Advances in Empirical Bayes Modeling and Bayesian Computation

Advances in Empirical Bayes Modeling and Bayesian Computation

A dissertation presented

by

Nathan Mathes Stein

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

May 2013

Dissertation Advisor: Professor Xiao-Li Meng Nathan Mathes Stein

Advances in Empirical Bayes Modeling and Bayesian Computation

Abstract

Chapter 1 of this thesis focuses on accelerating perfect sampling algorithms for a Bayesian hierarchical model. A discrete data augmentation scheme together with two different parameterizations yields two Gibbs samplers for sampling from the posterior distribution of the hyperparameters of the Dirichlet-multinomial hierarchical model under a default prior distribution. The finite-state space nature of this data augmentation permits us to construct two perfect samplers using bounding chains that take advantage of monotonicity and anti-monotonicity in the target posterior distribution, but both are impractically slow. We demonstrate however that a composite algorithm that strategically alternates between the two samplers' updates can be substantially faster than either individually. We theoretically bound the expected time until coalescence for the composite algorithm, and show via simulation that the theoretical bounds can be close to actual performance.

Chapters 2 and 3 introduce a strategy for constructing scientifically sensible priors in complex models. We call these priors *catalytic priors* to suggest that adding such prior information catalyzes our ability to use richer, more realistic models. Because they depend on observed data, catalytic priors are a tool for empirical Bayes modeling. The overall perspective is data-driven: catalytic priors have a pseudo-data interpretation, and the building blocks are alternative plausible models for observations, yielding behavior similar to hierarchical models but with a conceptual shift away from distributional assumptions on parameters. The posterior under a catalytic prior can be viewed as an optimal approximation to a target measure, subject to a constraint on the posterior distribution's predictive implications. In Chapter 3, we apply catalytic priors to several familiar models and investigate the performance of the resulting posterior distributions. We also illustrate the application of catalytic priors in a preliminary analysis of the effectiveness of a job training program, which is complicated by the need to account for noncompliance, partially defined outcomes, and missing outcome data.

Contents

To my parents and my brother

AUTHOR LIST

Xiao-Li Meng contributed to Chapter 1.

S. C. Samuel Kou and Donald B. Rubin contributed to Chapters 2 and 3.

Acknowledgments

I would like to thank my advisor Xiao-Li Meng for his invaluable guidance and support. His wisdom, energy, and empathy are a constant inspiration. His sharp questions have made me a better researcher and teacher, and I could not be more grateful to count him as a mentor and a friend.

I am grateful to my committee members Samuel Kou and Donald Rubin, both for their insights on research and for continuously pushing me to think more deeply about how to communicate new ideas. Their encouragement and advice have been immensely important to me.

I want to thank my astrostatistics and astronomy collaborators, David van Dyk, Ted von Hippel, Bill Jefferys, and Vinay Kashyap. I also want to take a moment to remember Alanna Connors, with whom I am enormously grateful that I had the opportunity to work.

I have been incredibly fortunate in the friends I have made while in this program. My time in the Statistics Department has been enriched by the opportunity to talk, study, and laugh with Xiaojin Xu, Sam Wong, Valeria Espinosa, and Alex Blocker, among many, many others. I am especially grateful for the friendship of Jen Sinnott, Ravi Goyal, and Stacey Alexeeff, and for our many hours in libraries and coffee shops around Cambridge.

PREFACE

This thesis focuses on methodology and computation for Bayesian and empirical Bayes models. The major theme running through this work is the benefits that can come from combining models and methods with complementary strengths. In perfect sampling, combining bounding chain algorithms with complementary strengths can yield a composite algorithm that outperforms either individual algorithm. Intuitively, some bounding chains may be fast "marginally" and others may be fast "conditionally," but for a valid and practical perfect sampler, we want an algorithm that is fast "jointly." In the Dirichlet-multinomial model, we demonstrate this idea by constructing a discrete data augmentation that leads to one bounding chain algorithm that is fast on the augmented-data margin, and another that is fast to coalesce on the parameter, once we have already coalesced on the augmented data. Individually, both algorithms are impractically slow to return a draw from the correct stationary distribution, but by strategically alternating between them, the time until coalescence can be reduced dramatically.

Catalytic priors work by trading off between the strengths of a more complicated and a simpler model. The more complicated model better reflects the complexity of the real world, but may result in noisy, unreliable estimates. The simpler model does not capture all of the relevant features of the phenomenon we are investigating, but it can be reliably estimated. The intuition behind catalytic priors is to impute a small amount of data under the simpler model, but analyze them alongside the observed data under the complicated model. These imputed data shrink the estimates under the complicated model away from unreasonable regions of parameter space, resulting in inferences that are more scientifically sensible—both more realistic than estimates under only the simpler model, and more reliable than estimates under only the complicated model.

1

Practical perfect sampling using composite bounding chains

1.1 THE DIRICHLET-MULTINOMIAL MODEL

The multinomial model and its conjugate Dirichlet prior distribution are common building blocks of more elaborate models for categorical data, with applications from topic modeling (Blei *et al.*, 2003) to biology (Holmes *et al.*, 2012). Despite the model's popularity, there is room to improve algorithms for Bayesian inference. Although fast Newton-Raphson itera-

tions can find maximum likelihood estimates, as demonstrated in the 2003 technical report by T. P. Minka, 'Estimating a Dirichlet distribution,' sampling from the posterior distribution of the parameters in a Bayesian setting is more challenging, especially in high dimensions.

In this paper we present a data augmentation scheme that yields a practical Gibbs sampler and facilitates perfect sampling algorithms based on composite bounding chains. The central idea of the composite perfect sampling algorithm is to strategically combine two or more bounding chains, such that the composite chain has a much faster coalescence time than the individual samplers. Our main theoretical result is a bound on the expected running time of the composite algorithm, and we prove it using somewhat general language about bounding chains, suggesting that similar speed gains may be possible for other bounding chain algorithms. The dramatic increase in speed we achieve is a small but encouraging step on the road toward perfect samplers that can be routinely used in practice for Bayesian computation.

Let y be an $N \times k$ matrix of observed counts with ith row $y_i^\top = (y_{i1}, \ldots, y_{ik})$, and let μ be an $N \times k$ matrix with ith row $\mu_i^\top = (\mu_{i1}, \ldots, \mu_{ik})$, a k-dimensional probability vector with $\mu_{ij} \geq 0$, $\sum_{j=1}^k \mu_{ij} = 1$. Conditioned on μ, the vectors y_i are independent multinomial random variables with probabilities μ_i and fixed sample sizes n_i:

$$p(y_i \mid \mu) = \frac{n_i!}{\prod_{j=1}^k y_{ij}!} \prod_{j=1}^k \mu_{ij}^{y_{ij}}, \quad i = 1, \ldots, N.$$

The probabilities μ_i in turn are independent draws from a Dirichlet($\alpha_1, \ldots, \alpha_k$) distribution:

$$p(\mu_i \mid \alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k \mu_{ij}^{\alpha_j - 1}, \quad i = 1, \ldots, N.$$

Integrating over μ, the posterior of the hyperparameters under a prior $\pi(\alpha_1, \ldots, \alpha_k)$ is

$$p(\alpha_1, \ldots, \alpha_k \mid y) \propto \pi(\alpha_1, \ldots, \alpha_k) \prod_{i=1}^{N} \left\{ \frac{\Gamma(\sum_{j=1}^{k} \alpha_j)}{\Gamma(\sum_{j=1}^{k} \alpha_j + n_i)} \prod_{j=1}^{k} \frac{\Gamma(\alpha_j + y_{ij})}{\Gamma(\alpha_j)} \right\}. \tag{1.1}$$

Constructing a practical perfect sampler for (1.1) is the main subject of this paper.

1.2 A DISCRETE DATA AUGMENTATION STRATEGY AND TWO GIBBS SAMPLERS

For reasons that will soon be clear, we shall use both the parameterization $\alpha = (\alpha_1, \ldots, \alpha_k)$ and the parameterization $\theta = (\omega, \lambda)$, where $\omega = \sum_{j=1}^{k} \alpha_j$ is a concentration parameter and $\lambda = (\lambda_1, \ldots, \lambda_k)$ is a mean vector with $\lambda_j = \alpha_j/\omega$. Throughout this paper, we assume that the prior distribution on λ and ω factors into independent priors

$$\lambda \sim \text{Dirichlet}(\delta_1, \ldots, \delta_k), \quad \omega \sim \pi_0. \tag{1.2}$$

Conditioning on the observed data y and the parameter θ, we construct our data augmentation scheme. If $y_{ij} > 0$, we define conditionally independent Bernoulli random variables $v_{ij} = (v_{ij,1}, \ldots, v_{ij,y_{ij}})$ with

$$\Pr(v_{ij,m} = 1 \mid y, \theta) = \frac{\omega\lambda_j}{\omega\lambda_j + m - 1}, \quad m = 1, \ldots, y_{ij};$$

and if $y_{ij} = 0$, then v_{ij} is defined as empty.

This data augmentation strategy has a nice interpretation in terms of a double-replacement sampling scheme. The distribution of y given θ, integrating over μ, is equivalent to supposing that there are N urns that initially have α_j balls of color j for $j = 1, \ldots, k$, and we draw balls independently from each urn following a double-replacement scheme. That is, when we draw a ball of color j, we replace it and add another ball of the same color before sampling the next

ball. The observation y_{ij} is the number of balls sampled from urn i with color j, and the data augmentation $v_{ij,m}$ indicates for the mth ($m = 1, \ldots, y_{ij}$) sampled ball whether it was drawn from the original pool, in which case $v_{ij,m} = 1$, or from the balls that were added to the urns through the double-replacement scheme, in which case $v_{ij,m} = 0$. Thus, it makes sense that if $y_{ij} = 0$, then v_{ij} is empty because there is nothing to indicate, and if $y_{ij} > 0$, then $v_{ij,1} = 1$ because the first sampled ball must have come from the original pool.

Letting $v = \{v_{ij}\}_{i,j}$, we then obtain the complete-data likelihood $p(y, v \mid \theta)$ as the product $p(v \mid y, \theta)p(y \mid \theta)$, clearly preserving the margin of interest $p(y \mid \theta)$. Since $\Gamma(x + 1) = x\Gamma(x)$,

$$p(y, v \mid \theta) = \prod_{i=1}^{N} \frac{\Gamma(\omega)}{\Gamma(\omega + n_i)} \prod_{j=1}^{k} \left\{ \prod_{m=1}^{y_{ij}} (\omega\lambda_j)^{v_{ij,m}} (m-1)^{1-v_{ij,m}} \right\}, \tag{1.3}$$

where we take the term in braces equal to 1 if $y_{ij} = 0$. If the mean and concentration parameters are independent in their prior distributions, which we will assume throughout, then they will also be independent in their complete-data posterior distribution, since terms involving ω and λ factor in (1.3). More intuition about (1.3) appears at the end of this section.

This data augmentation scheme yields a Gibbs sampler that is easy to implement. Denoting the sufficient statistics of the augmented data $z_j = \sum_{i=1}^{N} \sum_{m=1}^{y_{ij}} v_{ij,m}$, where we take $\sum_{m=1}^{y_{ij}} v_{ij,m} = 0$ if $y_{ij} = 0$, we transition from (z, θ) to (z', θ') by alternating between updating the parameters given the complete data by drawing θ' from

$$p(\theta' \mid z, y) \propto \pi(\omega', \lambda')\, \omega'^{\sum_{j=1}^{k} z_j} \left\{ \prod_{i=1}^{N} \frac{\Gamma(\omega')}{\Gamma(\omega' + n_i)} \right\} \left\{ \prod_{j=1}^{k} \lambda'^{z_j}_{j} \right\} \tag{1.4}$$

and updating the missing data given the parameters by drawing independently, for $j = 1, \ldots, k$,

$$z'_j \mid \theta', y \sim \sum_{i=1}^{N} \sum_{m=1}^{y_{ij}} \text{Bernoulli}\left(\frac{\omega'\lambda'_j}{\omega'\lambda'_j + m - 1} \right). \tag{1.5}$$

We will call equations (1.4)–(1.5) the standard Gibbs sampler. The discreteness of $z =$

$(z_1, \ldots, z_k)$ is a major advantage of this algorithm, as it enables perfect samplers to coalesce in finite time. Another benefit of this algorithm is that using a Dirichlet prior distribution on λ leads to a conjugate Dirichlet update for the multivariate λ, and the only step that requires special attention is sampling the univariate ω, for which many standard methods are available, including grid-based and rejection methods. Additionally, the density $p(\omega \mid z, y)$ arises in the context of Dirichlet process mixture models, and Escobar and West (1995) suggest a Gibbs sampler for this distribution.

Following Craiu and Meng (2011), we say that z' given θ' and y in (1.5) has a nonhomogeneous binomial distribution. In general, a nonhomogeneous binomial random variable $x \sim \text{NhBin}(N; (p_1, \ldots, p_N))$ can be represented as the sum of N independent Bernoulli random variables $b_i \sim \text{Bernoulli}(p_i)$, each with its own success probability p_i. The nonhomogeneous binomial distribution can be easily generalized to the nonhomogeneous multinomial distribution, which is familiar from the traditional data augmentation approach to fitting finite mixture models. To illustrate, suppose data $y = (y_1, \ldots, y_N)$ are independent and identically distributed according to the mixture model

$$p(y_i \mid \eta) = \sum_{j=1}^{k} \eta_j f_j(y_i), \tag{1.6}$$

where for simplicity we can suppose that the densities $f_j(\cdot)$ are fully known and the mixture weights $\eta = (\eta_1, \ldots, \eta_k)$ are of interest. The usual model-fitting approach augments y with a nonhomogeneous multinomial variable $z = (z_1, \ldots, z_N)$ such that

$$\Pr(z_i = j \mid y, \eta) = \frac{\eta_j f_j(y_i)}{\eta_1 f_1(y_i) + \cdots + \eta_k f_k(y_i)}, \quad j = 1, \ldots, k.$$

The complete-data likelihood is therefore $p(y, z \mid \eta) = \prod_{i=1}^{N} \prod_{j=1}^{k} \{\eta_j f_j(y_i)\}^{1(z_i=j)}$, which as a product can lead to straightforward Gibbs sampling and Expectation-Maximization al-

gorithms that are much easier to work with than the sums in (1.6). See Hobert *et al.* (1999), Murdoch and Meng (2001), Casella *et al.* (2002), and Mukhopadhyay and Bhattacharya (2012) for examples in the context of perfect sampling. The nonhomogeneous binomial augmentation in the Dirichlet-multinomial model plays the same role as in finite mixture models: it turns sums into products. In the Dirichlet-multinomial model, we can rewrite (1.1) as

$$p(\omega, \lambda \mid y) \propto \left\{ \pi(\omega, \lambda) \prod_{i=1}^{N} \frac{\Gamma(\omega)}{\Gamma(\omega + n_i)} \right\} \left\{ \prod_{i=1}^{N} \prod_{j=1}^{k} \prod_{m=1}^{y_{ij}} (\omega \lambda_j + m - 1) \right\}. \tag{1.7}$$

The z augmentation turns the sums $\omega\lambda_j + (m-1)$ on the right hand side of (1.7) into products as shown in (1.3), which are much more convenient for sampling.

Before developing our perfect samplers, it is helpful to introduce another Gibbs sampler based on this same data augmentation strategy. However, instead of using the θ parameterization, we use the original α parameterization and alternate between draws from

$$p(\alpha'_j \mid \alpha_{[-j]}, z, y), \quad p(z'_j \mid \alpha'_j, \alpha_{[-j]}, z_{[-j]}, y) \tag{1.8}$$

for $j = 1, \ldots, k$, where $\alpha_{[-j]} = (\alpha_1, \ldots, \alpha_{j-1}, \alpha_{j+1}, \ldots, \alpha_k)$ and $z_{[-j]}$ is similarly defined. We will see in Section 1.3.2 that while this algorithm sacrifices the factorization between the multivariate λ density and the univariate ω density, it offers convenient monotonicity properties that help in designing bounding chains.

1.3 Perfect sampling using bounding chains

1.3.1 Bounding chains

Propp and Wilson (1996) introduced coupling from the past to obtain exact draws from the stationary distribution of a Markov chain in finite time. The key idea is to run coupled Markov

chains from the past to the present with the same transition probabilities but starting from different states. The construction guarantees that if the chains couple by time 0, then the value at time 0 is an exact draw from the stationary distribution of the chains. Propp and Wilson (1996) recognized that the computation is much simpler when the updates of the chain are monotone with respect to some partial order on the state space.

In general, it can be difficult to construct monotone chains. Bounding chains were therefore introduced in Huber (1998) and Häggström and Nelander (1999) and developed in Huber (2004), among others. They provide a solution for perfect sampling without requiring the monotonicity used by Propp and Wilson (1996). A bounding chain for a Markov chain x_t on a state space Ω is a set-valued Markov chain X_t on the set of all subsets of Ω, such that the current state $x_t \in X_t$, for every starting value that could have been used for the x_t chain; see Huber (2004) for discussion and a slightly more general definition. When bounding chains are used in the context of coupling from the past, coalescence is detected and the algorithm returns a draw from the stationary distribution when X_0 is a singleton.

Specifically, if the original x_t chain can be written as a stochastic recursive sequence

$$x_{t+1} = \phi(x_t, u_t),$$

where ϕ is a deterministic function and u_t is a random input, then the set-valued chain

$$X_{t+1} = \Phi(X_t, u_t) \tag{1.9}$$

is a valid bounding chain if $\phi(x_t, u_t) \in \Phi(X_t, u_t)$ for every $x_t \in X_t$, as illustrated in Figure 1.1(a). Coupling from the past for this bounding chain proceeds as follows: first, set $T = T_0$ for some fixed T_0 such as 1, and $X_{-T} = \Omega$. Then, run the chain forward from X_{-T} by repeatedly calling (1.9) until we obtain X_0, as illustrated in Figure 1.1(b). If X_0 is a singleton $\{x\}$, then x is an exact draw from the stationary distribution of the chain defined by ϕ. Other-

wise, set $T_{\text{old}} = T$, set $T = T_{\text{new}} > T_{\text{old}}$, where typically $T_{\text{new}} = 2T_{\text{old}}$, and repeat the procedure, drawing new random inputs $u_{-T}, \ldots, u_{-T_{\text{old}}-1}$ and reusing the sequence $u_{-T_{\text{old}}}, \ldots, u_{-1}$.

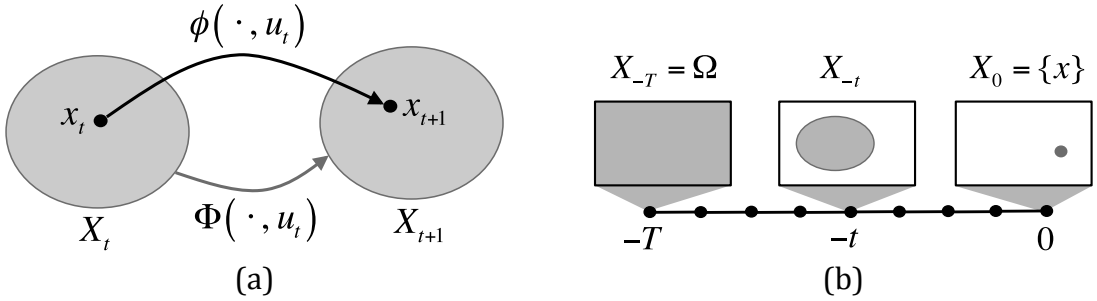


Figure 1.1: (a) The relationship of the bounding chain $\Phi(\cdot, u_t)$ to the underlying Markov chain $\phi(\cdot, u_t)$. (b) Illustration of coupling from the past using bounding chains.

To define the bounding sets used in our perfect samplers, we use the partial orders $z \preceq \tilde{z}$ when $z_j \leq \tilde{z}_j$, and $\alpha \preceq \tilde{\alpha}$ when $\alpha_j \leq \tilde{\alpha}_j$, for all $j = 1, \ldots, k$. We say that $\theta \preceq \tilde{\theta}$ when $\alpha \preceq \tilde{\alpha}$, and $(z, \theta) \preceq (\tilde{z}, \tilde{\theta})$ when $z \preceq \tilde{z}$ and $\theta \preceq \tilde{\theta}$. These partial orders admit a minimum and maximum state for z, but only a minimum state for α:

$$z^{\min} = \left(\sum_{i=1}^{N} 1(y_{i1} > 0), \ldots, \sum_{i=1}^{N} 1(y_{ik} > 0) \right), \quad z^{\max} = \sum_{i=1}^{N} y_i, \quad \alpha^{\min} = (0, \ldots, 0).$$

It may seem surprising that $z^{\min}$ is not 0, but recall that if $y_{ij} > 0$, then $v_{ij,1} = 1$, so $z_j^{\min}$ is the number of nonzero entries in the jth column of y. These partial orders allow us to construct bounding sets, the details of which appear below in the context of the two algorithms we use.

1.3.2 Component-wise algorithm

For our first algorithm, we work with the parameterization $(\alpha_1, \ldots, \alpha_k)$ instead of (ω, λ). To guarantee the necessary monotonicity, it is sufficient for the prior density on ω to satisfy the following property.

Property 1. The ratio

$$\frac{\pi_0(\alpha_j + \tilde{s}_j)}{\pi_0(\alpha_j + s_j)} \left(\frac{\alpha_j + \tilde{s}_j}{\alpha_j + s_j} \right)^{1-\sum_{j=1}^k \delta_j} \tag{1.10}$$

is increasing in α_j whenever $\tilde{s}_j \geq s_j$, where $\delta_1, \ldots, \delta_k$ are the Dirichlet parameters in (1.2).

From the conditional density

$$p(\alpha_j \mid y, z, \alpha_{[-j]}) \propto \pi_0 \left(\alpha_j + s_j\right) \left(\alpha_j + s_j\right)^{1-\sum_{j=1}^k \delta_j} \alpha_j^{\delta_j + z_j - 1} \prod_{i=1}^N \frac{\Gamma\left(\alpha_j + s_j\right)}{\Gamma(\alpha_j + s_j + n_i)}, \tag{1.11}$$

where $s_j = \sum_{\ell \neq j} \alpha_\ell$, we can show that for any prior satisfying Property 1, the conditional density (1.11) will enable monotone updates of α_j. One such prior is $\omega \sim \text{Gamma}(b_0, b_1)$ for $b_0 \leq \sum_{j=1}^k \delta_j$ and any b_1.

The following lemma, which is a slightly more general version of Lemma 1 in Møller (1999), can be used to demonstrate why Property 1 enables monotone updates for α_j.

Lemma 1. *Suppose X and Y are univariate random variables with densities f and g, respectively, with respect to the same measure ν. Let $S_X = \{x : f(x) > 0\}$ and $S_Y = \{x : g(x) > 0\}$. If the function h defined below is increasing for $x \in S_X \cup S_Y$,*

$$h(x) = \begin{cases} 0 & x \in S_X^C \cap S_Y, \\ f(x)/g(x) & x \in S_X \cap S_Y, \\ \infty & x \in S_X \cap S_Y^C, \end{cases} \tag{1.12}$$

then X stochastically dominates Y, that is, $\Pr(X \leq c) \leq \Pr(Y \leq c)$ for all c.

Proof. For $h(x)$ to be increasing in x, it must be true that if $x_1 \in S_X^C \cap S_Y$, $x_2 \in S_X \cap S_Y$, and $x_3 \in S_X \cap S_Y^C$, then $x_1 < x_2 < x_3$. Therefore, if $c \in S_X^C \cap S_Y$, then $\Pr(X \leq c) = 0 \leq \Pr(Y \leq c)$. Similarly, if $c \in S_X \cap S_Y^C$, then $\Pr(Y \leq c) = 1 \geq \Pr(X \leq c)$.

Now consider the case $c \in S_X \cap S_Y$. Let $A = \{X \in S_Y\}$. Since $1(Y \leq c)$ is decreasing in Y

and $h(Y)$ is increasing,

$$0 \geq \text{cov}\left\{1(Y \leq c), h(Y)\right\} = \Pr(\{X \leq c\} \cap A) - \Pr(Y \leq c)\Pr(A),$$

whence $\Pr(\{X \leq c\} \cap A) \leq \Pr(Y \leq c)$. But $\Pr(\{X \leq c\} \cap A^C) = 0$, since if $c \in S_X \cap S_Y$ and $x \in S_X \cap S_Y^C$, then $x > c$. Thus, $\Pr(X \leq c) \leq \Pr(Y \leq c)$ for all c. □

If $\tilde{\alpha}_{[-j]} \succeq \alpha_{[-j]}$ and $\tilde{z} \succeq z$, then the conditional distribution of α_j given $(y, \tilde{z}, \tilde{\alpha}_{[-j]})$ will stochastically dominate the distribution of α_j given $(y, z, \alpha_{[-j]})$. This enables a monotone update using inverse transform sampling by coupling upper and lower chains, which serve as upper and lower bounds for our bounding sets, with the same uniform random input. We can update z given (α, y) by drawing Uniform$(0, 1)$ random variables u_{im} for $i = 1, \ldots, N$ and $m = 1, \ldots, y_{ij} - 1$, and then setting

$$z_j = \sum_{i=1}^{N} 1(y_{ij} > 0) \left[1 + \sum_{m=1}^{y_{ij}-1} 1\left\{u_{im} \leq \frac{\alpha_j}{\alpha_j + m}\right\}\right]. \tag{1.13}$$

To update $\tilde{z}_j$, we use the same random inputs u but replace α_j by $\tilde{\alpha}_j$, which makes clear that if $\tilde{\alpha}_j \geq \alpha_j$, then the updated $\tilde{z}_j$ satisfies $\tilde{z}_j \geq z_j$.

It is natural to ask why any other algorithms are needed, since perfect sampling using monotone coupling from the past is well established. Two problems make this algorithm impractical. First, since the update for α_j conditions on $\alpha_{[-j]}$, this is not a two-step Gibbs sampler, and hence the marginal sequence $\{z_t, t = 1, \ldots\}$ does not form a Markov chain, where t indexes iterations rather than components. Therefore, if we only checked the coalescence on z, we would ignore the possibility that θ may be different in the lower and upper chains, which can in future steps allow z to uncoalesce. Unfortunately, updating α_j via inverse transform sampling guarantees that θ will almost surely never coalesce, because $\Pr(\alpha_j^{\text{U}} > \alpha_j^{\text{L}}) = 1$ at every time t, where U and L denote respectively the upper and lower chains; a similar situation

occurs in Møller (1999).

Second, to make the problem worse, because there is no natural maximum state for α, we do not even have a method to draw the initial values α^{U} and α^{L}, conditioning on $z = z^{\max}$ in the upper chain and $z = z^{\min}$ in the lower chain respectively, that would guarantee $\alpha^{\mathrm{U}}_{[-j]} \succeq \alpha^{\mathrm{L}}_{[-j]}$ for all j. The coordinate-wise approach using (1.11) cannot initialize the full parameter vector with this guarantee because it only updates each α_j conditional on all the other values $\alpha_{[-j]}$.

1.3.3 VECTOR ALGORITHM

Our vector algorithm solves both problems above, but introduces a new one. It first updates the entire parameter vector given the complete data, and then updates the augmented data given the parameter. All random draws in the upper and lower chains are coupled by using probability integral transform sampling with the same random quantiles. First, we draw

$$\omega^{\mathrm{L}} \sim p(\omega \mid y, z^{\mathrm{L}}), \quad \omega^{\mathrm{U}} \sim p(\omega \mid y, z^{\mathrm{U}}). \tag{1.14}$$

Here, the actual implementation of (1.14) will depend on the choice of π_0, which we assume satisfies Property 1. Then, for $j = 1, \ldots, k$, we draw

$$\gamma_j^{\mathrm{L}} \sim \text{Gamma}(\delta_j + z_j^{\mathrm{L}}, 1), \quad \gamma_j^{\mathrm{U}} \sim \text{Gamma}(\delta_j + z_j^{\mathrm{U}}, 1). \tag{1.15}$$

If we were to set $\alpha_j^{\mathrm{L}} = \omega^{\mathrm{L}}\gamma_j^{\mathrm{L}} / \sum_{\ell=1}^k \gamma_\ell^{\mathrm{L}}$ and $\alpha_j^{\mathrm{U}} = \omega^{\mathrm{U}}\gamma_j^{\mathrm{U}} / \sum_{\ell=1}^k \gamma_\ell^{\mathrm{U}}$, these would be valid draws from the complete-data posterior distributions $p(\alpha \mid y, z^{\mathrm{L}})$ and $p(\alpha \mid y, z^{\mathrm{U}})$, as given in (1.4). Unfortunately, this would not guarantee that $\alpha^{\mathrm{L}} \preceq \alpha^{\mathrm{U}}$. Even if $\gamma_j^{\mathrm{L}} \leq \gamma_j^{\mathrm{U}}$ for all j, it is possible for $\gamma_j^{\mathrm{L}} / \sum_{\ell=1}^k \gamma_\ell^{\mathrm{L}} > \gamma_j^{\mathrm{U}} / \sum_{\ell=1}^k \gamma_\ell^{\mathrm{U}}$ for some j, and the difference may be large enough that $\omega^{\mathrm{L}}\gamma_j^{\mathrm{L}} / \sum_{\ell=1}^k \gamma_\ell^{\mathrm{L}} > \omega^{\mathrm{U}}\gamma_j^{\mathrm{U}} / \sum_{\ell=1}^k \gamma_\ell^{\mathrm{U}}$, even though $\omega^{\mathrm{L}} < \omega^{\mathrm{U}}$.

However, we can achieve a valid bounding chain algorithm that preserves the order $\alpha^{\mathrm{L}} \preceq \alpha^{\mathrm{U}}$

by dividing by the sum of γ_ℓ in the opposite chain:

$$\alpha_j^{\mathrm{L}} = \frac{\omega^{\mathrm{L}}\gamma_j^{\mathrm{L}}}{\sum_{\ell=1}^k \gamma_\ell^{\mathrm{U}}}, \quad \alpha_j^{\mathrm{U}} = \frac{\omega^{\mathrm{U}}\gamma_j^{\mathrm{U}}}{\sum_{\ell=1}^k \gamma_\ell^{\mathrm{L}}}. \tag{1.16}$$

If $z^{\mathrm{L}} = z^{\mathrm{U}}$ immediately before this parameter update, then $\alpha_j^{\mathrm{L}} = \alpha_j^{\mathrm{U}}$ for all $j = 1, \ldots, k$, so that it is possible to coalesce on z and α.

To see that alternating between updating α using (1.14)–(1.16) and then updating z given α and y yields a valid bounding chain algorithm, suppose $(z_t, \theta_t) \in Z_t \times \Theta_t$, where $Z_t = \{z : z_t^{\mathrm{L}} \preceq z \preceq z_t^{\mathrm{U}}\}$ and $\Theta_t = \{\theta : \theta_t^{\mathrm{L}} \preceq \theta \preceq \theta_t^{\mathrm{U}}\}$. Then, we can draw $\omega \sim p(\omega \mid y, z_t)$ and $\gamma_j \sim \text{Gamma}(\delta_j + z_{t,j}, 1)$, where $z_{t,j}$ is the jth component of z at time t, and draw $\omega^{\mathrm{L}}, \omega^{\mathrm{U}}, \gamma^{\mathrm{L}}, \gamma^{\mathrm{U}}$ using (1.14) and (1.15), guaranteeing that $\omega^{\mathrm{L}} \leq \omega \leq \omega^{\mathrm{U}}$ and $\gamma_j^{\mathrm{L}} \leq \gamma_j \leq \gamma_j^{\mathrm{U}}$. Therefore,

$$\frac{\omega^{\mathrm{L}}\gamma_j^{\mathrm{L}}}{\sum_{\ell=1}^k \gamma_\ell^{\mathrm{U}}} \leq \frac{\omega\gamma_j}{\sum_{\ell=1}^k \gamma_\ell} \leq \frac{\omega^{\mathrm{U}}\gamma_j^{\mathrm{U}}}{\sum_{\ell=1}^k \gamma_\ell^{\mathrm{L}}},$$

so that $\theta_{t+1} \in \Theta_{t+1}$. The z draw (1.13) ensures that $z_{t+1} \in Z_{t+1}$.

The new problem is that this bounding chain is loose, making it impractical even on low-dimensional, low-count data sets; see Section 1.5. However, strategically alternating between the vector algorithm and the component-wise algorithm can yield a much speedier composite algorithm, as we demonstrate empirically in Section 1.5, after a theoretical investigation in Section 1.4.

1.4 Theoretical results: Composite bounding chains

As a general setting, suppose we wish to sample from $p(\theta \mid y)$, and we have an augmented-data model $p(z, \theta \mid y)$. To detect coalescence, it is helpful for z to be discrete, but see Murdoch and Green (1998) for perfect samplers on continuous state spaces. We assume two underlying Markov chains with stochastic recursive sequences $x_{t+1} = \phi(x_t, u_t)$ and $x_{t+1} = \psi(x_t, v_t)$,

where u_t and v_t are random inputs and $x_t = (z_t, \theta_t)$. We use bounding chains $X_t \ni x_t$ with associated stochastic recursive sequences $X_{t+1} = \Phi(X_t, u_t)$ and $X_{t+1} = \Psi(X_t, v_t)$. We assume that the bounding sets can be written as $X_t = Z_t \times \Theta_t$, with $z_t \in Z_t$ and $\theta_t \in \Theta_t$.

Since our goal is to draw samples of θ, a perfect sampler will not terminate before Θ_t is a singleton. Thus, if Φ can quickly reduce Z_t to a singleton and if Ψ can reduce Θ_t to a singleton once Z_t is a singleton, then alternating between Φ and Ψ can be faster than either individually. To take an extreme case, an alternating algorithm can coalesce on (z, θ) even if Φ can never coalesce on θ and Ψ can never coalesce on z.

Our composite algorithm alternates between $(M-1)$-fold compositions $\Phi \circ \cdots \circ \Phi = \Phi^{M-1}$ for some pre-chosen $M > 1$, and single instances of Ψ; more general combination strategies are of course possible. That is, in notation, the composite stochastic recursion function is $\Psi \circ \Phi^{M-1}$.

Recall that a stochastic recursive sequence for a monotone Markov chain satisfies $x \preceq \tilde{x} \Rightarrow \psi(x, v) \preceq \psi(\tilde{x}, v)$ for all random inputs v. For bounding chains, we can use the subset relationship as a partial order. That is, a bounding chain $\mathcal{B}$ is monotone if for all its random inputs v, $\mathcal{B}(X, v) \subseteq \mathcal{B}(\tilde{X}, v)$ whenever $X \subseteq \tilde{X}$.

Lemma 2. *If we let Φ denote the component-wise algorithm of Section 1.3.2 and Ψ denote the vector algorithm of Section 1.3.3, then both Φ and Ψ are monotone bounding chains.*

Proof. This is easy to verify directly once we recognize that if $z \preceq \tilde{z}$ and $\alpha_{[-j]} \preceq \tilde{\alpha}_{[-j]}$, then (i) $p(\alpha_j \mid y, \tilde{z}, \tilde{\alpha}_{[-j]})$ stochastically dominates $p(\alpha_j \mid y, z, \alpha_{[-j]})$, (ii) $p(\omega \mid y, \tilde{z})$ stochastically dominates $p(\omega \mid y, z)$, and (iii) Gamma$(\delta_j + \tilde{z}_j, 1)$ stochastically dominates Gamma$(\delta_j + z_j, 1)$. Then, the partial ordering $\tilde{x}^{\mathrm{L}} \preceq x^{\mathrm{L}} \preceq x^{\mathrm{U}} \preceq \tilde{x}^{\mathrm{U}}$ is preserved by using inverse transform sampling with the same uniform random inputs to draw from the distributions in Sections 1.3.2 and 1.3.3. □

To bound the expected running time until coalescence for $\Psi \circ \Phi^{M-1}$, we assume that if Z_t

is a singleton, then Ψ causes $Z_{t+1} \times \Theta_{t+1}$ to be a singleton:

$$\Psi(\{z\} \times \Theta, v) = \{z'\} \times \{\theta'\} \tag{1.17}$$

for any z, Θ, v. We call this property the ability to *conditionally induce coalescence* on θ. While this may seem restrictive, our focus is on Bayesian computation where θ is the parameter of interest and z is part of an augmented data model. In such settings it is often possible to develop bounding chain algorithms that can conditionally induce coalescence by including draws of the full parameter vector θ given the observed and augmented data.

The time until coalescence is $\tau = \min(t : X_t = \{x\})$, where t is incremented each time either Φ or Ψ is called, that is, we define the time index t via

$$X_{t+1} = \begin{cases} \Phi(X_t, u_t), & \text{if } t+1 \not\equiv 0; \\ \Psi(X_t, v_t), & \text{if } t+1 \equiv 0. \end{cases} \pmod{M} \tag{1.18}$$

Note that we check for coalescence only after each time Ψ is called. To initialize the chain, we pass the state space to Ψ:

$$X_0 = \Psi(\Omega_x, v_0). \tag{1.19}$$

In a coupling-from-the-past implementation, we would initialize $X_{-T} = \Psi(\Omega_x, v_{-T})$ for $T > 0$. Incrementing t each time we use either update, rather than with each block, guarantees that different values of τ corresponding to different choices of M can be compared directly, in that smaller τ roughly corresponds to faster computation. We then have the following theorem.

Theorem 1. *Suppose Φ and Ψ define monotone bounding chains, and that Ψ conditionally induces coalescence on θ as in (1.17). Then, letting Z_{M-1} be the bounding set for z at the end of the first sequence Φ^{M-1} and letting $|Z_{M-1}|$ be its cardinality, the expected time until coalescence of*

the composite algorithm (1.18) and (1.19) satisfies

$$M \leq E(\tau) \leq \frac{M}{\Pr(|Z_{M-1}| = 1)}. \tag{1.20}$$

In our composite algorithm for the Dirichlet-multinomial model, we let Φ be the component-wise algorithm of Section 1.3.2 and Ψ be the vector algorithm of Section 1.3.3. It is then straightforward to show that the assumptions of Theorem 1 apply. Moreover, if we let ϕ be a stochastic recursive sequence for the Gibbs sampler in (1.8) and ψ correspond to the sampler in (1.4)–(1.5), then the composite chain $\psi \circ \phi^{M-1}$ has the correct stationary distribution $p(z, \theta \mid y)$, and $\Psi \circ \Phi^{M-1}$ is a valid bounding chain for this composite Gibbs sampler. Thus, a perfect sampler using the composite updates $\Psi \circ \Phi^{M-1}$ will return a genuine draw from $p(z, \theta \mid y)$.

Proof of Theorem 1. The lower bound is trivially true because we check for coalescence only at the end of a block, and it takes M steps to reach the end of the first block.

For the upper bound, we will show that τ is stochastically dominated by a scaled geometric random variable with expectation $M/\Pr(|Z_{M-1}| = 1)$. Consider a modified version of Ψ, called Ψ_0, that first resets $X = \Omega_x = \Omega_z \times \Omega_\theta$ and then updates Ω_x via Ψ. That is, $\Psi_0(X, v) = \Psi(\Omega_x, v)$ for any X. Because we use Ψ_0 to set the algorithm's starting values, we can write the composite algorithm as

$$\cdots \circ \Psi \circ \Phi^{M-1} \circ \Psi \circ \Phi^{M-1} \circ \Psi_0.$$

Since Ψ conditionally induces coalescence on θ, if Z has coalesced at the end of Φ^{M-1}, then Ψ will cause Θ to coalesce in the next step. Thus we can group the composite algorithm's updates

$$\cdots \circ (\Phi^{M-1} \circ \Psi) \circ (\Phi^{M-1} \circ \Psi) \circ (\Phi^{M-1} \circ \Psi_0) \tag{1.21}$$

and check whether Z has coalesced at the end of each block, though in practice we must be

sure to still call Ψ after Z has coalesced to obtain a valid draw (z, θ). By using the same sequence of random inputs, we couple this composite algorithm to a modified algorithm that substitutes Ψ_0 for every Ψ, and we similarly group the sequence of updates in the modified algorithm:

$$\cdots \circ (\Phi^{M-1} \circ \Psi_0) \circ (\Phi^{M-1} \circ \Psi_0) \circ (\Phi^{M-1} \circ \Psi_0). \tag{1.22}$$

Assuming that the random inputs to the stochastic recursive sequences at different steps are independent, then since Ψ_0 deterministically resets X ignoring the current states, the blocks in the modified algorithm are independent, and the indicators I_m for Z-coalescence in block m are independent and identically distributed, with $\Pr(I_m = 1) = \Pr(|Z_{M-1}| = 1)$. Therefore the number of steps until the first $I_m = 1$ is a scaled geometric random variable with mean $M/\Pr(|Z_{M-1}| = 1)$, where the scaling is necessary because there are M steps per group.

Finally, since Φ and Ψ are monotone bounding chains, when we couple the composite and modified algorithm by using the same sequence of random inputs, Z-coalescence in the modified algorithm implies Z-coalescence in the composite algorithm, yielding the upper bound in (1.20). □

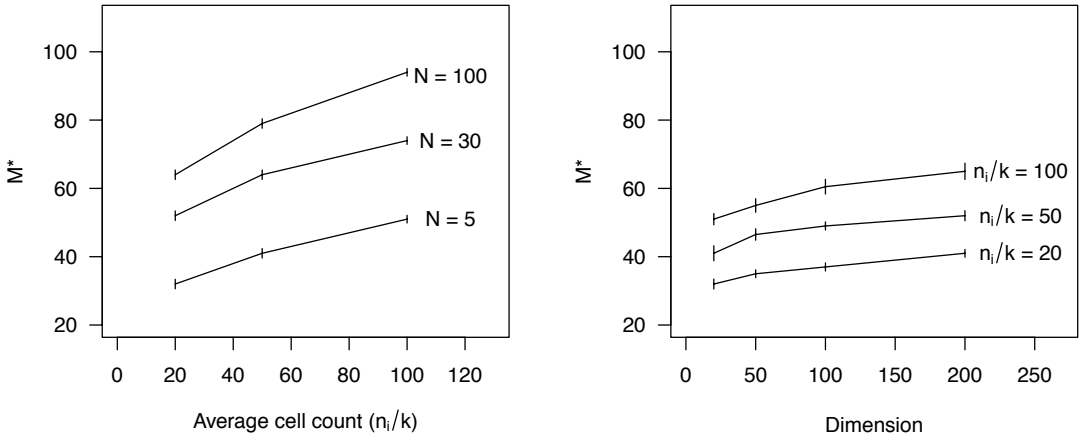
Since both M and $\Pr(|Z_{M-1}| = 1)$ are monotonically increasing functions of M, Theorem 1 suggests an intuitively sensible trade-off that occurs when choosing M, the block size. If M is too small, then $\Pr(|Z_{M-1}| = 1)$ is also small, and the upper bound on the expected time until coalescence blows up, so that we have much weaker guarantees on the expected running time. However, if M is so large that $\Pr(|Z_{M-1}| = 1) \approx 1$ and Z typically coalesces well before M steps, then much of the within-block computation will be wasted. In the next section, we will see from simulation that minimizing the upper bound in (1.20) can lead to a choice of M that is not far from optimal; see Figure 1.4.

1.5 NUMERICAL ILLUSTRATION

To explore how the composite algorithm scales with the dimension k, the row total n_i, and the number of rows N, we simulated data sets under various conditions. In Figure 1.2(a), we fixed the dimension $k = 20$ and simulated data sets with number of rows $N = 5, 30$, and 100, and with average cell count $n_i/k = 20, 50$, and 100. For Figure 1.2(b), we fixed the number of rows $N = 5$ and simulated data sets with $k = 20, 50, 100$, and 200, and with average cell count $n_i/k = 20, 50$, and 100, for all i. For each choice of N, n_i, and k, we simulated and fit 100 data sets under independent exponential priors with mean 1 on $\alpha_1, \ldots, \alpha_k$. Figure 1.2 plots the results for running the component-wise algorithm on each data set. The vertical axes are estimates of M^*, the block size such that $\Pr(|Z_{M^*-1}| = 1) = 0.5$, so that if the block size $M = M^*$, then the expected time until coalescence for the composite algorithm satisfies $M^* \leq E(\tau) \leq 2M^*$. Encouragingly, M^* appears to increase slower than linearly with N, k, and n_i/k, and our method is not restricted to small samples, as we are able to fit data sets with up to 200,000 total counts.

However, our implementation leaves room for improvement. With $N = 100$, each iteration took approximately 14 seconds on a 2.80 GHz CPU. We do not view this as a fundamental limitation of our method, however, since we made no attempt to optimize our code, and further numerical and computational work could reduce the computation time for the draws from (1.11) and (1.14), just as there are now fast algorithms to evaluate the inverses of common cumulative distribution functions, for instance. Of course, the computation time per iteration must also increase linearly with k, since the component-wise algorithm must cycle through each dimension.

To illustrate the speed gains in greater detail, we fit two simple artificial datasets, given in Table 1.1, assuming independent exponential prior distributions on α_1 and α_2. Figure 1.3



(a) (b)

Figure 1.2: Plot of the dependence of estimates of M^*, where $\Pr(|Z_{M^*-1}| = 1) = 0.5$, on the number of rows N, the dimension k, and the average cell count n_i/k. In (a), the dimension is fixed at 20, and in (b), the number of rows is fixed at 5. Estimates of M^* are based on 100 simulated data sets under each condition, and error bars are ± 2 standard deviations estimated via bootstrap.

shows a histogram of coalescence times for the vector algorithm of Section 1.3.3 applied to the first artificial data set. The 37% of runs that did not coalesce in 10^5 iterations are omitted. We also applied the vector algorithm to the second artificial data set, but we stopped it at 10^6 iterations without observing coalescence. Thus, it can take an extremely long time for the vector algorithm to return a value, even on low-dimensional, low-count problems with few observed rows. By itself, the vector algorithm is useless in practice, as is the component-wise algorithm, since the latter almost surely never returns a sample.

However, the composite algorithm (1.18) is much faster. Figure 1.4 shows the times until coalescence for the composite algorithm on both data sets, using various M. The gains are substantial. Whereas the vector algorithm could not return a single sample for the second artificial data set in 10^6 iterations, typical coalescence times for the composite algorithm are in the dozens. Also plotted are the lower and estimated upper bounds in (1.20) on the expected coalescence times. The denominator $\Pr(|Z_{M-1}| = 1)$ in the upper bound is estimated as the

Table 1.1: Artificial data sets used to illustrate the speed gains of the composite algorithm

	Artificial data set 1		Artificial data set 2	
	Class 1	Class 2	Class 1	Class 2
Observation 1	5	10	10	20
Observation 2	6	9	12	18

proportion with $|Z_{M-1}| = 1$ of 200 runs of the component-wise algorithm. On a 2.66 GHz CPU, code for the composite algorithm took on average 16.0 and 16.1 milliseconds per iteration, with standard deviations 0.2 milliseconds per iteration, for the low- and high-count artificial data sets, respectively.

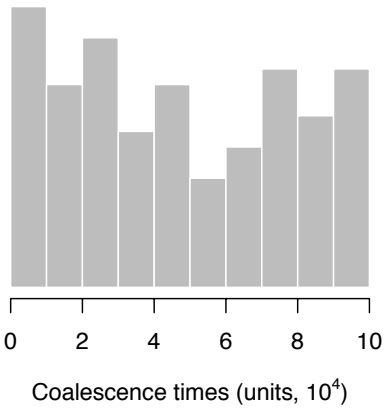
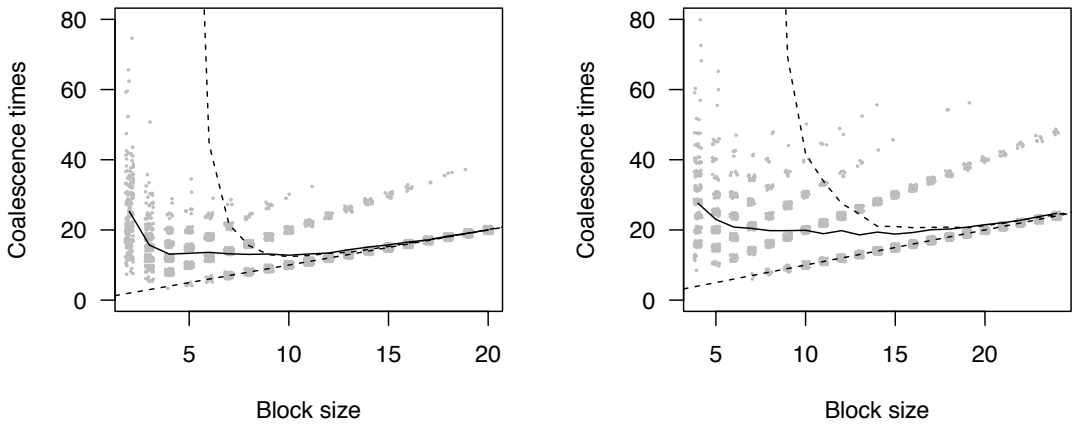


Figure 1.3: Histogram of coalescence times τ for the vector algorithm of Section 1.3.3 applied to the first artificial data set in Table 1.1. Only shown are the 63% of runs with coalescence times less than 10^5 iterations.



(a) (b)

Figure 1.4: Plots of the dependence of the composite algorithm's expected time until coalescence on the block size M, for the (a) lower-count and (b) higher-count artificial data sets in Table 1.1. Grey points are jittered values of the number of iterations until coalescence for different runs of the composite algorithm. Dashed lines are the lower bound and estimated upper bound in (1.20), and the solid lines connect average coalescence times for each block size.

1.6 Discussion

Our data augmentation strategy was inspired by the algorithm of Kou and McCullagh (2009) for approximating the weighted matrix permanent, defined as

$$\text{per}_{\omega}(A) = \sum_{\sigma \in \Pi_n} \omega^{\text{cyc}(\sigma)} \prod_{i=1}^{n} A_{i,\sigma(i)},$$

where $A_{j,k}$ is the (j,k)th entry of an $n \times n$ matrix A; $\text{cyc}(\sigma)$ is the number of cycles of the permutation σ; and Π_n is the set of all permutations of $\{1,\ldots,n\}$. The weighted permanent is the density function for permanental processes, a class of Cox processes that can be used for classification (McCullagh and Møller, 2006; McCullagh and Yang, 2006). The sequential importance sampling algorithm of Kou and McCullagh (2009) approximates the weighted permanent by drawing ordered partitions of $\{1,\ldots,n\}$, which can be mapped to permutations of

$\{1, \ldots, n\}$ by viewing each block of the partition as a cycle, with the order of elements in each cycle determined by their order in the partition. If $A_{j,k} = 1$ for all j and k, then the weighted permanent is $\Gamma(\omega + n)/\Gamma(\omega)$. The algorithm of Kou and McCullagh (2009) can then be interpreted as first randomly ordering the elements of $\{1, \ldots, n\}$, and then partitioning them by drawing

$$z_i \sim \text{Bernoulli}\left(\frac{\omega}{\omega + i - 1}\right), \quad i = 1, \ldots, n-1,$$

where $z_i = 0$ means that elements i and $i+1$ are in the same block.

In the aforementioned technical report by Minka (2003), an exponential-family approximation is used to interpret the Dirichlet-multinomial model as a "multinomial with 'damped' counts." Our method offers another perspective on that statement. Let us consider the likelihood for one observation $y = (y_1, \ldots, y_k)$:

$$p(y \mid \omega, \lambda) = \frac{\Gamma(\omega)}{\Gamma(\omega + n)} \prod_{j=1}^{k} \frac{\Gamma(\omega\lambda_j + y_j)}{\Gamma(\omega\lambda_j)}. \tag{1.23}$$

In (1.23), we observe the class assignments $a_1, \ldots, a_n$ and count the number of individuals assigned to each class $y_j = \sum_{i=1}^{n} 1(a_i = j)$. Unlike in Section 1.2, here we introduce an augmented variable σ via $p(y, \sigma \mid \theta) = p(\sigma \mid \theta)\, p(y \mid \sigma, \theta)$. We therefore must verify that $p(y \mid \theta) = \sum_\sigma p(\sigma \mid \theta)\, p(y \mid \sigma, \theta)$ is the same as (1.23).

We first draw σ, a permutation of $\{1, \ldots, n\}$, from

$$p(\sigma \mid \theta) = \frac{\Gamma(\omega)}{\Gamma(\omega + n)} \omega^{\text{cyc}(\sigma)}.$$

This distribution depends only on ω, not λ, thus isolating the role of the parameter ω and giving us a natural interpretation. This permutation then partitions the individuals $\{1, \ldots, n\}$ according to the cycles of the permutation. We let $C_i(\sigma) \subseteq \{1, \ldots, n\}$ denote the set of elements included in the ith cycle of σ, where $i = 1, \ldots, \text{cyc}(\sigma)$, and the order can be chosen, for

example, according to the smallest element in each cycle. We let B_{ij} denote the event that all of the individuals in cycle i are assigned to class j; that is, $B_{ij} = \{a_m = j \text{ for all } m \in C_i(\sigma)\}$. To ensure that class assignments are consistent within cycles of σ, we also define the event $D = \cap_{i=1}^{\text{cyc}(\sigma)} \cup_{j=1}^{k} B_{ij}$. Then, given σ, individuals are assigned to classes according to the multinomial distribution

$$p(y \mid \sigma, \theta) = 1_D \prod_{i=1}^{\text{cyc}(\sigma)} \prod_{j=1}^{k} \lambda_j^{1_{B_{ij}}},$$

where 1_D is the indicator for the event D. This distribution depends only on λ, not on ω. We can then verify that marginalizing $p(\sigma \mid \theta)\, p(y \mid \sigma, \theta)$ over σ yields (1.23).

This augmented-data model quantifies several well-known and intuitive features of the Dirichlet-multinomial model. First, no approximation is required to interpret the Dirichlet-multinomial model as a "multinomial with 'damped' counts." A draw from the Dirichlet-multinomial model can be exactly generated by first partitioning the data according to a permutation of individuals, and then by classifying individuals by a multinomial distribution acting on the blocks of the partition. The dampening is therefore a result of blocking observations according to the permutation σ.

Second, it is commonly known that small values of ω induce sparsity in the observed data, while the mean vector λ controls the expected frequencies of each class. Sparsity means that there may be many classes with zero observations and a few classes with many observations. Thus, sparsity is associated with high variance. In the permutation augmentation, small values of ω favor permutations with few cycles, so that most individuals are partitioned into just a few blocks. The multinomial allocation to classes then necessarily results in sparse observations, because the class assignment operates on blocks, not individuals. Conversely, large values of ω favor permutations with many cycles, corresponding to fine partitions of individuals and a lower chance of observing sparsity.

1.7 CONCLUDING REMARKS

The block structure of our composite algorithm is reminiscent of read-once coupling from the past (Wilson, 2000), which allows exact sampling without requiring the storage of random inputs necessary for traditional coupling from the past. However, in Wilson (2000), each block restarts with the maximal bounding set if the chain fails to coalesce in the previous block. Thus, using the notation of Section 1.4, a read-once implementation of our sampler could be based on blocks $\Psi \circ \Phi^{M-1} 4 \circ \Psi_0$.

On the limitation side, our approach assumes that the prior π on θ satisfies (1.2) and that the prior π_0 on ω satisfies Property 1. While this is limiting, importance sampling can be used if another prior is desired. Suppose we wish to estimate the expectation of a function $h(\theta)$ under the posterior distribution assuming the prior π_1. We can generate exact samples $\theta_{(1)}, \ldots, \theta_{(L)}$ from the posterior distribution assuming the prior π and use the importance sampling estimator

$$\hat{E}\left\{h(\theta) \mid y\right\} = \frac{\sum_{\ell=1}^{L} \frac{\pi_1(\theta_{(\ell)})}{\pi(\theta_{(\ell)})} h(\theta_{(\ell)})}{\sum_{\ell=1}^{L} \frac{\pi_1(\theta_{(\ell)})}{\pi(\theta_{(\ell)})}}.$$

An avenue for future work is to investigate the tradeoff between using this importance sampling estimator based on independent exact samples under π, and using an approximate algorithm for sampling directly under π_1, such as any non-exact Markov chain sampler. A further challenge is to extend our approach to models that account for covariates.

Despite the limitations of our specific algorithm, we believe the idea of using divide-and-conquer composite bounding chains is suitably general, and it may provide an important building block toward our ultimate goal of making perfect sampling a workhorse in statistical computing.

2

Catalytic priors: General methodology

2.1 Introduction

2.1.1 Our goals

This article proposes a general strategy for constructing prior distributions in complex models. In complicated data analysis settings, the primary model under consideration is often so flexible that additional regularization is necessary. One way to provide such regularization is through informative or weakly informative prior distributions, but in multilevel models it is often not clear how to build prior distributions on parameters that may be many levels

removed from the observed data, and bad choices of priors may lead to problems such as unrealistic predictions under the fitted model.

However, a simpler model for the observations is often available. While this simpler model does not capture all of the scientifically relevant characteristics of the data or the data-generating process, its simplicity means that it can be reliably estimated. The strategy we propose builds a prior that shrinks the complex model toward the estimated simpler model. One advantage of our approach is that it can be applied when the simpler model is not a sub-model of the more complicated one, and in fact even when there is no relationship between the parameters of the two models. Thus, our approach enables behavior similar to hierarchical models in settings in which it is not clear how to enforce such behavior through direct distributional assumptions on parameters. We call the priors resulting from our approach *catalytic priors* to suggest that adding such prior information catalyzes our ability to use richer, more realistic models.

The remainder of this paper is organized as follows. We review some of the literature on noninformative priors in Section 2.1.2 and the recent developments on weakly informative priors in Section 2.1.3. Section 2.2 introduces catalytic priors, which are defined in Section 2.2.3. We provide two separate justifications of catalytic priors, one from the perspective of missing data imputation in Section 2.3, and the other based on information theory in Section 2.4. Finally, Section 2.5 concludes with a discussion of some related work.

2.1.2 Default prior distributions

Before introducing catalytic priors, we briefly review some popular default priors. We refer the reader to Kass and Wasserman (1996) for a more in-depth discussion of many of the principles that have been suggested to guide the choice of default priors. After uniform priors, Jeffreys's prior (Jeffreys, 1961) is perhaps the most common default choice. Jeffreys's prior is $\pi_{\text{Jeff}}(\theta) \propto \{\det I(\theta)\}^{1/2}$, where $I(\theta)$ is the expected Fisher information. Jeffreys's

prior is often justified by its invariance to reparameterization. If we transform from θ to ψ and let J be the Jacobian matrix of partial derivatives with ijth element $J_{ij} = \partial\theta_i/\partial\psi_j$, then $\pi_{\text{Jeff}}(\psi) = \pi_{\text{Jeff}}(\theta)|\det J| \propto \{\det I(\psi)\}^{1/2}$. Thus, Jeffreys's prior takes the same form regardless of the choice of parameterization, which is attractive if we want to interpret Jeffreys's prior as noninformative. Intuitively, if we have no information about θ, then we also have no information about $\psi(\theta)$, and it would make little sense for a noninformative prior for θ to be informative for $\psi(\theta)$.

Jeffreys's prior also has attractive asymptotic frequentist properties, including bias reduction (Firth, 1993) and posterior intervals with coverage correct to $O(n^{-1})$ when θ is a scalar parameter (Welch and Peers, 1963; Nicolaou, 1993). Brown *et al.* (2001) showed that when used to construct confidence intervals for binomial proportions, Jeffreys's prior performs well in terms of coverage and interval length both in small samples and in "unlucky" large samples in which the usual Wald intervals do not perform well.

In multidimensional settings, however, Jeffreys's prior has well-known disadvantages. Consider $y_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$. If only μ is unknown, then Jeffrey's prior is uniform; if only σ^2 is unknown, then $\pi_{\text{Jeff}}(\sigma^2) \propto (\sigma^2)^{-1}$; but if both μ and σ^2 are unknown, then $\pi(\mu, \sigma^2) \propto (\sigma^2)^{-3/2}$. In fact, if there are k unknown means $\mu_1, \ldots, \mu_k$ and a common variance, then the usual rule for Jeffreys's prior yields $\pi(\mu_1, \ldots, \mu_k, \sigma^2) \propto (\sigma^2)^{-1-k/2}$. The marginal posterior distribution of an individual mean, say μ_1, is a t distribution with n degrees of freedom, where n is the total number of observations across all groups. Thus, the degrees of freedom do not depend on the number of means being estimated, which Jeffreys found unacceptable (Jeffreys, 1961). Under the prior $\pi(\mu_1, \ldots, \mu_k, \sigma^2) \propto (\sigma^2)^{-1}$, the corresponding degrees of freedom would be $n - k$, reflecting the extra uncertainty due to estimating k different means. Jeffreys addressed this by proposing to give location parameters uniform priors, and to apply the rule $\{\det I(\phi)\}^{1/2}$ to the remaining parameters, where $I(\phi)$ is the expected Fisher information for the non-location parameters ϕ, holding the location param-

eters fixed. Of course, this remedy sacrifices the invariance that was the original justification for this prior (although Jeffreys (1961, p. 183) argued that it maintains invariance for the only transformations we should care about).

Bernardo's reference prior is another default prior with appealing properties (Bernardo, 1979; Berger and Bernardo, 1992; Berger *et al.*, 2009, among others). The idea is to maximize for a given likelihood the amount of expected missing information, so that the prior is least informative in that it leaves the maximum amount of information to be learned from the data. In many circumstances, the reference prior reduces to Jeffreys's prior. Other principles that can be used to construct noninformative or default priors include maximizing entropy (Jaynes, 1957, 1968) and guaranteeing that the frequentist coverage of posterior intervals is approximately equal to the posterior mass contained within those intervals (see Fraser *et al.*, 2010; Datta and Mukerjee, 2004, for recent examples). Deriving matching priors that ensure correct coverage in complex models can be difficult in practice, or even theoretically impossible. We will discuss maximum entropy priors more in Section 2.4.

A more fundamental problem with most approaches to objective or default priors is that a diffuse prior can be very informative in high dimensions. Kass and Wasserman (1996) discuss this issue in depth. Their conclusion is that in large-sample settings in which the likelihood is sharply peaked, many default methods (most of which are based on asymptotic arguments) such as Jeffreys's prior will work quite well, but so may other naive diffuse priors. In fact, improper posteriors may even give reasonable results if model-fitting is concentrated around the peak of the likelihood. On the other hand, in small samples, conclusions may be extremely sensitive to the choice of prior. Then, none of the default methods may be reliable, and much care is needed. If the model is complicated enough, we may effectively be in a small sample setting even when there are many observations. The question then becomes whether there are good general strategies for guaranteeing that inferences are scientifically sensible, even when the likelihood is not sharply peaked. Providing such a strategy is one goal of weakly

informative priors, so we discuss them next.

2.1.3 Weakly informative prior distributions

Recently, weakly informative prior distributions have been proposed and investigated in Gelman (2006), Gelman *et al.* (2008), Fúquene *et al.* (2009), and Polson and Scott (2012), among others. These priors are motivated by a desire to find a robust middle ground between a noninformative prior and a subjective, informative prior. In modern statistical problems, investigators often have a wealth of subject-matter expertise. While some of this expertise is undoubtedly used to build a reasonable likelihood or hierarchical model, rarely does that model capture all of an investigator's scientific knowledge. A combination of statistical and subject-matter knowledge may also be available; for instance, Gelman *et al.* (2008) scaled their prior based on the plausible size of logistic regression coefficients when covariates have been standardized. The goal of a weakly informative prior distribution is to inject enough of this additional knowledge into the model to prevent unreliable behavior and scientifically unreasonable inferences, while leaving enough flexibility in the model so that the prior has little influence on the final conclusions if there is strong information in the data to contradict the prior.

The goals of our approach are similar to those of weakly informative priors. The literature on weakly informative priors has focused on developing and evaluating default weakly informative priors for typical models, one class of models at a time, as in Gelman (2006) and Polson and Scott (2012) for hierarchical variance parameters and Gelman *et al.* (2008) for logistic regression models. In contrast, the approach we take here is to develop a general strategy for building into the final inferences additional scientific knowledge in the form of a plausible simpler model that we can reliably fit using the data at hand. To be clear, we are aiming neither for noninformative priors nor for elicitation of subjective prior knowledge. Rather, we hope to provide a guide for the incorporation of additional weak prior information.

Finally, we emphasize that while Gelman (2006), Gelman *et al.* (2008), Fúquene *et al.* (2009), and Polson and Scott (2012) all focus on heavy-tailed priors such as the Cauchy, our approach does not: it yields conjugate priors which may have light tails. We do not view these as conflicting points of view because our framework makes it possible to combine these approaches, using a heavy-tailed prior as an underlying default prior, and altering it into a catalytic prior that effectively builds in hierarchical structure. Investigating such a combined approach is beyond the scope of the work here.

2.2 DEFINING CATALYTIC PRIORS

2.2.1 NOTATION

To avoid confusion among the several models at play, we will use the same letter to denote different densities within the same Bayesian model, allowing the arguments to specify the intended distribution. We let f denote the full, flexible model that needs additional regularization. The likelihood of our full model is $f(y \mid \theta)$, where $\theta \in \mathbb{R}^{d_f}$, and we assume we have a default prior density $f(\theta)$ with respect to Lebesgue measure. However, the posterior distribution under the full model may not be proper given observations y. That is, we may have $\int f(y \mid \theta) f(\theta) d\theta = \infty$. If $\int f(y \mid \theta) f(\theta) d\theta < \infty$, then we denote the posterior density under the full model by $f(\theta \mid y) = f(y \mid \theta) f(\theta) / f(y)$, where $f(y) = \int f(y \mid \theta) f(\theta) d\theta$.

We also have a simpler, estimable model g. For reasons that will soon become clear, we call g the *prior generating model*. The likelihood under the prior generating model is $g(y \mid \phi)$, where $\phi \in \mathbb{R}^{d_g}$, and we also have a default prior density $g(\phi)$. We assume that the posterior distribution $g(\phi \mid y)$ is proper, as is the posterior predictive distribution $g(\tilde{y} \mid y)$ for future data $\tilde{y}$.

We denote by π the final model used for inference and prediction. Thus, $\pi(\theta \mid y)$ is the operational posterior distribution on θ, the parameter of the full model, and $\pi(\tilde{y} \mid y)$ is the

operational posterior predictive distribution for $\tilde{y}$.

2.2.2 SETTING THE STAGE

We motivate our definition of catalytic priors with two examples.

Example 2.2.1. Suppose $y = (y_1, \ldots, y_n)$ consists of n independent and identically distributed draws from an exponential family with sufficient statistic $T(y)$ and natural parameter $B(\theta)$, so that

$$f(y \mid \theta) = e^{n\{T(y)'B(\theta) - C(\theta)\}} f_0(y).$$

A conjugate prior for θ is

$$\pi(\theta)d\theta \propto e^{n_0\{\mu_0' B(\theta) - C(\theta)\}} d\theta.$$

The posterior density under this likelihood and prior is therefore

$$\begin{aligned} \pi(\theta \mid y) &\propto f(y \mid \theta)\pi(\theta) \\ &\propto e^{\{n_0\mu_0 + nT(y)\}'B(\theta) - (n_0+n)C(\theta)}, \end{aligned}$$

which as a function of θ is identical to the likelihood if we had observed an additional n_0 observations with sufficient statistics μ_0. This is the well-known pseudo-data interpretation of conjugate priors.

Example 2.2.2. Consider a logistic regression model

$$z_j \mid \theta, m_j, x_j \sim \text{Binomial}\left(m_j, \frac{e^{x_j'\theta}}{1 + e^{x_j'\theta}}\right), \quad j = 1, \ldots, n.$$

Suppose the predictors x_j are discrete with finitely many distinct values, and suppose that for some but not all values of j, m_j may be 0, in which case $z_j = 0$. To prevent complete separation and infinite maximum likelihood estimates in this model, even when the observations are

sparse, Clogg *et al.* (1991) considered a data-dependent conjugate prior

$$\pi(\theta) \propto \prod_{j=1}^{n} \left(\frac{e^{x_j'\theta}}{1+e^{x_j'\theta}} \right)^{k\hat{p}/n} \left(\frac{1}{1+e^{x_j'\theta}} \right)^{k(1-\hat{p})/n},$$

where k is the dimension of θ and $\hat{p} = \sum_{j=1}^{n} z_j / \sum_{j=1}^{n} m_j$ is the marginal proportion of "successes" in the data. The posterior under this prior is identical to the likelihood if we had observed an additional $k\hat{p}/n$ successes and $k(1-\hat{p})/n$ failures for each possible predictor vector x_j. As long as $0 < \hat{p} < 1$, which is obviously necessary for this logistic regression model to be sensible, there will be positive values in every cell of the full $n \times 2$ contingency table formed by the data and the pseudo-data, guaranteeing finite posterior modes. Clogg *et al.* (1991) partly justified this data-dependent prior by drawing an analogy with James–Stein estimation, in which the mean squared error for estimating a multivariate normal mean vector can be improved by shrinking the maximum likelihood estimate toward the mean of the observed components.

2.2.3 CATALYTIC PRIORS

Our definition of catalytic priors provides a generalization of the approach of Clogg *et al.* (1991) that can be implemented in more complicated settings. A catalytic prior will shrink the full model estimates toward an estimated simpler model, in a manner that will be made more precise in Section 2.4.

Definition 1. A *catalytic prior* for a Bayesian model $\{f(y \mid \theta), f(\theta)\}$ given a prior generating model $\{g(y \mid \phi), g(\phi)\}$ is

$$\pi_{\tau,y}(\theta) \propto f(\theta) \exp\left\{ \tau \int g(y_{\mathrm{p}} \mid y) \log f(y_{\mathrm{p}} \mid \theta) dy_{\mathrm{p}} \right\}, \tag{2.1}$$

where $g(y_{\mathrm{p}} \mid y) = \int g(y_{\mathrm{p}} \mid \phi) g(\phi \mid y) d\phi$ is the posterior predictive distribution for y_{p}, and τ

is a tuning parameter. (The subscript 'p' is meant to indicate that y_{p} is used to create pseudo-data.)

To perform inference under a catalytic prior, we combine it with the likelihood of the full model to obtain the posterior distribution

$$\pi_\tau(\theta \mid y) \propto f(y \mid \theta)\pi_{\tau,y}(\theta).$$

Because of the appearance of $g(y_{\mathrm{p}} \mid y)$ in (2.1), catalytic priors depend on the observed data, and so depart from strict Bayesian principles. We believe it is still useful to refer to them as "priors" because of the role they play in regularizing complicated models. However, viewed as prior distributions, they do not result in coherent inferences, in that if the data y are partitioned into (y_1, y_2), observing first y_1 and then y_2 will give different results than observing first y_2 and then y_1, which will in turn differ from observing (y_1, y_2) together. These give different results because different data will be used to fit the prior generating model g.

Because of the inclusion of the factor $f(\theta)$ in (2.1), catalytic priors are invariant to one-to-one reparameterizations. Letting J be the Jacobian matrix of the transformation from θ to ψ,

$$\begin{aligned}
\pi_{\tau,y}(\psi) &= \pi_{\tau,y}\{\theta(\psi)\}|\det J| \\
&\propto \exp\left[\tau \int g(y_{\mathrm{p}} \mid y) \log f\{y_{\mathrm{p}} \mid \theta(\psi)\} dy_{\mathrm{p}}\right] f\{\theta(\psi)\}|\det J| \\
&= \exp\left\{\tau \int g(y_{\mathrm{p}} \mid y) \log f(y_{\mathrm{p}} \mid \psi) dy_{\mathrm{p}}\right\} f(\psi).
\end{aligned}$$

The propriety of $\pi_{\tau,y}(\theta)$ is often especially easy to verify in exponential families. Under mild regularity conditions, if $f(y \mid \theta)$ is a natural exponential family with sufficient statistic $T(y)$, then Theorem 1 of Diaconis and Ylvisaker (1979) implies that when $\tau > 0$ and $E_g\{T(y_{\mathrm{p}})|y\} =$

$\int g(y_{\text{p}} \mid y)T(y_{\text{p}})dy_{\text{p}}$ is not on the boundary of the sample space, we have

$$\int \exp\left\{\tau \int g(y_{\text{p}} \mid y) \log f(y_p \mid \theta)dy_{\text{p}}\right\} d\theta < \infty.$$

2.2.4 EXAMPLES OF CATALYTIC PRIORS

Example 2.2.1 (continued). Suppose the default prior $f(\theta) \propto 1$, and suppose $\log f(y_{\text{p}} \mid \theta) = T(y_{\text{p}})'B(\theta) - C(\theta)$, ignoring terms that do not depend on θ. If we set $\tau = n_0$, then any prior generating model such that $E_g\{T(y_{\text{p}}) \mid y\} = \mu_0$ for all y (for example, we may choose g so that $g(y_{\text{p}} \mid y) = g(y_{\text{p}})$) will yield the conjugate prior

$$\pi(\theta) \propto e^{n_0\{\mu_0' B(\theta) - C(\theta)\}}$$

as the catalytic prior $\pi_{\tau,y}(\theta)$. For example, to obtain a Beta(a, b) prior for a Binomial(n, θ) observation, we could let $f(y_{\text{p}} \mid \theta)$ be Bernoulli(θ), set $\tau = a + b$, and choose g so that $E_g(y_{\text{p}}|y) = a/(a + b)$.

Example 2.2.2 (continued). Again, suppose the default prior $f(\theta) \propto 1$, and also suppose that under the prior generating model the likelihood is

$$z_j \mid \phi, m_j, x_j \sim \text{Binomial}(m_j, \phi),$$

where ϕ is a scalar probability of success, so that $g(z_j \mid \phi, m_j, x_j) = g(z_j \mid \phi, m_j)$. Further, consider the prior $g(\phi) \propto \phi^{-1}(1 - \phi)^{-1}$. Under g, the posterior distribution of ϕ is Beta$(\sum_{j=1}^n z_j, \sum_{j=1}^n (m_j - z_j))$. Now, suppose $z_{\text{p}} = (z_{\text{p}1}, \ldots, z_{\text{p}n})$ is a vector of binary pseudo-responses. Thus,

$$\log f(z_{\text{p}} \mid \theta, x) = \text{const} + \sum_{j=1}^{n} z_{\text{p}j} x_j'\theta - \sum_{j=1}^{n} \log(1 + e^{x_j'\theta}),$$

and, for all j,

$$\begin{aligned} E_g(z_{\mathrm{p}j} \mid z, m) &= E_g(\phi \mid z, m) \\ &= \sum_{j=1}^{n} z_j / \sum_{j=1}^{n} m_j. \end{aligned}$$

Therefore, the approach of Clogg *et al.* (1991) can be viewed as a catalytic prior with the prior generating model given above and $\tau = k/n$.

Example 2.2.3. Suppose the full model for $y = (y_1, \ldots, y_n)$ is

$$y \mid \theta \sim N(\theta, \sigma^2 I_n), \tag{2.2}$$

where $\theta = (\theta_1, \ldots, \theta_n)$, I_n is the $n \times n$ identity matrix, and the variance σ^2 is for now assumed known. The celebrated James–Stein estimator

$$\hat{\theta}_i^{\mathrm{JS}} = \bar{y} + \left(1 - \frac{\sigma^2(n-3)}{\sum_{j=1}^{n}(y_j - \bar{y})^2}\right)(y_i - \bar{y}),$$

where $\bar{y}$ is the sample mean, has lower quadratic risk than the maximum likelihood estimate $\hat{\theta}_i^{\mathrm{MLE}} = y_i$ if $n \geq 4$, and has a nice empirical Bayes justification; see James and Stein (1961), Stein (1962, particularly Lindley's discussion), and Efron and Morris (1973), among many others.

To derive a catalytic prior for (2.2) that can be directly compared with the James–Stein estimator, we assume that the prior generating model is $y_{\mathrm{p}} \sim N(\mu_g \mathbf{1}_n, \sigma_g^2 I_n)$, where $\mathbf{1}_n$ is the n-dimensional vector of ones. For simplicity, we will treat σ_g^2 as known (assumptions we make on σ_g^2 do not affect the catalytic prior, so it does not matter if we assume σ_g^2 to be known or unknown), and we will let $g(\mu_g) \propto 1$. Because we assume that σ^2 in the full model is known, the sufficient statistic for a pseudo-dataset y_{p} under the full model is just y_{p}. Under the prior

generating model, the posterior predictive distribution y_{p} is $y_{\text{p}} \mid y \sim N(\bar{y}\mathbf{1}_n, \sigma_g^2(1+1/n)I_n)$. Therefore, assuming $f(\theta) \propto 1$, the catalytic prior is

$$\pi_{\tau,y}(\theta) \propto \exp\left\{-\frac{\tau}{2\sigma^2}\sum_{i=1}^{n}(\theta_i - \bar{y})^2\right\},$$

corresponding to $\theta \sim N(\bar{y}\mathbf{1}_n, \tau^{-1}\sigma^2 I_n)$. The posterior mean of θ under this model is

$$E_\tau(\theta \mid y) = \bar{y}\mathbf{1}_n + \left(1 - \frac{\tau}{1+\tau}\right)(y - \bar{y}\mathbf{1}_n),$$

so that the posterior mean agrees with the James–Stein estimator if

$$\tau = \left\{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sigma^2(n-3)} - 1\right\}^{-1}.$$

2.3 Imputation perspective

Viewing prior distributions as pseudo-data inspires us to treat prior specification as a missing data problem. We envision a complete dataset

$$\{(y_1, w_1), \ldots, (y_n, w_n), (y_{\text{p}1}, w_{\text{p}1}), \ldots, (y_{\text{p}m}, w_{\text{p}m})\},$$

where $y = (y_1, \ldots, y_n)$ are observations given weights $w = (w_1, \ldots, w_n)$, and the missing values $y_{\text{p}} = (y_{\text{p}1}, \ldots, y_{\text{p}m})$ are pseudo-observations given weights $w_{\text{p}} = (w_{\text{p}1}, \ldots, w_{\text{p}m})$. We will assume that $w_i = 1$ and $w_{\text{p}j} < 1$ for all i and j, downweighting the pseudo-data so that they have less influence on our final inferences. One simple way to incorporate the weights w_i is to set $f(y_i \mid w_i, \theta) = \{f(y_i \mid \theta)\}^{w_i}$, where $f(y_i \mid \theta)$ is the likelihood for y_i assuming unit weight. If we had observed the pseudo-data y_{p}, we could use the likelihood $f(y \mid w, \theta)$ of our

full model and a default prior $f(\theta)$ to obtain the posterior distribution

$$f(\theta \mid \{y, w\}, \{y_{\text{p}}, w_{\text{p}}\}) \propto \left\{ \prod_{i=1}^{n} f(y_i \mid w_i, \theta) \right\} \left\{ \prod_{j=1}^{m} f(y_{\text{p}j} \mid w_{\text{p}j}, \theta) \right\} f(\theta). \tag{2.3}$$

Because the pseudo-data y_{p} are not observed, we generate them under the prior generating model g. A single realization of y_{p} would correspond to a fixed prior distribution on θ, proportional to $f(\theta) \prod_{j=1}^{m} f(y_{\text{p}j} \mid w_{\text{p}j}, \theta)$. Here, we average over multiple realizations of y_{p}, drawn from the posterior predictive distribution $g(y_{\text{p}} \mid y)$ under the prior generating model. That is, we average $f(\theta \mid \{y, w\}, \{y_{\text{p}}, w_{\text{p}}\})$ over $g(y_{\text{p}} \mid y)$, yielding a posterior distribution

$$p(\theta \mid y) = \int f(\theta \mid \{y, w\}, \{y_{\text{p}}, w_{\text{p}}\}) g(y_{\text{p}} \mid y) dy_{\text{p}}. \tag{2.4}$$

This procedure is similar to multiple imputation (Rubin, 1987; Little and Rubin, 2002). In the multiple imputation literature, the roles of the imputer and the analyst are distinct. For example, given a survey with missing responses, the imputer may fill in these missing values multiple times using draws from a probabilistic model. Then, multiply-imputed surveys can be released to the analyst, who can analyze each of them with standard complete-data methods and combine the results using rules that lead to valid inferences. Here, we view y_{p} as missing data and multiply impute it under the prior generating model $g(y_{\text{p}} \mid y)$. However, there are at least two differences between this setting and traditional multiple imputation for missing data. First, the pseudo-data are merely a useful construction, not real missing observations. Second, we intentionally choose a prior generating (imputation) model g that is simpler than the full (analysis) model f in order to constrain the inferences under the full model. In multiple imputation for missing data, if the imputer assumes a simpler model than the analyst, the analyst may be unable to obtain valid inferences (Meng, 1994). Here, without the additional constraints provided by the observations imputed under a simpler model, the

analyst may be unable to fit the full model at all.

Of course, (2.4) does not agree with the catalytic prior of Definition 1. The derivation leading to (2.4) assumed that the pseudo-dataset had a fixed size m, but it is not clear how to choose a good value of m. We can consider a high-resolution version of (2.4) by letting $m \to \infty$. Of course, to avoid overwhelming the observed data, the weights given to a dataset with larger m should be reduced. We can do this by setting $w_{\mathrm{p}j} = \tau/m$ for all j. As suggested above, we can incorporate the weights by setting

$$f(y_{\mathrm{p}j} \mid w_{\mathrm{p}j}, \theta) = \{f(y_{\mathrm{p}j} \mid \theta)\}^{w_{\mathrm{p}j}} = \{f(y_{\mathrm{p}j} \mid \theta)\}^{\tau/m}.$$

Then, by the law of large numbers,

$$\begin{aligned} \log \prod_{j=1}^{m} f(y_{\mathrm{p}j} \mid w_{\mathrm{p}j}, \theta) &= \frac{\tau}{m} \sum_{j=1}^{m} \log f(y_{\mathrm{p}j} \mid \theta) \\ &\to \tau \int g(y_{\mathrm{p}} \mid y) \log f(y_{\mathrm{p}} \mid \theta) dy_{\mathrm{p}} \end{aligned}$$

as $m \to \infty$, assuming the $y_{\mathrm{p}j}$ are independent and identically distributed draws from $g(y_{\mathrm{p}} \mid y)$. This inspires the posterior distribution

$$\pi_\tau(\theta \mid y) \propto f(y \mid \theta) f(\theta) \exp\left\{\tau \int g(y_{\mathrm{p}} \mid y) \log f(y_{\mathrm{p}} \mid \theta) dy_{\mathrm{p}}\right\}. \tag{2.5}$$

which in turn motivates our definition of catalytic priors.

Clogg *et al.* (1991) used catalytic priors in logistic regression to multiply impute 1980 U.S. Census industry and occupation codes for 1970 public use samples. They wanted to use a flexible imputation model so that a wide range of valid analyses could be performed on the imputed datasets. Thus, rather than dropping covariates and using more parsimonious logistic regression models, they retained all plausibly relevant predictors and introduced weak

regularization through catalytic priors. Because their application was multiple imputation, their goal was prediction; in particular, they wanted to obtain predictive distributions that were appropriately diffuse when the data provided little information. In the next section, we will investigate how catalytic priors constrain the predictive consequences of the final posterior predictive distribution $\pi_\tau(\tilde{y} \mid y)$.

2.4 Information perspective

We can rewrite the logarithm of (2.5) as

$$\log \pi_\tau(\theta \mid y) = \text{const} + \log f(y \mid \theta) + \log f(\theta) - \tau D_y(\theta), \tag{2.6}$$

where

$$D_y(\theta) = \text{KL}\{g(y_\text{p} \mid y), f(y_\text{p} \mid \theta)\} = \int g(y_\text{p} \mid y) \log \frac{g(y_\text{p} \mid y)}{f(y_\text{p} \mid \theta)} dy_\text{p}$$

is the Kullback–Leibler divergence (Kullback and Leibler, 1951) from the posterior predictive distribution under the prior generating model to the likelihood of the full model. Thus, (2.5) can be interpreted as adding a penalty to the original log posterior that keeps the likelihood under the full model from diverging too far from the estimated predictive distribution under the simpler model.

Defining $\hat{\theta}_\tau \equiv \arg\max_\theta \log \pi_\tau(\theta \mid y)$ and observing that

$$\begin{aligned}
\hat{\theta}_\tau &= \arg\max_\theta \{\log f(y \mid \theta) + \log f(\theta) - \tau D_y(\theta)\} \\
&= \arg\max_\theta \left\{ \frac{1}{1+\tau} \log f(y \mid \theta) + \frac{1}{1+\tau} \log f(\theta) - \frac{\tau}{1+\tau} D_y(\theta) \right\},
\end{aligned}$$

we immediately see that $\hat{\theta} \equiv \lim_{\tau \to \infty} \hat{\theta}_\tau = \arg\min_\theta D_y(\theta)$. Under regularity conditions,

$$\frac{\partial}{\partial \theta} D_y(\theta)\bigg|_{\theta=\hat{\theta}} = E_{g(y_\mathrm{p}|y)} \left\{ -\frac{\partial \log f(y_\mathrm{p} \mid \theta)}{\partial \theta}\bigg|_{\theta=\hat{\theta}} \right\} = 0,$$

assuming it is valid to interchange differentiation and integration. Under further conditions, as τ increases, the posterior distribution of $\tau^{1/2}(\theta - \hat{\theta})$ converges to a normal distribution with mean zero and inverse covariance matrix

$$\frac{\partial^2}{\partial \theta \partial \theta'} D_y(\theta)\bigg|_{\theta=\hat{\theta}} = E_{g(y_\mathrm{p}|y)} \left\{ -\frac{\partial^2 \log f(y_\mathrm{p} \mid \theta)}{\partial \theta \partial \theta'}\bigg|_{\theta=\hat{\theta}} \right\}.$$

See Johnson (1967) for a rigorous development of the normal asymptotics in the case of a scalar parameter. Thus, in the limit $\tau \to \infty$, the posterior shrinks to a point mass on $\hat{\theta}$, and $f(y_\mathrm{p} \mid \hat{\theta})$, plugging in $\hat{\theta}$ for θ, is the Kullback–Leibler projection of the prior generating model's estimated predictive distribution onto the full model family. This is reminiscent of the well-known result that the maximum likelihood estimate under the wrong model converges to the parameter value that minimizes the Kullback–Leibler divergence from the true model to the parameterized family under which the MLE is calculated.

We now consider the posterior predictive distribution of future data $\tilde{y}$ under catalytic priors. Denoting this posterior predictive distribution

$$\pi_\tau(\tilde{y} \mid y) = \int f(\tilde{y} \mid \theta) \pi_\tau(\theta \mid y) d\theta,$$

by Jensen's inequality we have

$$\mathrm{KL}\left\{h(\tilde{y}), \pi_\tau(\tilde{y} \mid y)\right\} \leq E_{\pi_\tau(\theta|y)} \left[\mathrm{KL}\left\{h(\tilde{y}), f(\tilde{y} \mid \theta)\right\}\right]. \tag{2.7}$$

If we replace $h(\tilde{y})$ by $g(\tilde{y} \mid y)$, and if $\tilde{y}$ has the same sampling distribution as y_p under both f

and g, we obtain

$$\text{KL}\left\{g(\tilde{y} \mid y), \pi_\tau(\tilde{y} \mid y)\right\} \leq E_{\pi_\tau(\theta|y)}\left\{D_y(\theta)\right\}. \tag{2.8}$$

On the other hand, if $\tilde{y}$ is not restricted to have the same sampling distribution as y_{p}, and if we view $h(\tilde{y})$ in (2.7) as the true distribution generating future data $\tilde{y}$, then (2.7) is an argument in favor of estimating $h(\tilde{y})$ with the posterior predictive distribution, rather than the plug-in estimate $f(\tilde{y} \mid \theta^*)$, for θ^* drawn from $\pi_\tau(\theta \mid y)$ (Akaike, 1978). Of course, there may still exist some θ^* such that $\text{KL}\left\{h(\tilde{y}), f(\tilde{y} \mid \theta^*)\right\} < \text{KL}\left\{h(\tilde{y}), \pi_\tau(\tilde{y} \mid y)\right\}$. However, Aitchison (1975) showed that for multivariate normal observations, the posterior predictive distribution under a flat prior on the mean uniformly dominates in Kullback–Leibler risk the distribution obtained by plugging in maximum likelihood estimates.

Aitchison (1975) also showed that a posterior predictive distribution $p(\tilde{y} \mid y)$ minimizes the Bayes risk of estimating the true distribution $f(\tilde{y} \mid \theta_{\text{true}})$ under Kullback–Leibler loss, where the Bayes risk averages over the prior distribution used to obtain $p(\tilde{y} \mid y)$. Aitchison (1975) interpreted this as the unsurprising result that if we really believe our prior, we should be Bayesian. However, here we do not necessarily believe our prior $\pi_{\tau,y}(\theta)$, but we would like to know whether the posterior distribution $\pi_\tau(\theta \mid y)$ and the posterior predictive distribution $\pi_\tau(\tilde{y} \mid y)$ are optimal under any other considerations. This turns out to be true, and the result is closely related to maximum entropy methods as in Jaynes (1968) and to variational Bayesian approximations.

Theorem 2. *Suppose* $\int f(y \mid \theta) f(\theta) e^{-\tau D_y(\theta)} d\theta < \infty$ *if* $\tau > 0$. *Then the posterior* $\pi_{\tau(c)}(\theta \mid y)$ *under the catalytic prior is the solution to the constrained optimization problem*

$$\textit{minimize} \quad \int \pi(\theta \mid y) \log \frac{\pi(\theta \mid y)}{f(y \mid \theta) f(\theta)} d\theta \quad \textit{subject to} \quad E_{\pi(\theta|y)}\{D_y(\theta)\} = c.$$

The constraint implies that

$$KL\{g(y_p \mid y), \pi_{\tau(c)}(y_p \mid y)\} \leq c. \tag{2.9}$$

Proof. See, for example, Kullback (1968, p. 37–38), but note that the proof is valid even if $\int f(y \mid \theta) f(\theta) d\theta = \infty$. Also, see Shore and Johnson (1981). Finally, (2.9) is a consequence of (2.8). □

It is easy to show that

$$\frac{\partial}{\partial \tau} E_{\pi_\tau(\theta|y)}\{D_y(\theta)\} = -\text{Var}_{\pi_\tau(\theta|y)}\left\{D_y(\theta)\right\},$$

so that if $\text{Var}_{\pi_\tau(\theta|y)}\left\{D_y(\theta)\right\} > 0$ for all τ, then $E_{\pi_\tau(\theta|y)}\{D_y(\theta)\}$ is strictly decreasing in τ. Thus, $c(\tau) = E_{\pi_\tau(\theta|y)}\{D_y(\theta)\}$ is a one-to-one function of τ. Valid choices of c must lie between $\lim_{\tau\to\infty} c(\tau) = D_y(\hat{\theta})$ and $\lim_{\tau\to 0} c(\tau)$, to guarantee that the corresponding $\tau = \tau(c)$ will fall within $(0, \infty)$.

Theorem 2 means that if $f(y) = \int f(y \mid \theta) f(\theta) d\theta < \infty$, then the posterior under the catalytic prior can be viewed as the best approximation of the posterior

$$f(\theta \mid y) = f(y \mid \theta) f(\theta) / f(y),$$

subject to a constraint on the expectation of the Kullback–Leibler divergence from the posterior predictive distribution $g(y_{\text{p}} \mid y)$ to the full model likelihood $f(y_{\text{p}} \mid \theta)$. On the other hand, if the posterior distribution under the full model is improper, meaning that $\int f(y \mid \theta) f(\theta) d\theta = \infty$, then we are finding the best probability distribution to approximate the measure $\mu(A) = \int_A f(y \mid \theta) f(\theta) d\theta$, again subject to a predictive constraint. Although we interpret this constraint as restricting candidate posteriors to a family of models that lead to predictive distributions not too far from a simple but reliably estimated predictive distribution, we

do not optimize under an inequality constraint on $\text{KL}\{g(y_\text{p} \mid y), \pi_\tau(y_\text{p} \mid y)\}$. This latter expression, as a divergence from one posterior predictive distribution to another, may be easier to interpret than the posterior expectation $E_{\pi_\tau(\theta|y)}\{D_y(\theta)\}$. By optimizing instead under the stricter constraint on $E_{\pi_\tau(\theta|y)}\{D_y(\theta)\}$, we both obtain a tractable solution that can be implemented in complex models and guarantee a bound on the more interpretable divergence $\text{KL}\{g(y_\text{p} \mid y), \pi_\tau(y_\text{p} \mid y)\}$.

We can derive the catalytic prior analogously by trying to approximate the prior $f(\theta)$ subject to a constraint on the prior expectation $E_{\pi_{\tau,y}(\theta)}\{D_y(\theta)\}$. The proof of the following result is essentially identical to that of Theorem 2. Of course, in general, $E_{\pi_{\tau,y}(\theta)}\{D_y(\theta)\} \neq E_{\pi_\tau(\theta|y)}\{D_y(\theta)\}$, so for the same value of τ, the values of c in Theorems 2 and 2* will differ.

Theorem 2*. *Suppose $\int f(\theta)e^{-\tau D_y(\theta)}d\theta < \infty$ if $\tau > 0$. The catalytic prior $\pi_{\tau(c),y}(\theta)$ is the solution to the constrained optimization problem*

$$\textit{minimize} \quad \int \pi(\theta) \log \frac{\pi(\theta)}{f(\theta)} d\theta \quad \textit{subject to} \quad E_{\pi(\theta)}\{D_y(\theta)\} = c.$$

The constraint implies that $\text{KL}\{g(y_p \mid y), m_{\tau(c),y}(y_p)\} \leq c$, where

$$m_{\tau(c),y}(\cdot) = \int f(\cdot \mid \theta)\pi_{\tau(c),y}(\theta)d\theta \tag{2.10}$$

is the prior predictive density or marginal likelihood.

The optimizations in Theorems 2 and 2* are closely related to variational Bayesian methods in which an approximation is sought of a computationally intractable posterior distribution $p(\theta \mid y)$. A family $\mathcal{Q}$ of more tractable densities $q(\theta)$ is chosen, and we approximate the posterior distribution of interest via

$$p^*(\theta \mid y) = \arg\min_{q \in \mathcal{Q}} \text{KL}\{q(\theta), p(\theta \mid y)\}.$$

Though they minimize the same objective function, catalytic priors have a different aim. The computational tractability of the full-model posterior $f(\theta \mid y)$ is not the issue; instead, we are concerned about possible posterior impropriety, noisy behavior, and unreasonable parameter estimates, all of which can result in predictive distributions for future data that conflict with our scientific knowledge. Thus, catalytic priors are designed to restrict the family of candidate approximate posteriors for the purpose of regularization.

The form of the catalytic prior is very similar to maximum entropy priors derived by Jaynes (1968) under constraints on prior expectations of functions of θ, where for continuous θ the entropy is defined as $-\int \pi(\theta) \log\{\pi(\theta)/\pi_0(\theta)\} d\theta$, with $\pi_0(\theta)$ an "invariant measure" intended to be a noninformative prior for θ. This is somewhat ambiguous, since it is often not clear how to choose a noninformative prior $\pi_0(\theta)$ (Berger, 1985, p. 92). Catalytic priors do not suffer from this ambiguity because no attempt is made to create a formal rule leading to an objective prior.

Bernardo's prior (Bernardo, 1979; Berger *et al.*, 2009) also originates in information theoretic considerations, but there are several crucial differences between Bernardo's prior and catalytic priors. Bernardo's prior is motivated by an attempt to maximize the expected Kullback–Leibler divergence from the posterior to the prior under the same model, where the expectation is taken over the marginal distribution of the data, while the catalytic prior attempts to shrink toward an estimated predictive distribution under a simpler model. There is only one likelihood at play in the Bernardo prior, while the catalytic prior relies on two different likelihoods, one for the full model and one for the simpler prior generating model. In contrast to the catalytic prior, Bernardo's prior has no tuning parameters. Finally, unlike Bernardo's prior (see Berger and Bernardo, 1992), the catalytic prior does not involve a distinction between parameters of interest and nuisance parameters.

Statistical physicists call the normalizing constant $Z(\beta) = \int f(\theta) e^{\beta D_y(\theta)} d\theta$ the partition function, where we have substituted $\beta = -\tau$ to agree with standard notation in physics, and

it is well known and easy to show that

$$E_{\pi_{\tau,y}(\theta)}\{D_y(\theta)\} = \frac{d}{d\beta}\log Z(\beta).$$

This immediately leads to the following result, which is perhaps more statistically interpretable and is closely related to path sampling for computing ratios of normalizing constants (Gelman and Meng, 1998).

Result 1. *Suppose* $\int f(\theta)e^{-\tau D_y(\theta)}d\theta < \infty$*. Then the difference between the posterior and prior expectations of the Kullback–Leibler divergence* $D_y(\theta)$ *satisfies*

$$E_{\pi_\tau(\theta|y)}\{D_y(\theta)\} - E_{\pi_{\tau,y}(\theta)}\{D_y(\theta)\} = -\frac{\partial}{\partial\tau}\log m_{\tau,y}(y), \tag{2.11}$$

where the derivative is evaluated at the value of τ *used in the catalytic prior, and* $m_{\tau,y}(\cdot)$ *is the prior predictive density defined in (2.10).*

Proof. Assuming we can interchange integration and differentiation, (2.11) can be verified directly by differentiating

$$\log m_{\tau,y}(y) = \log\int f(y \mid \theta)f(\theta)e^{-\tau D_y(\theta)}d\theta - \log\int f(\theta)e^{-\tau D_y(\theta)}d\theta.$$

□

Intuitively, if increasing τ decreases $\log m_{\tau,y}(y)$, then the observed data y do not favor the values of θ that make $f(y_{\text{p}} \mid \theta)$ close to $g(y_{\text{p}} \mid y)$. It is therefore unsurprising that, in such settings, the posterior expectation of the divergence of $f(y_{\text{p}} \mid \theta)$ from $g(y_{\text{p}} \mid y)$ is larger than the prior expectation.

2.5 Discussion

The power priors of Ibrahim and Chen (2000), Ibrahim *et al.* (2003), and Chen and Ibrahim (2006), among others, have some similarities to our approach. Power priors are proportional to the likelihood of a historical dataset, raised to a power a_0 between 0 and 1. Ibrahim and Chen (2000) suggest a hierarchical specification in which a_0 is given a prior distribution. The posterior (2.3) that conditions on a particular realized pseudo-dataset $\{y_\text{p}, w_\text{p}\}$ is similar to a power prior with fixed a_0, if we interpret y_p as historical data and we set $w_{\text{p}j} = a_0$ for all j. Of course, y_p are not actually historical data and must be simulated from a model in our approach. Ibrahim and Chen (2000, Section 7) briefly discuss the scenario in which historical data $D_0 = (n_0, y_0, X_0)$ are not available, suggesting that prior prediction can be used to obtain y_0. The focus on prior prediction differs from our approach of using posterior predictions under another fitted model.

As mentioned in Section 2.2.3, our definition of catalytic priors was inspired by Clogg *et al.* (1991). Heinze and Schemper (2002) and Galindo-Garre *et al.* (2004) compared Clogg *et al.*'s approach to alternative methods, including Jeffreys's prior. Heinze and Schemper (2002) found that the penalized likelihood method of Firth (1993) had lower bias than Clogg *et al.*'s method, but this is unsurprising since Firth's method is explicitly designed to reduce bias. We are not willing to dismiss Clogg *et al.*'s approach on the basis of this study because in practice we are often interested in criteria other than bias. In the Monte Carlo simulations of Galindo-Garre *et al.* (2004), Jeffreys's prior and the Clogg *et al.*'s approach performed almost identically in terms of median squared error, interval coverage, and interval width, and the authors concluded that these two priors "may be the most recommendable in general settings" (Galindo-Garre *et al.*, 2004, p. 113). Heinze (2006) also discussed methods for separation in logistic regression, recommending penalized profile likelihood, but did not include a comparison to Clogg *et al.*'s approach. Zorn (2005, p. 162) mentioned Clogg *et al.*'s method

but dismissed it in an apparent misreading of Galindo-Garre *et al.* (2004):

> In addition to the ad hoc nature of [Clogg *et al.*'s] solution, both Heinze and Schemper (2002, p. 2413–2414) and Galindo-Garre *et al.* (2004) demonstrate conclusively via Monte Carlo simulations that Clogg *et al.*'s approach is inferior to other available alternatives; accordingly, I do not discuss it any further here.

This statement is not supported by the results in the papers cited. We believe it is important to push back against this claim because we do not want it to become conventional wisdom, leading practitioners to ignore Clogg *et al.*'s approach. In Chapter 3, we will show through simulations that Clogg *et al.*'s catalytic priors often perform extremely well. In addition, we believe that the work in this thesis, by placing catalytic priors on a principled foundation, should reassure practitioners tempted to dismiss Clogg *et al.*'s approach as ad hoc.

This chapter has proposed a general strategy for constructing prior distributions that ensure shrinkage toward an estimated simpler model. We expect this strategy to be most useful in complex problems in which a simpler plausible model is available but other methods for constructing priors are either intractable or lead to ill-behaved posterior distributions. While we have focused on the case in which the simpler model is a Bayesian model that leads to a posterior predictive distribution $g(y_{\text{p}} \mid y)$, it would be possible to consider a more general set-up in which the expectation of $\log f(y_{\text{p}} \mid \theta)$ is calculated with respect to any estimated predictive distribution $\hat{g}(y_{\text{p}})$, not necessarily a Bayesian posterior predictive distribution. For instance, instead of using $g(y_{\text{p}} \mid y)$, it may sometimes be much more tractable to use $g(y_{\text{p}} \mid \hat{\phi})$, plugging in the maximum likelihood estimate $\hat{\phi}$ under $g(y \mid \phi)$ for the parameter ϕ, and in some cases there may be little practical difference between the results under the plug-in and posterior predictive approaches. Investigating these issues more closely is left to future work.

3

Catalytic priors: Specific models

3.1 Introduction

In this chapter, we develop and evaluate catalytic priors for a variety of parametric models. In Section 3.2, we will examine a simple two-state Markov process. In Section 3.3, we explore how to apply catalytic priors to regression models that include covariates, applying this approach to logistic regression in Section 3.4 and to linear regression in Section 3.5. To illustrate how to apply catalytic priors when the observed-data model is a latent variable model, we examine latent variable and multilevel models in Section 3.6, and Section 3.7 extends this

to a more complicated mixture model in a preliminary analysis of the effectiveness of a job training program.

3.2 TWO-STATE MARKOV PROCESS

Suppose we observe a two-state continuous-time Markov process $x(t)$ at regular time intervals. Such a process can arise in biophysics, for example when studying single-molecule dynamics (Kou *et al.*, 2005). Let the states of the chain be 0 and 1 and the interval between observations be δ. We can write the infinitesimal generator of the chain as

$$\begin{pmatrix} -k_{01} & k_{01} \\ k_{10} & -k_{10} \end{pmatrix}.$$

Sojourn times in states 0 and 1 are independently exponentially distributed with means k_{01}^{-1} and k_{10}^{-1}, respectively, and the stationary distribution of this chain has probability $\mu = k_{01}/(k_{01} + k_{10})$ of being in state 1. Letting $\lambda = k_{01} + k_{10}$, the transition probabilities $P_{ij,t}(\mu, \lambda) \equiv \Pr(x(t) = j \mid x(0) = i)$ are

$$\begin{aligned} P_{00,t}(\mu, \lambda) &= (1 - \mu) + \mu e^{-\lambda t} \\ P_{01,t}(\mu, \lambda) &= \mu(1 - e^{-\lambda t}) \\ P_{10,t}(\mu, \lambda) &= (1 - \mu)(1 - e^{-\lambda t}) \\ P_{11,t}(\mu, \lambda) &= \mu + (1 - \mu)e^{-\lambda t}\,, \end{aligned}$$

where the notation is chosen to emphasize the dependence of the transition probabilities on the parameters μ and λ. We observe $y_i = x(i\delta)$, $i = 0, 1, \ldots, n$, which can be viewed as a realization of a discrete-time Markov chain with transition probabilities that are functions of

the parameters of interest. Conditioning on y_0, we obtain the full likelihood

$$f(y \mid \mu, \lambda) = \prod_{(i,j) \in \mathcal{A}} P_{ij,\delta}(\mu, \lambda)^{N_{ij}} ,$$

where $\mathcal{A} = \{0, 1\} \times \{0, 1\}$ and N_{ij} is the number of observed switches from state i to state j.

If δ is too large, then it may be possible to accurately infer μ, but the likelihood will be relatively flat in λ. Indeed, as $\delta \to \infty$, the process becomes a sequence of independent and identically distributed Bernoulli(μ) observations, with no information on λ. On the other hand, if the time interval $n\delta$ is too short relative to k_{01}^{-1} and k_{10}^{-1}, we will not observe many switches between states, which can make it hard to reliably estimate both μ and λ via maximum likelihood. In fact, if no switches between states are observed, then both maximum likelihood and posterior estimates under catalytic priors will fail. This is analogous to trying to fit a binary logistic regression when all responses are equal.

Figure 3.1 displays the probability of observing at least one switch between states 0 and 1 as a function of the total observation time $n\delta$, fixing $k_{01} = 0.1$ and $k_{10} = 0.9$. In general, this probability is

$$1 - \mu \left\{P_{11,\delta}(\mu, \lambda)\right\}^n - (1 - \mu) \left\{P_{00,\delta}(\mu, \lambda)\right\}^n ,$$

assuming y_0 is drawn from the stationary distribution. Although in the likelihood we do not condition on observing at least one switch, in practice we would not analyze data with zero switches (assuming we only observe one dataset, so there is no opportunity for pooling information across multiple datasets). As seen in Figure 3.1, when $n\delta$ is too small, observing any switches is unlikely. In this regime, avoiding explicitly conditioning on observing at least one switch may lead to estimation problems because we will almost certainly observe very few switches, and the model will not properly account for the strength of evidence in such data for extremely low rates.

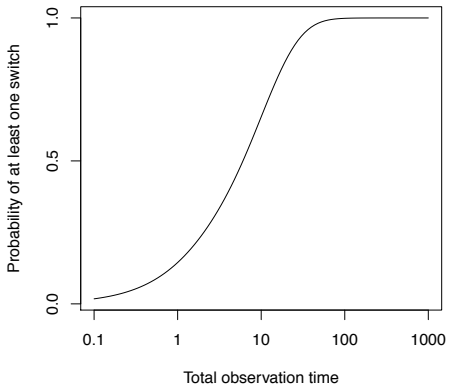


Figure 3.1: The probability of at least one switch between states in $n = 100$ observations as a function of the total observation time $n\delta$, assuming that $k_{01} = 0.1$ and $k_{10} = 0.9$.

For moderate δ, we can use a catalytic prior. For our prior generating model, we fix $\mu = 1/2$ as a plausible default choice, so we can write the likelihood under the prior generating model as $g(y \mid \lambda)$, depending only on λ. To compute the expected log likelihood of the full model under the posterior predictive distribution of this null model, we begin with the expectations $E_g(N_{\mathrm{p}ij} \mid \lambda)$ of the sufficient statistics with respect to $g(y_{\mathrm{p}} \mid \lambda)$, where $N_{\mathrm{p}ij}$ denotes the number of transitions from state i to state j in a pseudo-dataset of size n_{p}. The expected number of transitions from state i to state j can be computed by constructing the Markov chain on the augmented states $v_\ell = (y_\ell, y_{\ell+1})$. The stationary distribution of this chain is

$$\Pr\{v_\ell = (i, j)\} = \mu^i (1 - \mu)^{1-i} P_{ij,\delta}(\mu, \lambda).$$

Consequently,

$$E_g(N_{\mathrm{p}ij} \mid \lambda) = \begin{cases} \frac{n_{\mathrm{p}}}{4} \left(1 + e^{-\delta\lambda}\right), & i = j, \\ \frac{n_{\mathrm{p}}}{4} \left(1 - e^{-\delta\lambda}\right), & i \neq j. \end{cases}$$

Assuming a uniform default prior in the full model, so that $f(\mu, \lambda) \propto 1$, and letting

$$a_{ij} = \int E_g(N_{\text{p}ij} \mid \lambda) g(\lambda \mid y) d\lambda, \tag{3.1}$$

the catalytic prior yields the posterior

$$\pi_\tau(\mu, \lambda \mid y) \propto \prod_{(i,j) \in \mathcal{A}} \{P_{ij,\delta}(\mu, \lambda)\}^{N_{ij} + \tau a_{ij}}. \tag{3.2}$$

Evaluating the integral in (3.1) may not be straightforward. However, if the sample size is large enough, then under reasonable choices of the prior $g(\lambda)$ for the prior generating model, a_{ij} can be well approximated by

$$\tilde{a}_{ij} = E_g(N_{\text{p}ij} \mid \lambda_g = \hat{\lambda}_g), \tag{3.3}$$

where $\hat{\lambda}_g$ maximizes the likelihood $g(y \mid \lambda)$. Using $\tilde{a}_{ij}$ means we are shrinking toward the plug-in estimate of the predictive distribution, $g(y_{\text{p}} \mid \hat{\lambda}_g)$, rather than toward the posterior predictive distribution $g(y_{\text{p}} \mid y)$. This may be reasonable in many settings, particularly when the uncertainty in the maximum likelihood estimate is low and when the full posterior predictive distribution is intractable.

Conveniently, fixing $\mu = 1/2$ leads to a simple expression for the maximum likelihood estimate $\hat{\lambda}_g$:

$$\hat{\lambda}_g = -\frac{1}{\delta} \log \left(\frac{T - S}{n} \right),$$

where $S = N_{01} + N_{10}$ is the number of observed switches and $T = n - S$ is the number of transitions to the same state. This expression demonstrates that we will only run into problems when either $S = 0$, in which case we have no hope of fitting this model because we have observed no switching between states, or $T \leq S$, in which case the likelihood $g(y \mid \lambda)$ is

Table 3.1: Comparison of the simple plug-in estimate $\tilde{a}_{ij}$ with values of a_{ij} obtained via numerical integration under two different priors, the Jeffreys prior and λ^{-2}. Under the null model that sets $\mu = 1/2$, the pseudo-observations $a_{11} = a_{00}$ and $a_{10} = a_{01}$.

	a_{00}			a_{01}		
(T, S)	Plug-in	Jeffreys	λ^{-2}	Plug-in	Jeffreys	λ^{-2}
$(75, 25)$	37.50	37.38	37.96	12.50	12.62	12.04
$(90, 10)$	45.00	44.80	45.50	5.00	5.20	4.50
$(98, 2)$	49.00	48.76	49.50	1.00	1.24	0.50

monotonically increasing in λ and, because we have observed at least as much switching as persistent states, the model itself is perhaps not a sensible choice.

Table 3.1 compares the plug-in $\tilde{a}_{ij}$ to the a_{ij} given in (3.1) under the full posterior predictive distributions for two choices of the prior distribution $g(\lambda)$, for a range of choices of T and S and with n_{p} chosen to equal $n = 100$. Because $\lim_{\lambda \to \infty} g(y \mid \lambda) = 2^{-n} > 0$, the improper prior $g(\lambda) \propto \lambda^{-1}$, which might seem a natural choice, leads to an improper posterior distribution. Thus, we compare the prior λ^{-2} to the Jeffreys prior $e^{-\delta\lambda}(1 - e^{-2\delta\lambda})^{-1/2}$. The approximations $\tilde{a}_{ij}$ fall between the values under the two priors.

Figure 3.2 displays results from fitting this model to data simulated under the model, for a range of time intervals δ. The process was simulated using $k_{01} = 0.1$, $k_{10} = 0.9$, and $n = 100$. We set $\tau = 1/n_{\text{p}}$, equivalent to adding one total pseudo-count split among the four possible transitions. The likelihood was parameterized in terms of $(\log k_{01}, \log k_{10})$. Estimates were found via maximum likelihood and by maximizing the posterior (3.2), and standard errors were estimated from the Hession of the log likelihood or log posterior at the mode. Figure 3.2 compares the coverage and width of the intervals $\hat{\theta} \pm z_{\alpha/2}\hat{\text{se}}(\hat{\theta})$, where $z_{\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution and θ denotes either $\log k_{01}$ or $\log k_{10}$. When the total observation time is approximately in the range $(25, 350)$, the coverage of both intervals is satisfactory, but the posterior intervals are almost always shorter. At the lower end of this range, posterior intervals are shorter than likelihood-based intervals by an order of

magnitude or more, and the posterior modes have much lower mean squared error than the maximum likelihood estimates.

When $n\delta$ is larger than about 350, the coverage of both likelihood-based and posterior intervals drops sharply, with the posterior intervals doing notably worse, but as discussed above, when δ is too large, there is very little information about λ. If $n\delta$ is smaller than about 25, then we may observe few switches between states. While maximum likelihood estimates usually exist in these cases, they may perform quite poorly. For these cases, likelihood-based and posterior intervals have lower than nominal coverage, but in this regime, the posterior modes dominate the maximum likelihood estimates in mean squared error.

3.3 INTRODUCING COVARIATES

We often want to condition on observed covariates. To apply catalytic priors in this context, we assume that the data $y = (z, X)$, where z is a response vector and X is a matrix of predictors. Suppose the full model is a regression model and we are only interested in the conditional distribution of z given X. We assume that

$$f(z, X \mid \theta, \xi) = f(z \mid X, \theta) f(X \mid \xi),$$

where ξ are nuisance parameters not of scientific interest, and θ and ξ are distinct in the sense of Rubin (1976), that is, independent in the default prior, $f(\theta, \xi) = f(\theta) f(\xi)$, with no links implied by parameter space restrictions. The prior generating model is a joint model on z and X:

$$g(z, X \mid \phi) = g(z \mid X, \phi) g(X \mid \phi).$$

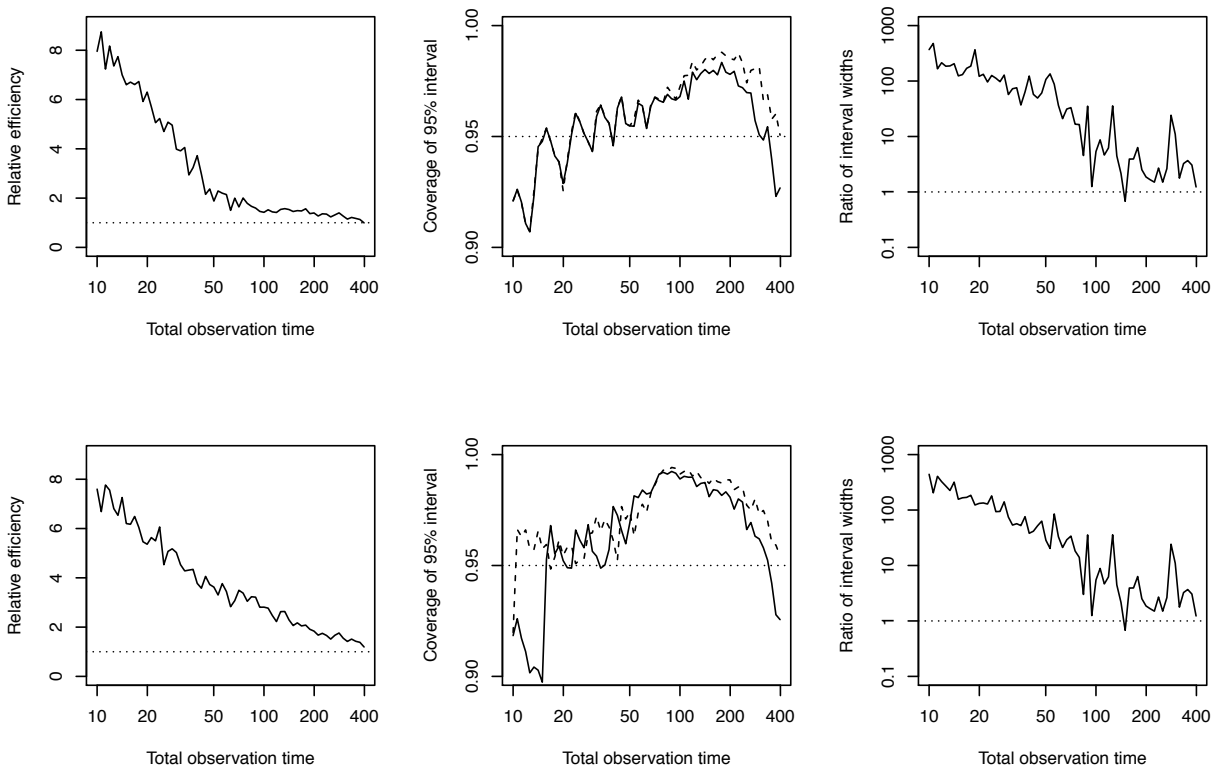


Figure 3.2: Simulation results for a continuous-time Markov process observed at regular time intervals. True values were $k_{01} = 0.1$ and $k_{10} = 0.9$. The process was simulated 5000 times for each value of the time δ between observations using $n = 100$ intervals. The top row gives results for $\log k_{01}$, the bottom row for $\log k_{10}$. The first column is the ratio of mean squared error for the maximum likelihood estimates to the mean squared error using the catalytic prior. The second column displays coverage of nominal 95% intervals, with the dashed line corresponding to intervals based on maximum likelihood, the solid line corresponding to posterior intervals using the catalytic prior. The third column gives the ratio of the average width of the likelihood-based intervals to the average width of the posterior intervals. Dotted lines are provided for reference.

Then we can construct the catalytic prior

$$\begin{aligned}\pi_{\tau,y}(\theta) &\propto f(\theta)\exp\left\{\tau\int g(z_{\mathrm{p}},X_{\mathrm{p}}\mid y)\log f(z_{\mathrm{p}},X_{\mathrm{p}}\mid\theta,\xi)dz_{\mathrm{p}}dX_{\mathrm{p}}\right\}\\ &\propto f(\theta)\exp\left\{\tau\int g(z_{\mathrm{p}},X_{\mathrm{p}}\mid y)\log f(z_{\mathrm{p}}\mid X_{\mathrm{p}},\theta)dz_{\mathrm{p}}dX_{\mathrm{p}}\right\},\end{aligned}\tag{3.4}$$

where the second line follows because $\log f(X_{\mathrm{p}}\mid\xi)$ does not involve the parameter of interest θ. Requiring a prior generating model for X_{p} even when we are only interested in the parameters of the conditional distribution of z given X forces us to be explicit about the predictive distribution toward which the catalytic prior will shrink the final model.

Depending on the structure of the data, we might evaluate the expectation over X_{p} in different ways. If we have a fully-specified joint (z,X) prior generating model, we may be able to evaluate it analytically, but this will probably be rare in practice. If we cannot evaluate the expectation over $g(X_{\mathrm{p}}\mid y)$ analytically, we have several options.

3.3.1 Discrete predictors with few combinations

Let n_{p} denote the number of rows X_{p}, the covariate matrix to be used in constructing a catalytic prior. Suppose the predictors are discrete, and there are few enough of them that the number c_{p} of distinct possible predictor vectors is not too large. Let X_{p} be the matrix of all possible distinct predictor vectors (so that $n_{\mathrm{p}}=c_{\mathrm{p}}$), and consider a prior generating model that places equal probability on each predictor vector, so that $g(z_{\mathrm{p}},X_{\mathrm{p}}\mid\phi)=g(z_{\mathrm{p}}\mid X_{\mathrm{p}},\phi)g(X_{\mathrm{p}})$. Further, let $x'_{\mathrm{p}i}$ be the ith row of X_{p}, and suppose that under the full model, responses are conditionally independent given the covariates and θ, that is, $f(z_{\mathrm{p}}\mid X_{\mathrm{p}},\theta)=\prod_{i=1}^{n_{\mathrm{p}}}f(z_{\mathrm{p}i}\mid x_{\mathrm{p}i},\theta)$.

Assuming that $g(z_{\mathrm{p}}, X_{\mathrm{p}} \mid \phi, y) = g(z_{\mathrm{p}}, X_{\mathrm{p}} \mid \phi)$, we have

$$\begin{aligned} g(z_{\mathrm{p}}, X_{\mathrm{p}} \mid y) &= g(X_{\mathrm{p}}) \int g(z_{\mathrm{p}} \mid X_{\mathrm{p}}, \phi) g(\phi \mid y) d\phi \\ &= g(X_{\mathrm{p}}) g(z_{\mathrm{p}} \mid X_{\mathrm{p}}, y). \end{aligned}$$

Then, taking the expectation over $g(X_{\mathrm{p}})g(z_{\mathrm{p}} \mid X_{\mathrm{p}}, y)$, the prior (3.4) becomes

$$\begin{aligned} \pi_{\tau,y}(\theta) &\propto f(\theta) \exp\left\{ \tau \int g(z_{\mathrm{p}} \mid X_{\mathrm{p}}, y) \frac{1}{c_{\mathrm{p}}} \sum_{i=1}^{c_{\mathrm{p}}} \log f(z_{\mathrm{p}i} \mid x_{\mathrm{p}i}, \theta) dz_{\mathrm{p}} \right\} \\ &= f(\theta) \exp\left\{ \frac{\tau}{c_{\mathrm{p}}} \int g(z_{\mathrm{p}} \mid X_{\mathrm{p}}, y) \log f(z_{\mathrm{p}} \mid X_{\mathrm{p}}, \theta) dz_{\mathrm{p}} \right\}. \end{aligned} \tag{3.5}$$

Clogg *et al.* (1991) took exactly this approach, letting X_{p} be the matrix of all possible distinct predictor vectors. Their choice of weight corresponds to $\tau = k$, where k is the number of parameters in the full likelihood. By formulating the catalytic prior through a joint expectation for $(z_{\mathrm{p}i}, x_{\mathrm{p}i})$, we therefore obtain a natural justification for making the pseudo-counts in Clogg *et al.*'s method inversely proportional to c_{p}. (Note that in Example 2.2.2 (continued), we did not consider a taking an expectation over $x_{\mathrm{p}i}$, and we accordingly set $\tau = k/c_{\mathrm{p}}$, in the notation of this section.)

The prior (3.5) also has an appealing invariance to coarsening or refining the predictors. Consider two models for the covariates, $g(X_{\mathrm{p}})$ and $\tilde{g}(\tilde{X}_{\mathrm{p}})$, where $\tilde{X}_{\mathrm{p}}$ collapses some of the categories of X_{p} together so that $\tilde{c}_{\mathrm{p}} < c_{\mathrm{p}}$, where $\tilde{c}_{\mathrm{p}}$ and c_{p} are the numbers of distinct predictors in the coarse and fine models, respectively. Fixing τ to a constant ensures that the pseudo-dataset is roughly the same size compared to the observed data under either covariate model, because we are simply calculating expectations over different covariate models in (3.4).

3.3.2 CONTINUOUS PREDICTORS

Now suppose either that there are continuous predictors or that the predictors are discrete but c_{p} is too large for it to be computationally feasible to enumerate all of the distinct possible predictor vectors. Then we can approximate the expectation by sampling n_{p} different predictor vectors from an appropriate model. The prior (3.4) becomes

$$\pi_{\tau,y}(\theta) \propto f(\theta) \exp\left\{ \frac{\tau}{n_{\mathrm{p}}} \int g(z_{\mathrm{p}} \mid X_{\mathrm{p}}, y) \log f(z_{\mathrm{p}} \mid X_{\mathrm{p}}, \theta) dz_{\mathrm{p}} \right\}. \tag{3.6}$$

Design principles can also be used instead of random sampling for choosing predictor vectors for the pseudo-data.

3.3.3 EMPIRICAL DISTRIBUTION

Finally, in some settings we may want to set $X_{\mathrm{p}} = X$, so that the covariates for the pseudo-data are the same as for the observed data. This is equivalent to using the empirical distribution as $g(X_{\mathrm{p}} \mid y)$. In complicated regression analyses with many observations and predictors, this may be much more practical than trying to build a reasonable model $g(X_{\mathrm{p}} \mid y)$ or trying to fill in a large contingency table of discretized predictors. It also may make more sense than building a model if a particular experimental design led to the original X. Here, the expression for the catalytic prior will also be (3.6).

3.4 LOGISTIC REGRESSION

3.4.1 GENERAL APPROACH

We have briefly discussed logistic regression in Examples 2.2.2 and 2.2.2 (continued), but we now examine it in more detail. First, we give the general formulation. In Section 3.4.2 we will re-analyze a simulated example from Clogg *et al.* (1991), comparing Clogg *et al.*'s ap-

proach both to maximum likelihood and to the scaled Cauchy priors advocated by Gelman *et al.* (2008). In Section 3.4.3, we look at a more complicated simulation in which some interactions are important, and which gives us the opportunity to investigate the choice of prior generating model.

Following the development of Section 3.3, the data $y = (z, X)$ consist of responses z and covariates X. Suppose the full model $f(z \mid X, \theta)$ is a non-hierarchical logistic regression model

$$z_j \mid m_j, x_j, \theta \sim \text{Binomial}(m_j, p(x_j'\theta)), \quad (j = 1, \ldots, n),$$

where $x_j = (x_{j1}, \ldots, x_{jk})$, $\theta = (\theta_1, \ldots, \theta_k)$, and $p(x_j'\theta) = 1/\{1+\exp(-x_j'\theta)\}$. As a default prior, we use $f(\theta) \propto 1$. If we have many predictors, we will often see complete separation, leading to infinite maximum likelihood estimates for components of θ. To constrain these estimates for θ, we use a simpler logistic regression model

$$z_j \mid m_j, x_j, \phi \sim \text{Binomial}(m_j, p(x_j'\phi)),$$

as our prior generating model $g(z \mid X, \phi)$, where $\phi = (\phi_1, \ldots, \phi_{k_g}, 0, \ldots, 0)$. That is, the prior generating model uses only the first $k_g < k$ predictors. Clogg *et al.* (1991) used the intercept-only model with $\phi = (\phi_1, 0, \ldots, 0)$. Under the prior generating model, we can use a flat prior on the components of ϕ that are not fixed at zero, assuming that we have enough data that the posterior distribution $g(\phi \mid y)$ under this prior is well-behaved. Of course, we could also use other choices of $g(\phi)$.

In regression problems, we need to specify covariates for the pseudo-data:

$$x_{\text{p}j} = (x_{\text{p}j1}, \ldots, x_{\text{p}jk}), \quad j = 1, \ldots, n_{\text{p}}.$$

In Clogg *et al.* (1991), because the predictors were discrete, the natural choice was to include

a covariate vector for every possible combination of predictor values, so that $n_{\mathrm{p}} = c_{\mathrm{p}}$; see Section 3.3.1. In logistic regression, the sufficient statistics are linear in the observations, so for the catalytic prior, we must compute

$$\begin{aligned} a_j &\equiv \int E_g(y_{\mathrm{p}} \mid x_{\mathrm{p}j}, \phi) g(\phi \mid y) d\phi \\ &= m_{\mathrm{p}j} \int p(x'_{\mathrm{p}j}\phi) g(\phi \mid y) d\phi, \end{aligned} \tag{3.7}$$

where we fix $m_{\mathrm{p}j}$ to a convenient value. For simplicity, we will assume that $m_{\mathrm{p}j} = 1$ for all j. We still have flexibility in choosing the weight τ assigned to the pseudo-data, but setting $m_{\mathrm{p}j} = 1$ for all j means that the same weight is assigned to each row of predictors $x_{\mathrm{p}i}$.

Under these choices for f and g, the posterior distribution under the catalytic prior becomes

$$\pi_\tau(\theta \mid y) \propto \left[\prod_{j=1}^{n} \left\{p(x'_j\theta)\right\}^{y_j} \left\{1 - p(x'_j\theta)\right\}^{m_j - y_j}\right] \left[\prod_{j=1}^{n_{\mathrm{p}}} \left\{p(x'_{\mathrm{p}j}\theta)\right\}^{\tau a_j / n_{\mathrm{p}}} \left\{1 - p(x'_{\mathrm{p}j}\theta)\right\}^{\tau(1-a_j)/n_{\mathrm{p}}}\right].$$

After choosing τ and calculating a_j, this is easy to fit using standard logistic regression software.

3.4.2 AN EXAMPLE FROM CLOGG *ET AL.* (1991)

As discussed in Example 2.2.2 (continued), the approach of Clogg *et al.* (1991) can be interpreted as a catalytic prior under a specific choice of prior generating model g and tuning parameter τ. Clogg *et al.* (1991) compared the performance of their method to maximum likelihood for simulated data in a setting in which it is likely that finite maximum likelihood estimates will not exist. In Table 3.2, we reproduce a $2 \times 2 \times 2$ contingency table with two sampling zeros from Clogg *et al.* (1991, Table 6). We want to fit the additive logit model with

Table 3.2: A $2 \times 2 \times 2$ contingency table with two sampling zeros, reproduced from Table 6 of Clogg *et al.* (1991).

Predictors		Response	
x_{j1}	x_{j2}	y_j	$m_j - y_j$
1	1	0	3
−1	1	9	4
1	−1	6	3
−1	−1	5	0
Total		20	10

linear predictor

$$\eta_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2},$$

but, as Clogg *et al.* (1991) note, even though there are observations in each row of the contingency table and the sufficient statistics are nonzero, the maximum likelihood estimates are not finite. Fitting the data in Table 3.2 using Clogg *et al.*'s method results in the parameter estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (0.73, -1.29, -1.23)$. For this example, $a_j = 2/3$ for all j and, because the full model has three parameters, $\tau = 3$. Multiplying the fitted probabilities in each cell of the table by the corresponding sum $m_j + \tau/n_{\text{p}}$ yields expected frequencies for samples of size $\sum_{j=1}^{n}(m_j + \tau/n_{\text{p}}) = 33$. Table 3.3 gives expected frequencies for samples of size $N = \sum_{j=1}^{n} m_j = 30$. Clogg *et al.* (1991) simulated datasets with $N = 30$ and $N = 100$ using the corresponding expected frequencies from this fitted model to evaluate the repeated sampling properties of their suggested approach and of maximum likelihood.

We have repeated the analysis of Clogg *et al.* (1991), adding a comparison with the scaled Cauchy priors that Gelman *et al.* (2008) recommend as a default for logistic regression. Table 3.4 shows mean squared error, interval coverage, and interval width for the maximum likelihood estimates, posterior modes under Gelman *et al.*'s priors, and posterior modes under Clogg *et al.*'s priors. Posterior intervals were obtained from normal approximations to

Table 3.3: Expected frequences of the $2 \times 2 \times 2$ contingency table when $n = 30$, used for a simulation in Clogg *et al.* (1991).

Predictors		Expected frequencies	
x_{j1}	x_{j2}	$E(y_j)$	$E(m_j - y_j)$
1	1	0.5	2.9
-1	1	8.6	3.9
1	-1	5.9	3.0
-1	-1	5.0	0.2
Total		20	10

the posterior distribution. For $N = 30$, maximum likelihood estimates existed in less than half of the simulated datasets; for $N = 100$, they existed 91% of the time. Of course, because Gelman *et al.*'s priors are proper and because Clogg *et al.*'s priors add positive values to each cell of the contingency table, estimates always exist under both priors. The coverage probabilities for maximum likelihood-based intervals are conditional on the existence of maximum likelihood estimates. Unconditionally, the coverage for these intervals is reduced by a factor of 0.48 or 0.91, depending on the sample size. Overall, with the exception of the intercept term for small samples, Clogg *et al.*'s prior performs the best in terms of mean squared error. Moreover, Gelman *et al.*'s prior leads to under-coverage, which is less of a problem for Clogg *et al.*'s prior. We also performed a simulation fixing the row totals m_j to the values in Table 3.2 when $N = 30$ and to $(10, 44, 30, 16)$ (from the top row to the bottom row in Table 3.2) when $N = 100$, instead of using multinomial sampling as in Clogg *et al.* (1991). Appendix A.1 presents those results, which do not change our conclusion that Clogg *et al.*'s prior leads to superior performance in terms of both mean squared error and interval coverage.

For completeness, we now describe how we obtained point estimates and posterior intervals using Gelman *et al.*'s prior, as implemented in the `bayesglm` routine in the R package `arm`. Gelman *et al.* (2008) advocate scaled Cauchy priors for all logistic regression coefficients, with a scale of 10 for the intercept and a scale of 2.5 for all other predictors. However, this prior is

Table 3.4: Comparison of maximum likelihood (ML), scaled Cauchy (SC) priors (Gelman *et al.*, 2008), and Clogg *et al.*'s priors (CP) for a simulated $2 \times 2 \times 2$ table with high probability of sampling zeros. 5000 simulated data sets were generated. Compare to Table 8 of Clogg *et al.* (1991).

	$N = 30$			$N = 100$		
	ML	SC	CP	ML	SC	CP
1. Fraction of samples where estimates exist	0.48	1	1	0.91	1	1
2. Mean squared error (given existence)						
(a) $\hat{\beta}_0$	.23	.22	.29	.07	.06	.06
(b) $\hat{\beta}_1$	.27	.30	.26	.11	.12	.11
(c) $\hat{\beta}_2$	.28	.31	.27	.12	.13	.12
3. Coverage of 95% intervals for coefficients (percent of samples where estimates exist)						
(a) β_0	96.4	96.9	96.8	95.5	95.7	95.3
(b) β_1	94.9	89.7	94.4	96.9	92.9	95.0
(c) β_2	93.8	89.6	93.9	96.9	92.2	94.8
4. Coverage of 95% intervals for logits (percent of samples where estimates exist)						
(a) $(x_{j1}, x_{j2}) = (-1, -1)$	93.5	88.3	93.0	96.8	92.0	94.4
(b) $(1, -1)$	97.3	97.1	97.8	95.4	95.5	95.6
(c) $(-1, 1)$	97.0	96.6	97.2	95.7	95.8	95.8
(d) $(1, 1)$	93.8	89.6	93.9	96.6	92.6	94.4
5. Width of 95% intervals for logits (given existence)						
(a) $(x_{j1}, x_{j2}) = (-1, -1)$	4.81	4.17	4.90	3.26	2.81	3.02
(b) $(1, -1)$	2.98	2.82	2.91	1.53	1.49	1.50
(c) $(-1, 1)$	2.55	2.38	2.47	1.31	1.28	1.29
(d) $(1, 1)$	4.30	3.89	4.31	3.02	2.62	2.78

designed for models with standardized regression inputs. Binary inputs are assumed to have mean zero and range one. To use Gelman *et al.*'s prior, we standardized the inputs for each simulated dataset, found the point estimates and estimated covariance matrix of regression coefficients on the standardized scale, and transformed the results to the original scale using the appropriate linear transformation. This will be valid as long as the normal approximation to the posterior distribution is adequate.

3.4.3 Simulations with interactions

To investigate the performance of different priors under a more realistic logistic regression setting, we simulated data with eight binary covariates in a full factorial design. In order to directly use the scaled Cauchy priors of Gelman *et al.* (2008), we coded these covariates as ± 0.5. Figure 3.3 shows the four coefficient vectors used. Figures 3.4–3.6 show results from fitting 1000 simulated datasets with $m_j = 1$ for all j, using maximum likelihood and the posterior modes under scaled Cauchy priors and catalytic priors with varying weights and different prior generating models. Figure 3.4 displays mean squared error of estimates for the linear predictors $x_j'\beta$, but results for the coefficients β were qualitatively similar. There are $n_{\mathrm{p}} = c_{\mathrm{p}} = 2^8$ linear predictors, corresponding to each row of covariates, so box plots are given for each model. Results are shown for catalytic priors over a range of values of τ when the prior generating model only includes the intercept term, as in Clogg *et al.* (1991), and when the prior generating model is the model that includes the intercept and all main effects. Figures 3.5 and 3.6 show coverage and width, respectively, of 95% intervals for linear predictors.

In practice, we may not collect observations for every covariate combination. We also simulated data by sampling $N = 120$ units with covariates given values ± 0.5 with probability 0.5; see Figures 3.7–3.9. In our simulation, 100 rows of the full covariate matrix were observed. We fixed this observed covariate matrix and simulated 1000 response vectors under each of

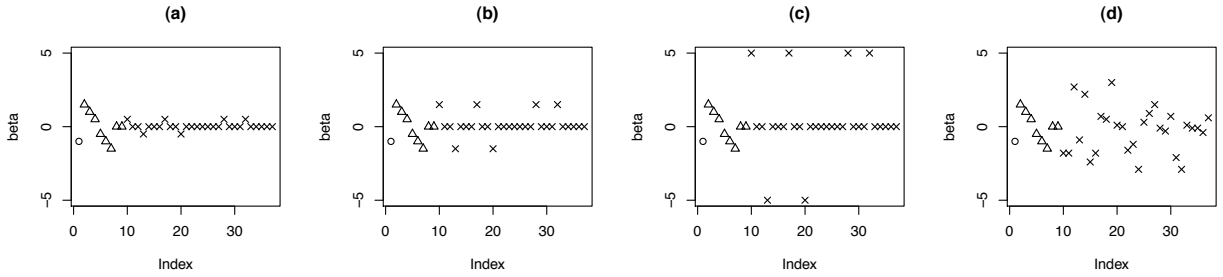


Figure 3.3: Four coefficient vectors used in regression simulations, with (a) small sparse interactions, (b) medium sparse interactions, (c) large sparse interactions, and (d) medium dense interactions. The circles denote the intercept terms, the triangles the main effects, and the $\times$'s the two-way interactions.

the four coefficient vectors used above in the full factorial simulation. As in the full factorial design, we set X_{p} to the full factorial predictor matrix, so that $n_{\mathrm{p}} = c_{\mathrm{p}} = 2^8$. Because of the smaller sample size, complete separation is a problem for maximum likelihood. For some of the cases, the R routine `glm` failed to converge. For others, `glm` satisfied its convergence criterion, but estimated standard errors were on the order of 10^6 or higher for some covariates. To make our comparisons as favorable to maximum likelihood as possible, we (somewhat arbitrarily) disregarded simulations in which standard errors for any coefficients exceeded 1000 when we evaluated the mean squared error, coverage, and interval width. Table 3.5 gives the fraction of simulated datasets for which maximum likelihood either failed to converge or numerically converged but had excessively large standard errors. It is clear that maximum likelihood struggles with the smaller sample. Especially striking is that only 29 of the 1000 small-sample simulations with large, sparse two-way interactions led to well-behaved maximum likelihood-based intervals. Even with the full factorial design, however, the coefficient vector with large, sparse two-way interactions causes problems.

It is clear from these simulations that the catalytic priors, with an appropriate choice of weight, can do substantially better than maximum likelihood. Catalytic priors are competitive with the scaled Cauchy priors of Gelman *et al.* (2008). When the true interactions are sparse and not too large, the catalytic priors under the intercept-only prior generating model are

clearly better than the other models, with lower mean squared error, higher coverage, and shorter intervals. In this setting, catalytic priors under the main-effect prior generating model perform roughly the same as scaled Cauchy priors, with some slight under-coverage. With too high a tuning parameter, interval coverage for the catalytic priors under both prior generating models unsurprisingly drops.

Scaled Cauchy priors do better than catalytic priors when the true model has large, sparse two-way interactions. One way to interpret this is that if the truth is heavy-tailed, we will do better using a heavy-tailed prior. In these cases, interval coverage for some of the linear predictors under the catalytic priors can be much too low if we choose the weight according to the recommendation of Clogg *et al.* (1991), although a slightly lower tuning parameter would enable the catalytic priors to perform comparably with the scaled Cauchy priors. This suggests that adaptively choosing the catalytic prior tuning parameter could lead to improved performance, if a good adaptive procedure were developed.

When the true model has dense interactions, catalytic priors and scaled Cauchy priors perform similarly. When the observed covariates are a full factorial design, the catalytic priors yield slightly better mean squared error than the Cauchy priors, but under the intercept-only prior generating model, there is more under-coverage than under the main-effect prior generating model or under Cauchy priors. When we only observe a sample of the full table of covariates, catalytic priors yield lower mean squared error but also slightly more under-coverage than Cauchy priors.

Overall, the original procedure of Clogg *et al.* (1991) (intercept-only prior generating model and $\tau = k$) does remarkably well, only clearly deficient when the true model has large, sparse interactions and our criterion is coverage. In applications, context should help us decide whether large, sparse interactions are scientifically plausible and thus should be a concern. If no such guidance is available, robust priors may be a better option. However, it is important to keep the analysis goals in mind, as the catalytic priors can yield substantially lower mean

Table 3.5: The fraction of simulated datasets out of 1000 for which the maximum likelihood software failed to converge and for which it nominally converged but returned standard errors greater than 1000, for four different true coefficient vectors. *Small*, *medium*, *large*, *sparse*, and *dense* refer to the magnitudes and number of nonzero coefficients for two-way interactions.

	Full factorial ($N = 256$)		Smaller sample ($N = 120$)	
	Not converged	SE > 1000	Not converged	SE > 1000
Small, sparse	0	.003	.351	.062
Medium, sparse	0	.003	.583	.053
Large, sparse	.005	.033	.959	.012
Medium, dense	0	0	.796	.028

squared error than scaled Cauchy priors, even in settings in which catalytic priors struggle to achieve nominal coverage.

3.5 LINEAR REGRESSION

3.5.1 GENERAL APPROACH

Consider a linear regression of z on X, where z is a vector of length n and X is an $n \times k$ matrix of covariates. The full model $f(z \mid X, \theta)$ is

$$z \mid X, \theta \sim N(X\beta, \sigma^2 I),$$

where $\theta = (\beta, \sigma^2)$. Let X_{p} be $n_{\text{p}} \times k$. We assume that if we had observed the responses z_{p} corresponding to X_{p}, we could reliably estimate the regression parameters for a full dataset that included both (z, X) and $(z_{\text{p}}, X_{\text{p}})$. In particular, we must choose X_{p} so that $X'X + (\tau/n_{\text{p}})X'_{\text{p}}X_{\text{p}}$ is full rank when $\tau > 0$.

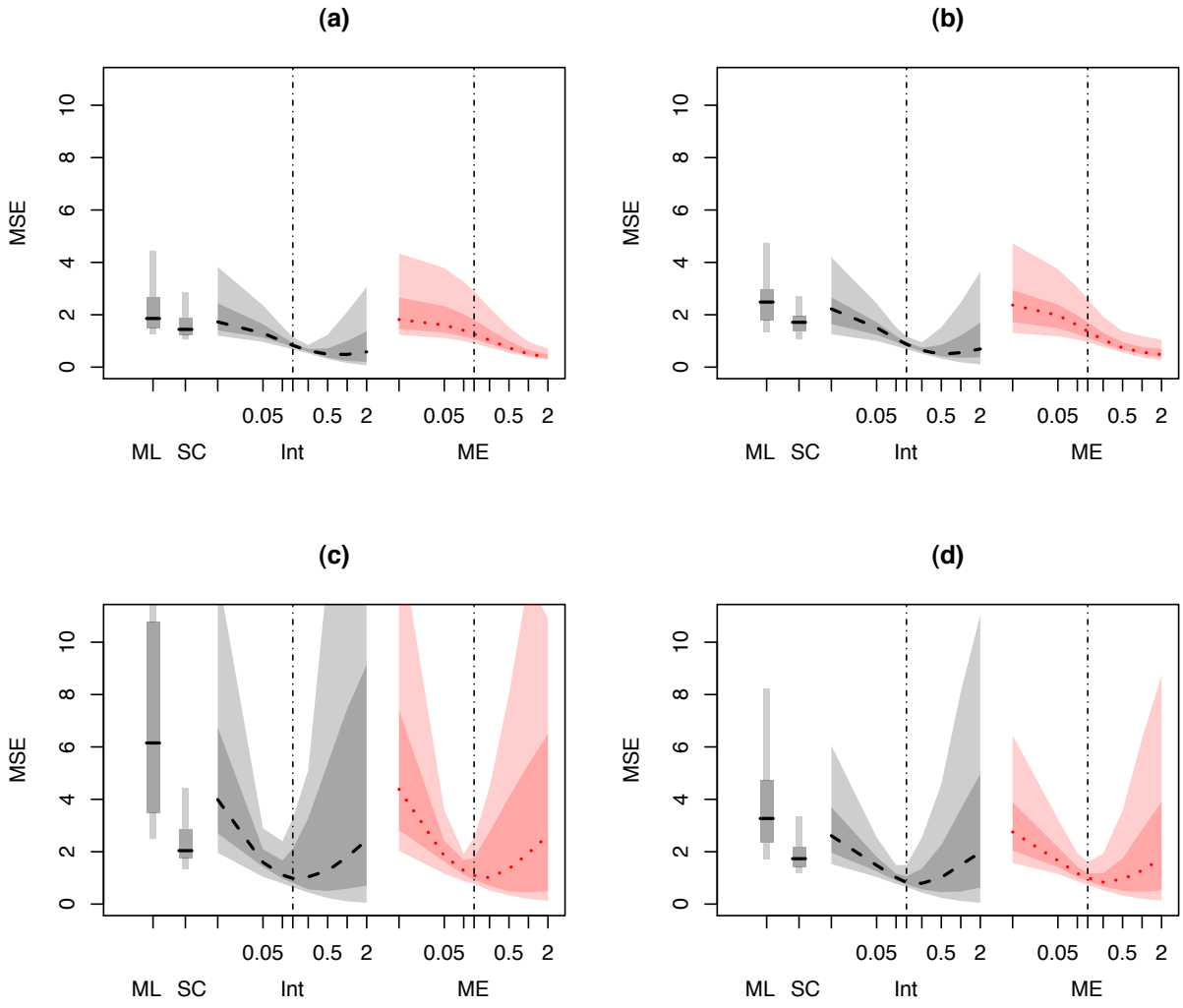


Figure 3.4: Simulation results for a logistic regression with eight binary covariates in a full factorial design. (a)–(d) correspond to the coefficient vectors shown in Figure 3.3(a)–(d), respectively. Box plots of the mean squared errors for estimating the true linear predictors $x_j'\beta$, $j = 1, \ldots, 2^8$, for maximum likelihood (ML), scaled Cauchy (SC) priors, catalytic priors with intercept-only prior generating model (Int, black dashed line) and with prior generating model that included all main effects (ME, red dotted line). The shaded regions for catalytic priors correspond to varying τ. A vertical line is included at $\tau = k$, the weight suggested by Clogg *et al.* (1991), where $k = 1 + 8 + \binom{8}{2}$ is the number of parameters in the full model that includes all two-way interactions.

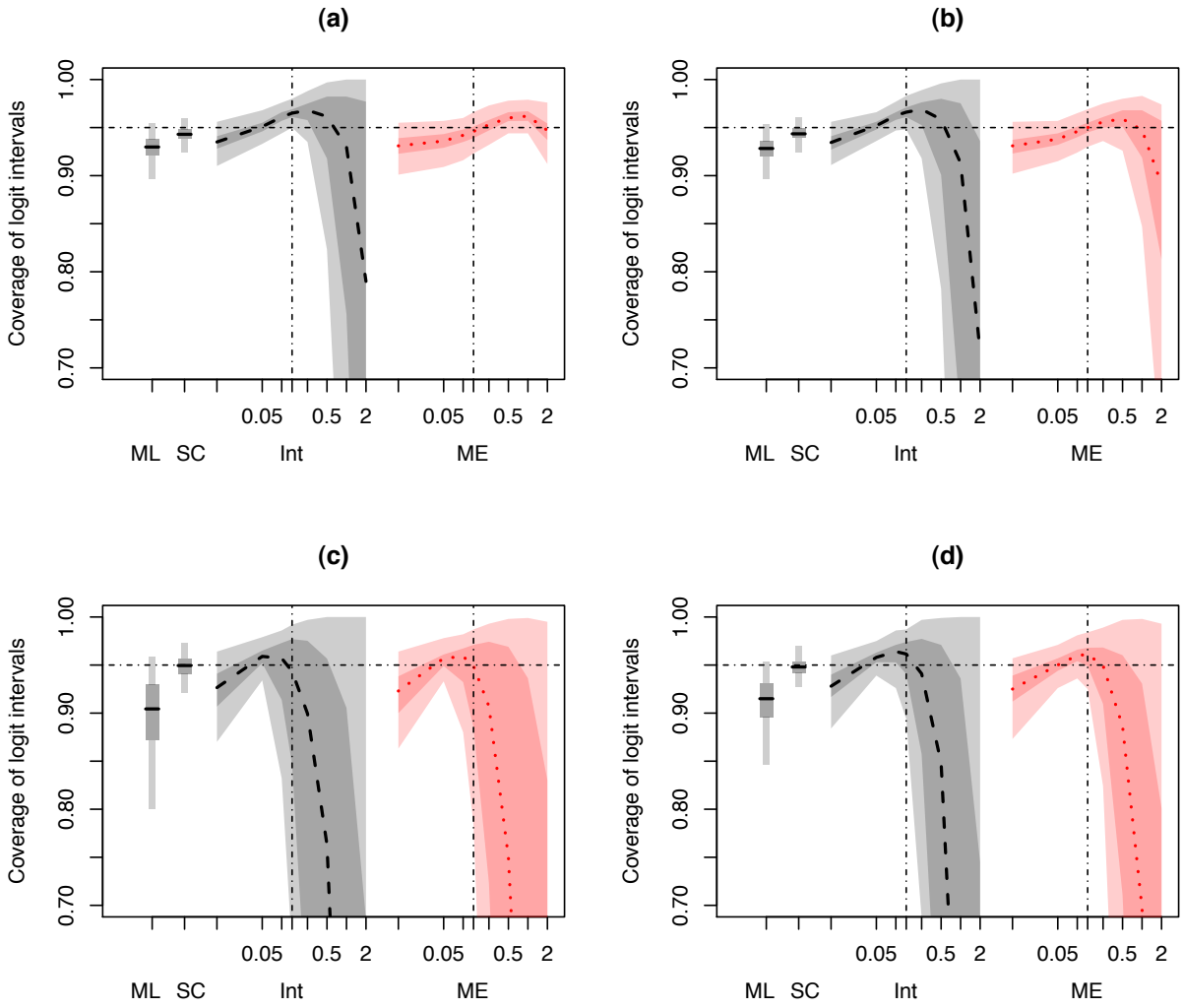


Figure 3.5: Coverage of logit intervals in a logistic regression simulation with eight binary covariates in a full factorial design.

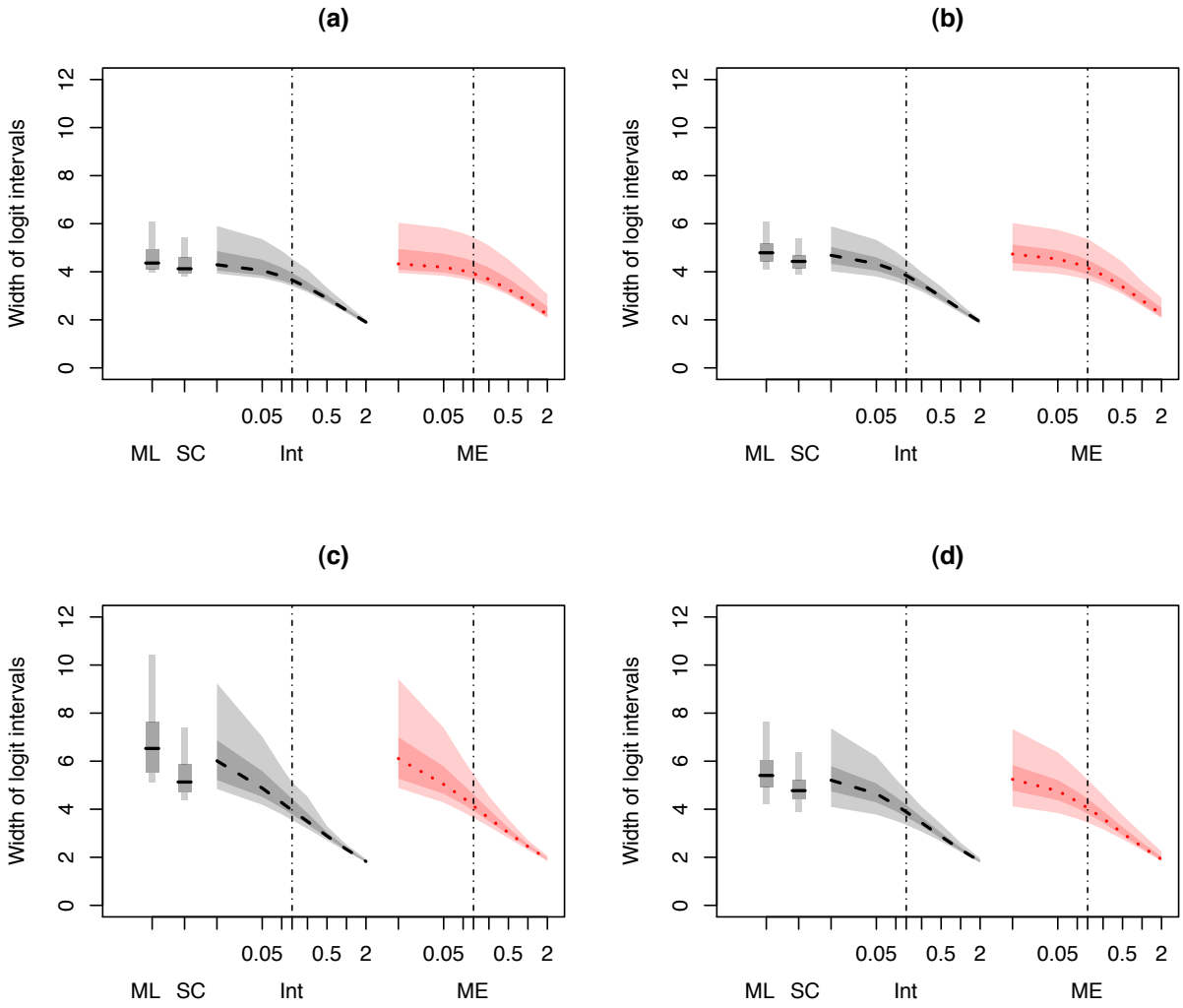


Figure 3.6: Width of logit intervals in a logistic regression simulation with eight binary covariates in a full factorial design.

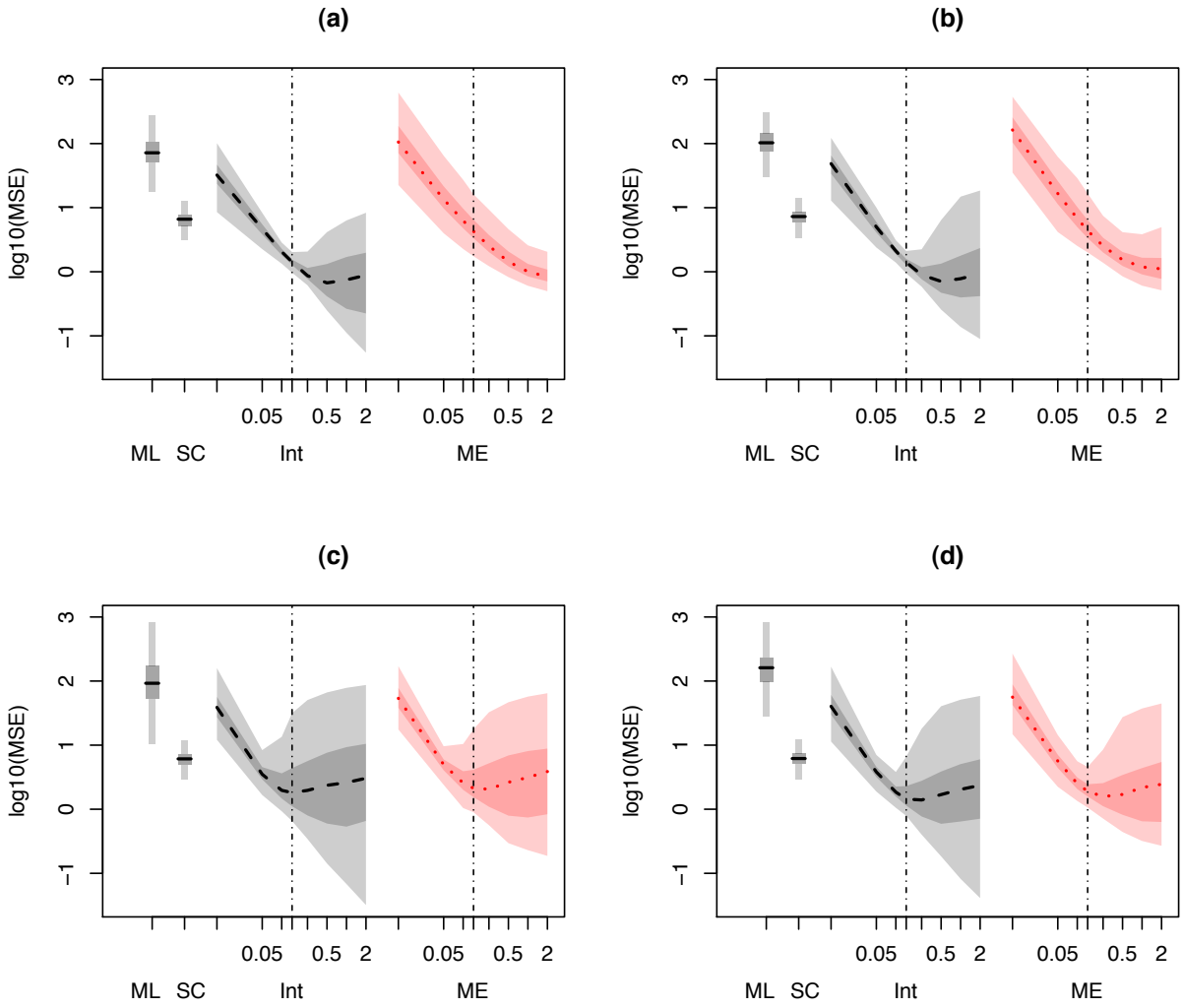


Figure 3.7: Simulation results for a logistic regression with eight binary covariates and $N = 120$ total observations. Mean squared error is plotted on a logarithmic scale.

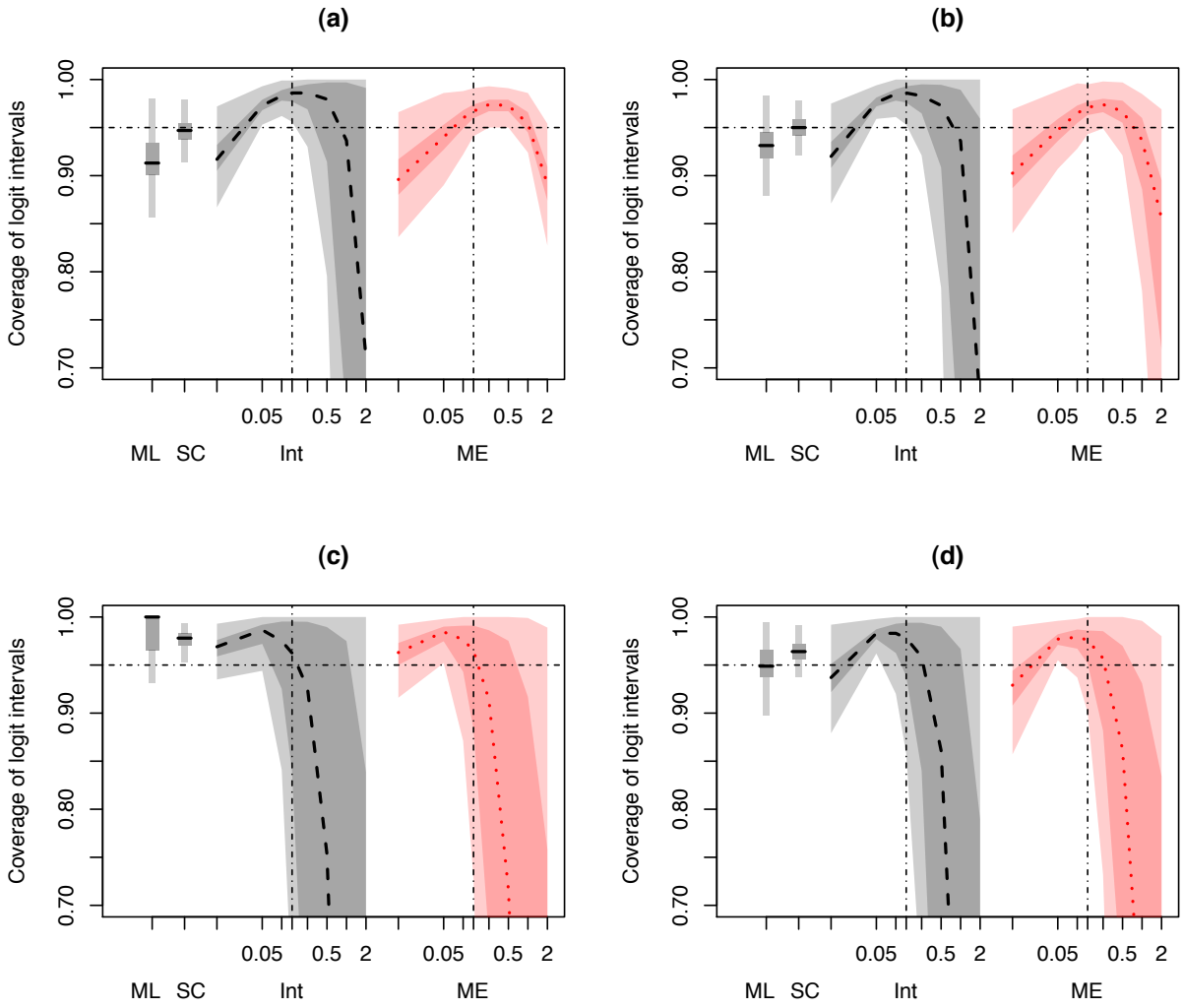


Figure 3.8: Coverage of intervals for logits in a logistic regression simulation with eight binary covariates and $N = 120$ total observations.

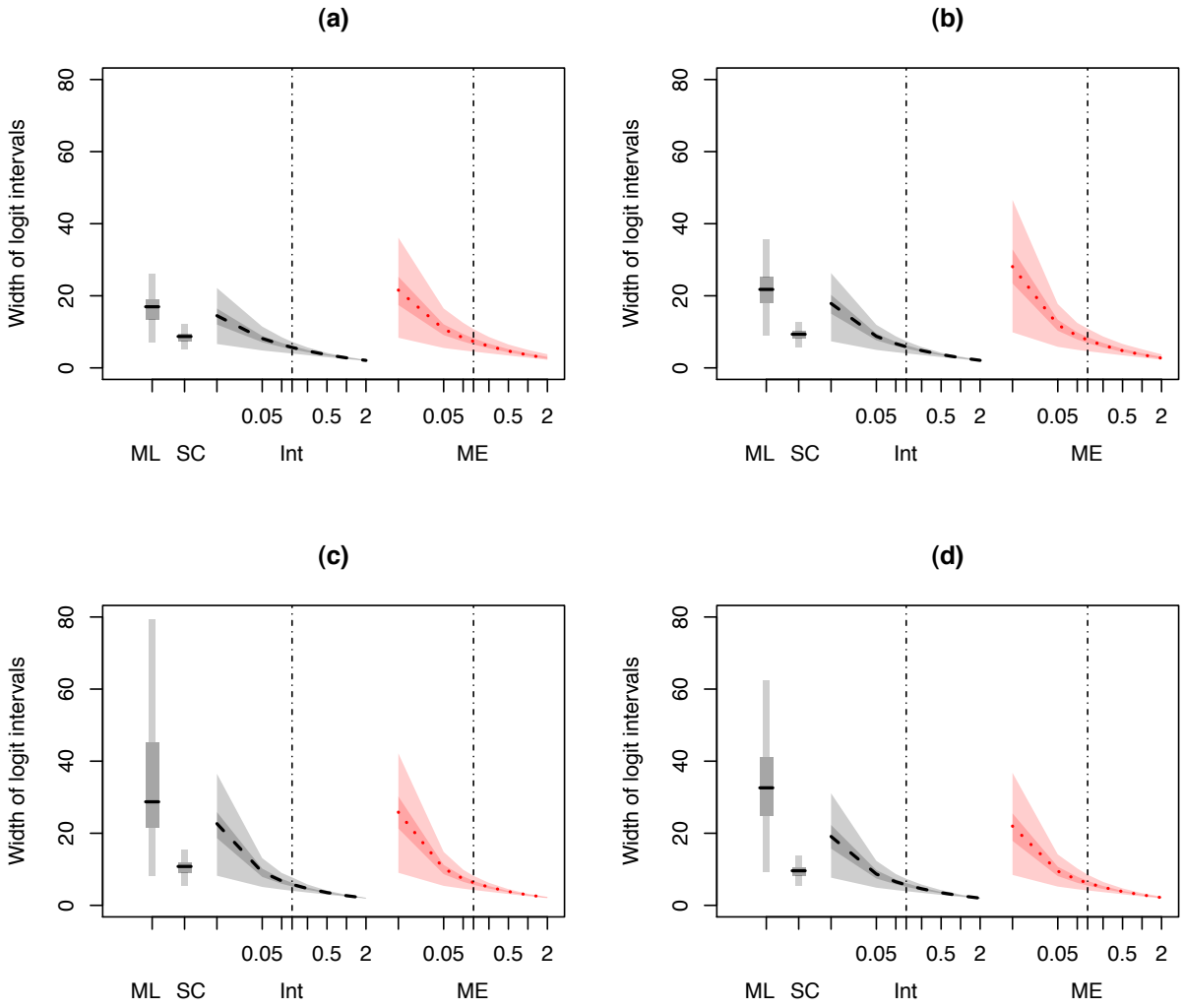


Figure 3.9: Width of intervals for logits in a logistic regression simulation with eight binary covariates and $N = 120$ total observations.

For the prior generating model $g(y \mid \phi)$, we assume that

$$z \mid X, \phi \sim N(X_g \beta_g, \sigma_g^2 I)$$

where $\phi = (\beta_g, \sigma_g^2)$ and X_g is the $n \times k_g$ matrix formed by selecting k_g of the columns of X. If u_j is the jth column of X, then, say, $X_g = (u_1 \cdots u_{k_g})$, where k_g is small enough that $X_g' X_g$ is full rank. Under the default prior $g(\beta_g, \sigma_g^2) \propto 1/\sigma_g^2$, the posterior distribution of σ_g^2 is

$$\sigma_g^2 \mid y \sim \frac{\text{RSS}_g}{\chi^2_{n-k_g}},$$

where $\text{RSS}_g = ||z - X_g \hat{\beta}_g||^2$, $||\cdot||$ denotes Euclidean norm, and $\hat{\beta}_g = (X_g' X_g)^{-1} X_g' z$ is the usual least squares estimator. Conditioning on σ_g^2, the posterior predictive distribution of z_{p} is

$$z_{\text{p}} \mid \sigma_g^2, y \sim N(X_{\text{p}g} \hat{\beta}_g, \sigma_g^2 \{I_{n_{\text{p}}} + X_{\text{p}g} (X_g' X_g)^{-1} X_{\text{p}g}'\}),$$

where $X_{\text{p}g}$ is the matrix of the first k_g columns of X_{p}.

Let $\hat{\gamma}_g = (\hat{\beta}_g, 0, \ldots, 0)$ be the k-vector formed by filling in zeros for all coefficients of predictors not included in X_g, and let $k_g^* = \text{tr}\{(X_g' X_g/n)^{-1} (X_{\text{p}g}' X_{\text{p}g}/n_{\text{p}})\}$, so that if the covariate vectors for the data and pseudo-data have approximately equal sample means and covariances, then $k_g^* \approx k_g$. The prior (3.6) leads to the posterior

$$\begin{aligned} \sigma^2 \mid y &\sim \frac{\tau\{1 + k_g^*/n\} E_g(\sigma_g^2 \mid y) + \text{RSS}_f}{\chi^2_{n+\tau-k}} \\ \beta \mid \sigma^2, y &\sim N(\hat{\beta}, \sigma^2 V_\beta), \end{aligned}$$

where

$$
\begin{aligned}
E_g(\sigma_g^2 \mid y) &= \text{RSS}_g/(n - k_g - 2), \\
V_\beta &= \{X'X + (\tau/n_\text{p})X_\text{p}'X_\text{p}\}^{-1}, \\
\hat{\beta} &= V_\beta\{X'z + (\tau/n_\text{p})X_\text{p}'X_\text{p}\hat{\gamma}_g\}, \\
\text{RSS}_f &= ||z - X\hat{\beta}||^2 + (\tau/n_\text{p})||X_\text{p}\hat{\gamma}_g - X_\text{p}\hat{\beta}||^2.
\end{aligned}
$$

Thus, $\tau > k - n$ is necessary for posterior propriety. This accords with the interpretation of τ as the sample size of a pseudo-dataset, as it means that if there are more regression coefficients than observations, then we must add enough pseudo-observations that the total size $n+\tau$ is larger than the number of coefficients to be estimated. Although the recommendation of Clogg *et al.* (1991) to use $\tau = k$ was for logistic regression, not linear regression, it would obviously guarantee this.

3.5.2 Simulation

We applied our general linear regression approach using the same true coefficient vectors as in the simulation in Section 3.4.3. We sampled $n = 30$ rows without replacement from the 2^8 rows of the full factorial covariate matrix, and simulated 1000 different datasets for each true coefficient vector. Thus, there were enough observations to fit the main effect model with nine parameters but not enough to fit the model with all two-way interactions with 37 parameters via maximum likelihood. For the catalytic priors, we used the full factorial covariate matrix for X_p. Figures 3.10–3.13 summarize the mean squared error, coverage, and interval width for estimates of the linear predictors $x_j'\beta$ for maximum likelihood using only the main effects, catalytic priors with intercept-only prior generating model, and catalytic priors with prior generating model that included all main effects. We also compare to ridge regression for a variety of ridge tuning parameters. More precisely, the model for ridge regression is a fully

Bayesian model with prior

$$p(\sigma^2, \beta_1) \propto 1,$$
$$(\beta_{[-1]} \mid \sigma^2, \beta_1) \sim N(0, \sigma^2 \lambda^{-1} (X'_{[-1]} X_{[-1]})^{-1}),$$

where $\beta_{[-1]} = (\beta_2, \ldots, \beta_k)$, $X_{[-1]}$ is the matrix that omits the first column of X, and λ is the ridge regression tuning parameter.

When the model including only main effects is close to true, maximum likelihood under this wrong model can do extremely well in terms of mean squared error; see Figures 3.10(a) and (b). Similarly, catalytic priors with a prior generating model including main effects can do extremely well in these settings. However, perhaps because of the discrepancy between the truth and the assumed model, maximum likelihood with only main effects leads to some under-coverage under all true parameter vectors, with the problem becoming more severe when there are more non-negligible true interaction terms; see Figures 3.11 and 3.12. For small choices of the catalytic prior tuning parameter τ, the catalytic priors lead to over-coverage under all true coefficient vectors used in the simulation. However, this seems preferable to the potentially severe under-coverage that maximum likelihood under a too-simple model and ridge regression can exhibit, as in Figures 3.11(d). Of course, catalytic priors achieve this over-coverage by inflating the width of posterior intervals, leading to imprecise inferences.

Interestingly, for the three coefficient vectors with sparse true interactions, catalytic priors with the intercept-only prior generating model always yielded wider intervals than with the main-effects prior generating model. For the coefficient vector with dense true interactions, this relationship was reversed, and the intercept-only prior generating model always yielded

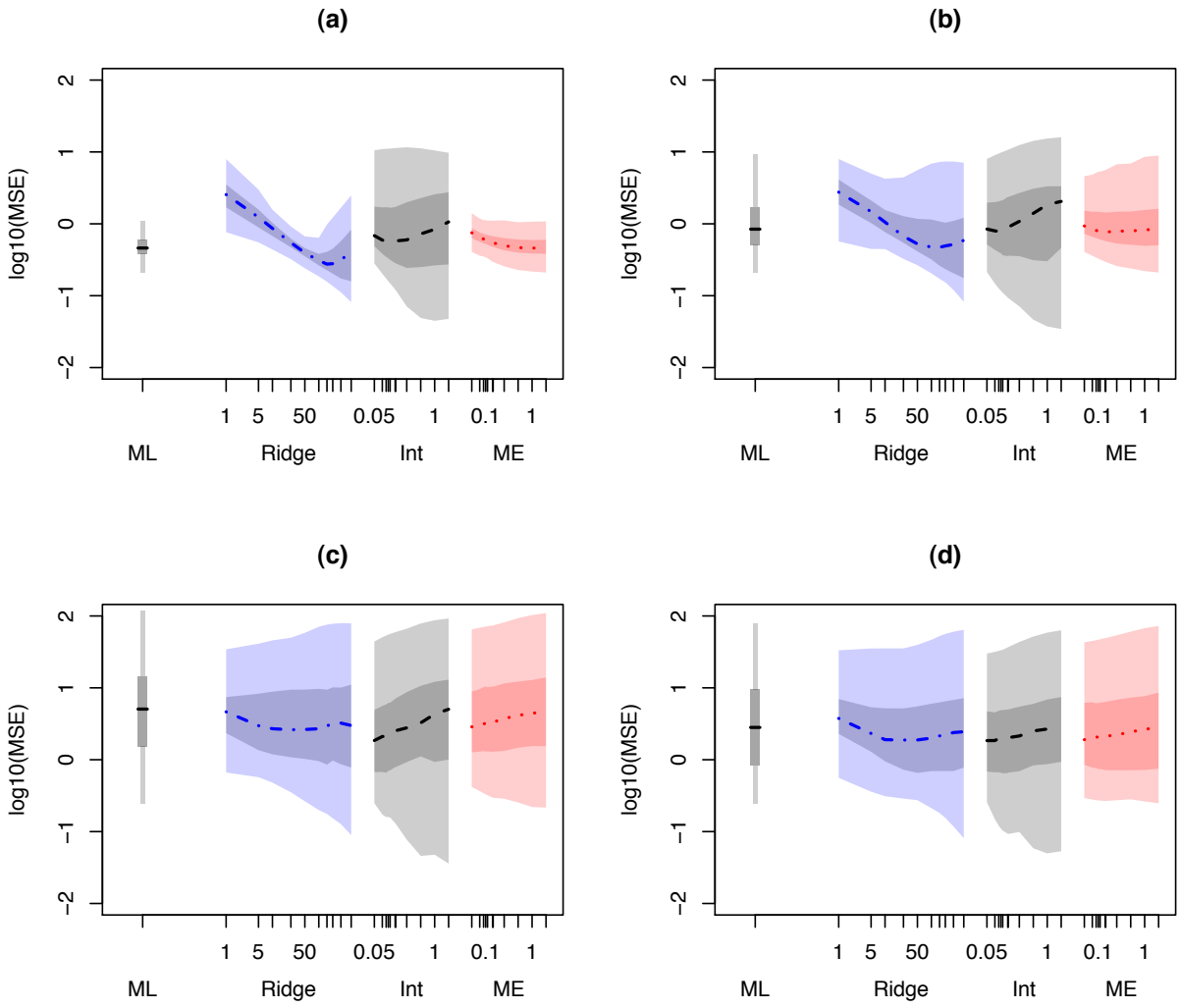


Figure 3.10: Mean squared error for linear regression with $n = 30$, eight binary predictors, and all two-way interactions. Maximum likelihood (ML) uses only the main effects. Also shown are results for the catalytic priors using prior generating models with intercept only (Int) and with all main effects (ME). (a)–(d) use the coefficient vectors of Figures 3.3(a)–(d), respectively. Ridge regression results (Ridge) are shown for a range of tuning parameters. The results under catalytic priors are plotted for varying choices of τ/n_p.

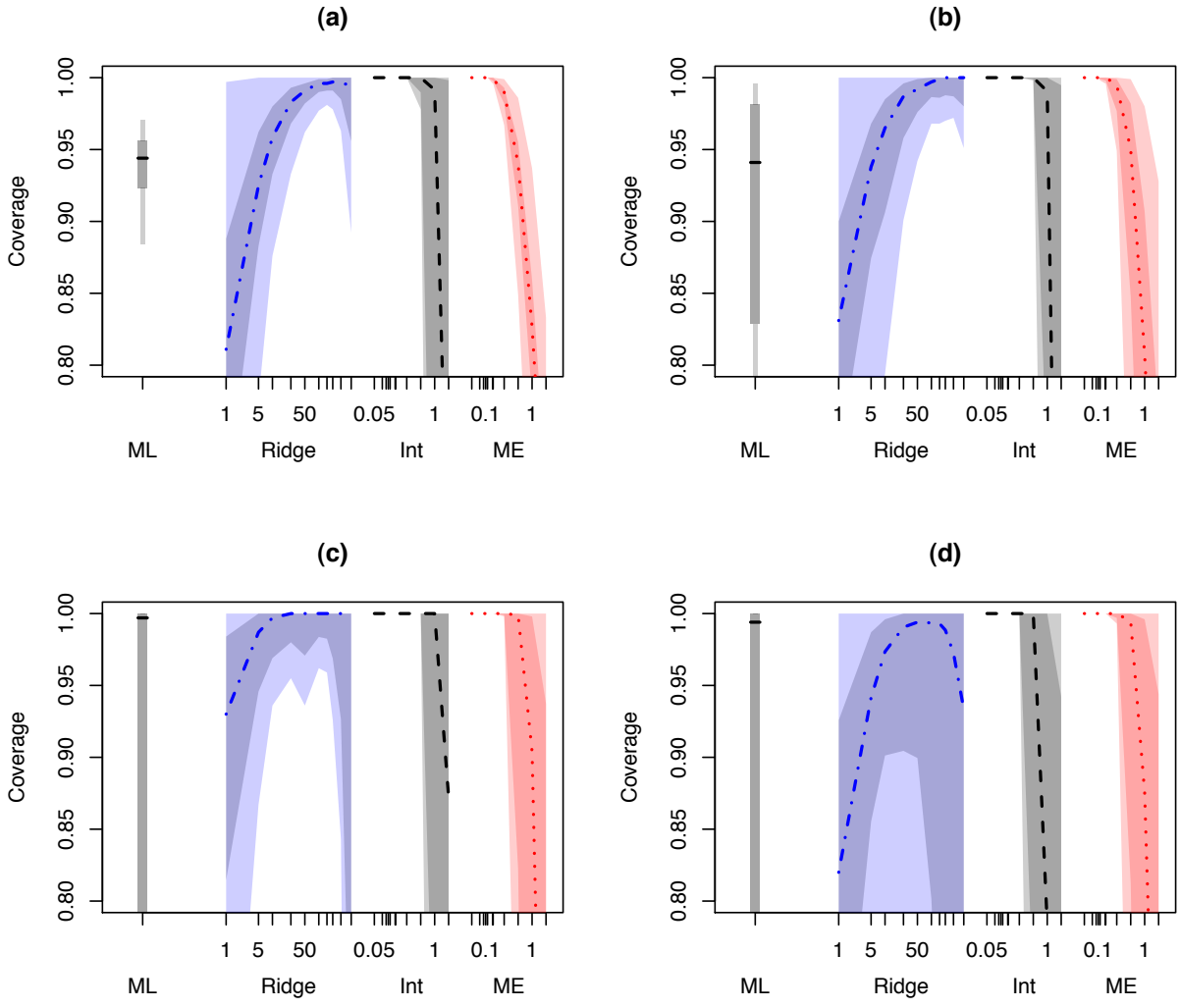


Figure 3.11: Coverage of 95% intervals for linear predictors $x_j'\beta$ in a linear regression simulation.

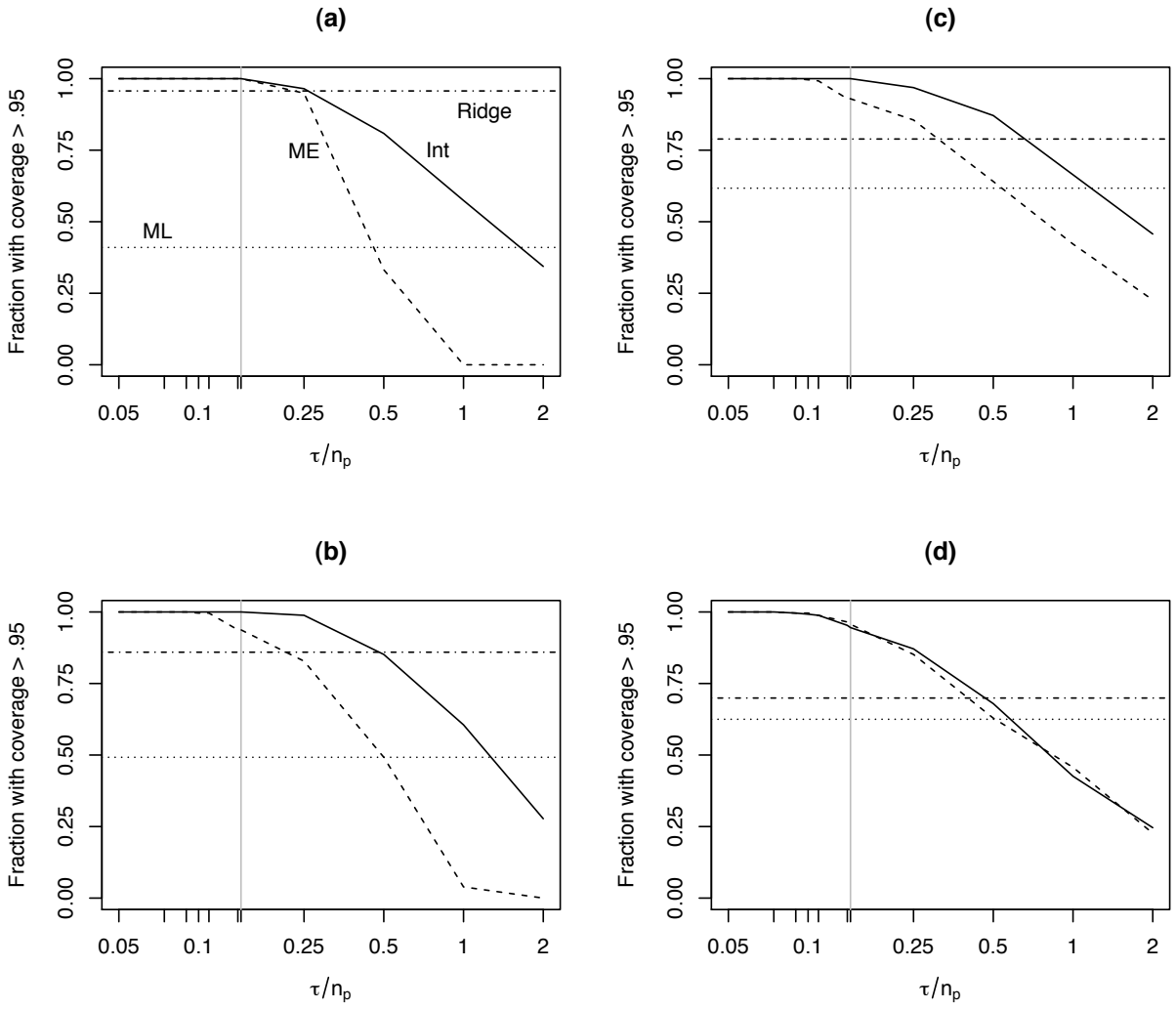


Figure 3.12: Fraction of 95% intervals for linear predictors $x_j'\beta$ that have coverage at least 95%. The gray vertical line is at $\tau = k$, the suggestion in Clogg *et al.* (1991) for logistic regression. The black solid line is for the catalytic priors with intercept-only prior generating model; the dashed line is for the catalytic priors with prior generating model that includes all main effects; the dotted line is for maximum likelihood using only the main effects; and the dashed-dotted line is the maximum fraction achieved with ridge regression, for the choices of ridge tuning parameter considered.

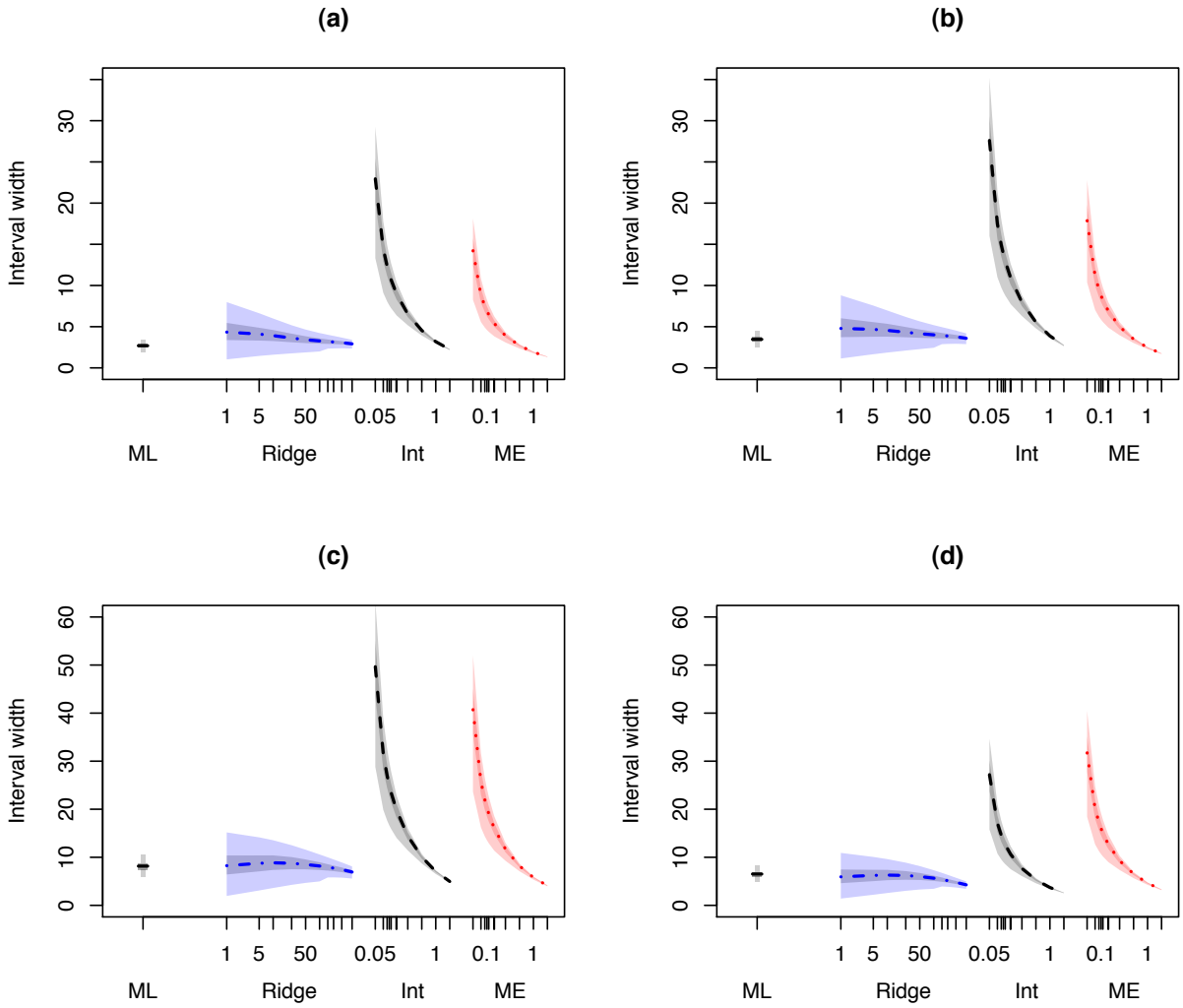


Figure 3.13: Width of 95% intervals for linear predictors $x_j'\beta$ in a linear regression simulation. The y-axes for the top and bottom rows have different scales.

shorter intervals.

3.6 LATENT VARIABLE AND MULTILEVEL MODELS

Suppose the full model for y given parameters θ is the marginal distribution of a joint distribution of observations y and latent variables v:

$$f(y \mid \theta) = \int f(y, v \mid \theta) dv.$$

Many complicated multilevel models can be formulated in this way. It is often more convenient to develop a catalytic prior using the joint full model $f(y, v \mid \theta)$ and a corresponding joint prior generating model $g(y, v \mid \phi)$, yielding

$$\pi_{\tau,y}(\theta) \propto f(\theta) \exp\left\{\tau \int g(y_{\mathrm{p}}, v_{\mathrm{p}} \mid y) \log f(y_{\mathrm{p}}, v_{\mathrm{p}} \mid \theta) dy_{\mathrm{p}} dv_{\mathrm{p}}\right\}. \tag{3.8}$$

3.6.1 FINITE MIXTURE MODELS

Suppose the model for the observed data is the J-component mixture model

$$f(y_i \mid \theta) = \sum_{j=1}^{J} \lambda_j h(y_i \mid \beta_j),$$

where $\theta = (\lambda, \beta)$, $\lambda = (\lambda_1, \ldots, \lambda_J)$ is a vector of probabilities such that $\sum_{j=1}^{J} \lambda_j = 1$, $\beta = (\beta_1, \ldots, \beta_J)$, and $h(y_i \mid \beta_j)$ is a density with parameter β_j. This model admits the latent variable formulation

$$\begin{aligned} v_i \mid \theta &\sim \text{Multinomial}(1, \lambda) \\ y_i \mid v_i, \theta &\sim h(\cdot \mid \beta_{v_i}), \end{aligned}$$

where v_i is a latent allocation variable indicating which mixture component was responsible for generating y_i.

We can view the latent allocation variables as missing data. The observed-data likelihood averages over these allocations and consequently can be computationally intractable. The complete-data likelihood, on the other hand, factors into separate likelihoods for each mixture component, where each factor only includes the subset of the data assigned to that component. Letting $y = (y_1, \ldots, y_n)$ and $v = (v_1, \ldots, v_n)$, the complete-data likelihood is

$$f(y, v \mid \theta) = \prod_{i=1}^{n} \prod_{j=1}^{J} \{\lambda_j h(y_i \mid \beta_j)\}^{1(v_i = j)},$$

where $1(v_i = j)$ is the indicator function that equals 1 if $v_i = j$ and 0 otherwise. If y_{p} consists of conditionally independent values $y_{\mathrm{p}i}$, $i = 1, \ldots, n_{\mathrm{p}}$, then (3.8) becomes

$$\pi_{\tau,y}(\theta) \propto f(\theta) \prod_{j=1}^{J} \lambda_j^{\tau \sum_{i=1}^{n_{\mathrm{p}}} p_{ij}} \exp\left\{\tau \sum_{i=1}^{n_{\mathrm{p}}} p_{ij} C(\beta_j)\right\}, \tag{3.9}$$

where

$$p_{ij} = \Pr_g(v_{\mathrm{p}i} = j \mid y)$$
$$C(\beta_j) = E_g\left\{\log h(y_{\mathrm{p}i} \mid \beta_j) \mid y\right\}.$$

3.6.2 Inferring latent processes on a network

Blocker and Airoldi (2011) and Airoldi and Blocker (2013) considered the problem of estimating point-to-point traffic flows on a network when only aggregate traffic is observed. They developed a non-conjugate multilevel state-space model to account for the complex noise characteristics in the observations, but they found that their model needed additional regularization. In particular, under a naive prior on the top-level parameters, their estimates

of the latent time series of interest were worse than competing approaches on datasets for which ground truth was available.

However, they were able to use a simpler model to obtain rough estimates of the latent point-to-point traffic. They then used these estimates to set the top-level parameters in their more complicated dynamic multilevel model, partly justifying their method by noting its similarity to that of Clogg *et al.* (1991). Airoldi and Blocker (2013) found that their dynamic multilevel model, when regularized using estimates from an appropriate simpler model, outperformed competing state-of-the-art methods. The choice of an appropriate simpler model depended on, for instance, the structure of the network.

Airoldi and Blocker (2013) also found that their dynamic multilevel model often did not significantly improve on the performance of the simpler model used for regularization. The catalytic prior framework offers one possible explanation for this behavior. Their strategy of plugging in estimates for top-level parameters, where those esimates were obtained under a simpler model, can be viewed as a posterior distribution under a catalytic prior with $\tau = \infty$, which as discussed in Section 2.4 corresponds to the Kullback–Leibler projection of the fitted simpler model onto the full model family. Since Airoldi and Blocker's simpler model was not a submodel of their dynamic multilevel model, this projection does not trivially reduce to the fitted simpler model. An interesting question for further work is whether Airoldi and Blocker's results can be improved upon by using a smaller value of τ and thus imposing less regularization on their dynamic multilevel model.

3.7 Evaluating a job training program

3.7.1 Introduction

Frumento *et al.* (2012) analyzed a randomized experiment to evaluate the Job Corps job training program for disadvantaged youths aged 16 to 24. The causal estimands of interest were

the effects of training on employment and wages, but several features of the experiment complicated the analysis. First, while encouragement to enroll in Job Corps was randomized, not everyone who was encouraged actually enrolled. In fact, only 68% of those assigned treatment enrolled within the first semester after assignment. Second, one of the outcomes of interest, wages, is only defined when individuals are employed. Finally, some of the outcome data was missing. To address these complications, Frumento *et al.* (2012) used principal stratification. Their model was a finite mixture of regression models, where the latent mixture components corresponded to the principal strata, and they used a likelihood-based analysis. Here, we explore choices of catalytic priors to obtain posterior distributions under this model. We compare estimates and estimated uncertainties obtained via posterior distributions under different catalytic priors with the maximum likelihood estimates of Frumento *et al.* (2012).

We will first give a brief overview of the model, referring the reader to Frumento *et al.* (2012) for more details. Our sample consists of $n = 13\,794$ individuals. We observe a covariate matrix X with $k = 23$ columns, including an intercept term, design weights, sex, age, race (white or not), and covariates describing each individual's family, education, and employment history. The predictors X also include some missingness indicators and two-way interactions.

We know each individual's treatment assignment Z_i, where $Z_i = 1$ if individual i was encouraged to enroll in Job Corps and $Z_i = 0$ otherwise. Let $D_i(1)$ denote the compliance status if individual i was offered treatment, where $D_i(1) = 1$ if the subject would enroll in Job Corps within six months after being offered and $D_i(1) = 0$ otherwise. In this study, if an individual was not offered participation in Job Corps, then it was not possible to enroll; thus $D_i(0) = 0$ by definition. We also observe the potential employment status $S_i(z)$ ($S_i(z) = 1$ if employed, 0 otherwise), potential wage $W_i(z)$, and potential missingness indicator $M_i(z)$ under treatment z. If the individual is not employed, then wages are not well defined, so we use the special symbol $W_i(z) = *$ if $S_i(z) = 0$.

The latent principal strata are determined by compliance behavior and by potential em-

ployment status. Compliers are those individuals with $D_i(1) = 1$, and noncompliers are those with $D_i(1) = 0$. There are four goups of individuals determined by their potential employment status: the always-employed, who would be employed under either treatment assignment; those who would be employed only if assigned treatment; those who would be employed only if assigned control; and those who would not be employed under either treatment assignment.

We assume two exclusion restrictions for noncompliers, one for employment status and one for wages. That is, if individual i would not enroll in Job Corps regardless of treatment assignment, then neither employment status nor wages will be affected by treatment assignment, so that $S_i(0) = S_i(1)$ and $W_i(0) = W_i(1)$. The exclusion restriction on employment status eliminates two principal strata because it removes the possibility that noncompliers could be employed only under treatment or only under control. Thus, we have $J = 6$ principal strata:

1. always-employed compliers,
2. compliers only employed under treatment,
3. compliers only employed under control,
4. never-employed compliers,
5. always-employed noncompliers, and
6. never-empoyed noncompliers.

Following the notation of Section 3.6.1, we let $v_i = j$ if individual i is assigned to the jth latent principal stratum. We model the conditional distribution of v_i given covariates as a multinomial logistic regression:

$$\Pr(v_i = j \mid x_i, \alpha) = \frac{\exp(x_i'\alpha_j)}{\sum_{\ell=1}^{J} \exp(x_i'\alpha_\ell)} \equiv p_j(x_i, \alpha) \tag{3.10}$$

where $x_i = (x_{i1}, \ldots, x_{ik})$ is the covariate vector for individual i, $\alpha = (\alpha_1, \ldots, \alpha_J)$, and $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{jk})$. We take never-employed noncompliers as the baseline and set $\alpha_J = (0, \ldots, 0)$.

We model employed individuals' wages by a linear regression of log wages on the covariates X. The exclusion restriction on wages implies that always-employed noncompliers must have the same distribution of wages regardless of treatment assignment. This leaves five groups, each with its own regression parameters:

1. always-employed compliers, assigned treatment;
2. always-employed compliers, assigned control;
3. compliers employed only under treatment, assigned treatment;
4. compliers employed only under control, assigned control; and
5. always-employed noncompliers.

We let this set of groups be denoted $\mathcal{E}$ (for "earners"), so that we have

$$\log\{W_i(Z_i)\} \sim N(x_i'\beta_h, \sigma_h^2) \tag{3.11}$$

if individual i is in group h, where $h \in \mathcal{E}$.

3.7.2 Catalytic priors

In the model of Section 3.7.1, the complete data include the principal stratum membership indicators $v = (v_1, \ldots, v_n)$ and the wages W. In this model, there is one complication that has not appeared in our previous discussions of regression and mixture models: how to treat the treatment assignment indicators for the pseudo-data. Inspired by Hirano *et al.* (2000), we assume that each individual in the pseudo-dataset has probability $1/2$ of being assigned

treatment, and this probability does not depend on the individual's covariates $x_{\mathrm{p}i}$ or principal stratum membership $v_{\mathrm{p}i}$. Thus, ignoring constant factors, the complete-data full-model likelihood for the pseudo-dataset is

$$f(v_{\mathrm{p}}, u_{\mathrm{p}}, Z_{\mathrm{p}} \mid X_{\mathrm{p}}, \theta) = \prod_{i=1}^{n_{\mathrm{p}}} \left[\prod_{j=1}^{J} \{p_j(x_{\mathrm{p}i}, \alpha)\}^{1(v_{\mathrm{p}i}=j)} \right] \left[\prod_{h \in \mathcal{E}} \left\{ f(u_{\mathrm{p}i} \mid x_{\mathrm{p}i}, \beta_h, \sigma_h^2) \right\}^{1(i \in h)} \right], \quad (3.12)$$

where $v_{\mathrm{p}} = (v_{\mathrm{p}1}, \ldots, v_{\mathrm{p}n_{\mathrm{p}}})$, $u_{\mathrm{p}i} = \log\{W_{\mathrm{p}i}(Z_{\mathrm{p}i})\}$, $u_{\mathrm{p}} = (u_{\mathrm{p}1}, \ldots, u_{\mathrm{p}n_{\mathrm{p}}})$, $f(u_{\mathrm{p}i} \mid x_{\mathrm{p}i}, \beta_h, \sigma_h^2)$ is the density of a normal random variable with mean $x'_{\mathrm{p}i}\beta_h$ and variance σ_h^2, and $1(i \in h)$ is the indicator for whether $(v_{\mathrm{p}i}, Z_{\mathrm{p}i})$ agrees with earner group h. For our pseudo-dataset, we set $n_{\mathrm{p}} = n$ and $X_{\mathrm{p}} = X$, which we believe is reasonable because the original study was a designed experiment.

Equation (3.12) defines the full model likelihood $f(y_{\mathrm{p}} \mid X_{\mathrm{p}}, \theta)$ that appears in the catalytic prior (3.6). To finish specifying a catalytic prior we need to choose a prior generating model g, evaluate the expectation of $\log f(y_{\mathrm{p}} \mid X_{\mathrm{p}}, \theta)$ under $g(y_{\mathrm{p}} \mid y)$, and choose the tuning parameter τ. However, to simplify our approach, instead of finding and averaging over the full posterior predictive distribution $g(y_{\mathrm{p}} \mid y)$, we will use the simpler approach of finding the maximum likelihood estimate $\hat{\phi}$ under $g(y \mid \phi)$ and evaluating the expectation of $\log f(y_{\mathrm{p}} \mid X_{\mathrm{p}}, \theta)$ under $g(y_{\mathrm{p}} \mid \hat{\phi})$, plugging in the MLE.

Ignoring terms that do not involve θ, we have

$$\begin{aligned} &E\left\{\log f(v_{\mathrm{p}}, u_{\mathrm{p}}, Z_{\mathrm{p}} \mid X_{\mathrm{p}}, \theta)\right\} = \\ &\quad \sum_{i=1}^{n_{\mathrm{p}}} \left(\sum_{j=1}^{J} \Pr(v_{\mathrm{p}i} = j) \log p_j(x_{\mathrm{p}i}, \alpha) \right. \\ &\qquad \left. + \sum_{h \in \mathcal{E}} \Pr(i \in h) \left[-\frac{1}{2} \log \sigma_h^2 - \frac{1}{2\sigma_h^2} E\left\{ (u_{\mathrm{p}i} - x'_{\mathrm{p}i}\beta_h)^2 \mid i \in h \right\} \right] \right), \end{aligned}$$

where the probabilities and expectations are calculated under $g(y_{\mathrm{p}} \mid \hat{\phi})$.

The catalytic prior factors into independent priors on α and on (β_h, σ_h^2). We will treat these priors separately, which is helpful because of different constraints on the tuning parameters, as we will see below. First, let τ_α be the tuning parameter for the catalytic prior on α. Then, assuming a default prior $f(\alpha) \propto 1$, we have

$$\pi_{\tau,y}(\alpha) \propto \prod_{i=1}^{n_\mathrm{p}} \prod_{j=1}^{J} \{p_j(x_{\mathrm{p}i}, \alpha)\}^{(\tau_\alpha/n_\mathrm{p}) \Pr(v_{\mathrm{p}i}=j)}, \tag{3.13}$$

where the factor $1/n_\mathrm{p}$ in the exponent arises from the same arguments as in Section 3.3. A small value of τ_α relative to $n_\mathrm{p} = 13\,794$ will lead to a diffuse prior. It is not clear how to directly apply the recommendations of Clogg *et al.* (1991) for the choice of tuning parameter in this model because the pseudo-data are constructed for the complete-data likelihood, not the observed-data likelihood.

Next, we consider catalytic priors for (β_h, σ_h^2). Consider the terms

$$\sum_{i=1}^{n_\mathrm{p}} \Pr(i \in h) \left[-\frac{1}{2} \log \sigma_h^2 - \frac{1}{2\sigma_h^2} E\left\{ (u_{\mathrm{p}i} - x'_{\mathrm{p}i}\beta_h)^2 \mid i \in h \right\} \right]$$

in the expectation of the complete-data log likelihood. Recall that in Section 3.3, if we wanted to use the empirical distribution of the covariates, the catalytic prior for a regression model would take the form

$$\pi_{\tau,y}(\theta) \propto f(\theta) \exp\left\{ \frac{\tau}{n_\mathrm{p}} \int g(z_\mathrm{p} \mid X_\mathrm{p}, y) \log f(z_\mathrm{p} \mid X_\mathrm{p}, \theta) dz_\mathrm{p} \right\}.$$

Here, the factor $\Pr(i \in h)$ can be viewed as a weight assigned to the ith individual, so rather than dividing τ by n_p, we divide by $\sum_{i=1}^{n_\mathrm{p}} \Pr(i \in h)$. Letting τ_h be the tuning parameter for the

catalytic prior on (β_h, σ_h^2) and letting $r_{ih} = \Pr(i \in h)$, we have

$$\pi_{\tau,y}(\beta_h, \sigma_h^2) \propto f(\beta_h, \sigma_h^2)(\sigma_h^2)^{-\tau_h/2} \times \exp\left[-\frac{\tau_h}{2\sigma_h^2 \sum_{i=1}^{n_\mathrm{p}} r_{ih}} \sum_{i=1}^{n_\mathrm{p}} r_{ih} \left\{\hat{\sigma}_h^2 + ||x'_{\mathrm{p}i}(\beta_h - \hat{\beta}_h)||^2\right\}\right], \tag{3.14}$$

where $\hat{\sigma}_h^2$ and $\hat{\beta}_h$ are estimates under the prior generating model. If $\tau_h > k$ and $f(\beta_h, \sigma_h^2) \propto 1/\sigma_h^2$, we can express the catalytic prior $\pi_{\tau,y}(\beta_h, \sigma_h^2)$ as

$$\sigma_h^2 \sim \frac{\tau_h \hat{\sigma}_h^2}{\chi^2_{\tau_h - k}},$$
$$(\beta_h \mid \sigma_h^2) \sim N\left(\hat{\beta}_h, \tau_h^{-1}\sigma_h^2 (X'_{\mathrm{p}} \Lambda_h X_{\mathrm{p}})^{-1}\right),$$

where $\Lambda_h = \mathrm{diag}(r_{1h}, \ldots, r_{n_\mathrm{p}h}) / \sum_{i=1}^{n_\mathrm{p}} r_{ih}$.

3.7.3 Choice of prior generating model

We will compare two prior generating models. The first has the same structure as the full model, but with fewer covariates. We will only include the constant term and the design weights as predictors in both the multinomial logistic regression for principal stratum memberships and the linear regressions for log wages. One possible problem with this model is that discarding so many covariates makes the assumption that data are missing at random less plausible. The second prior generating model includes all of the covariates, but assumes that the parameters do not change over time. We have observations at three different points in time, at 52, 130, and 208 weeks after treatment assignment. Frumento *et al.* (2012) analyzed these as three separate datasets, but for our second prior generating model we analyze all three weeks together.

An alternative prior generating model could simplify the full model by imposing additional restrictions, such as monotonicity of employment, no effect of assignment on employment

for compliers, or no affect of assignment on wages for always-employed compliers (Frumento *et al.*, 2012). We find this less attractive than discarding covariates because if we make such meaningful restrictions in our prior generating model, then if the final inferences seem to support such restrictions, it will be less clear if this can be attributed to the data or just to the prior.

3.7.4 CHOICE OF τ

Finally, we need to choose values of the tuning parameters in the catalytic priors. We choose $\tau_\alpha = 5$, which is much less than $n = 13\,794$, and hence should lead to a very diffuse prior on α relative to the likelihood. Using the sums of $\Pr(i \in h)$ to scale the tuning parameters in the wage parameter priors ensures that setting $\tau_h = \tau_{\beta,\sigma}$ for all h can be interpreted as adding the approximately same number of pseudo-observations to each wage stratum. One subtlety is that $\tau_{\beta,\sigma}$ must be strictly greater than k, the number of columns in X, to guarantee a proper posterior distribution. Here, $k = 23$, so we can try $\tau_{\beta,\sigma} = 24$. We restrict $\tau_{\beta,\sigma} > k$ because it is possible for some of the latent groups in $\mathcal{E}$ to have no actual observations assigned to them (see Frumento *et al.*, 2012, Table 3). Thus, to ensure that the wage regressions on k predictors are estimable even if these latent strata have no observations, we must add greater than k pseudo-observations. The restriction $\tau_{\beta,\sigma} > k$ is an argument for using different tuning parameters in the catalytic priors on α and on (β_h, σ_h^2) because there is no such restriction on τ_α.

Interpreting τ_α and $\tau_{\beta,\sigma}$ as prior sample sizes is less straightforward here because the pseudo-dataset includes information not available in the observed dataset, namely the stratum memberships. It is tempting to directly compare these tuning parameters with the observed sample size $13\,794$ and conclude that the pseudo-data should contribute, say, roughly $\tau_\alpha/(\tau_\alpha + 13\,794)$ of the information in the final inferences on α, but we do not know how much more an individual with fully specified stratum memberships will constrain the likelihood

than one with memberships only partially specified. Nonetheless, it is appealing that we can interpret the tuning parameters as the number of fully stratified pseudo-observations we have added to the dataset.

3.7.5 Results

We have analyzed the data for week 52 under the catalytic prior of Section 3.7.2, with $\tau_\alpha = 5$ and $\tau_{\beta,\sigma} = 24$. Figure 3.14 displays estimated posterior densities of the following causal effects:

- $\Delta^{(ZS)}$, the average effect of treatment assignment on employment;
- $\Delta^{(DS)}$, the average treatment effect on employment for compliers; and
- $\Delta^{(DW)}$, the average treatment effect on wages for always-employed compliers.

Substantively, these posteriors largely agree with the estimates of Frumento *et al.* (2012). However, using ratios of maximized likelihoods, Frumento *et al.* (2012) concluded that the restriction $\Delta^{(DS)} = 0$, or no assignment effect on employment for compliers, was supported by the data. In contrast, the posterior interval for $\Delta^{(DS)}$ under this catalytic prior does not contain zero, suggesting the opposite conclusion. We note that there is very little difference between the posterior distributions under the two prior generating models. We stress that this is a preliminary analysis. We have included it to demonstrate how to develop catalytic priors for a realistically complex model.

3.8 Conclusion

In this chapter, we have investigated the application of catalytic priors in models of increasing complexity. In relatively simple models, catalytic priors often perform extremely well. In logistic regression, the original approach of Clogg *et al.* (1991) turns out to be hard to beat in

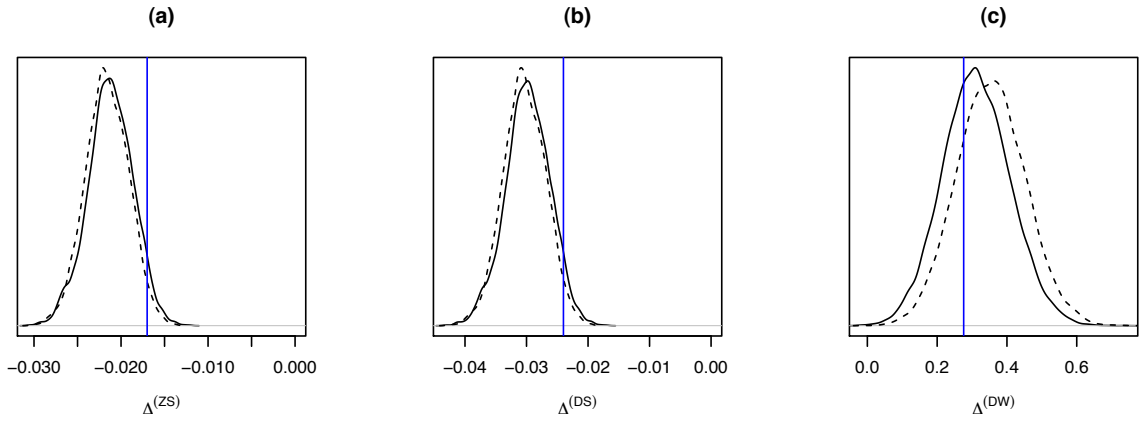


Figure 3.14: Posterior distributions of average effects of (a) assignment to treatment on employment, (b) receipt of treatment on employment for compliers, and (c) receipt of treatment on wages for always-employed compliers, for the data at week 52. Solid vertical lines give maximum likelihood estimates reported in Frumento *et al.* (2012). The solid density curve is the estimated posterior distribution under the prior generating model that only includes an intercept and the design weights as covariates; the dashed curve is under the prior generating that pools the data across time.

a wide range of circumstances. As for more complex models, Section 3.7 demonstrates that catalytic priors can be developed for models for causal inference that account for real-world complications, and Section 3.6.2 draws a link between catalytic priors and the approach of Airoldi and Blocker (2013), which outperforms competing methods on benchmark datasets.

There is much room for future work. It would be interesting and useful to develop more theoretical guidance for the choice of prior generating model g and tuning parameter τ. More attempts to apply catalytic priors to complex models in practical settings would also be valuable, both in order to compare the performance of catalytic priors with competing approaches and to indicate where there is a need for extensions to the methodology developed here.

A

Appendix

A.1 Supplementary tables and figures

Table A.1: Comparison of maximum likelihood (ML), scaled Cauchy (SC) priors (Gelman *et al.*, 2008), and Clogg *et al.*'s priors (CP) for a simulated $2 \times 2 \times 2$ table with high probability of sampling zeros. 5000 simulated data sets were generated. In contrast to Table 3.4, here the row totals m_j were fixed for all simulations.

	$N = 30$			$N = 100$		
	ML	SC	CP	ML	SC	CP
1. Fraction of samples where estimates exist	0.47	1	1	0.89	1	1
2. Mean squared error (given existence)						
(a) $\hat{\beta}_0$	.24	.19	.25	.06	.05	.06
(b) $\hat{\beta}_1$	.26	.28	.24	.11	.13	.12
(c) $\hat{\beta}_2$	.29	.29	.26	.11	.13	.12
3. Coverage of 95% intervals for coefficients (percent of samples where estimates exist)						
(a) β_0	97.3	97.2	97.5	95.9	96.4	96.0
(b) β_1	93.9	91.2	93.9	97.1	92.2	95.5
(c) β_2	91.6	88.0	93.2	96.9	92.8	95.4
4. Coverage of 95% intervals for logits (percent of samples where estimates exist)						
(a) $(x_{j1}, x_{j2}) = (-1, -1)$	92.1	87.8	92.8	97.0	92.0	95.0
(b) $(1, -1)$	97.3	95.8	97.9	95.3	95.5	95.7
(c) $(-1, 1)$	97.1	96.9	97.2	95.7	95.7	95.9
(d) $(1, 1)$	93.7	90.5	91.6	96.8	92.3	94.5
5. Width of 95% intervals for logits (given existence)						
(a) $(x_{j1}, x_{j2}) = (-1, -1)$	4.76	4.09	4.81	3.34	2.86	3.09
(b) $(1, -1)$	2.85	2.69	2.77	1.51	1.47	1.48
(c) $(-1, 1)$	2.43	2.28	2.36	1.28	1.25	1.26
(d) $(1, 1)$	4.35	3.88	4.32	3.12	2.69	2.87

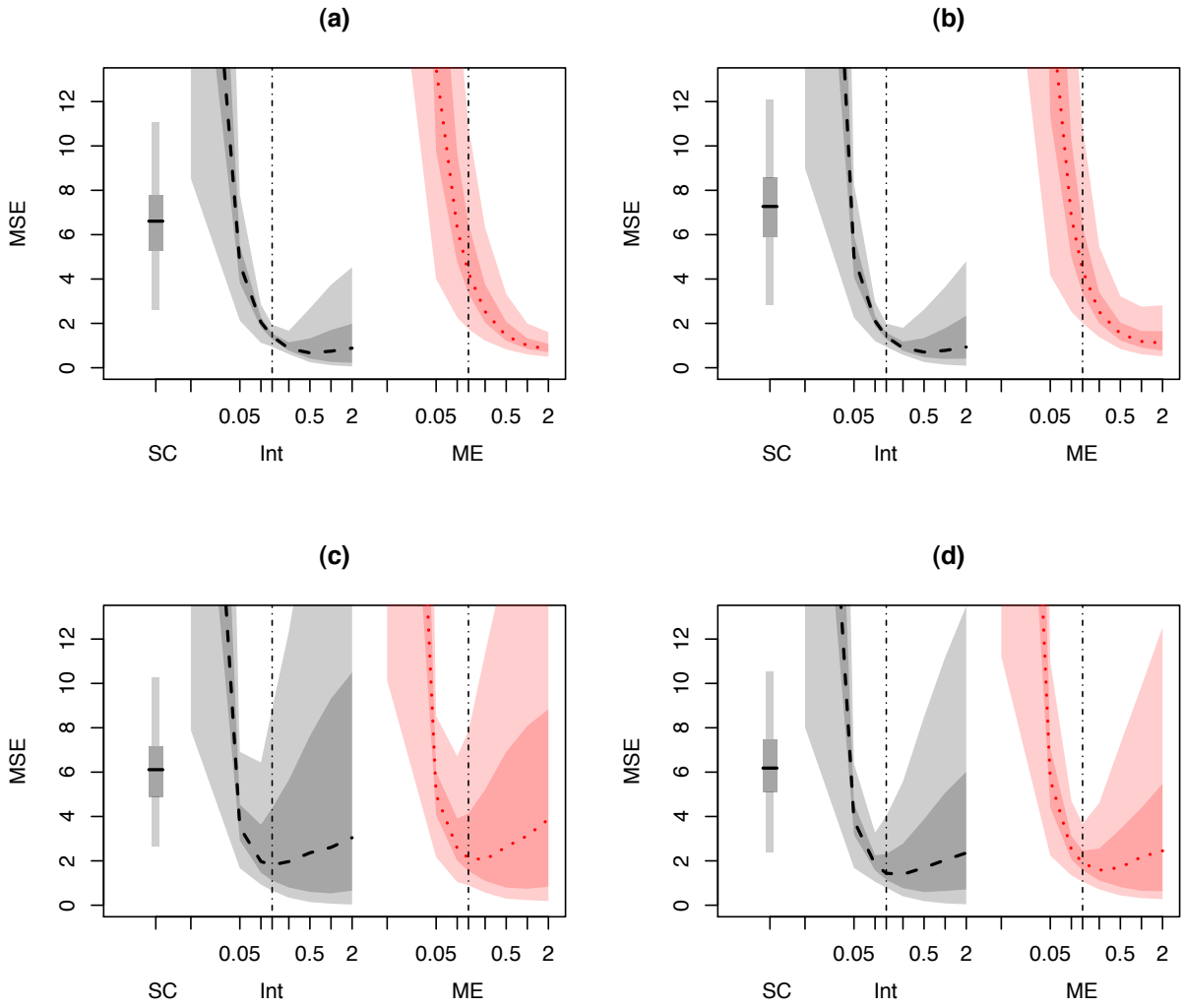


Figure A.1: Simulation results for a logistic regression with eight binary covariates and $N = 120$ total observations. These are the same results as in Figure 3.7, except that mean squared error is plotted on a raw scale over a smaller range, and maximum likelihood results are omitted because their medians are out of range.

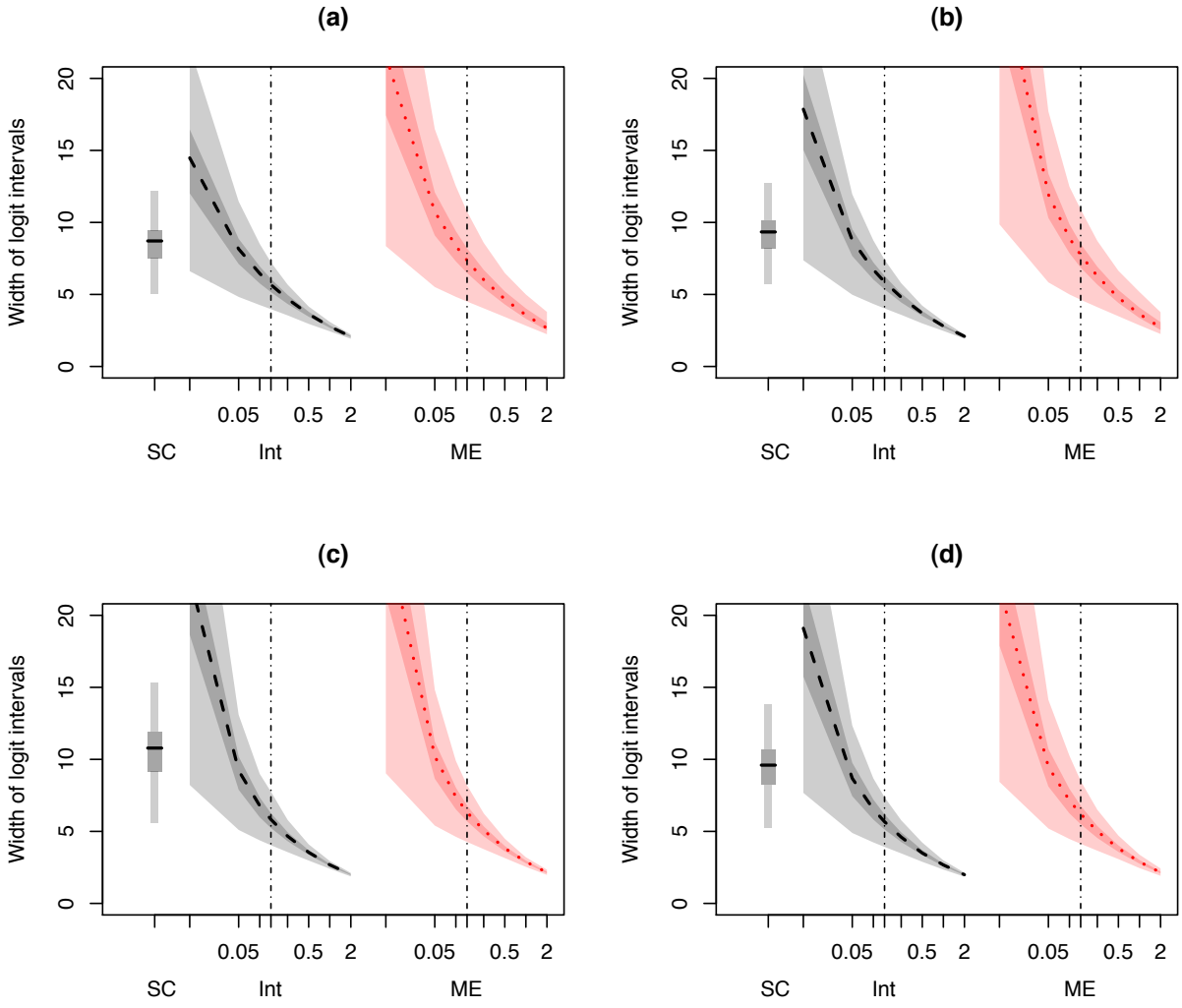


Figure A.2: Width of intervals for logits in a logistic regression simulation with eight binary covariates and $N = 120$ total observations. A zoomed-in version of Figure 3.9, excluding the results for maximum likelihood.

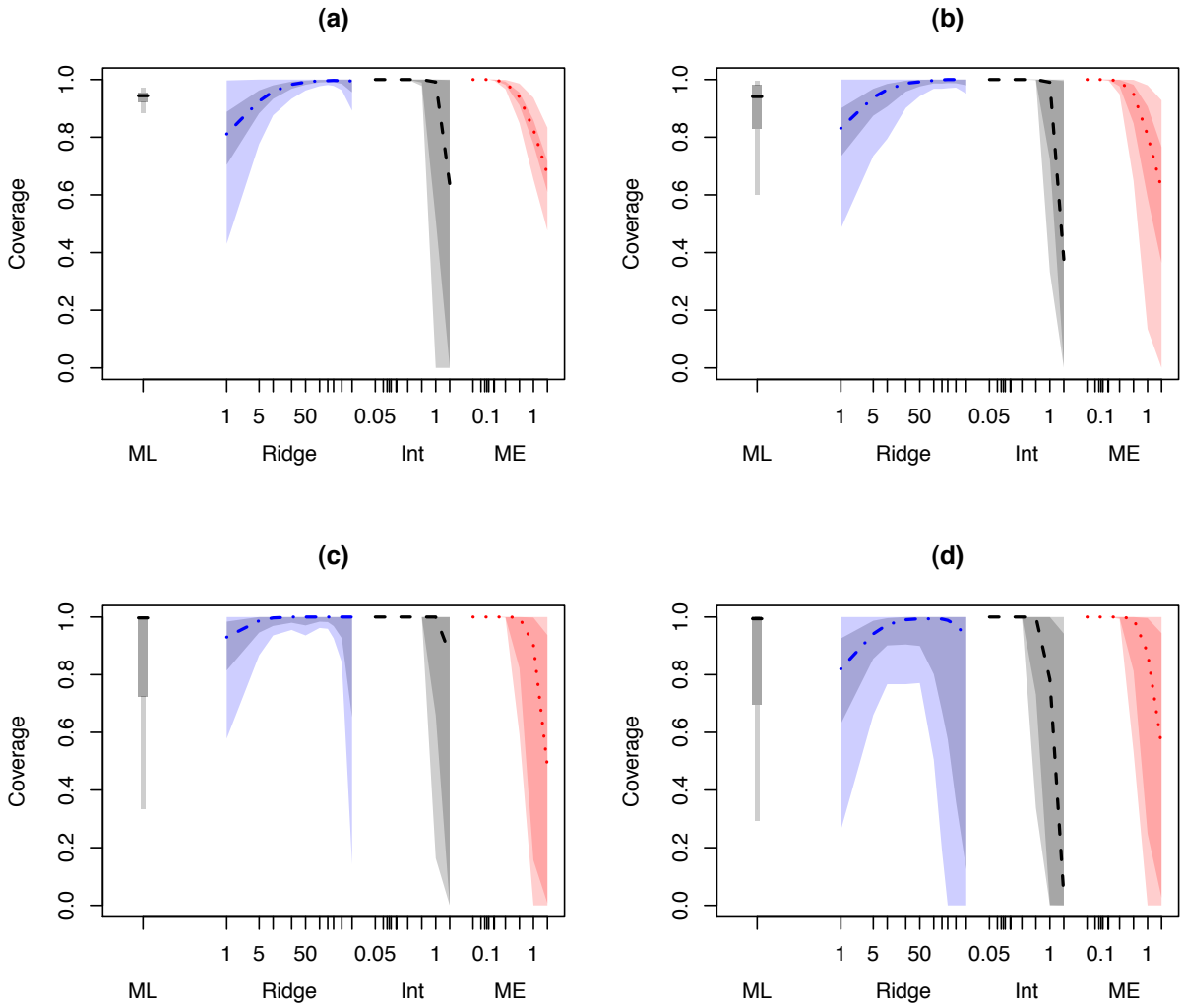


Figure A.3: Coverage of 95% intervals for linear predictors $x_j'\beta$ in the linear regression simulation described in Section 3.5.2. A zoomed-out version of Figure 3.11.

A.2 COMPARING A CATALYTIC PRIOR WITH A HIERARCHICAL MODEL

It is natural to ask how a posterior under a catalytic prior will behave differently than a fully Bayesian hierarchical model designed to shrink toward a simpler model with the same structure as g. To begin answering that question, we can compare catalytic priors in the multivariate normal model to a simple hierarchical model. We will not assume that the variance in the full model is known. Instead, we assume that the full model likelihood is $y_i \mid \theta \sim N(\beta_i, \sigma^2)$, independently for $i = 1, \ldots, n$, where $\theta = (\beta_1, \ldots, \beta_n, \sigma^2)$. For the catalytic prior, we use the prior generating model $y_i \mid \phi \sim N(\mu, V)$, where $\phi = (\mu, V)$ so that both the mean and variance are unknown. Using default priors, we use the following full model and prior generating model:

$$\begin{aligned} f(y \mid \theta): \quad & y_i \mid \theta \sim N(\beta_i, \sigma^2) \quad \text{independently, } i = 1, \ldots, n, \\ & f(\theta) \propto 1/\sigma^2 \\ g(y \mid \phi): \quad & y_i \mid \phi \sim N(\mu, V) \quad \text{independently, } i = 1, \ldots, n, \\ & g(\phi) \propto 1/V. \end{aligned} \tag{A.1}$$

We will compare our approach with the hierarchical model

$$\begin{aligned} & y_i \mid \theta, \mu \sim N(\beta_i, \sigma^2) \quad \text{independently, } i = 1, \ldots, n, \\ & \beta \mid \mu, \sigma^2 \sim N(\mu \mathbf{1}_n, \tau^{-1}\sigma^2 I_n) \\ & p(\mu, \sigma^2) \propto 1/\sigma^2, \end{aligned} \tag{A.2}$$

where $\beta = (\beta_1, \ldots, \beta_n)$. In both models, τ will be considered fixed (although τ does not appear in our specification (A.1), it still denotes the tuning parameter of the catalytic prior). We have attempted to construct a fair comparison by ensuring that both models shrink estimates

of β_i toward the grand mean of the data, and τ governs the amount of shrinkage.

We are concerned with the posterior distribution of β given the observations y, under a catalytic prior and under the hierarchical model (A.2). Under the hierarchical model, we have

$$\begin{aligned}\beta \mid \sigma^2, y &\sim N\left(\frac{1}{1+\tau}y + \frac{\tau}{1+\tau}\bar{y}\mathbf{1}_n, \frac{\sigma^2}{1+\tau}I_n + \frac{\sigma^2}{n}\frac{\tau}{1+\tau}\mathbf{1}_n\mathbf{1}_n'\right)\\ \sigma^2 \mid y &\sim \frac{\tau}{1+\tau}\sum_{i=1}^{n}(y_i-\bar{y})^2\frac{1}{\chi^2_{n-1}}.\end{aligned}$$

Under the catalytic prior that uses our choices of f and g, we have

$$\begin{aligned}\beta \mid \sigma^2, y &\sim N\left(\frac{1}{1+\tau}y + \frac{\tau}{1+\tau}\bar{y}\mathbf{1}_n, \frac{\sigma^2}{1+\tau}I_n\right)\\ \sigma^2 \mid y &\sim \left\{\frac{\tau}{1+\tau} + \frac{\tau(1+n^{-1})}{n-3}\right\}\sum_{i=1}^{n}(y_i-\bar{y})^2\frac{1}{\chi^2_{\tau n}}.\end{aligned}$$

Thus, under both models, we have the same posterior means of β_i. However, under the catalytic prior, the individual β_i's are conditionally independent of each other, while they are correlated under the hierarchical model. If τ is small relative to n, this difference will be small. One key difference between the two models is that the degrees of freedom for the inverse-χ^2 posterior distribution for σ^2 are fixed as a function of τ under the hierarchical model, but depend on τ under the catalytic prior. If we interpret the catalytic prior as pseudo-data, it makes sense that we should no longer be able to estimate σ^2 well as τ as approaches 0.

The posterior variances of β_i, given by

$$\begin{aligned}\text{Var}_{\text{hier}}(\beta_i|y) &= \left\{\frac{1}{1+\tau} + \frac{\tau}{n(1+\tau)}\right\}\frac{\tau}{1+\tau}\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-3},\\ \text{Var}_{\text{cat}}(\beta_i|y) &= \left\{\frac{n-3}{\tau n-2}\left(\frac{1}{1+\tau}\right) + \frac{\tau(1+n^{-1})}{\tau n-2}\right\}\frac{\tau}{1+\tau}\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-3},\end{aligned}$$

exhibit more surprising behavior, as shown in Figure A.4. If $\tau < n/(n-2)$, then under the hierarchical model, the posterior variance of β_i *decreases* as τ decreases. This is because the

hierarchical model implies that the distribution of y_i, conditional on σ^2 and μ but marginalized over β, is $N(\mu, \sigma^2(1+\tau^{-1}))$. Intuitively, if τ is too small, the model compensates by decreasing σ^2 so that this marginal variance is closer to the observed variation around $\bar{y}$. Under the catalytic prior, β_i only has a posterior variance if $\tau > 2/n$, but when the variance exists, it increases as τ decreases, which accords with our intuition that a smaller prior sample size should constrain final inferences less.

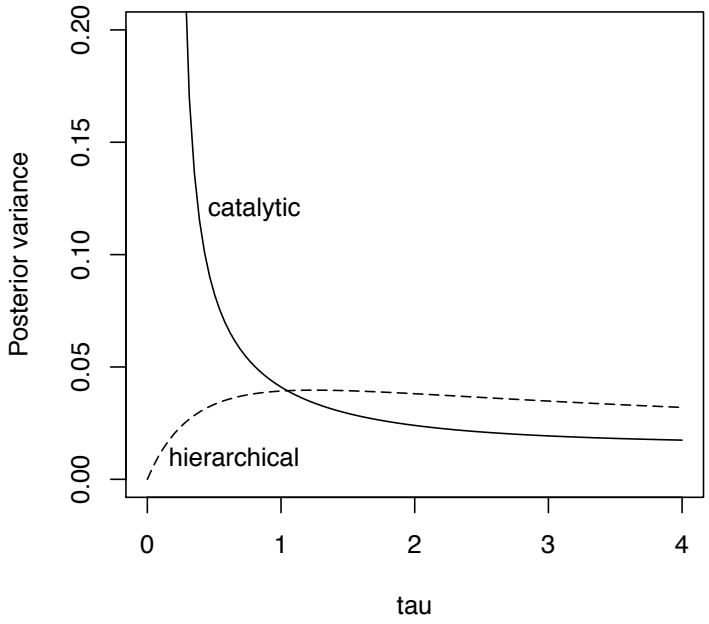


Figure A.4: The posterior variance Var$(\beta_i|y)$ versus τ, when $n = 10$ and $\sum_{i=1}^{n}(y_i - \bar{y})^2 = n - 3$. The solid line is for the posterior under the catalytic prior, the dashed line under the hierarchical model. The posterior variance under the catalytic prior does not exist if $\tau \leq 2/n$.

We do not explore imposing a model on τ and fitting it from the data, though that is an interesting question for future work. This example suggests that models that work well for top-level variance parameters in hierarchical models may not be appropriate for catalytic prior tuning parameters, even though much of our original intuition might lead us to believe, mis-

takenly, that they should behave similarly.

Bibliography

Airoldi, E. M. and Blocker, A. W. (2013). Estimating latent processes on a network from indirect measurements. *Journal of the American Statistical Association* **108**, 501, 149–164.

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62**, 3, 547–554.

Akaike, H. (1978). A new look at the Bayes procedure. *Biometrika* **65**, 1, 53–59.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 3rd edn.

Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., *Bayesian Statistics 4*, 35–60. Oxford University Press, Oxford, UK.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics* **37**, 2, 905–938.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)* **41**, 2, 113–147.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022.

Blocker, A. W. and Airoldi, E. M. (2011). Deconvolution of mixing time series on a graph. *Proceedings of the 27th conference on uncertainty in artificial intelligence (UAI)* 51–60.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion (with discussion). *Statistical Science* **16**, 2, 101–133.

Casella, G., Mengersen, K. L., Robert, C. P., and Titterington, D. M. (2002). Perfect samplers for mixtures of distributions. *Journal of the Royal Statistical Society: Series B (Methodological)* **64**, 4, 777–790.

Chen, M.-H. and Ibrahim, J. G. (2006). The relationship between the power prior and hierarchical models. *Bayesian Analysis* **1**, 3, 551–574.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association* **86**, 413, 68–78.

Craiu, R. V. and Meng, X.-L. (2011). Perfection within reach: Exact MCMC sampling. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, eds., *Handbook of Markov Chain Monte Carlo*, 205–232. Chapman & Hall/CRC.

Datta, G. S. and Mukerjee, R. (2004). *Probability matching priors: Higher order asymptotics*. Springer, New York.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of Statistics* **7**, 2, 269–281.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors: An empirical Bayes approach. *Journal of the American Statistical Association* **68**, 341, 117–130.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the Americal Statistical Association* **90**, 430, 577–588.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 1, 27–38.

Fraser, D. A. S., Reid, N., Marras, E., and Yi, G. Y. (2010). Default priors for Bayesian and frequentist inference. *Journal of the Royal Statistical Society: Series B (Methodological)* **72**, 5, 631–654.

Frumento, P., Mealli, F., Pacini, B., and Rubin, D. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association* **107**, 498, 450–466.

Fúquene, J. A., Cook, J. D., and Pericchi, L. R. (2009). A case for robust Bayesian priors with applications to clinical trials. *Bayesian Analysis* **4**, 4, 817–846.

Galindo-Garre, F., Vermunt, J. K., and Bergsma, W. P. (2004). Bayesian posterior estimation of logit parameters with small samples. *Sociological Methods and Research* **33**, 1, 88–117.

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

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* **2**, 4, 1360–1383.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science* **13**, 2, 163–185.

Häggström, O. and Nelander, K. (1999). On exact simulation from Markov random fields using coupling from the past. *Scand. J. Statist.* **26**, 395–411.

Heinze, G. (2006). A comparative investigation of the methods for logistic regression with separated or nearly separated data. *Statistics in Medicine* **25**, 4216–4226.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21**, 2409–2419.

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1**, 69–88.

Hobert, J. P., Robert, C. P., and Titterington, D. M. (1999). On perfect simulation for some mixtures of distributions. *Statist. Comp.* **9**, 287–298.

Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *Public Library of Science One* **7**, 2, e30126.

Huber, M. (1998). Exact sampling and approximate counting techniques. In *Proc. 30th Sympos. Theory Comp.*, 31–40. ACM, New York.

Huber, M. (2004). Perfect sampling using bounding chains. *The Annals of Applied Probability* **14**, 2, 734–753.

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science* **15**, 1, 46–60.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2003). On optimality properties of the power prior. *Journal of the American Statistical Association* **98**, 461, 204–213.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 361–379. University of California Press.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *The Physical Review* **106**, 4, 620–630.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics* **SSC-4**, 3, 227–241.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, 3rd edn.

Johnson, R. A. (1967). Asymptotic expansions associated with the nth power of a density. *The Annals of Mathematical Statistics* **38**, 4, 1266–1272.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**, 435, 1343–1370.

Kou, S. C. and McCullagh, P. (2009). Approximating the α-permanent. *Biometrika* **96**, 635–644.

Kou, S. C., Xie, X. S., and Liu, J. S. (2005). Bayesian analysis of single-molecule experimental data (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**, 3, 469–506.

Kullback, S. (1968). *Information theory and statistics*. Dover, New York.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 1, 79–86.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, Hoboken, New Jersey, 2nd edn.

McCullagh, P. and Møller, J. (2006). The permanental process. *Adv. Appl. Prob.* **38**, 873–888.

McCullagh, P. and Yang, J. (2006). Stochastic classification models. In M. Sanz-Solé, J. Soria, J. L. Varona, and J. Verdera, eds., *Proc. Int. Cong. Math.*, vol. 3, 669–686. European Mathematical Society, Zurich.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science* **9**, 4, 538–573.

Møller, J. (1999). Perfect simulation of conditionally specified models. *Journal of the Royal Statistical Society: Series B (Methodological)* **61**, 1, 251–264.

Mukhopadhyay, S. and Bhattacharya, S. (2012). Perfect simulation for mixtures with known and unknown number of components. *Bayesian Analysis* **7**, 3, 675–714.

Murdoch, D. and Meng, X.-L. (2001). Towards perfect sampling for Bayesian mixture priors. In E. George, ed., *Proc. ISBA 2000*, 381–390. Eurostat.

Murdoch, D. J. and Green, P. J. (1998). Exact sampling from a continuous state space. *Scand. J. Statist.* **25**, 3, 483–502.

Nicolaou, A. (1993). Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. *Journal of the Royal Statistical Society: Series B (Methodological)* **55**, 2, 377–390.

Polson, N. G. and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis* **7**, 2, 1–16.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Struct. Algor.* **9**, 1&2, 223–252.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Shore, J. E. and Johnson, R. W. (1981). Properties of cross-entropy minimization. *IEEE Transactions on Information Theory* **IT-27**, 4, 472–482.

Stein, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)* **24**, 2, 265–296.

Welch, B. L. and Peers, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *Journal of the Royal Statistical Society: Series B (Methodological)* 318–329.

Wilson, D. B. (2000). How to couple from the past using a read-once source of randomness. *Random Struct. Algor.* **16**, 85–113.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis* **13**, 157–170.