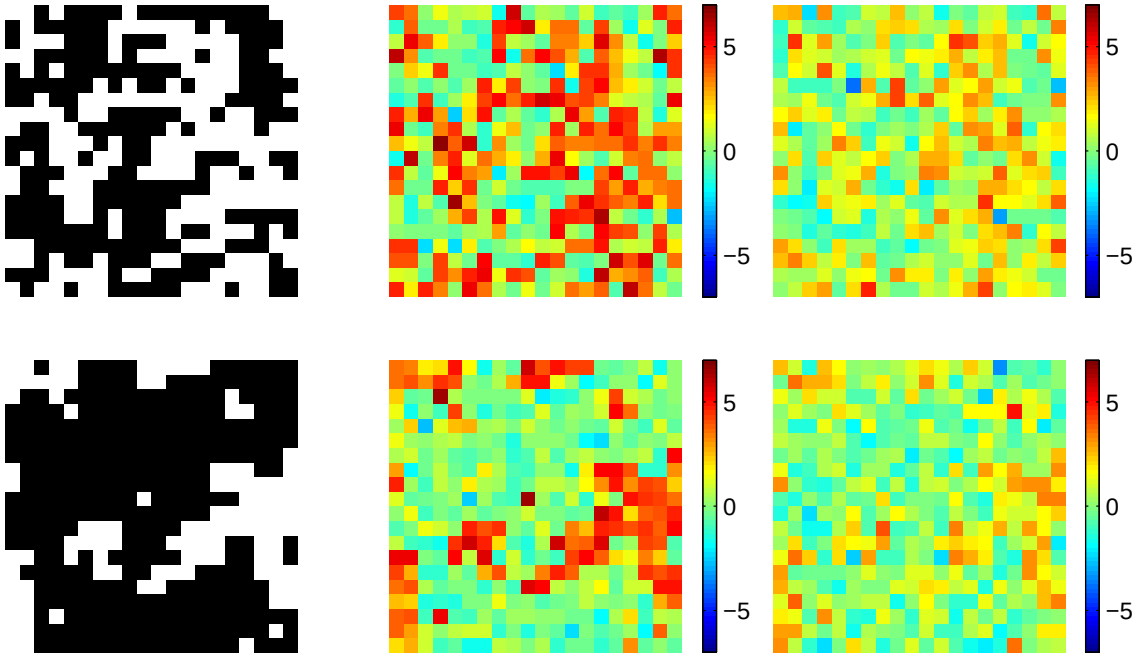


Figure 4.4: Binary non-null patterns and data arrays for $\beta = .2$ (top row) and $\beta = .45$ (bottom row) in the Ising model. The non-null distribution is $N(4, 1)$ in the middle panels and $N(2.1, 1)$ in the right panels.

Our simulation results demonstrate that both the SB model and our proposed model incorporate the number of tests in evaluating posterior non-null probabilities. The multiple testing penalty is stronger under the SB model, though, making it gen-

Table 4.3: False discovery proportion (FDP), false non-discovery proportion (FNP), and overall misclassification proportion (MP) for both models with $\beta = .2$ in the Ising model. The left column shows the mean of the non-null density.

	SB			CAR		
Mean	FDP	FNP	MP	FDP	FNP	MP
$\theta = 4$	.0412	.0049	.0225	0.0000	.1050	.0625
$\theta = 2.1$	0.0000	.4675	.4675	0.0143	.3576	.2975

Table 4.4: False discovery proportion (FDP), false non-discovery proportion (FNP), and overall misclassification proportion (MP) for both models with $\beta = .45$ in the Ising model. The left column shows the mean of the non-null density.

	SB			CAR		
Mean	FDP	FNP	MP	FDP	FNP	MP
$\theta = 4$	.0631	0.0000	.0175	.0727	.0069	.0250
$\theta = 2.1$	0.0000	.2600	.2600	.0196	.1547	.1375

erally less powerful with weak to moderate non-null signal strengths. The relative performance of the two models appears to depend on the amplitude of the activations. With extremely strong signals, the SB model may be superior to our own. The presence of such strong signals would make many multiple testing procedures acceptable, though, obviating the need for sensitivity considerations. The typically weaker signal-to-noise ratio in fMRI data make both sensitivity and specificity important in evaluating testing procedures. Truly activated voxels can exhibit weak activation compared to the noise. Indeed, our purpose for using the weaker non-null mean here is to simulate more realistic functional neuroimaging data (Hartvig and Jensen, 2000). The CAR model is clearly much more sensitive at detecting true activation with a weaker signal, where the more conservative SB model can fail to select any locations at all. The theoretical properties of our model merit deeper investigation in future work, but it is apparent that incorporating spatial dependence into a Bayesian testing procedure is appropriate for the analysis of fMRI data.

Chapter 5

Data Analysis

Within fMRI studies as they are conducted on human brain activation, both single-subject and groupwise analyses are important. Single-subject analyses are useful for identifying regions of a specific brain that may be associated with certain behaviors or conditions. This approach allows for measuring the consistency of physiological processes or statistical procedures across multiple individuals. It also allows researchers to identify individual differences, or variability across subjects. Groupwise analyses, on the other hand, are appropriate when researchers want to draw conclusions concerning an entire group of individuals or to generalize results to a broader population. Combining information across brains also tends to make statistical procedures more powerful. General research questions about individual variation versus groupwise analyses may be found in, e.g., Huettel et al. (2009). Statistical issues associated with combining information across brains are explored in Lazar et al. (2002) and Lazar (2008).

To illustrate our procedure, we use the data described in Chapter 3. We analyze them two different ways. In Section 5.1 below, we evaluate each participant's slice individually, using only that brain's observations to derive the posterior distribution. We present results from analyzing all participants' brains together in Section 5.2. In each analysis, we treat as interesting cases those voxels that show an increase in the BOLD signal associated with the antisaccade task. This means that we restrict our attention to only the positive alternative to $\theta_j = 0$. At each location, we estimate the probability of task-related activation, $P(\theta_j > 0 \mid \mathbf{y})$, with $\hat{p}_j = (1/N)\sum_{k=1}^{N} I(\theta_j^{(k)} > 0)$, where $\theta_j^{(k)}$ is the product of the k^{th} draws of γ_j and

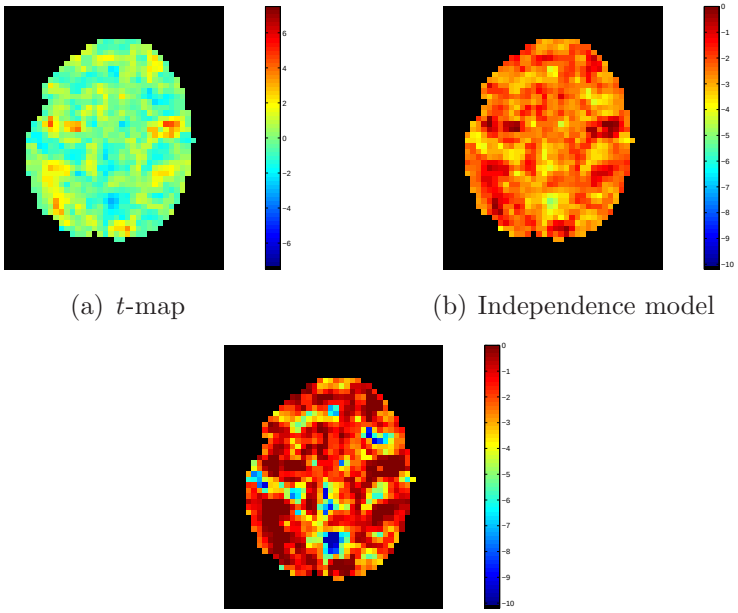
μ_j from the joint posterior distribution and $N = 10,000$ is the posterior sample size. When analyzing each brain separately, we use both the Scott-Berger (SB) model under the independence assumption and our model incorporating the CAR dependence structure. For the analysis using the information in all participants' brains together, we use only the CAR model. We do this to study the differences obtained by using one posterior distribution with much more information as opposed to evaluating fourteen separate posterior distributions, each with considerably less information. For both the individual and group analyses, we apply an FDR thresholding procedure and compare its performance to the Bayesian models. To accommodate dependence in the data, we use the modified FDR algorithm of Benjamini and Yekutieli (2001), which is appropriate regardless of the dependence structure in the p-values. In addition, we perform a sensitivity analysis to assess the impact of the prior on the model hypervariance and the shape parameter in the prior for p.

5.1 Individual Analysis

We implement the SB model using a Gibbs sampler entirely in `R`. See Appendix A for a derivation of the algorithm. After an initial 50,000 iterations to achieve approximate convergence, we perform another 100,000 iterations, retaining every tenth draw to reduce autocorrelation. The CAR model is implemented with `WinBUGS` and the included `GeoBUGS` package. `WinBUGS` uses Gibbs sampling with adaptive rejection sampling (Gilks, 1992) and slice sampling (Neal, 1997) steps for the non-conjugate distributions. We use the same number of iterations and samples as in the independence case.

In the prior for p, we take $\alpha = 1$ for the independence case, $p \sim \text{Uniform}(0, 1)$, and $\alpha = 108$ for the CAR model, $p \sim \text{Beta}(108, 1)$. The uniform prior for the independence case is suggested in Scott and Berger (2006). The second value follows from recognizing that $P(\gamma_j = 1 \mid \alpha) = \alpha \int_0^1 (1-p)p^{\alpha-1}dp = 1/(\alpha+1)$ and using

an empirical Bayes estimate of α. The method of moments yields $\tilde{\alpha} = (\hat{P}(\gamma_j = 1 \mid \alpha))^{-1} - 1$, where an *ad hoc* estimator of the marginal probability of activation is $\hat{P}(\gamma_j = 1 \mid \alpha) = \sum_{m=1}^{14} \sum_{j=1}^{J_m} I(y_{mj} > 4) / \sum_{m=1}^{14} J_m$. Here y_{mj} denotes the observed value at the j^{th} voxel of subject m and 4 is a conservative threshold that gives considerable evidence for activation in a single one-sided t-test with large degrees of freedom.



(a) t-map

(b) Independence model

(c) CAR model

Figure 5.1: Comparative results for the individual slice from Subject 6. The upper left panel is the t-map, the upper right panel shows estimated non-null probabilities from the Scott-Berger independence model, and the bottom panel shows estimated non-null probabilities from the CAR model. Probabilities are displayed on the log scale for improved resolution.

Figure 5.1 shows the t-map and posterior probability maps from the independence and CAR models for one subject. The probabilities are presented on the log scale for enhanced resolution between areas of differing probability. Warmer colors of red and orange indicate areas where the estimated probability of task-related activation is higher. Cooler shades of green and blue represent low estimated probabilities of activation. The maps are presented in the axial view. For reference, a structural slice from a single brain is presented in Figure 5.2. Superimposed on this map is the neural

circuitry expected to be associated with antisaccade tasks (Dyckman et al., 2007). The frontal eye fields (FEF), supplementary eye field (SEF), and posterior parietal cortex (PPC) are labeled in the Figure. The pattern serves as a benchmark against which the results from each testing model can be judged. The depicted regions are based on the average of activation patterns of participants in this study. We would not expect any single brain to be this well-defined.

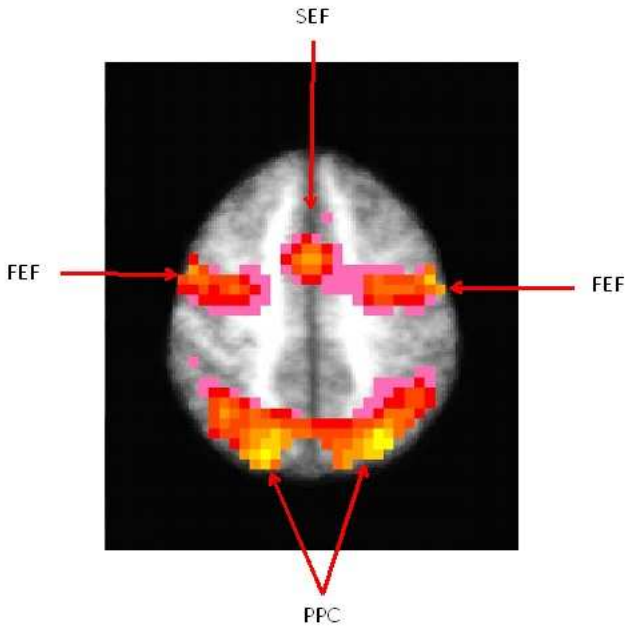
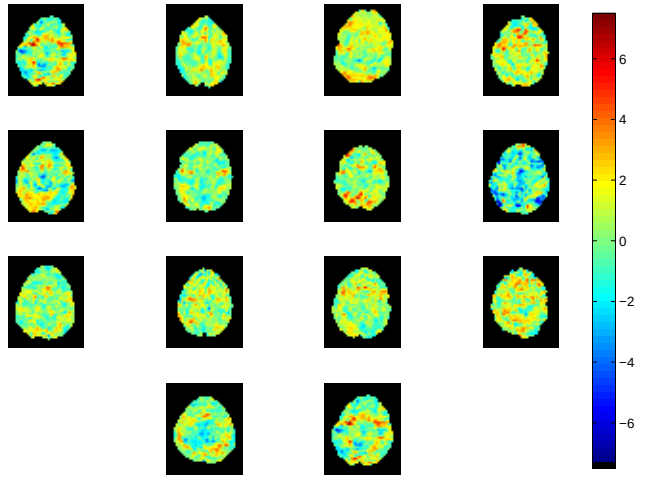


Figure 5.2: Activation patterns expected to be associated with the antisaccade task (Dyckman et al., 2007). This image is generated using the same participants from the study described in Section 3.1.

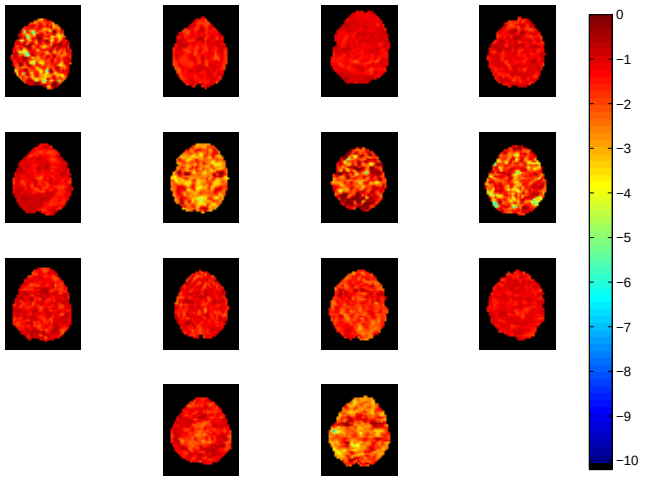
For the participant in Figure 5.1, assuming independence leads to a probability map that does not very clearly delineate regions of the brain. While there is some clear correspondence between the posterior probabilities and the t-map, it would be difficult for a researcher to look at this image and distinguish supportive neural circuitry from regions unrelated to the antisaccade task. By sharing information across voxels, our model incorporating the CAR spatial dependence structure allows clearer discernment of areas that are likely to be part of the associated neural circuitry. We can see that

the relatively shallow 'peaks' and 'valleys' of probability in the independence map are enhanced in the CAR map, leading to more pronounced shapes of the regions we would expect, based on the t-map. This difference is the result of the CAR structure allowing each t statistic to be analyzed in the context of its immediate neighbors. While the independence assumption only judges each voxel on its own merit, incorporating spatial dependence allows the voxels to borrow strength from one another through the sharing of information. In other words, a particular t statistic in and of itself may not give overwhelming evidence of activation, but the voxels around it may also provide a fair amount of evidence in favor of (or against) activation. Neighbors work together to boost the probability that the voxel in question is task-related (or not). Figure 5.3 displays the t-maps of all fourteen participants along with the corresponding probability maps using both the independence assumption and the CAR dependence structure. It can be seen that, in general, stronger contrasts between regions of interest occur as a result of the CAR assumption.

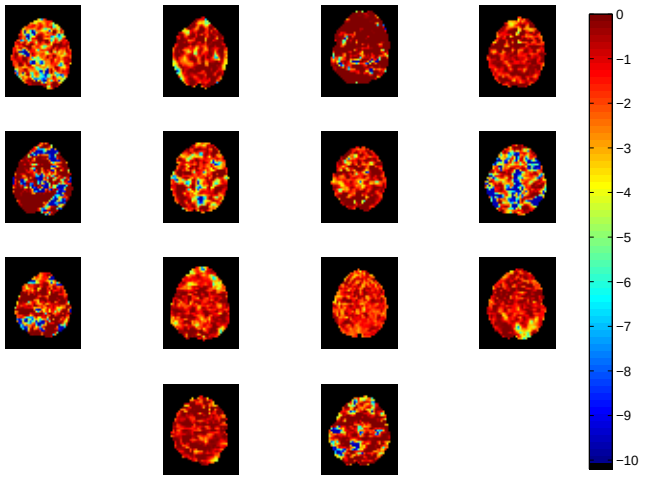
The difference between accounting for spatial dependence in the CAR model and assuming independence of the data is again elucidated by looking at the thresholded activation map for the previously discussed subject in Figure 5.4. We threshold at an estimated probability of task-related activation of at least .95, marking voxels that attain or exceed that level with the value one with the others set to zero. For this participant, we can see that very few voxels are selected as likely to be task-related under the independence model. By contrast, the activation map from the CAR model is much more liberal in determining which voxels are likely activated during the experiment. In particular, it becomes easier to identify the FEFs. This is indicative of the loss of sensitivity of the SB model demonstrated in our simulation study in Chapter 4. What may be relatively weak activations in this map would be missed entirely without allowing neighboring voxels to borrow strength from each other.



(a) t-maps



(b) Independence model probabilities

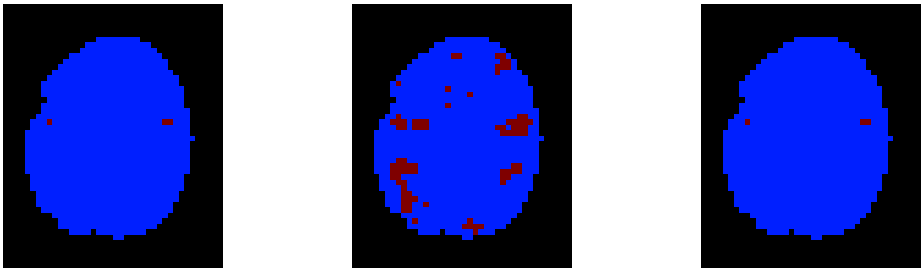


(c) CAR model probabilities

Figure 5.3: t-maps and probability maps for fourteen healthy participants under both the Scott-Berger and CAR models. The probabilities are shown on the log scale for enhanced resolution.

We further evaluate the performance of our Bayesian CAR testing model by comparing its results with those obtained by thresholding to control the false discovery rate. Thresholding posterior probability maps at .95 means that the voxels identified as being activated are such that $P(\theta_j = 0 \mid \mathbf{y}) \leq .05$. This is analogous to controlling the proportion of discoveries that are uninteresting cases to be no more than .05 (Friston and Penny, 2003). Indeed, when assuming as we do here that the statistics arise from a density of the form $f(y) = pf_0(y) + (1-p)f_1(y)$, the q threshold used in the FDR algorithm can be expressed as an estimate of the Bayes probability that a rejected hypothesis is incorrectly selected (Efron, 2010, Chapter 4) . Thus, we take q to be .05 for the results below to be a fair comparison.

Figure 5.4(c) displays results from thresholding with the FDR step-up procedure. The independence Bayes, CAR-structured Bayes, and FDR approaches all seem to agree on the selected voxels with the strongest changes in BOLD signal. The CAR model, however, is still the most liberal of the three approaches, with the FDR procedure exhibiting performance closer to that of the independence Bayes model. Again, the CAR structure yields results that expand the selected regions of activation compared to FDR control so that more voxels are included in contiguous clusters identified as supportive of the antisaccade task.



(a) Independence (b) CAR model (c) FDR control

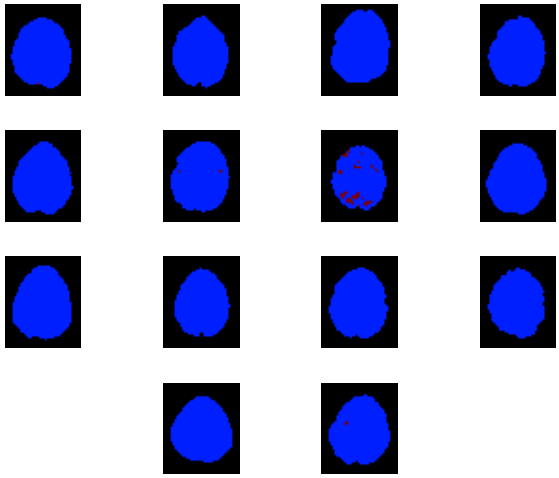
Figure 5.4: Thresholded activation maps for Subject 6. Voxels are selected with estimated non-null probability of at least .95. The false discovery rate (FDR) results are obtained using the Benjamini and Yekutieli (2001) algorithm, controlling at a rate of .05.

Figure 5.5 displays thresholded activation maps for all fourteen participants using the independence model, our model, and FDR control. We can see that both the independence model and FDR procedure are much more conservative than the CAR model, in general. Table 5.1 reports the distances between the corresponding maps as a measure of similarity. The distance between two images is found by expressing them as vectors $\mathbf{x}$, $\mathbf{y}$ and calculating the usual Euclidean norm of the difference, $((\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}))^{1/2}$. The Table shows that the independence model results and FDR thresholding are more comparable to each other than either is to the CAR model. Results obtained under our model are substantially different from either of the other two approaches.

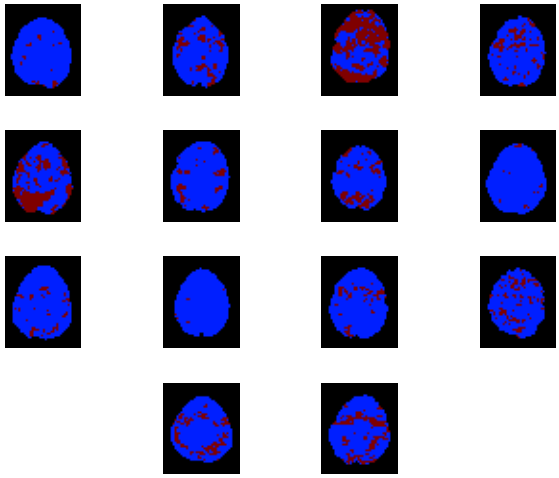
Table 5.1: Euclidean distances between corresponding vectorized threshold images under independence, CAR, and FDR thresholding, for each subject.

Subject	SB-FDR	SB-CAR	CAR-FDR
1	1.00	5.66	5.74
2	1.00	11.58	11.53
3	2.24	22.65	22.54
4	5.20	11.71	10.49
5	2.65	17.97	17.78
6	0.00	9.11	9.11
7	4.36	8.83	9.85
8	1.73	4.47	4.12
9	0.00	8.37	8.37
10	2.00	2.65	1.73
11	2.00	9.00	8.78
12	3.46	11.70	11.18
13	1.41	12.33	12.25
14	4.90	13.71	12.80
Average	2.28	10.70	10.45

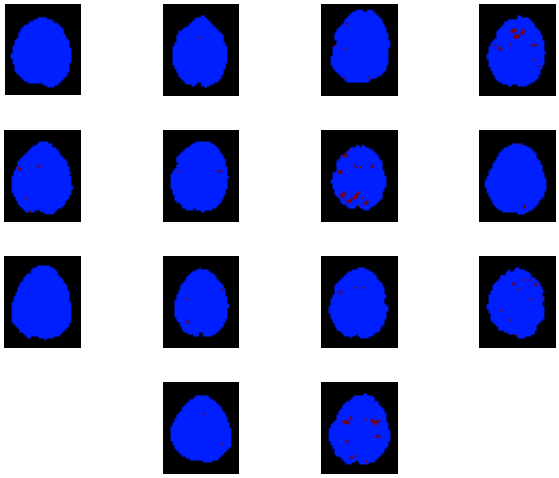
We find that the increased sensitivity of the Bayesian CAR model compared to the Bayesian model under independence seems to hold in general. Figure 5.6 displays the thresholded activation maps averaged over all of the participants. Each voxel in the image is the average of the zeros and ones appearing in the fourteen maps. The maps only retain points for which the corresponding voxels are non-masked



(a) Independence model



(b) CAR model



(c) FDR thresholding

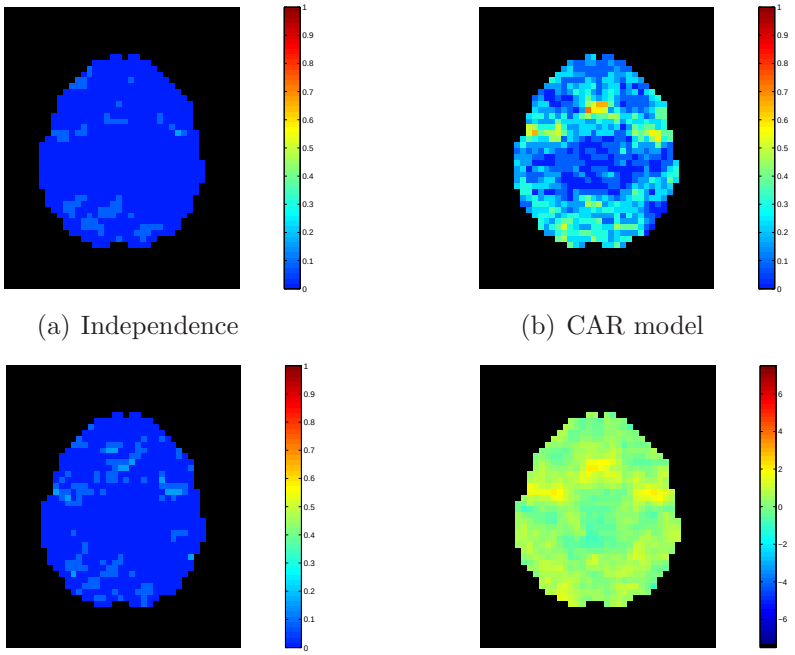
Figure 5.5: Thresholded activation maps for all fourteen participants using the Scott-Berger model, the CAR model, and Benjamini and Yekutieli (2001) FDR control.

over all participants so that the summary thresholds are displayed on a common intersection. For a clear picture of where the strongest BOLD signal changes are observed in general, a map of the t statistics averaged at each voxel is also included. This highlights where the common centers of activation tend to be for the participants and is largely in agreement with Figure 5.2. The average t-map also displays only the intersection of non-masked voxels over all subjects. Combining brains in this way is problematic when doing formal inference (Lazar et al., 2002), but it is useful here for illustrative purposes.

We see that the Bayesian independence model is extremely conservative, perhaps over-correcting for multiplicity like the more traditional corrective procedures. In this case, we come back to the same problem as with the classical approaches. By modeling spatial dependence in the data, we have increased the power, making physiologically meaningful regions associated with the task easier to identify. This brings us closer to the desired delineation of task-related neural circuitry. By harnessing the dependence to our advantage instead of treating it as a burden, we improve our ability to select regions of interest that may be further investigated in subsequent analyses.

Figure 5.6 also displays the summary map of the thresholds calculated using the general FDR procedure. The algorithm is applied to each participant separately so that there are fourteen sets of p-values. Zero-one indicators are used as before to depict selected voxels. The thresholded images for individual slices, displayed in Figure 5.5, are averaged over the fourteen participants at each location. Only the intersection of voxels that are non-masked in each brain is retained in the final image.

These results underscore the power that can be gained with modeling spatial dependence in SPM analysis as opposed to Bayesian testing under independence or FDR control with arbitrary dependence structure. The Bayesian testing model under independence is indeed effective in adjusting for multiplicity. The adjustment may be too strong for the purposes of fMRI, though. In allowing for an arbitrary distribution



(a) Independence (b) CAR model

(c) FDR control (d) Average t-map

Figure 5.6: Threshold maps averaged over participants for each model and the average t-map. The value at each voxel is the proportion of times it is selected as active over the fourteen subjects. Each image is the intersection of non-masked voxels over all participants.

of the p-values, the FDR-controlling algorithm also sacrifices power. This algorithm is closer to the overly conservative Bayesian independence results than those obtained under the CAR structure.

5.2 Analysis With All Participants Simultaneously

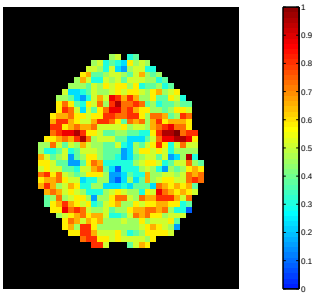
In addition to treating each of the fourteen participants separately, we perform a combined analysis. Here, all the observed test statistics are considered simultaneously as one data set rather than as fourteen separate sets. As in the single participant case, we view each test statistic y_j as arising from the spike-CAR mixture with common error and signal variances. That is, we suppose that $y_j \sim N(\theta_j, \sigma^2)$, $j = 1, \ldots, J$, where $J = \sum_m J_m = 11,958$ is the total number of observations included over all participants, after masking. We assign the means of the datum-specific distributions

the mixture with continuous component given by (3.1). This is the same model as that which was used for each participant individually; there are just more observations and thus more parameters included in the evaluation of the posterior distribution. The information separating one brain from another for modeling CAR dependence is totally contained in the neighborhood structure via the adjacency matrix.

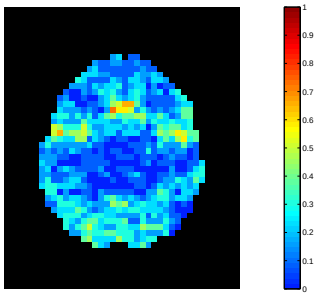
While the spatial structure is preserved, this is different from treating each participant separately in that much more information is being shared throughout the model. The posterior distributions of σ and τ^{-2} are updated using the information from the approximately $12,000$ test statistics simultaneously. When we treat each participant separately, the posteriors only use data contained in the particular subject's slice being analyzed. The intrinsic Bayesian multiplicity adjustment uses a much greater number of tests in the correction for the combined set, affecting the posterior distribution of p and thus the marginal probabilities of inactivation at each observation, p_j, $j = 1, \ldots, J$. We discuss differences in the posteriors below.

Figure 5.7 shows the averaged threshold maps obtained from treating each participant separately as well as simultaneously. In both analyses, we use a common shape parameter of $\alpha = 108$ in the prior for p. We see that using the same model for the much larger combined data set results in probability estimates that are more sensitive to small perturbations from zero. This is likely due to the fact that the greater number of observations considered has a stronger effect on the posterior distribution. A more influential shape parameter in the prior for p is necessary in the presence of the additional data points to enforce sparsity, as discussed in Section 2.2. We investigate this sensitivity on these data in the next section.

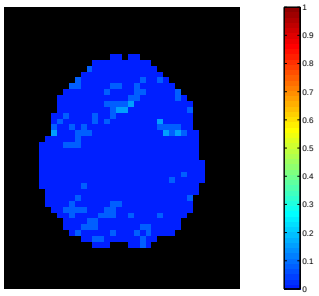
Also included in Figure 5.7 are results from two other procedures for which researchers may opt when attempting to combine information over all participants. The first is using the general FDR procedure with $J = \sum_{m=1}^{14} J_m$ in place of the subject-specific J_m. We separate out by slice the indicators of selected voxels so that each



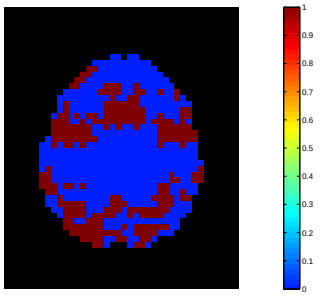
(a) Bayesian CAR model - all participants simultaneously



(b) Bayesian CAR model - each participant separately



(c) FDR control with all participants simultaneously



(d) FDR control on Fisher's combined p-values

Figure 5.7: Comparison of threshold maps combining information in all participants' slices. For panels (a)-(c), the value at each voxel is the proportion of times it was selected over the fourteen subjects. Panel (b) is the same as the CAR results presented in Figure 5.6. Panel (d) is the result of applying FDR control to p-values obtained from Fisher's method.

point in the summary map is the average of indicators over the corresponding voxels in each participant's slice. The second procedure we consider is Fisher's p-value method (Fisher, 1950). This method creates a new map of p-values instead of defining a rule for thresholding the original ones. The p-value at location j is found by calculating $T_j = -2\sum_{m=1}^{l_j} \log p_{mj}$, where p_{mj} is the p-value observed at the j^{th} voxel of subject m and l_j is the number of terms used in evaluating the sum at voxel j. Each location in the new map may be considered a summary voxel with value T_j, which follows a chi-squared distribution with $2l_j$ degrees of freedom. We use the general FDR procedure on this new map of p-values and indicate selected summary voxels. Since this results in only one map of p-values, there is no averaging. To maintain uniformity

between the maps, we include in the final map only voxels that are non-masked over all participants.

Using FDR control over all participants' slices simultaneously yields similar results to those we find by treating each participant individually. The combined p-value method is clearly much more liberal, even more so than the CAR analysis on each participant individually. This gain in power is desirable, but there is the risk of it being too liberal. It smooths regions together and selects areas that may not be of interest. By adding the natural logarithms of the p-values across subjects, it only takes one extreme observation to make the test statistic large. One unusual participant could cause a voxel to be declared significant despite the majority of others who exhibited no interesting BOLD signal changes at all (McNamee and Lazar, 2004).

5.3 Sensitivity Analysis

Two aspects of our model that may seem arbitrary are the prior on the hypervariance, τ^2, and the specification of the shape parameter in the prior for p. The influence of prior specifications on any posterior calculations are usually of interest in practice (Gelman et al., 2004). We thus perform a sensitivity analysis to determine the robustness of these modeling assumptions. We explore effects of both the prior on τ^2 (equivalently, $\tau = \sqrt{\tau^2}$) and the value of α below.

Robustness to the Prior on τ

It is suggested in Gelman (2006) that the priors specified for the variance parameters in hierarchical models may have an undesirable effect on posterior inference. Priors used for scale hyperparameters that are theoretically noninformative may, in fact, have a considerable influence on the posterior distribution. To address this possibility, we fit our model to the experimental fMRI data with different priors and visually compare the results. Following Gelman (2006), the three priors we use are Gamma,

$\tau^{-2} \sim \text{Ga}(.001, .001)$, a vague but proper Uniform prior, $\tau \sim \text{U}(0, 1000)$, and a *folded-t*, or half-t, distribution. This distribution can be expressed simply as that of $|Z|W^{-1/2}$, where Z is a standard normal random variable and W follows a chi-squared distribution with ν degrees of freedom. The density is given by

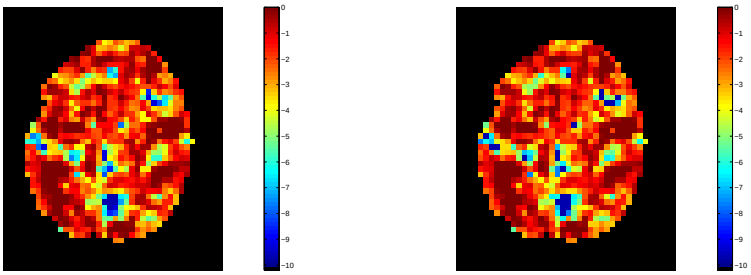
$$\pi_{|t|}(\tau) \propto \left(1 + \frac{1}{\nu}\left(\frac{\tau}{C}\right)^2\right)^{-(\nu+1)/2}, \quad \tau > 0, \tag{5.1}$$

where C is a scale parameter. Interested readers are referred to Johnson and Kotz (1972) or Gelman (2006) for more details. Here, we take $C = 1$ and $\nu = 2$, which we denote as $\tau \sim |t_2|$. For the analysis of the priors on τ, we keep α fixed at 108 in the prior for p.

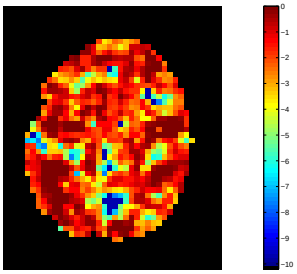
Figure 5.8 displays the log-scale estimated posterior probability maps for Subject 6 using each prior on τ. The three maps are virtually indistinguishable from each other. We also see this similarity when examining the thresholded activation maps of the same brain, displayed in Figure 5.9. The only noticeable differences in selected voxels are in the area of the SEF. Otherwise, the results from the three model variants closely agree.

To better ascertain the effect of the prior, we display the averaged threshold maps that result from using each prior in Figure 5.10. The results are largely in agreement with each other, although it appears that the Uniform prior is slightly more conservative compared to the other two. There are no major inconsistencies across the three alternatives, indicating that our CAR testing model is robust to the prior used on the scale hyperparamter.

Our exploration of the effects of different priors on the hypervariance demonstrates that choosing one over another results in negligible differences. This is likely because the effect of the variance prior is mainly a concern when the number of groups, and hence the number of parameters being modeled in the prior, is small, as noted in

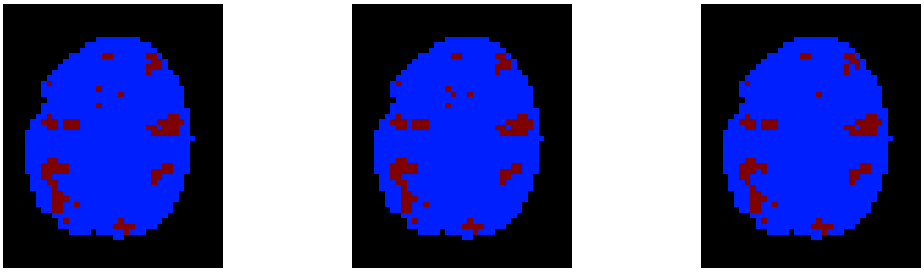


(a) Ga(.001, .001) (b) U(0, 1000)



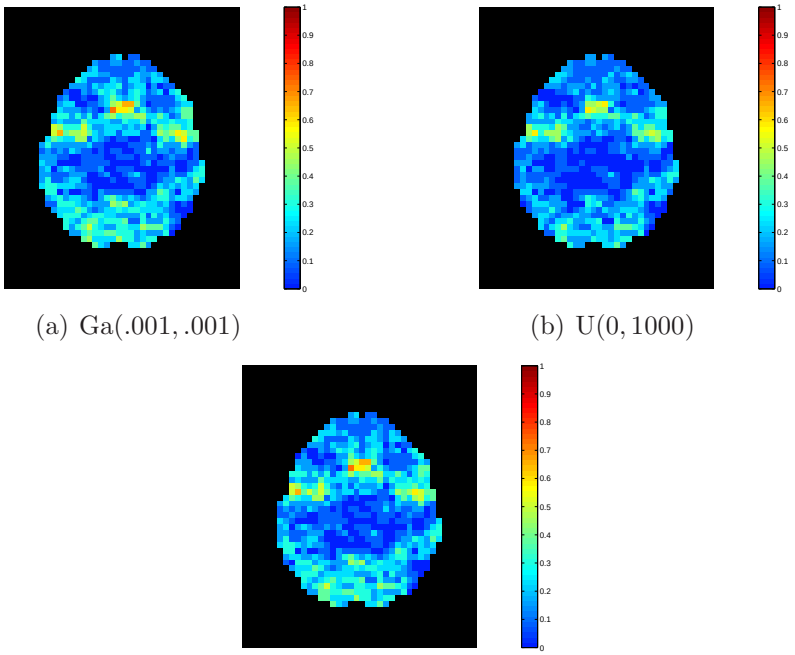
(c) $|t_2|$

Figure 5.8: Comparative posterior probability maps of activation for the individual slice from Subject 6 with different hyperpriors for the scale hyperparameter, τ (τ^{-2} in Panel a). The three distributions compared are the same as those considered in Gelman (2006). Probabilities are displayed on the log scale for improved resolution.



(a) Ga(.001, .001) (b) U(0, 1000) (c) $|t_2|$

Figure 5.9: Comparative thresholded activation maps for the individual slice from Subject 6 with different hyperpriors for the scale hyperparameter. The distributional classes compared are the same as those considered in Gelman (2006).



(a) Ga(.001, .001)

(b) U(0, 1000)

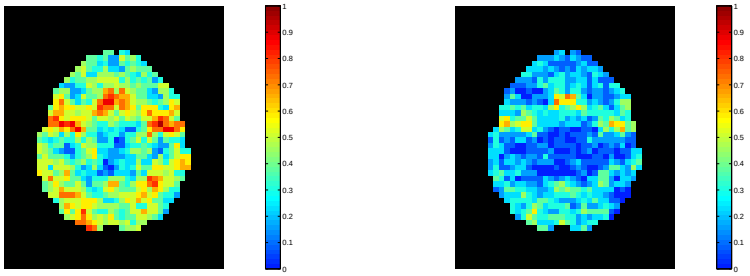
(c) $|t_2|$

Figure 5.10: Threshold maps averaged over participants for each hyperprior on the scale hyperparameter. The value at each voxel is the proportion of times it is selected as active over the fourteen subjects. Each image is the intersection of non-masked voxels over all participants.

Gelman (2006). In our case, each statistic is considered as arising from its own class, meaning that the number of groups is the same as the number of tests being performed. The sensitivity is thus drastically reduced for large J_m. It is worth noting the similarity between $\pi_{\tau^2|\sigma^2}(\tau^2 \mid \sigma^2)$ in Model 2 and the half-t density (5.1), which is suggested by Gelman as a "weakly informatve" alternative to the usual Inverse Gamma. We believe, though, that in the absence of any compelling reason to choose a specific prior (e.g. strong prior knowledge about the data to be analyzed), a researcher may choose the most computationally convenient prior and still maintain the integrity of inferences.

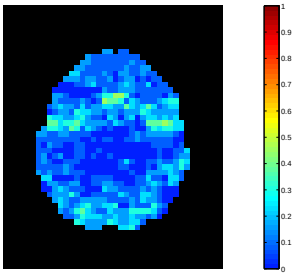
Sensitivity of the Shape Parameter in the Prior for p

We investigate the sensitivity of the shape parameter in $\pi_p(\cdot)$ by repeatedly implementing the model over a range of α values. We determine the values over which to run the simulations by applying our empirical Bayes estimation method over a range of suitable thresholds. Section 5.1 shows how this estimator relies on estimating the probability of activation with the proportion of t statistics exceeding a certain threshold. For example, a threshold of 4 results in $\tilde{\alpha} = 108$. For the usual one-sided z test, a .01 significance level has a critical value of $z' \approx 2.33$, whereas the corresponding one-sided t test with few degrees of freedom uses a critical value of $t' \approx 4$. Therefore, to determine a grid of reasonable values, we calculate $\hat{P}(\gamma_j = 1 \mid \alpha) = \sum_{m=1}^{14} \sum_{j=1}^{J_m} I(y_{mj} > t') / \sum_{m=1}^{14} J_m$ for a sequence of t' between 2.3 and 4.3, finding $\tilde{\alpha}$ from each resulting estimate.



(a) $\alpha = 25$

(b) $\alpha = 94$



(c) $\alpha = 121$

Figure 5.11: Threshold maps averaged over participants for selected values of α in the p prior, $\pi_p(p) = \alpha p^{\alpha-1}$. The value at each voxel is the proportion of times that voxel was selected over the fourteen subjects using the CAR model. Each image is on the intersection of non-masked voxels in the slices.

Figure 5.11 displays the threshold maps averaged over subjects for $\alpha = 25$, $\alpha = 94$, and $\alpha = 121$, corresponding to critical values of 2.94, 3.91, and 4.11, respectively. We observe the effect of increasing α in the decreasing number of voxels selected as task-related and concomitant shrinking of the neuronal regions identified. This reflects the fact that higher α values lower a researcher's estimated probability of activation at any particular voxel, strengthening the criterion which must be met for selection. High α values should be specified when the activity is thought to be sparse, in which case only the most extreme voxels will be identified as task-related. In this sense, the parameter may be used to adjust the threshold for multiplicity. This could potentially lead a researcher into a problem similar to choosing the threshold for p-values in the classical hypothesis test setting. Hierarchical modeling, though, may make it possible to avoid this issue by placing a prior on the α parameter to reflect more uncertainty. The additional uncertainty in the prior has the effect of allowing more freedom in the data to determine the plausible values. We elaborate on this point in the next chapter.

Chapter 6

Concluding Remarks

6.1 Discussion

The analysis of fMRI data requires performing inference on a massive scale. The data with which we are concerned in this dissertation are types of SPMs arising from fMRI studies. As such, we assume that there is no temporal component to the data being analyzed; each voxel contains one value quantifying the observed changes over time. Beginning with these types of maps has the advantage that they are already popular in neuroimaging. A researcher can use preprocessing steps to construct the maps with standard software such as AFNI (Cox, 1996), SPM (Wellcome Department of Cognitive Neurology, London, UK), or FSL (Nuffield Department of Clinical Neurosciences, Oxford, UK).

In correcting for the thousands of simultaneous tests that must be performed, there is a risk of using an overly conservative procedure. *Ad hoc* threshold corrections can reduce or eliminate selected areas of activation associated with task performance in an fMRI study. We show that this risk is still present under the Bayesian approach to large-scale testing. In particular, assuming independence among the test statistics causes the Bayesian testing model to become extremely conservative. The self-calibration that is inherent in the procedure tends to over-correct much like other multiple testing procedures that have already proven to be problematic for high throughput data.

We demonstrate that a gain in power can be achieved by introducing a conditional autoregressive structure to account for local spatial dependence among the voxels. This model appears also to be more powerful than the Benjamini-Yekutieli procedure

under arbitrary p-value distributions. Rather than being treated as an inconvenience, we harness the spatial dependence present in brain images to improve our ability to detect task-related activation. By allowing voxels to borrow strength from those in their immediate neighborhoods, our model can increase the sensitivity of the Bayesian testing procedure. This draws attention to the inherent flexibility of Bayesian models to capture nuances in particular data structures that are otherwise difficult to account for. Any FDR procedure, on the other hand, is ultimately at the mercy of the null distribution used to calculate p-values. The correct distribution to use is something upon which theory and empirical evidence do not always agree. It is also worth noting that while the *expected* false discovery rate is controlled, there is no guarantee that the *actual* proportion of discoveries that are false for any particular data set is even close to the nominal rate.

One of the advantages of groupwise analyses is that common activation across subjects enables easier identification of true regions of activation. Intra-subject analyses, on the other hand, may have interesting areas that are more difficult to identify because of a lack of corroborating information. We explore our model's behavior both in the analysis of a single brain image and in combining slices across subjects. By increasing sensitivity, our model improves an investigator's ability to identify significant activation at the level of a single person. When analyzing groupwise results, our approach can avoid the disproportionate influence of one participant, unlike Fisher's p-value method.

We mention in Section 3.2 the simplifications made for the sake of computational implementation. Extending the CAR dependence structure to account for three-dimensional dependence in a whole-brain analysis is theoretically straightforward. One only needs expand the definition of a neighborhood to include the 26 voxels adjacent to and above and below each location. In practice, though, computational efficiency and prior parameter specification are two issues requiring investigator input.

For a fixed value of α in the prior for p, the results are seen to be less sensitive when considering a single slice versus multiple slices simultaneously. It is likely that this would also be the case when analyzing a three-dimensional volume from a single person.

We expect that the methodology presented in this paper can be readily extended to other imaging modalities besides fMRI. For example, our model could be applied to data sets obtained from positron emission tomography (PET). Bowman (2007) briefly commented on the similarity between fMRI and PET, noting that the design matrix is the primary difference between the GLMs used for their analyses. It is possible to construct an SPM corresponding to test statistics for summarizing the voxel-specific stimulus effects on cerebral blood flow, the response variable in PET data. The observations for the PET SPMs can then be treated in the same manner as they are in the fMRI case presented here. Indeed, we believe that any procedure that results in summary maps quantifying evidence of activation over voxels could incorporate this procedure, with appropriate modifications in the null distribution and other parameters.

6.2 Future Research

Extensions

The empirical results presented in this dissertation suggest an extreme sensitivity to the choice of prior on p in our CAR model, making careful selection crucial. The specification of appropriate priors is a question that merits further investigation. Several approaches for prior specification were addressed at length by Berger (1985, Chapter 3). The more fundamental question of whether to use an elicited prior or perhaps some noninformative or otherwise computationally simpler prior could also be addressed. Indeed, Carlin and Louis (2009, page 32) remarked that "the inherent difficulty of the elicitation task makes us lean toward its use only in situations where

anticipated data sample sizes are small, and the experts possess good reliable prior information ... on the subject at hand." Given the wealth of knowledge that has accumulated about certain simple cognitive tasks (such as eye movements), elicitation might be feasible in this case.

One possible direction for future research is to more directly address the sensitivity to the α shape parameter in the prior for the mixing parameter p. Rather than modeling p with a Beta$(\alpha, 1)$ distribution, it may be better to model the voxel-specific activation probabilities, $q_j \equiv 1 - p_j$, with a spike-Beta mixture, $q_j \sim (1-\xi)\delta_0 + \xi\text{Beta}(sr, s(1-r))$, and $\xi \sim \text{Beta}(a\nu, a(1-\nu))$, with hyperparameters specified so that ξ is likely to be small and the Beta part of the distribution of q_j places most of its probability mass near one. Lucas et al. (2006) and Carvalho et al. (2008) showed that introducing a prior of this form induces greater shrinkage in the posterior distribution, resulting in a stronger distinction between points declared null and non-null. In our model, perhaps this would allow the shape parameters to automatically adapt to the number of tests being performed, circumventing the problem of specifying α directly. Alternatively, data-driven approaches to selecting α could be explored. It may be possible to incorporate the number of tests J_m directly or otherwise use an empirical procedure to optimally choose α for a given data set.

A current need in the neuroscience community is for viable methods by which comparisons can be made between groups of people. For example, a difficult but important problem in mental health research is summarizing the differences between schizophrenia patients, their relatives, and a healthy population. After a suitable model has been chosen for a single brain, or a single group of brains, it would be useful to explore techniques with which the model can be extended to make groupwise comparisons possible, i.e. through hierarchical modeling. Currently, one of the most popular methods for comparing groups of subjects in neuroimaging studies is within the linear mixed model framework. While this has certainly proven to be a useful

tool, it is known to be quite conservative and sensitive to model misspecification (McNamee and Lazar, 2004; Lindquist, 2008).

Another promising avenue is extending the Bayesian multiple testing framework to distinguish between positive, task-related activation and task-induced deactivation associated with the so-called "default network" (Buckner et al., 2008). This network, which is the subject of increasing discussion among neuroscientists, exhibits a much weaker signal than task-related activation. An extension allowing for the classification of voxels into three or more cases could perhaps be accomplished with the Brown-Stein model (Brown, 1971; Stein, 1981; Efron, 2009). This is a model of the form

$$\begin{aligned} \delta &\sim g \\ y \mid \delta &\sim f_\delta, \end{aligned}$$

and is useful for modeling observations as arising from one of several classes, e.g. $y \mid \delta \sim N(\delta, \sigma^2)$ with g a finite mixture distribution. It was suggested by Muralidharan (2010) that taking g to have three components could allow the observations to be partitioned into null, positive, and negative effect cases.

Addressing the Dependence Structure

It would also be desirable to explore other dependence structures for incorporation into a Bayesian testing model. We use the CAR model here for both its simplicity in MCMC sampling and its interpretability. Our model could be modified by adapting the approach of Smith and Fahrmeir (2007). Letting the means $\boldsymbol{\mu}$ be conditionally independent, it could be assumed that the indicators $\boldsymbol{\gamma}$ follow the distribution determined by

$$\pi(\gamma_j \mid \boldsymbol{\gamma}_{(-j)}) \propto \exp\left(\sum_i \alpha(\gamma_i) + \sum_{j \sim i} \omega_{ji}\theta_{ji} I(\gamma_j = \gamma_i)\right),$$

where $\alpha(\cdot)$ is an "external field" and the second term is the interaction effect between γ_j and its neighbors. Since the estimate of the probability of activation at location j would then be influenced more by γ_j than μ_j, stronger clustering among the points determined to be active may be induced. This could allow for more clearly-defined regions of activation while avoiding oversmoothing.

Another possible alternative is modeling the dependence structure with a Gaussian process in the likelihood of a hierarchical model. For instance, letting (c_{1j}, c_{2j}, c_{3j}) denote the coordinate of voxel j, one could model the data from the m^{th} subject as $\mathbf{y}_m \sim N(\theta_m, \sigma_m^2 \kappa_\phi)$, $m = 1, \ldots, M$, where κ_ϕ is a Gaussian correlation matrix $\{\kappa_\phi(k,l)\}_{k,l=1}^{J_m}$ with $\kappa_\phi(k,l) = \exp\left(-\phi_1(c_{1k} - c_{1l})^2 - \phi_2(c_{2k} - c_{2l})^2 - \phi_3(c_{3k} - c_{3l})^2\right)$. For an entire brain volume, the computational burden of a fully Bayesian analysis would be unbearable for even modern computing technology. The range parameter $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3)$ would be particularly troublesome to estimate. One approach to this problem is empirical Bayes, i.e. to find an estimate of $\boldsymbol{\phi}$ and treat it as known. Qian and Wu (2008) assigned the components of $\boldsymbol{\phi}$ independent Gamma distributions and then treated the posterior mode of $\boldsymbol{\phi}$ as fixed. Alternatively, the optimization problem for estimating $\boldsymbol{\phi}$ could be simplified by using composite likelihood methods (Besag, 1977; Lindsay, 1988; Cox and Reid, 2004), approximating the full likelihood with the marginal distributions of local neighborhoods in the images (e.g. Nott and Ryden, 1999). See Santner et al. (2003) for an introduction to the Bayesian analysis of Gaussian processes.

Neuroimaging data are known to exhibit complex correlation. Research and experience suggest that the dependence structure among voxels in fMRI data do not exhibit a conventional spatial dependence structure determined by Euclidean distance. A different measure of distance, such as one based on neuronal pathways connecting regions of the brain, may be more appropriate for analyzing such data. Some sections of the brain seem to be correlated according to a *functional* dependence structure,

where they are associated according to a common functional behavior. Bowman (2007) incorporated such a distance metric into a spatiotemporal mixed model for analyzing PET scans and noted a straightforward extension to accommodate fMRI. While other work has been done on modeling complex measures of distance and covariance structures (e.g. Dryden et al., 2009), it has been limited, particularly with respect to fMRI. The nature of the true dependence structure is no doubt difficult to model, but necessary to facilitate more reliable inference.

Model Assessment and Selection

Model checking is a vastly underexplored area of neuroimaging (Lindquist, 2008). With competing models using different dependence structures, it would be natural to ask which one is "best". A standard Bayesian method for comparing models is the *Bayes factor* (BF). For two competing models M_1 and M_2, the Bayes factor is simply the posterior odds of M_1 to M_2 divided by the prior odds,

$$BF = \frac{P(M_1 \mid \mathbf{y})P(M_2)}{P(M_2 \mid \mathbf{y})P(M_1)} = \frac{P(\mathbf{y} \mid M_1)}{P(\mathbf{y} \mid M_2)}.$$

Pure Bayes factors are limited, though, in that they force a choice between only two models, assuming one is correct, and rely on proper priors to be well-defined. This has led to variants of the BF, such as intrinsic Bayes factors (Berger and Pericchi, 1996, 1998) and the fractional Bayes factor (O'Hagan, 1995). An overview of model selection from both Bayesian and frequentist perspectives may be found in Kadane and Lazar (2004).

A similar issue to model selection is that of model assessment. The appropriate measure of fit for Bayesian models is still a much-debated topic in the literature. Perhaps the most natural technique is to examine a set of residuals from a fitted model. Carlin and Louis (2009) suggested a cross-validation approach that, for each

observation, uses the posterior calculated from all the data less the observation in question to predict the missing datum and find the residual. That is, one can examine

$$d_i = \frac{y_i - E(Y_i \mid \mathbf{y}_{(-i)})}{\sqrt{\mathrm{Var}(Y_i \mid \mathbf{y}_{(-i)})}},$$

where the mean and variance are calculated with respect to the conditional predictive distribution, $f(y_i \mid \mathbf{y}_{(-i)}) = \int f(y_i \mid \theta, \mathbf{y}_{(-i)}) p(\theta \mid \mathbf{y}_{(-i)}) d\theta$. Much closer to traditionally frequentist ideas is the notion of a *Bayesian p-value* (Gelman et al., 2004). This is defined as $p_B := P(T(\mathbf{y}^{rep}, \theta) \geq T(\mathbf{y}, \theta) \mid \mathbf{y})$, where $T(\cdot)$ is a suitable test quantity and $\mathbf{y}^{rep}$ is a vector of 'future' observations drawn from $p(\theta, \mathbf{y}^{rep} \mid \mathbf{y})$. In other words, the Bayesian p-value is obtained by comparing a function of the predicted data to the same statistic calculated from the observed data. Using the tail-area probability, the Bayesian p-value functions like a classical p-value in that a small value is evidence against the working model. Combining notions from both the Bayesian and frequentist frameworks for calibrating the operating characteristics of Bayesian models was advocated by Little (2006, 2011), who cited Box (1980) and Rubin (1984) as seminal works in the area.

Additional Applications

Lastly, we note that the scope of possible applications of this work is not limited to brain imaging. Other applications are possible. There is an interest in genomics to identify gene pathways that are associated with certain diseases, including cancer (West, 2003) and mental illness. Such analyses involve performing inference among tens of thousands of genes simultaneously, often in the presence of considerable dependence among the observations (Efron, 2007). The emerging field of syndromic surveillance seeks to identify potential disease outbreaks through continuous monitoring of reported symptoms at hospitals in geographic regions (Rath et al., 2003;

Banks et al., 2012). This field is pertinent to epidemiology as well as domestic policy concerning bioterrorism. There are potential applications to imaging problems in astronomy, as well. Observatories are constantly collecting images containing massive amounts of information for learning about, e.g., the evolution of galaxies or planetary systems. These collections of observations can dwarf even the largest neuroimaging data sets (Liang et al., 2004; Efron, 2008).

It is our hope that our research will lead to models that are easily extendable to other types of data. Some models, such as the CAR model presented here, are contingent upon an areal structure, with clearly defined neighborhoods that are only a function of adjacency and not Euclidean distance. A Gaussian process model, on the other hand, explicitly incorporates the measurable distance between two points into the correlation structure. Such a model could thus be used to analyze not only areal data like fMRI, but more complex geostatistical data.

6.3 Conclusion

We live in The Information Age in which technology has evolved to the point where data are being collected on a scale unimaginable even fifteen years ago. Researchers in many fields now have access to massive amounts of information about a wide range of interests, including the shopping behavior of millions of online customers, financial trends in the stock market, disease patterns in geographic regions over time, real-time internet traffic, and detailed biomedical images, to name a few. Our ability to collect data has now far outpaced our knowledge of how to analyze them. Techniques for analyzing large-scale data are in demand as never before.

This work introduces a specific method for incorporating spatial dependence into a Bayesian thresholding framework for neuroimaging data. With hierarchical modeling, adaptive parameter specification, and tenable choices of dependence structures, an entire class of models may be developed for large-scale areal data. Such a class of

models could provide a unifying framework for automatic multiple testing adjustments in a variety of high throughput settings.

We are optimistic that Bayesian procedures will be made more accessible to a broader community of researchers; the biggest hindrance now being computational limitations. More research needs to be done toward finding ways to increase computational efficiency in MCMC and other routines for the analysis of massive data sets. Reducing the computational burden is critical for researchers to enjoy the full flexibility that hierarchical modeling provides. The beneficiaries of such advancements would not be limited to neuroimaging, but would include the entire scientific community.

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton: Chapman & Hall/CRC.

Banks, D., Datta, G., Karr, A., Lynch, J., Niemi, J., and Vera, F. (2012), "Bayesian CAR models for syndromic surveillance on multiple data streams: Theory and practice," *Information Fusion*, 13, 105–116.

Barbieri, M. M. and Berger, J. O. (2004), "Optimal predictive model selection," *Annals of Statistics*, 32, 870–897.

Bartle, R. G. (1976), *The Elements of Real Analysis*, Hoboken: John Wiley & Sons, 2nd ed.

Benjamini, Y. (2010), "Discovering the false discovery rate," *Journal of the Royal Statistical Society, Series B*, 72, 405–416.

Benjamini, Y. and Heller, R. (2007), "False discovery rates for spatial signals," *Journal of the American Statistical Association*, 102, 1272–1281.

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

Benjamini, Y. and Yekutieli, D. (2001), "The control of the false discovery rate in multiple testing under dependency," *Annals of Statistics*, 29, 1165–1188.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag, 2nd ed.

Berger, J. O. and Pericchi, L. R. (1996), "The intrinsic Bayes factor for model selection and prediction," *Journal of the American Statistical Association*, 91, 109–122.

— (1998), "Accurate and stable Bayesian model selection: The median intrinsic Bayes factor," *Sankhyā: The Indian Journal of Statistics, Series B*, 60, 1–18.

Bernardo, J. M. and Smith, A. F. M. (1994), *Bayesian Theory*, New York: Wiley.

Berry, D. A. and Hochberg, Y. (1999), "Bayesian perspectives on multiple comparisons," *Journal of Statistical Planning and Inference*, 82, 215–277.

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

— (1977), "Efficiency of pseudolikelihood estimation for simple Gaussian fields," *Biometrika*, 64, 616–618.

Besag, J. and Kooperberg, C. (1995), "On conditional and intrinsic autogressions," *Biometrika*, 82, 733–746.

Besag, J., York, J. C., and Mollié, A. (1991), "Bayesian image restoration, with two applications in spatial statistics (with discussion)," *Annals of the Institute of Statisical Mathematics*, 43, 1–59.

Bogdan, M., Ghosh, J. K., and Tokdar, S. T. (2008), "A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing," in *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, eds. Balakrishnan, N., Peña, E. A., and Silvapulle, M. J., Beachwood: Institute of Mathematical Statistics, pp. 211–230.

Bowman, F. D. (2007), "Spatiotemporal models for region of interest analyses of functional neuroimaging data," *Journal of the American Statistical Association*, 102, 442–453.

Bowman, F. D., Caffo, B., Bassett, S. S., and Kilts, C. (2008), "A Bayesian hierarchical framework for spatial modeling of fMRI data," *NeuroImage*, 39, 146–156.

Box, G. E. P. (1980), "Sampling and Bayes' inference in scientific modeling and robustness (with discussion)," *Journal of the Royal Statistical Society, Series A*, 143, 383–430.

Brook, D. (1964), "On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems," *Biometrika*, 51, 481–483.

Brown, L. D. (1971), "Admissable estimators, recurrent diffusions, and insoluble boundary value problems," *Annals of Mathematical Statistics*, 42, 855–903.

Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008), "The brain's default network," *Annals of The New York Academy of Sciences*, 1124, 1–38.

Camchong, J., Dyckman, K. A., Austin, B. P., Clementz, B. A., and McDowell, J. E. (2008), "Common neural circuitry supporting volitional saccades and its disruption in schizophrenia patients and relatives," *Biological Psychiatry*, 64, 1024–1050.

Carlin, B. and Banerjee, S. (2003), "Hierarchical multivariate CAR models for spatio-temporally correlated survival data," in *Bayesian Statistics 7*, eds. Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., Oxford: Oxford University Press, pp. 45–63.

Carlin, B. P. and Louis, T. A. (2009), *Bayesian Methods for Data Analysis*, Boca Raton: Chapman & Hall/CRC, 3rd ed.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008), "High-dimensional sparse factor modeling: Applications in gene expression genomics," *Journal of the American Statistical Association*, 103, 1438–1456.

Chib, S. and Greenberg, E. (1995), "Understanding the Metropolis-Hastings algorithm," *American Statistician*, 49, 327–335.

Chipman, H. (1996), "Bayesian variable selection with related predictors," *Canadian Journal of Statistics*, 24, 17–36.

Chumbley, J. R. and Friston, K. J. (2009), "False discovery rate revisited: FDR and topological inference using Gaussian random fields," *NeuroImage*, 44, 62–70.

Clyde, M. and George, E. I. (2004), "Model uncertainty," *Statistical Science*, 19, 81–94.

Cole, D. M., Smith, S. M., and Beckmann, C. F. (2010), "Advances and pitfalls in the analysis and interpretation of resting-state fMRI data," *Frontiers in Systems Neuroscience*, 4, 1–15.

Cox, D. R. and Reid, N. (2004), "A note on pseudolikelihood constructed from marginal densities," *Biometrika*, 91, 729–737.

Cox, R. W. (1996), "AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages," *Computers and Biomedical Research*, 29, 162–173.

Cressie, N. A. (1993), *Statistics for Spatial Data*, New York: John Wiley & Sons.

Dempster, A. P. (1972), "Covariance selection," *Biometrics*, 28, 157–175.

Donoho, D. L. and Johnstone, I. M. (1994), "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, 81, 425–455.

Dryden, I. L., Koloydenko, A., and Zhou, D. (2009), "Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging," *Annals of Applied Statistics*, 3, 1102–1123.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), "Multiple hypothesis testing in microarray experiments," *Statistical Science*, 18, 71–103.

Duncan, D. B. (1955), "Multiple range and multiple F tests," *Biometrics*, 11, 1–42.

Dunn, O. J. (1961), "Multiple comparisons among means," *Journal of the American Statistical Association*, 56, 52–64.

Dunnett, C. W. (1955), "A multiple comparison procedure for comparing several treatments with a control," *Journal of the American Statistical Association*, 50, 1096–1121.

Dyckman, K. A., Camchong, J., Clementz, B. A., and McDowell, J. E. (2007), "An effect of context on saccade-related behavior and brain activity," *NeuroImage*, 36, 774–784.

Efron, B. (1998), "R. A. Fisher in the 21st century," *Statistical Science*, 13, 95–122.

— (2004), "Large-scale simultaneous hypothesis testing: The choice of a null hypothesis," *Journal of the American Statistical Association*, 99, 96–104.

— (2007), "Correlation and large-scale simultaneous significance testing," *Journal of the American Statistical Association*, 102, 93–103.

— (2008), "Microarrays, empirical Bayes, and the two-groups model," *Statistical Science*, 23, 1–22.

— (2009), "Empirical Bayes estimates for large-scale prediction problems," *Journal of the American Statistical Association*, 104, 1015–1028.

— (2010), *Large Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, New York: Cambridge University Press.

Einot, I. and Gabriel, K. R. (1975), "A study of the powers of several methods of multiple comparisons," *Journal of the American Statistical Association*, 70, 574–583.

Fahrmeir, L. and Gössl, C. (2002), "Semiparametric Bayesian models for human brain mapping," *Statistical Modelling*, 2, 235–249.

Fisher, R. A. (1935), *The Design of Experiments*, London: Oliver & Boyd, Ltd.

— (1950), *Statistical Methods for Research Workers*, London: Oliver & Boyd, Ltd, 11th ed.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., and Noll, D. C. (1995), "Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold," *Magnetic Resonance in Medicine*, 33, 636–647.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. J. (1995), "Statistical parametric maps in functional imaging: A general linear approach," *Human Brain Mapping*, 2, 189–210.

Friston, K. J. and Penny, W. (2003), "Posterior probability maps and SPMs," *NeuroImage*, 19, 1240–1249.

Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398–409.

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

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Boca Raton: Chapman & Hall/CRC, 2nd ed.

Gelman, A., Su, Y.-S., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., and Zheng, T. (2012), *ARM: Data analysis using regression and multilevel/hierarchical models*, Comprehensive R Archive Network. http://cran.r-project.org/web/packages/arm/.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Genovese, C. R. (2000), "A Bayesian time-course model for functional magnetic resonance imaging (with discussion)," *Journal of the American Statistical Association*, 95, 691–703.

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002), "Thresholding of statistical maps in functional neuroimaging using the false discovery rate," *NeuroImage*, 15, 870–878.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006), "False discovery control with p-value weighting," *Biometrika*, 93, 509–524.

George, E. I. and McCulloch, R. E. (1993), "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

Geweke, J. (1989), "Bayesian inference in econometric models using Monte Carlo integration," *Econometrica*, 57, 1317–1339.

— (1996), "Variable selection and model comparison in regression," in *Bayesian Statistics 5*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., Oxford: Oxford University Press, pp. 609–620.

Gilks, W., Richardson, S., and Spiegelhalter, D. (eds.) (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.

Gilks, W. R. (1992), "Derivative-free adaptive rejection sampling for Gibbs sampling," in *Bayesian Statistics 4*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., Oxford: Oxford University Press, pp. 641–649.

Gössl, C., Auer, D. P., and Fahrmeir, L. (2001), "Bayesian spatiotemporal inference in functional magnetic resonance imaging," *Biometrics*, 57, 554–562.

Hammersley, J. M. and Clifford, P. (1971), "Markov fields on finite graphs and lattices," Unpublished.

Hammersley, J. M. and Handscomb, D. C. (1964), *Monte Carlo Methods*, London: Methuen.

Harrison, L. M., Penny, W., Daunizeau, J., and Friston, K. J. (2008), "Diffusion-based spatial priors for functional magnetic resonance images," *NeuroImage*, 41, 408–423.

Hartvig, N. V. and Jensen, J. L. (2000), "Spatial mixture modeling of fMRI data," *Human Brain Mapping*, 11, 233–248.

Hastings, W. K. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, 57, 97–109.

Higdon, D. (1994), "Spatial applications of Markov chain Monte Carlo for Bayesian inference," Unpublished doctoral thesis, University of Washington, Department of Statistics.

Hochberg, Y. (1988), "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, 75, 800–803.

Hoel, P. G., Port, S. C., and Stone, C. J. (1972), *Introduction to Stochastic Processes*, Boston: Houghton Mifflin Company.

Holm, S. (1979), "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, 6, 65–70.

Holmes, A. P., Blair, R. C., Watson, J. D. G., and Ford, I. (1996), "Nonparametric analysis of statistic images from functional mapping experiments," *Journal of Cerebral Blood Flow and Metabolism*, 16, 7–22.

Huettel, S. A., Song, A. W., and McCarthy, G. (2009), *Functional Magnetic Resonance Imaging*, Sunderland: Sinauer Associates, Inc., 2nd ed.

Jefferys, W. H. and Berger, J. O. (1992), "Ockham's razor and Bayesian analysis," *American Scientist*, 80, 64–72.

Johnson, N. L. and Kotz, S. (1972), *Distributions in Statistics*, New York: John Wiley & Sons.

Kadane, J. B. and Lazar, N. A. (2004), "Methods and criteria for model selection," *Journal of the American Statistical Association*, 99, 279–290.

Kang, J., Johnson, T. D., Nichols, T. E., and Wager, T. D. (2011), "Meta analysis of functional neuroimaging data via Bayesian spatial point processes," *Journal of the American Statistical Association*, 106, 124–134.

Lange, N., Strother, S. C., Anderson, J. R., Nielsen, F. A., Holmes, A. P., Kolenda, T., Savoy, R., and Hansen, L. K. (1999), "Plurality and resemblance in fMRI data analysis," *NeuroImage*, 10, 282–303.

Lazar, N. A. (2008), *The Statistical Analysis of Functional MRI Data*, New York: Springer Science+Business Media, LLC.

Lazar, N. A., Luna, B., Sweeney, J. A., and Eddy, W. F. (2002), "Combining brains: A survey of methods for statistical pooling of information," *NeuroImage*, 16, 538–550.

Liang, C., Rice, J., dePater, I., Alcock, C., Axelrod, T., Wang, A., and Marshall, S. (2004), "Statistical methods for detecting stellar occulations by Kuiper belt objects: The Taiwanese-American occulation survey," *Statistical Science*, 19, 265–274.

Lindley, D. V. and Smith, A. F. M. (1972), "Bayes estimates for the linear model," *Journal of the Royal Statistical Society, Series B*, 34, 1–41.

Lindquist, M. A. (2008), "The statistical analysis of fMRI data," *Statistical Science*, 23, 439–464.

Lindsay, B. G. (1988), "Composite likelihood methods," *Contemporary Mathematics*, 80, 221–239.

Little, R. J. A. (2006), "Calibrated Bayes: A Bayes/frequentist roadmap," *American Statistician*, 60, 213–223.

— (2011), "Calibrated Bayes, for statistics in general, and missing data in particular," *Statistical Science*, 26, 162–174.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. R., and West, M. (2006), "Sparse statistical modelling in gene expression genomics," in *Bayesian Inference for Gene Expression and Proteomics*, eds. Müller, P., Do, K., and Vannucci, M., Cambridge: Cambridge University Press.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000), "WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility," *Statistics and Computing*, 10, 325–337.

Marchini, J. and Presanis, A. (2004), "Comparing methods of analyzing fMRI statistical parametric maps," *NeuroImage*, 22, 1203–1213.

Margulies, D. S., Böttger, J., Long, X., Lv, Y., Kelly, C., Schäfer, A., Goldhahn, D., Abbushi, A., Milham, M. P., Lohmann, G., and Villringer, A. (2010), "Resting developments: A review of fMRI post-processing methodologies for spontaneous brain activity," *Magnetic Resonance Materials in Physics, Biology and Medicine*, 23, 289–307.

Martin, R. and Tokdar, S. T. (2012), "A nonparametric empirical Bayes framework for large-scale multiple testing," *Biostatistics*, 13, 427–439.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, New York: John Wiley and Sons.

McNamee, R. L. and Lazar, N. A. (2004), "Assessing the sensitivity of fMRI group maps," *NeuroImage*, 22, 920–931.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of state calculations by fast computing machines," *Journal of Chemical Physics*, 21, 1087–1091.

Mitchell, T. J. and Beauchamp, J. J. (1988), "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, 83, 1023–1032.

Morris, J. S., Baladandayuthapani, V., Herrick, R. C., Sanna, P., and Gutstein, H. (2011), "Automated analysis of quantitative image data using isomorphic functional mixed models with application to proteomics data," *Annals of Applied Statistics*, 5, 894–923.

Muralidharan, O. (2010), "An empirical Bayes mixture method for effect size and false discovery rate estimation," *Annals of Applied Statistics*, 4, 422–438.

Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., and Bandettini, P. A. (2009), "The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced?" *NeuroImage*, 44, 893–905.

Murphy, K. and Mahdaviani, M. (2005), *MATBUGS: A MATLAB interface to WinBUGS*, https://code.google.com/p/matbugs/.

Neal, R. M. (1997), "Markov chain Monte Carlo methods based on 'slicing' the density function," Technical Report No. 9722, University of Toronto, Department of Statistics.

Nichols, T. and Hayasaka, S. (2003), "Controlling the familywise error rate in functional neuroimaging: A comparative review," *Statistical Methods in Medical Research*, 12, 419–446.

Nichols, T. E. and Holmes, A. P. (2001), "Nonparametric permutation tests for functional neuroimaging: A primer with examples," *Human Brain Mapping*, 15, 1–25.

Nott, D. J. and Ryden, T. (1999), "Pairwise likelihood methods for inference in image models," *Biometrika*, 86, 661–676.

Oehlert, G. W. (2000), *A First Course in Design and Analysis of Experiments*, New York: W. H. Freeman and Company.

O'Hagan, A. (1995), "Fractional Bayes factors for model comparisons (with discussion)," *Journal of the Royal Statistical Society, Series B*, 57, 99–138.

Penny, W., Kiebel, S., and Friston, K. (2003), "Variational Bayesian inference for fMRI time series," *NeuroImage*, 19, 727–741.

Penny, W. D., Trujillo-Barreto, N. J., and Friston, K. J. (2005), "Bayesian fMRI time series analysis with spatial priors," *NeuroImage*, 24, 350–362.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012), "Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion," *NeuroImage*, 59, 2142–2154.

Qian, P. Z. G. and Wu, C. F. J. (2008), "Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments," *Techometrics*, 50, 192–204.

Qiu, X., Klebanov, L., and Yakovlev, A. (2005), "Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes," *Statistical Applications in Genetics and Molecular Biology*, 4, article 34.

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, New York: John Wiley & Sons, 2nd ed.

Rath, T. M., Carrerras, M., and Sebastiani, P. (2003), "Automated detection of influenza epidemics with hidden Markov models," in *Advances in Intelligent Data Analysis V*, eds. Berthold, M. R., Lenz, H.-J., Bradley, E., Kruse, R., and Borgelt, C., Berlin: Springer-Verlag Berlin Heidelberg.

Resnick, S. I. (1992), *Adventures in Stochastic Processes*, Boston: Birkhäuser.

Ripley, B. D. (1987), *Stochastic Simulation*, New York: John Wiley & Sons.

Robert, C. P. (2007), *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, New York: Springer Science+Business Media, LLC, 2nd ed.

Ross, S. M. (2007), *Introduction to Probability Models*, Oxford: Elsevier, Inc., 9th ed.

Rubin, D. B. (1984), "Bayesianly justifiable and relevant frequency calculations for the applied statistician," *Annals of Statistics*, 12, 1151–1172.

Ryan, T. A. (1960), "Significance tests for multiple comparisons of proportions, variances, and other statistics," *Psychological Bulletin*, 57, 318–328.

Saad, Z. S., Gotts, S. J., Murphy, K., Chen, G., Jo, H. J., Martin, A., and Cox, R. W. (2012), "Trouble at rest: How correlation patterns and group differences become distorted after global signal regression," *Brain Connectivity*, 2, 25–32.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer-Verlag.

Schabenberger, O. and Gotway, C. A. (2005), *Statistical Methods for Spatial Data Analysis*, Boca Raton: Chapman & Hall/CRC.

Scheffé, H. (1953), "A method for judging all contrasts in the analysis of variance," *Biometrika*, 40, 87–104.

Scott, J. G. and Berger, J. O. (2006), "An exploration of aspects of Bayesian multiple testing," *Journal of Statistical Planning and Inference*, 136, 2144–2162.

— (2010), "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem," *Annals of Statistics*, 38, 2587–2619.

Simes, R. J. (1986), "An improved Bonferroni procedure for multiple tests of significance," *Biometrika*, 73, 751–754.

Smith, M. and Fahrmeir, L. (2007), "Spatial Bayesian variable selection with application to functional magnetic resonance imaging," *Journal of the American Statistical Association*, 102, 417–431.

Smith, M. and Kohn, R. (1996), "Nonparametric regression using Bayesian variable selection," *Journal of Econometrics*, 75, 317–343.

Stein, C. M. (1981), "Estimation of the mean of a multivariate normal distribution," *Annals of Statistics*, 9, 1135–1151.

Storey, J. D. (2002), "A direct approach to false discovery rates," *Journal of the Royal Statistical Society, Series B*, 64, 479–498.

— (2003), "The positive false discovery rate: A Bayesian interpretation of the q-value," *Annals of Statistics*, 31, 2013–2035.

Strang, G. (1988), *Linear Algebra and Its Applications*, United States: Thomson Learning, Inc., 3rd ed.

Sun, W. and Cai, T. T. (2009), "Large-scale multiple testing under dependence," *Journal of the Royal Statistical Society, Series B*, 71, 393–424.

Talaraich, J. and Tournoux, P. (1988), *Co-Planar Stereotaxic Atlas of the Human Brain*, New York: Thieme.

Tukey, J. W. (1952), "Allowances for various types of error rates," Unpublished IMS address.

Waller, R. A. and Duncan, D. B. (1969), "A Bayes rule for the symmetric multiple comparisons problem," *Journal of the American Statistical Association*, 64, 1484–1503.

Welsch, R. E. (1977), "Stepwise multiple comparison procedures," *Journal of the American Statistical Association*, 72, 566–575.

West, M. (2003), "Bayesian factor regression models in the 'large p, small n' paradigm," in *Bayesian Statistics 7*, eds. Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., Oxford: Oxford University Press, pp. 723–732.

Westfall, P. H., Johnson, W. O., and Utts, J. M. (1997), "A Bayesian perspective on the Bonferroni adjustment," *Biometrika*, 84, 419–427.

Woolrich, M. W. (2012), "Bayesian inference in fMRI," *NeuroImage*, 62, 801–810.

Worsley, K. J. (1994), "Local maxima and the expected Euler characteristic of excursion sets of χ^2, F and t fields," *Advances in Applied Probability*, 26, 13–42.

— (2003), "Detecting activation in fMRI data," *Statistical Methods in Medical Research*, 12, 401–418.

Worsley, K. J., Evans, A. C., Marrett, S., and Neelin, P. (1992), "A three-dimensional statistical analysis for CBF activation studies in human brain," *Journal of Cerebral Blood Flow and Metabolism*, 12, 900–918.

Xu, L., Johnson, T. D., Nichols, T. E., and Nee, D. E. (2009), "Modeling inter-subject variability in fMRI activation location: A Bayesian hierarchical spatial model," *Biometrics*, 65, 1041–1051.

Appendix A

Derivation of a Gibbs Sampling Algorithm for the Scott-Berger Model

A.1 The Model

The multiple testing model of Scott and Berger (2006) (see Section 2.2) supposes that the test statistics are distributed as $y_j \sim N(\theta_j, \sigma^2)$, $j = 1, \ldots, J$. The competing hypotheses of $\theta_j = 0$ versus $\theta_j \neq 0$ at each location may be represented by parameterizing the means as $\theta_j = \gamma_j \mu_j$, where $\gamma_j \sim \text{Bernoulli}(1-p)$, $j = 1, \ldots, J$, are latent indicators of activation. In implementing their model, we follow their suggestion and take p to be uniform on the unit interval. The model may be summarized as follows:

- $y_j \overset{ind}{\sim} N(\gamma_j \mu_j, \sigma^2), \quad j = 1, \ldots, J$
- $\gamma_j \overset{ind}{\sim} \text{Bernoulli}(1-p), \quad j = 1, \ldots, J$
- $\mu_j \overset{ind}{\sim} N(0, \tau^2), \quad j = 1, \ldots, J$
- $\pi_{(\sigma^2,\tau^2)}(\sigma^2, \tau^2) = (\tau^2 + \sigma^2)^{-2}, \quad \sigma^2, \tau^2 > 0$
- $p \sim U(0,1)$

Averaging over μ_j for $\gamma_j = 0$ and $\gamma_j = 1$, we find the marginal distributions of the test statistics to be $y_j \mid \gamma_j = 1, \sigma^2, \tau^2 \sim N(0, \sigma^2 + \tau^2)$ and $y_j \mid \gamma_j = 0, \sigma^2 \sim N(0, \sigma^2)$ (Berger, 1985). Since the y_j are independent, the joint likelihood of $\mathbf{y}$ can be written

$$
\begin{aligned}
f(\mathbf{y} \mid \boldsymbol{\gamma}, \sigma^2, \tau^2) &= \prod_{j=1}^{J} f(y_j \mid \gamma_j, \sigma^2, \tau^2) \\
&\propto \left(\prod_{j:\gamma_j=1} (\sigma^2+\tau^2)^{-1/2} \exp\left(-\frac{y_j^2}{2(\sigma^2+\tau^2)} \right) \right) \\
&\quad \times \left(\prod_{j:\gamma_j=0} (\sigma^2)^{-1/2} \exp\left(-\frac{y_j^2}{2\sigma^2} \right) \right) \\
&= (\sigma^2+\tau^2)^{-\frac{\sum_j \gamma_j}{2}} \exp\left(-\frac{\sum_j \gamma_j y_j^2}{2(\sigma^2+\tau^2)} \right) \\
&\quad \times (\sigma^2)^{-\frac{(J-\sum_j \gamma_j)}{2}} \exp\left(-\frac{\sum_j (1-\gamma_j) y_j^2}{2\sigma^2} \right)
\end{aligned}
$$

The joint posterior distribution (up to a normalizing constant) can then be calculated as

$$
\begin{aligned}
\pi(\boldsymbol{\gamma}, \sigma^2, \tau^2, p \mid \mathbf{y}) &\propto f(\mathbf{y} \mid \boldsymbol{\gamma}, \sigma^2, \tau^2) \pi_{\boldsymbol{\gamma}|p}(\boldsymbol{\gamma} \mid p) \pi_p(p) \pi_{(\sigma^2,\tau^2)}(\sigma^2, \tau^2) \\
&= (\sigma^2+\tau^2)^{-\frac{\sum_j \gamma_j}{2}} \exp\left(-\frac{\sum_j \gamma_j y_j^2}{2(\sigma^2+\tau^2)} \right) \\
&\quad \times (\sigma^2)^{-\frac{(J-\sum_j \gamma_j)}{2}} \exp\left(-\frac{\sum_j (1-\gamma_j) y_j^2}{2\sigma^2} \right) \\
&\quad \times (1-p)^{\sum_j \gamma_j} p^{J-\sum_j \gamma_j} (\sigma^2+\tau^2)^{-2} \pi_p(p),
\end{aligned}
$$

where $\pi_p(p) = I(0 < p < 1)$. Gibbs sampling requires specification of the distribution of each parameter conditioned on all of the others. With the target (posterior) distribution specified, these can be derived.

A.2 Full Conditional Distributions

Let $x = \sigma^2$ and $w = \sigma^2 + \tau^2$ to simplify notation. For two-component mixtures of the form $f(y_j) = pf(y_j \mid \gamma_j = 0) + (1-p)f(y_j \mid \gamma_j = 1)$, the posterior for the indicators follows from the definition of conditional probability:

$$\begin{aligned} P(\gamma_j = 1 \mid \mathbf{y}) &= \frac{(1-p)f(y_j \mid \gamma_j = 1)}{pf(y_j \mid \gamma_j = 0) + (1-p)f(y_j \mid \gamma_j = 1)} \\ &= 1 - P(\gamma_j = 0 \mid \mathbf{y}). \end{aligned}$$

Thus, for $\boldsymbol{\gamma}$, the component-wise full conditional distributions are

$$\begin{aligned} P(\gamma_j = 1 \mid x, w, p, \mathbf{y}) &= \frac{(1-p)w^{-\frac{1}{2}} \exp\left(-\frac{y_j^2}{2w}\right)}{(1-p)w^{-\frac{1}{2}} \exp\left(-\frac{y_j^2}{2w}\right) + px^{-\frac{1}{2}} \exp\left(-\frac{y_j^2}{2x}\right)} \\ &=: p_j^* \\ &= 1 - P(\gamma_j = 0 \mid x, w, p, \mathbf{y}), \end{aligned}$$

for $j = 1, \ldots, J$. The inclusion probability p depends on the data only through $\boldsymbol{\gamma}$, which is a vector of Bernoulli random variables. The Beta-Binomial model is a conjugate pair, so the conditional posterior is

$$p \mid \boldsymbol{\gamma}, \sigma^2, \tau^2, \mathbf{y} \sim \text{Beta}\left(J - \sum_j \gamma_j + 1, \sum_j \gamma_j + 1\right).$$

For w, we note that

$$\begin{aligned} \pi(\tau^2 \mid p, \boldsymbol{\gamma}, x, \mathbf{y}) &\propto \pi(\boldsymbol{\gamma}, x, w, p \mid \mathbf{y}) \\ &\propto f(\mathbf{y} \mid \boldsymbol{\gamma}, \sigma^2, \tau^2)\pi_{\boldsymbol{\gamma}|p}(\boldsymbol{\gamma} \mid p)\pi_p(p)\pi_{(\sigma^2,\tau^2)}(\sigma^2, \tau^2) \end{aligned}$$

$$
\begin{aligned}
&= w^{-\frac{\sum_j \gamma_j}{2}} \exp\left(-\frac{\sum_j \gamma_j y_j^2}{2w}\right) \\
&\quad \times (\sigma^2)^{-\frac{(J-\sum_j \gamma_j)}{2}} \exp\left(-\frac{\sum_j (1-\gamma_j) y_j^2}{2\sigma^2}\right) \\
&\quad \times (1-p)^{\sum_j \gamma_j} p^{J-\sum_j \gamma_j} w^{-2} \pi_p(p) \\
&\propto (w^{2+\frac{1}{2}\sum_j \gamma_j})^{-1} \exp\left(-\frac{\sum \gamma_j y_j^2}{2w}\right),
\end{aligned}
$$

which is the density of an Inverse Gamma distribution. But $w = \sigma^2 + \tau^2 \Rightarrow w > \sigma^2 = x$, so it is actually a truncated distribution. That is,

$$
w \mid p, \boldsymbol{\gamma}, x, \mathbf{y} \sim \text{IG}\left(\frac{1}{2}\sum_j \gamma_j + 1, \frac{2}{\sum_j \gamma_j y_j^2}\right) I(w > x).
$$

For x, we have four cases to consider. In each case, the density can be obtained by examining the joint posterior density as we did for w.

1. $\sum_j \gamma_j = J$:
   $\pi(x \mid \sum \gamma_j = J, w, p, \mathbf{y}) \propto I(0 < x < w)$

2. $\sum_j \gamma_j = J - 1$:
   $\pi(x \mid \sum \gamma_j = J-1, w, p, \mathbf{y}) \propto x^{-\frac{1}{2}} \exp\left(-\frac{\sum_j (1-\gamma_j) y_j^2}{2x}\right) I(0 < x < w)$

3. $\sum_j \gamma_j = J - 2$:
   $\pi(x \mid \sum \gamma_j = J-2, w, p, \mathbf{y}) \propto x^{-1} \exp\left(-\frac{\sum_j (1-\gamma_j) y_j^2}{2x}\right) I(0 < x < w)$

4. $\sum_j \gamma_j \leq J - 3$:
   $\pi(x \mid \gamma, w, p, \mathbf{y}) \propto x^{-\frac{J-\sum_j \gamma_j}{2}} \exp\left(-\frac{\sum_j (1-\gamma_j) y_j^2}{2x}\right) I(0 < x < w)$

The Gibbs sampling algorithm iterates through each of these distributions, successively drawing from them until convergence. The algorithm is outlined below.

A.3 The Gibbs Sampler Algorithm

The Gibbs sampling algorithm is as follows:

1. Calculate initial estimates of all the parameters except $\boldsymbol{\gamma}$. The indicators are drawn as the first step of the sampler.

2. For $j = 1, \ldots, J$, draw γ_j from Bernoulli(p_j^*), where p_j^* is given above.

3. Draw $p \sim \text{Beta}(J - \sum_j \gamma_j + 1, \sum_j \gamma_j + 1)$

4. For τ^2, there are two cases:

   I: $\sum_j \gamma_j = 0$:

   (i) Draw $U \sim U(0, 1)$

   (ii) Set $\tau^2 = \frac{1-U}{U}\sigma^2$

   II: $\sum_j \gamma_j > 0$:

   (i) Draw $R \sim \text{Ga}\left(\frac{1}{2}\sum_j \gamma_j + 1, \frac{2}{\sum_j \gamma_j y_j^2}\right)$

   (ii) If $R < \frac{1}{\sigma^2}$, set $\tau^2 = \frac{1}{R} - \sigma^2$, else return to (i).

   Then set $W = \tau^2 + \sigma^2$.

5. There are four cases to consider for x.

   I: $\sum_j \gamma_j = J$:

   Draw $X \sim U(0, w)$

   II: $\sum_j \gamma_j = J - 1$:

   Note that if $U \sim U(0, 1)$, then $T = wU^2$ has density $f_{_T}(t) = (2\sqrt{tw})^{-1} I(t < w)$. So we can take $f_{_T}(x) = (2\sqrt{xw})^{-1} I(x < w)$ as a generating density for a rejection sampler. For $x < w$,

$$\begin{aligned} f(x)/g(x) &= 2\sqrt{w} \times \exp\left(-\frac{\sum_j (1-\gamma_j) y_j^2}{2x}\right) \\ &\leq 2\sqrt{w} \times \exp\left(-\frac{\sum_j (1-\gamma_j) y_j^2}{2w}\right) \\ &=: \Upsilon \end{aligned}$$

Therefore,

(i) Draw $x_0 = wU^2, U \sim U(0,1)$

(ii) Draw $Q \sim U(0,1)$. If $Q < f(x_0)/(\Upsilon g(x_0))$, accept x_0, else return to (i).

III: $\sum_j \gamma_j = J - 2$:

Take $g(x) = (2\sqrt{xw})^{-1} I(x < w)$. Then

$$f(x)/g(x) = 2\sqrt{\frac{w}{x}} \times \exp\left(-\frac{\sum_j (1-\gamma_j) y_j^2}{2x}\right) I(x < w),$$

so that

$$\sup_x f(x)/g(x) = 2e^{-1/2} \sqrt{\frac{2}{\sum_j (1-\gamma_j) y_j^2}} =: f(x^*)/g(x^*).$$

Thus, with $\Upsilon = \max\{\frac{f(x^*)}{g(x^*)}, \frac{f(w)}{g(w)}\}$,

(i) Draw $x_0 = wU^2, U \sim U(0,1)$.

(ii) Draw $Q \sim U(0,1)$. If $Q < f(x_0)/(\Upsilon g(x_0))$, accept x_0, else return to (i).

IV: $\sum_j \gamma_j \leq J - 3$:

(i) Draw $T \sim \text{Ga}\left(\frac{J - \sum_j \gamma_j}{2} - 1, \frac{2}{\sum_j (1-\gamma_j) y_j^2}\right)$.

(ii) If $1/T < w$, set $x = 1/T$, else return to (i).

6. Repeat steps (1) - (5) to convergence. Given samples of the other parameters, draws for $\mu_j, j = 1, \ldots, J$, can be made from $\mu_j \mid \boldsymbol{\gamma}, V, \sigma^2, p, \mathbf{y}$, taking advantage of normal-normal conjugacy.