Binomial likelihoods and the Polya-Gamma distribution

James Scott
University of Texas at Austin

Describing joint work with:
Jesse Windle (UT-Austin/Duke)
Nick Polson (Chicago)

Outline

- A quick review of latent-variable methods in Bayesian inference
- Logit models and the Polya-Gamma family
- Sampling the PG distribution
- Examples
- An interesting connection with variational inference

	Age	AgeGroup	Race	Completed	InsuranceType	Location	PracticeType
515	21	18to26	Black	0	Military	Odenton	FamilyPractice
423	21	18to26	Black	0	PrivatePayer	Odenton	FamilyPractice
388	17	11to17	White	0	PrivatePayer	Odenton	Pediatric
6	11	11to17	Black	0	Medicaid	Odenton	Pediatric
1104	19	18to26	Black	0	Medicaid	Bayview	Pediatric
1412	19	18to26	Black	0	Medicaid	JohnsHopkins	OBGYN
1354	24	18to26	White	0	PrivatePayer	JohnsHopkins	OBGYN
318	18	18to26	Black	1	Military	Odenton	FamilyPractice
768	24	18to26	White	1	PrivatePayer	Odenton	OBGYN
29	13	11to17	Other/Unknown	0	PrivatePayer	Odenton	FamilyPractice
1173	14	11to17	Hispanic	0	PrivatePayer	Bayview	Pediatric
799	24	18to26	White	0	PrivatePayer	Odenton	OBGYN
633	24	18to26	White	1	PrivatePayer	WhiteMarsh	OBGYN
111	13	11to17	Other/Unknown	0	Medicaid	Odenton	Pediatric
69	15	11to17	Black	0	PrivatePayer	Odenton	FamilyPractice
559	12	11to17	Black	0	Military	Odenton	Pediatric
1289	26	18to26	White	1	HospitalBased	Bayview	OBGYN
1127	18	11to17	White	0	Medicaid	Bayview	Pediatric
1250	18	11to17	Black	0	PrivatePayer	Bayview	Pediatric
1098	15	11to17	White	1	Medicaid	Bayview	Pediatric
378	12	11to17	White	1	Military	Odenton	FamilyPractice
702	26	18to26	White	0	PrivatePayer	WhiteMarsh	OBGYN
....							

	Age	AgeGroup	Race	Completed	InsuranceType	Location	PracticeType
515	21	18to26	Black	0	Military	Odenton	FamilyPractice
423	21	18to26	Black	0	PrivatePayer	Odenton	FamilyPractice
388	17	11to17	White	0	PrivatePayer	Odenton	Pediatric
6	11	11to17	Black	0	Medicaid	Odenton	Pediatric
1104	19	18to26	Black	0	Medicaid	Bayview	Pediatric
1412	19	18to26	Black	0	Medicaid	JohnsHopkins	OBGYN
1354	24	18to26	White	0	PrivatePayer	JohnsHopkins	OBGYN
318	18	18to26	Black	1	Military	Odenton	FamilyPractice
768	24	18to26	White	1	PrivatePayer	Odenton	OBGYN
29	13	11to17	Other/Unknown	0	PrivatePayer	Odenton	FamilyPractice
1173	14	11to17	Hispanic	0	PrivatePayer	Bayview	Pediatric
799	24	18to26	White	0	PrivatePayer	Odenton	OBGYN
633	24	18to26	White	1	PrivatePayer	WhiteMarsh	OBGYN
111	13	11to17	Other/Unknown	0	Medicaid	Odenton	Pediatric
69	15	11to17	Black	0	PrivatePayer	Odenton	FamilyPractice
559	12	11to17	Black	0	Military	Odenton	Pediatric
1289	26	18to26	White	1	HospitalBased	Bayview	OBGYN
1127	18	11to17	White	0	Medicaid	Bayview	Pediatric
1250	18	11to17	Black	0	PrivatePayer	Bayview	Pediatric
1098	15	11to17	White	1	Medicaid	Bayview	Pediatric
378	12	11to17	White	1	Military	Odenton	FamilyPractice
702	26	18to26	White	0	PrivatePayer	WhiteMarsh	OBGYN
....							

A classical analysis will likely begin with logistic regression.

(log-odds interpretation, existence of a minimal sufficient statistic, etc.)

	Age	AgeGroup	Race	Completed	InsuranceType	Location	PracticeType
515	21	18to26	Black	0	Military	Odenton	FamilyPractice
423	21	18to26	Black	0	PrivatePayer	Odenton	FamilyPractice
388	17	11to17	White	0	PrivatePayer	Odenton	Pediatric
6	11	11to17	Black	0	Medicaid	Odenton	Pediatric
1104	19	18to26	Black	0	Medicaid	Bayview	Pediatric
1412	19	18to26	Black	0	Medicaid	JohnsHopkins	OBGYN
1354	24	18to26	White	0	PrivatePayer	JohnsHopkins	OBGYN
318	18	18to26	Black	1	Military	Odenton	FamilyPractice
768	24	18to26	White	1	PrivatePayer	Odenton	OBGYN
29	13	11to17	Other/Unknown	0	PrivatePayer	Odenton	FamilyPractice
1173	14	11to17	Hispanic	0	PrivatePayer	Bayview	Pediatric
799	24	18to26	White	0	PrivatePayer	Odenton	OBGYN
633	24	18to26	White	1	PrivatePayer	WhiteMarsh	OBGYN
111	13	11to17	Other/Unknown	0	Medicaid	Odenton	Pediatric
69	15	11to17	Black	0	PrivatePayer	Odenton	FamilyPractice
559	12	11to17	Black	0	Military	Odenton	Pediatric
1289	26	18to26	White	1	HospitalBased	Bayview	OBGYN
1127	18	11to17	White	0	Medicaid	Bayview	Pediatric
1250	18	11to17	Black	0	PrivatePayer	Bayview	Pediatric
1098	15	11to17	White	1	Medicaid	Bayview	Pediatric
378	12	11to17	White	1	Military	Odenton	FamilyPractice
702	26	18to26	White	0	PrivatePayer	WhiteMarsh	OBGYN
....							

A classical analysis will likely begin with logistic regression.

(log-odds interpretation, existence of a minimal sufficient statistic, etc.)

But a Bayesian analysis will almost invariably use a probit model.

This is due to Albert and Chib's simple latent-variable method.

Bayesian Analysis of Binary and Polychotomous Response Data

JAMES H. ALBERT and SIDDHARTHA CHIB*

A vast literature in statistics, biometrics, and econometrics is concerned with the analysis of binary and polychotomous response data. The classical approach fits a categorical response regression model using maximum likelihood, and inferences about the model are based on the associated asymptotic theory. The accuracy of classical confidence statements is questionable for small sample sizes. In this article, exact Bayesian methods for modeling categorical response data are developed using the idea of data augmentation. The general approach can be summarized as follows. The probit regression model for binary outcomes is seen to have an underlying normal regression structure on latent continuous data. Values of the latent data can be simulated from suitable truncated normal distributions. If the latent data are known, then the posterior distribution of the parameters can be computed using standard results for normal linear models. Draws from this posterior are used to sample new latent data, and the process is iterated with Gibbs sampling. This data augmentation approach provides a general framework for analyzing binary regression models. It leads to the same simplification achieved earlier for censored regression models. Under the proposed framework, the class of probit regression models can be enlarged by using mixtures of normal distributions to model the latent data. In this normal mixture class, one can investigate the sensitivity of the parameter estimates to the choice of "link function," which relates the linear regression estimate to the fitted probabilities. In addition, this approach allows one to easily fit Bayesian hierarchical models. One specific model considered here reflects the belief that the vector of regression coefficients lies on a smaller dimension linear subspace. The methods can also be generalized to multinomial response models with $J > 2$ categories. In the ordered multinomial model, the J categories are ordered and a model is written linking the cumulative response probabilities with the linear regression structure. In the unordered multinomial model, the latent variables have a multivariate normal distribution with unknown variance–covariance matrix. For both multinomial models, the data augmentation method combined with Gibbs sampling is outlined. This approach is especially attractive for the multivariate probit model, where calculating the likelihood can be difficult.

KEY WORDS: Binary probit; Data augmentation; Gibbs sampling; Hierarchical Bayes modeling; Latent data; Logit model; Multinomial probit; Residual analysis; Student-t link function.

1. INTRODUCTION

Suppose that N independent binary random variables $Y_1, \ldots, Y_N$ are observed, where Y_i is distributed Bernoulli with probability of success p_i. The p_i are related to a set of covariates that may be continuous or discrete. Define the binary regression model as $p_i = H(\mathbf{x}_i^T\beta)$, $i = 1, \ldots, N$, where β is a $k \times 1$ vector of unknown parameters, $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ik})$ is a vector of known covariates, and $H(\quad)$ is a known cdf linking the probabilities p_i with the linear structure $\mathbf{x}_i^T\beta$. The probit model is obtained if H is the standard Gaussian cdf, whereas the logit model is obtained if H is the logistic cdf. (For general discussions of this class of models, see Cox 1971, Finney 1947, Nelder and McCullagh 1989, and Maddala 1983.)

Let $\pi(\beta)$, a proper or improper prior density, summarize our prior information about β. Then the posterior density of β is given by

$$\pi(\beta \mid \text{data}) = \frac{\pi(\beta)\prod_{i=1}^N H(\mathbf{x}_i^T\beta)^{y_i}(1 - H(\mathbf{x}_i^T\beta))^{1-y_i}}{\int \pi(\beta)\prod_{i=1}^N H(\mathbf{x}_i^T\beta)^{y_i}(1 - H(\mathbf{x}_i^T\beta))^{1-y_i}\, d\beta}, \quad (1)$$

which is largely intractable. Letting $N_k(\mu, \Sigma)$ denote the k-variate multivariate normal distribution with mean μ and variance-covariance matrix Σ, the usual asymptotic approximation is that β is distributed $N_k(\hat{\beta}, \mathbf{I}(\hat{\beta})^{-1})$, where $\hat{\beta}$ is the posterior mode and $\mathbf{I}(\hat{\beta})$ is the negative of the second derivative matrix evaluated at the mode. When a uniform prior is chosen for β, $\hat{\beta}$ is the maximum likelihood estimate (MLE) and $\mathbf{I}(\quad)$ is the observed information matrix. From a non-Bayesian viewpoint, Griffiths, Hill, and Pope (1987) found the MLE to have significant bias for small samples. Zellner and Rossi (1984), from a Bayesian approach, also commented on the inaccuracy of the normal approximation for small N. For a small number of parameters, they summarized the posterior using numerical integration. For large models (k large), they computed posterior moments by Monte Carlo integration with a multivariate Student's t importance function.

In this article we introduce a simulation-based approach for computing the exact posterior distribution of β. Suppose that the link function H is the standard Gaussian cdf (the probit case). The key idea is to introduce N independent latent variables $Z_1, \ldots, Z_N$ into the problem, where Z_i is distributed $N(\mathbf{x}_i^T\beta, 1)$, and define $Y_i = 1$ if $Z_i > 0$ and $Y_i = 0$ if $Z_i \le 0$. Observe that if the Z_i are known and a multivariate normal prior is chosen for β, then the posterior distribution for β can be derived using standard normal linear model results. The Z_i are of course unknown; however, given the data Y_i, the distribution of Z_i follows a truncated normal distribution. These principal observations, combined with the tool of Gibbs sampling, allow us to simulate from the exact posterior distribution of β. This approach is very similar to the data augmentation/Gibbs sampling framework used in censored regression models (Chib 1992; Wei and Tanner 1990).

* James H. Albert is Professor, Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403. Siddhartha Chib is Associate Professor of Economics at the Olin School of Business, Washington University, St. Louis, MO 63130. The work was completed while the second author held a joint visiting appointment with the Economics Department and the Olin School of Business at Washington University. The authors thank the editor, the associate editor, and two referees for many helpful comments.

Journal of the American Statistical Association
June 1993, Vol. 88, No. 422, Theory and Methods

$$y_i \sim \textsf{Bern}(w_i)\,, \quad w_i = \Phi(x_i^T\beta)$$

$$\Longleftrightarrow$$

$$y_i = \mathbb{I}_{z_i>0}$$

$$z_i = x_i^T\beta + \epsilon_i$$

$$\epsilon_i \sim N(0,1)$$

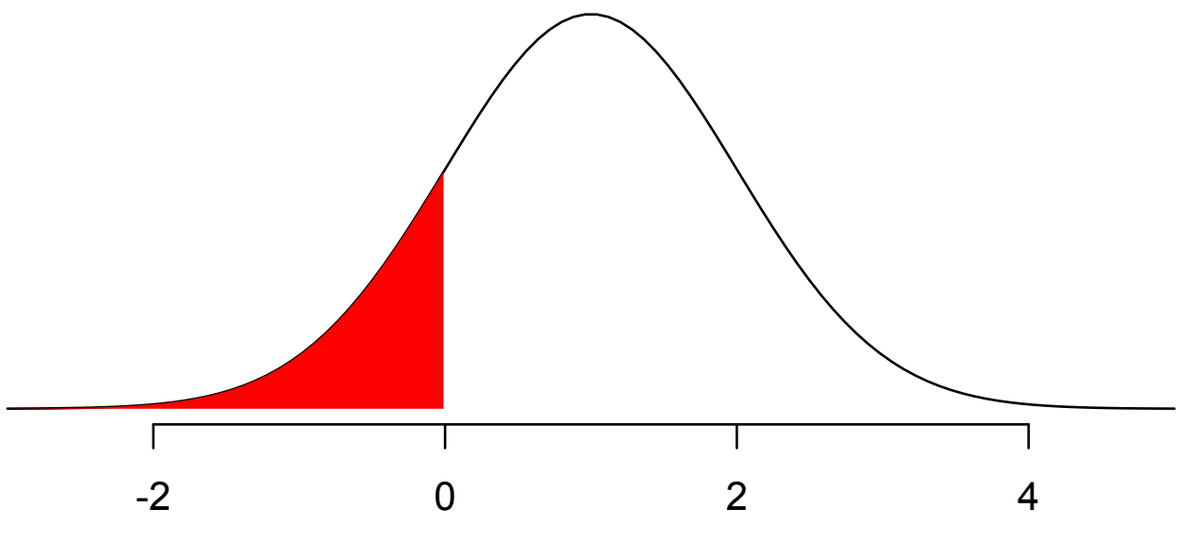
```
library(msm)

for(t in 1:1000)
{
 # Update latents
 z = rtnorm(N, X %*% beta, L, U)

 # Update regression coefficients
 V = solve( XtX + PrPrec )
 m = V %*% (PrPrec %*% PrM + t(X) %*% z)
 beta = t(rmvnorm(1,m,V))
 BetaSave[t,] = beta
}
```

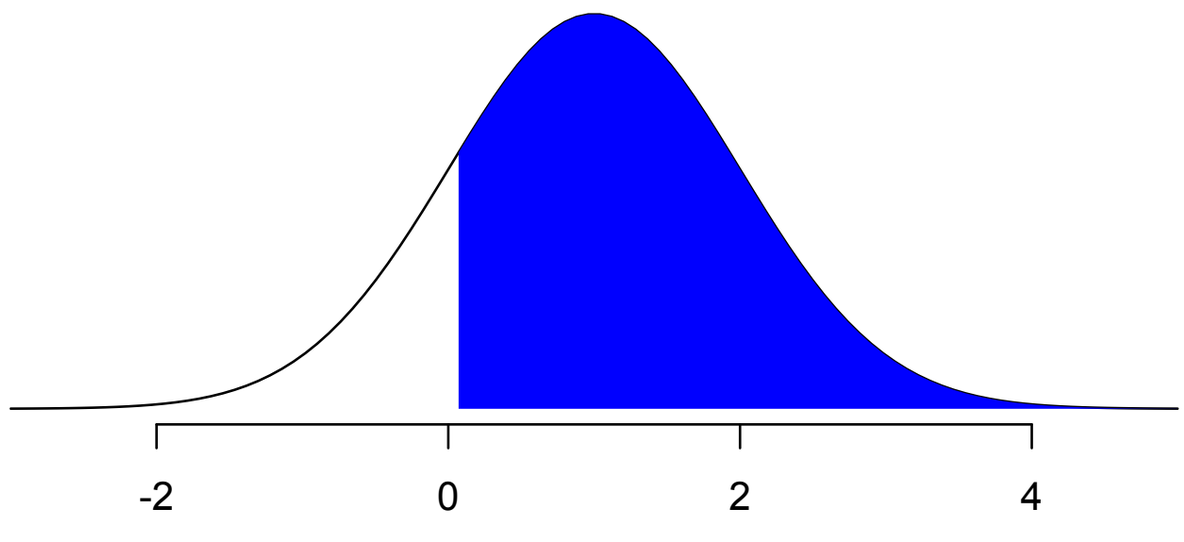
The likelihood in a probit model factorizes as

$$L(\beta) = \prod_{i=1}^{N} \{\Phi(x_i^T \beta)\}^{1-y_i} \cdot \{1 - \Phi(x_i^T \beta)\}^{y_i} .$$

Notice that each likelihood term $L_i(\beta)$ can be written as the integral of a simpler quantity:

$$\begin{aligned} L_i(\beta) &= \frac{1}{\sqrt{2\pi}} \int_{A_i} \exp\{-(z_i - x_i^T \beta)^2/2\} \, dz_i \\ A_i &= \begin{cases} (-\infty, 0) , & y_i = 0 \\ (0, \infty) , & y_i = 1 . \end{cases} \end{aligned}$$



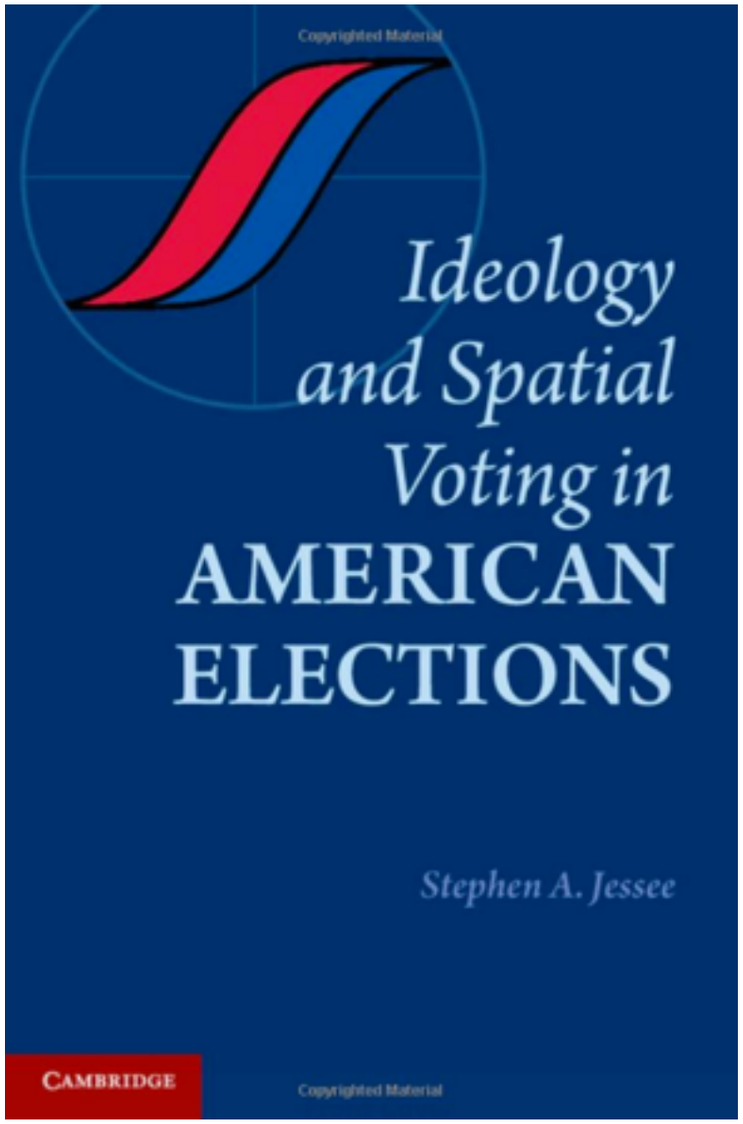


This allows us to write the posterior for β as the marginal of a higher-dimensional joint distribution:

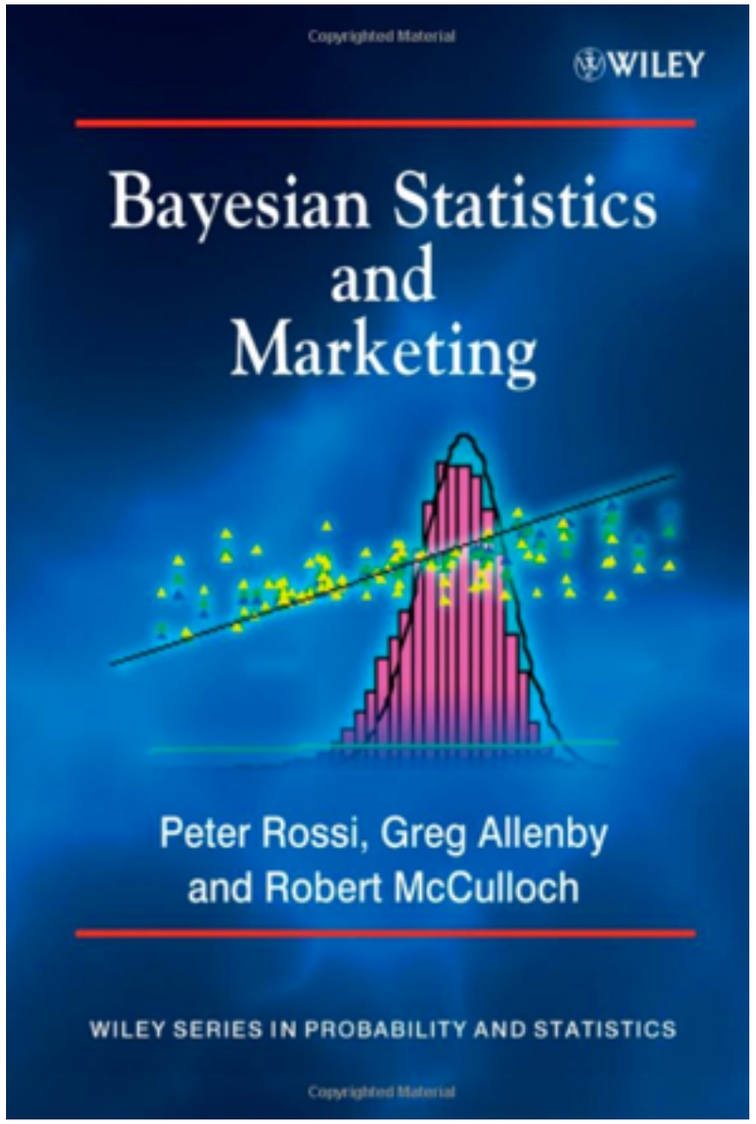
$$
\begin{aligned}
p(\beta \mid Y) &\propto p(\beta) L(\beta) \\
&= p(\beta) \prod_{i=1}^{N} \int_{A_i} \phi(z_i; x_i^T \beta, 1)\, dz_i \\
&= \int_{\mathbb{R}^n} \mathbb{I}\left\{ z \in \prod_{i=1}^{n} A_i \right\} p(\beta) \prod_{i=1}^{N} \phi(z_i; x_i^T \beta, 1)\, dz \\
&\propto \int_{R^n} p(\beta, z \mid Y) dz
\end{aligned}
$$

We can then sample from the joint distribution $p(\beta, z \mid Y)$ by iteratively sampling $p(\beta \mid z, Y)$ and $p(z \mid \beta, Y)$.

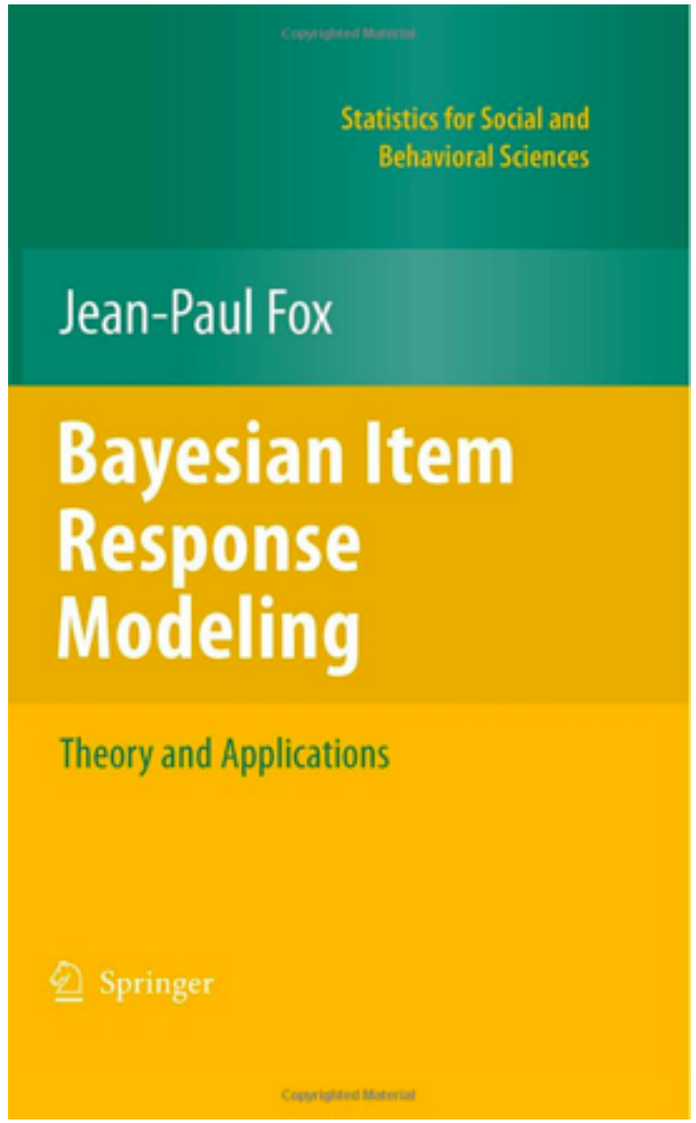
Because this method is simple and easily generalized, Bayesian probit models have become widely used by non-statisticians.



Political science
(probit factor models)



Market research
(discrete-choice models)



Psychometrics
(Rasch-like models)

The situation is very different for logistic regression.

$$y_i \sim \mathsf{Bern}(w_i)\,, \quad w_i = \frac{1}{1+\exp(-x_i^T\beta)}$$

Even the basic version is hard:

$$\begin{aligned} p(\beta \mid Y) &\propto p(\beta) \cdot \prod_{i=1}^{N} \frac{\{\exp(x_i^T\beta)\}^{y_i}}{1+\exp(x_i^T\beta)} \\ &\stackrel{?}{=} \int p(\beta, z \mid Y) dz\,. \end{aligned}$$

Until now, there has been no known, simple solution to this integral equation.

Should this matter to non-Bayesians?

Should this matter to non-Bayesians?

$$y_{ij} \sim \text{Bern}(w_{ij})$$

$$w_{ij} = \frac{1}{1 + \exp\{-(x_{ij}^T \beta + \alpha_i)\}}$$

$$\alpha_i \sim N(0, \tau^2)$$

Should this matter to non-Bayesians?

$$y_{ij} \sim \text{Bern}(w_{ij})$$

$$w_{ij} = \frac{1}{1 + \exp\{-(x_{ij}^T \beta + \alpha_i)\}}$$

$$\alpha_i \sim N(0, \tau^2)$$

The off-the-shelf classical fit:

```
Generalized linear mixed model fit by
the Laplace approximation
Formula: cbind(swim, set) ~ (1 | Dad)
   Data: coral
   AIC   BIC logLik deviance
 54.94 57.53 -25.47    50.94
Random effects:
 Groups Name        Variance   Std.Dev.
 Dad    (Intercept) 2.7e-19    5.2e-10
Number of obs: 27, groups: Dad, 9
```

Should this matter to non-Bayesians?

$$y_{ij} \sim \text{Bern}(w_{ij})$$

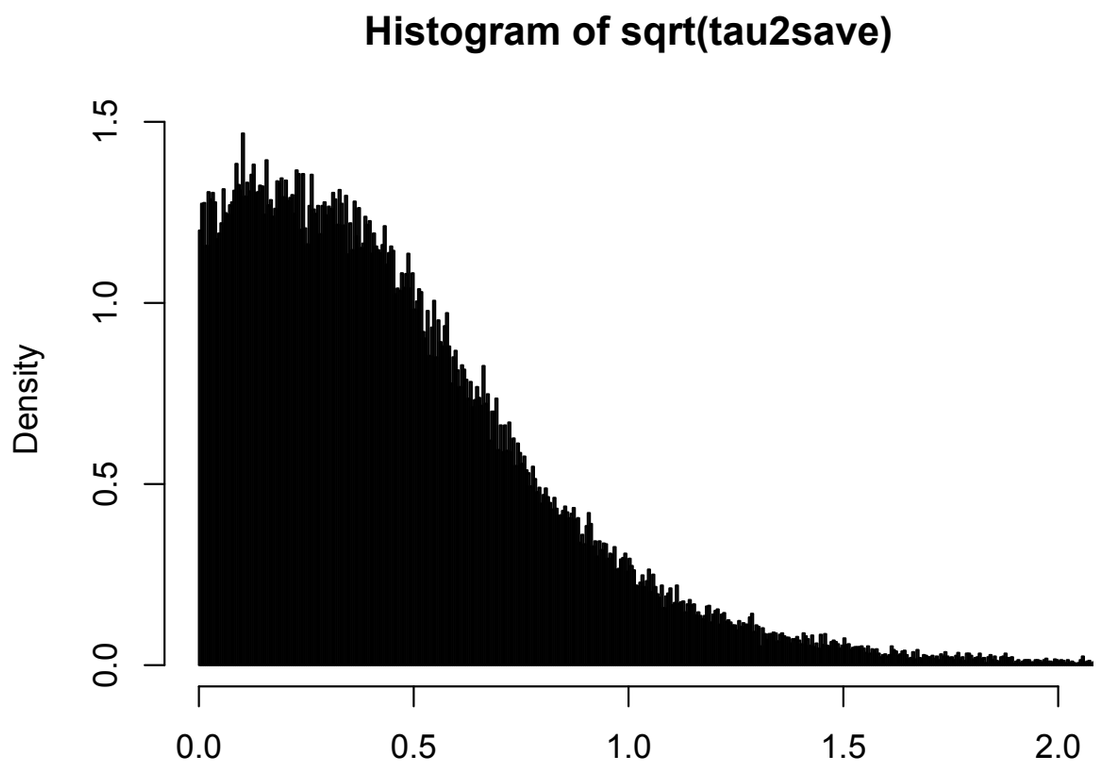
$$w_{ij} = \frac{1}{1 + \exp\{-(x_{ij}^T \beta + \alpha_i)\}}$$

$$\alpha_i \sim N(0, \tau^2)$$

The off-the-shelf classical fit:

```
Generalized linear mixed model fit by
the Laplace approximation
Formula: cbind(swim, set) ~ (1 | Dad)
   Data: coral
   AIC   BIC logLik deviance
 54.94 57.53 -25.47    50.94
Random effects:
 Groups Name        Variance   Std.Dev.
 Dad    (Intercept) 2.7e-19    5.2e-10
Number of obs: 27, groups: Dad, 9
```

The true marginal-likelihood function in tau:



Previous strategies for fitting the logit model gave relied on analytic approximations or brute-force algorithms.

Skene and Wakefield (1990); Carlin (1992); Gamerman (1997); Chib et al. (1998); S. Scott (2004)

There have also been attempts to mimic the Albert/Chib strategy (Holmes and Held, 2006; Fruhwirth-Schatter, 2010):

$$y_i \sim \text{Bern}(w_i)\,, \quad w_i = \frac{1}{1+\exp(x_i^T\beta)}$$

$$\Longleftrightarrow$$

$$y_i = \mathbb{I}_{z_i>0}$$
$$z_i = x_i^T\beta + \epsilon_i$$
$$\epsilon_i \sim \text{Logistic}(0,1)$$

But handling the logistic (error) term is tricky, and leads to complicated, slowly mixing samplers.

Our new approach outperforms all these methods.

It is non-obvious, yet is both simple and exact.

It works for any binomial model parametrized by log odds.

(Thus an equally simple treatment for the negative binomial.)

It is the true logit analogue of Albert and Chib's method.

Our new approach outperforms all these methods.

It is non-obvious, yet is both simple and exact.

It works for any binomial model parametrized by log odds.

(Thus an equally simple treatment for the negative binomial.)

It is the true logit analogue of Albert and Chib's method.

probit

```
library(msm)

for(t in 1:1000)
{
 # Update latents
 z = rtnorm(N, X %*% beta, L, U)

 # Update regression coefficients
 V = solve( XtX + PrPrec )
 m = V %*% (PrPrec %*% PrM + t(X) %*% z)
 beta = t(rmvnorm(1,m,V))
 BetaSave[t,] = beta
}
```

Our new approach outperforms all these methods.

It is non-obvious, yet is both simple and exact.

It works for any binomial model parametrized by log odds.

(Thus an equally simple treatment for the negative binomial.)

It is the true logit analogue of Albert and Chib's method.

probit

```
library(msm)

for(t in 1:1000)
{
 # Update latents
 z = rtnorm(N, X %*% beta, L, U)

 # Update regression coefficients
 V = solve( XtX + PrPrec )
 m = V %*% (PrPrec %*% PrM + t(X) %*% z)
 beta = t(rmvnorm(1,m,V))
 BetaSave[t,] = beta
}
```

logit

```
library(BayesLogit)

for(t in 1:1000)
{
 # Update latents
 om = rpg(N, rep(1,N), X %*% beta)

 # Update regression coefficients
 V = solve( t(X) %*% diag(om) %*% X + PrPrec )
 m = V %*% (PrPrec %*% PrM + t(X) %*% {Y-1/2})
 beta = t(rmvnorm(1,m,V))
 BetaSave[t,] = beta
}
```

The approach is based on a new distribution that we call the Polya-Gamma class: $X \sim \mathsf{PG}(b, c)$ if

$$
\begin{aligned}
X &\overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)} \\
g_k &\overset{iid}{\sim} \mathsf{Ga}(b, 1)
\end{aligned}
$$

The approach is based on a new distribution that we call the Polya-Gamma class: $X \sim \mathsf{PG}(b, c)$ if

$$
\begin{aligned}
X &\overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)} \\
g_k &\overset{iid}{\sim} \mathsf{Ga}(b, 1)
\end{aligned}
$$

Surprisingly, this provides the needed simple solution to our integral equation:

$$
\frac{(e^{\psi})^a}{(1 + e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^{\infty} e^{-\omega\psi^2/2} \, p(\omega) \, d\omega \, ,
$$

where $\kappa = a - b/2$ and $\omega \sim \mathsf{PG}(b, 0)$.

Details:

- y_i: the number of successes
- n_i: the number of trials
- $x_i = (x_{i1}, \ldots, x_{ip})$: the vector of regressors.
- $\kappa = (y_1 - n_1/2, \ldots, y_N - n_N/2)$
- Prior: $\beta \sim \mathsf{N}(b, B)$

The Polya-Gamma method has only two steps:

$$\begin{aligned} (\omega_i \mid \beta) &\sim \mathsf{PG}(n_i, x_i^T \beta) \\ (\beta \mid y, \omega) &\sim \mathsf{N}(m_\omega, V_\omega) \,, \end{aligned}$$

where

$$\begin{aligned} V_\omega &= (X^T \Omega X + B^{-1})^{-1} \\ m_\omega &= V_\omega (X^T \kappa + B^{-1} b) \\ \Omega &= \mathsf{diag}(\omega_1, \ldots, \omega_N) \,. \end{aligned}$$

Electronic Journal of Statistics
Vol. 7 (2013) 2054–2064
ISSN: 1935-7524
DOI: 10.1214/13-EJS837

The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic

Hee Min Choi and James P. Hobert

Department of Statistics
University of Florida
e-mail: heemin@stat.ufl.edu; jhobert@stat.ufl.edu

Abstract: One of the most widely used data augmentation algorithms is Albert and Chib's (1993) algorithm for Bayesian probit regression. Polson, Scott, and Windle (2013) recently introduced an analogous algorithm for Bayesian logistic regression. The main difference between the two is that Albert and Chib's (1993) truncated normals are replaced by so-called Polya-Gamma random variables. In this note, we establish that the Markov chain underlying Polson, Scott, and Windle's (2013) algorithm is uniformly ergodic. This theoretical result has important practical benefits. In particular, it guarantees the existence of central limit theorems that can be used to make an informed decision about how long the simulation should be run.

AMS 2000 subject classifications: Primary 60J27; secondary 62F15.
Keywords and phrases: Polya-Gamma distribution, data augmentation algorithm, minorization condition, Markov chain, Monte Carlo.

Received May 2013.

		Data set											
		Nodal	Diab.	Heart	AC	GC1	GC2	Sim1	Sim2	Sim1	Sim2	GP1	GP2
ESS	Pólya-Gamma	4860	5445	3527	3840	5893	5748	7692	2612	7646	3590	6309	6386
	Best RU-DA	1645	2071	621	1044	2227	2153	3031	574	719	915	1296	1157
	Best Metropolis	3609	5245	1076	415	3340	1050	4115	1388	749	764	—	—

There are two obvious questions here:

- Where did the Polya-Gamma representation come from?
- How do I simulate it?

There are two obvious questions here:

- Where did the Polya-Gamma representation come from?
- How do I simulate it?

As a brief answer to the second point, we can simulate PG random variates exactly, without truncating the infinite sum:

$$X \stackrel{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2 + c^2/(4\pi^2)} \,.$$

We do this using a simple, efficient rejection sampler.

- The proposal: exponential, uniform, and normal draws.
- Checking for acceptance: roughly like one IG density evaluation.
- Acceptance probability: in practice, usually better than 0.9998 …

There are two obvious questions here:

Where did the Polya-Gamma representation come from?

How do I simulate it?

As a brief answer to the second point, we can simulate PG random variates exactly, without truncating the infinite sum:

$$X \stackrel{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2 + c^2/(4\pi^2)} \,.$$

We do this using a simple, efficient rejection sampler.

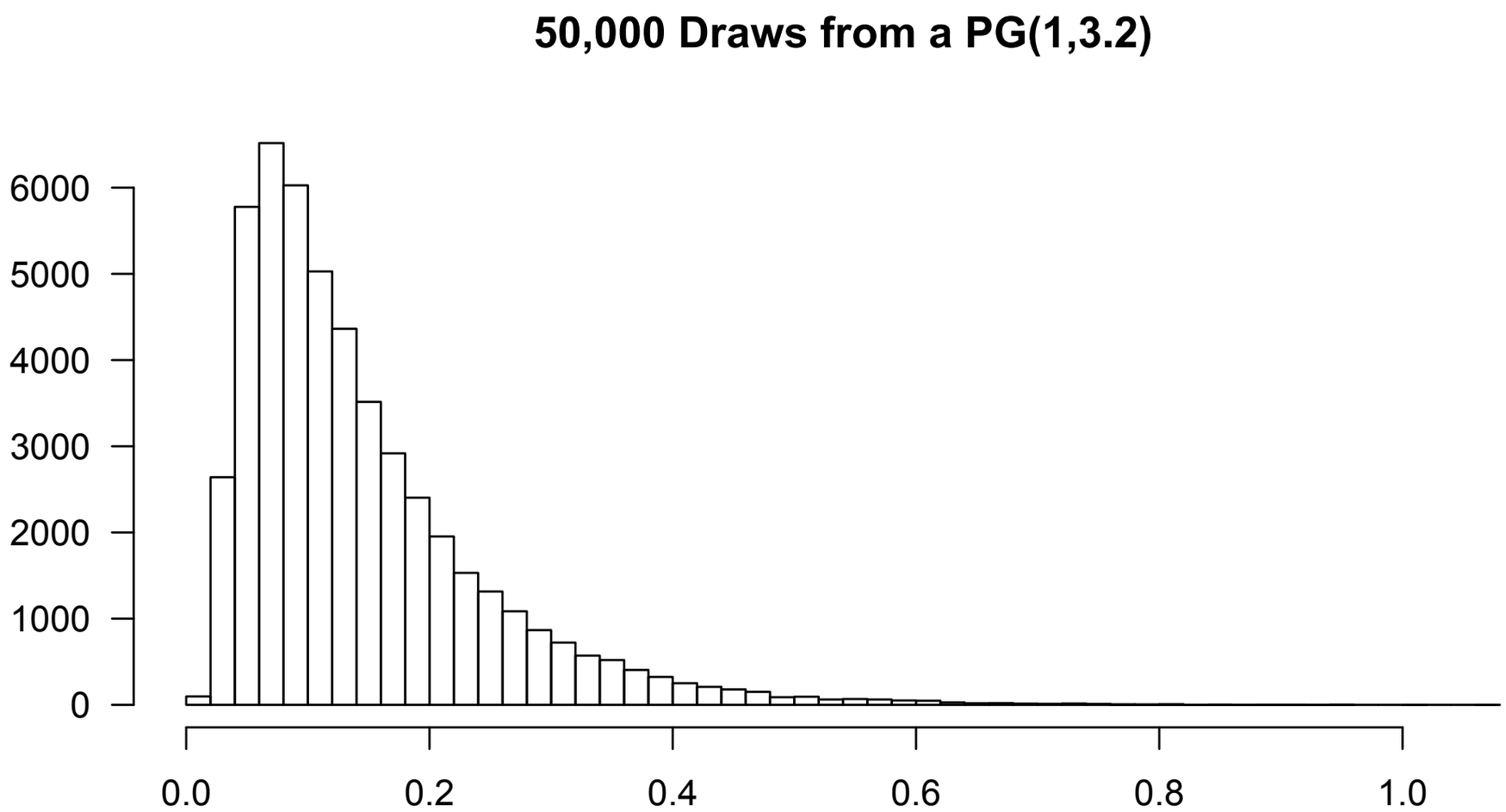
The proposal: exponential, uniform, and normal draws.

Checking for acceptance: roughly like one IG density evaluation.

Acceptance probability: in practice, usually better than 0.9998 …

and uniformly bounded below at 0.9992.

We provide this as a black box in our R package (BayesLogit):

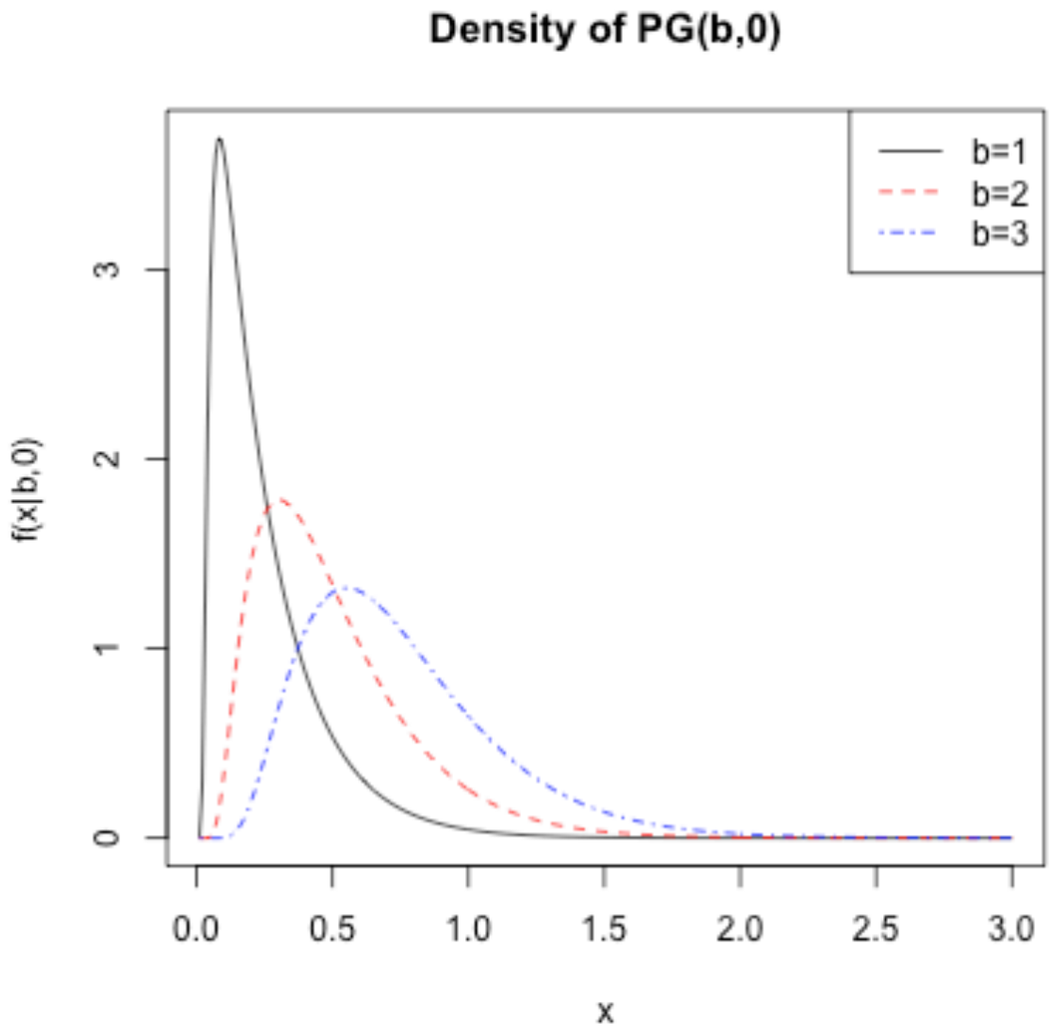


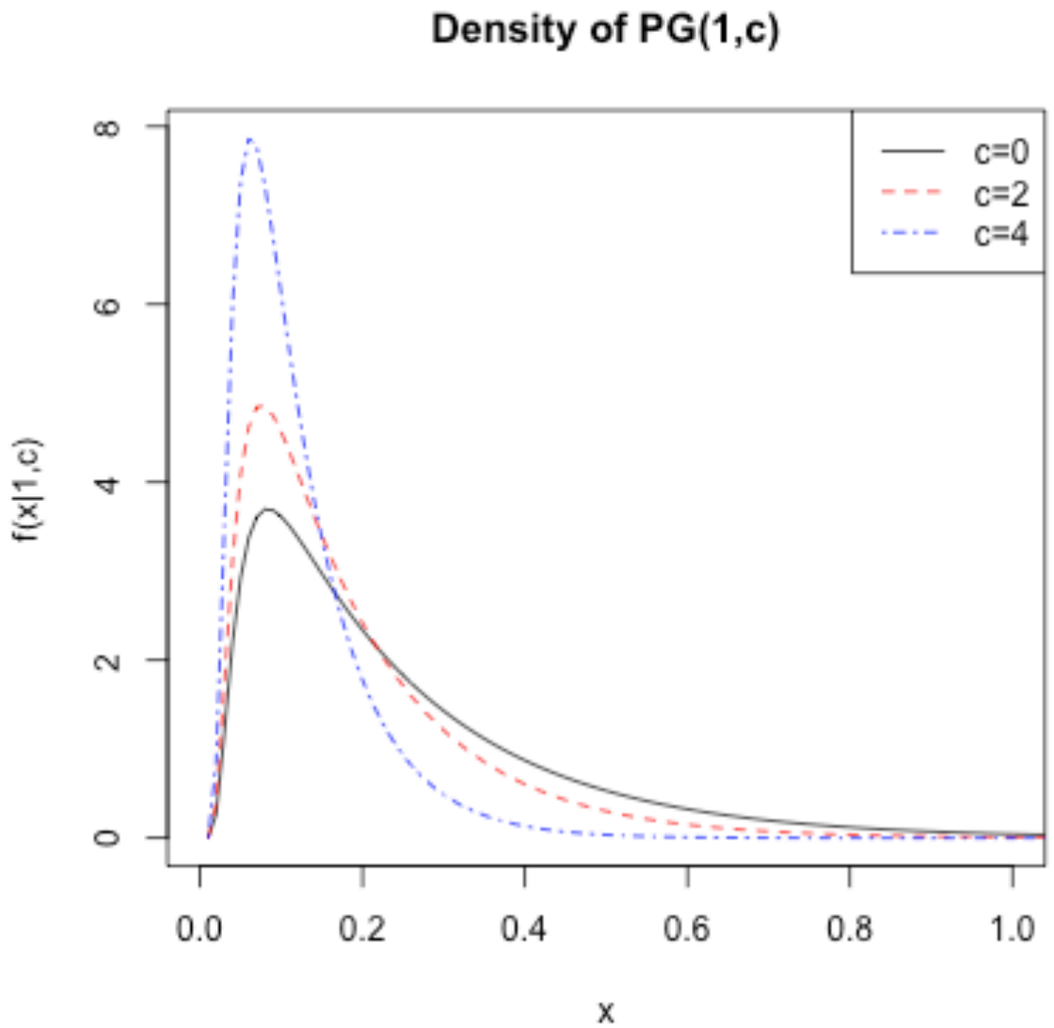

```
> system.time( rpg(50000,1,3.2) )
   user  system elapsed
  0.035   0.002   0.036
> system.time( rgamma(50000,1,3.2) )
   user  system elapsed
  0.010   0.000   0.009
> system.time( rtnorm(50000,0,1, lower=0, upper=Inf) )
   user  system elapsed
  0.110   0.031   0.141
```

First: more detail on the PG(b,c) distribution.

b large: more Gaussian

c large: more like a point mass at 1/c





We got the idea from Biane et. al. (2001), who survey many laws that connect analytic number theory and Brownian excursions.

One such distribution, which we denote by $J^*(b)$, has Laplace transform given by

$$\mathbb{E}[e^{-tJ^*(b)}] = \cosh^{-b}(\sqrt{t/2}) \, ;$$

Biane et. al. (2001) show that this distribution has a density $p(\omega; b)$. For technical reasons, our $J^*(b)$ is their $J^*(b)/4$.

We define the PG(b, z) as the random variable with density

$$p(\omega; b, z) = \cosh^b(z/2) \exp\left(-\frac{z^2}{2}\omega\right) p(\omega; b) \, .$$

Note that the normalizing constant comes by evaluating the Laplace transform of $J^*(b)$ at $z^2/2$.

The Laplace transform of a PG(b, c) distribution may be calculated using standard results on exponential tilts, along with Weierstrass factorization theorem:

$$
\begin{aligned}
E\left\{\exp\left(-\omega t\right)\right\} &= \frac{\cosh^b\left(\frac{c}{2}\right)}{\cosh^b\left(\sqrt{\frac{c^2/2+t}{2}}\right)} \\
&= \prod_{k=1}^{\infty}\left(\frac{1+\frac{c^2/2}{2(k-1/2)^2\pi^2}}{1+\frac{c^2/2+t}{2(k-1/2)^2\pi^2}}\right)^b \\
&= \prod_{k=1}^{\infty}(1+d_k^{-1}t)^{-b} \\
d_k &= 2\left(k-\frac{1}{2}\right)^2\pi^2+c^2/2\ .
\end{aligned}
$$

Thus the PG(b, c) is an infinite convolution of gammas.

These results are directly useful for the logit model.

First, we can write each likelihood term as

$$\begin{aligned}
\frac{(e^{\psi_i})^{y_i}}{(1+e^{\psi_i})^{n_i}} &= e^{\kappa_i \psi_i} \frac{1}{\{e^{-\psi_i/2} + e^{\psi_i/2}\}^{n_i}} \\
&= 2^{-n_i} e^{\kappa_i \psi_i} \cosh^{-n_i}(\psi_i/2) \\
&= 2^{-n_i} e^{\kappa_i \psi_i} \int_0^\infty e^{-\omega \psi^2/2} \, p(\omega) \, d\omega \, ,
\end{aligned}$$

where $\kappa_i = y_i - n_i/2$, and ω is a PG$(n_i, 0)$ random variable. Simple regression-type calculations lead to the conditionally Gaussian posterior for β.

Moreover, the conditional distribution for ω, given ψ_i, is just an exponential tilt of PG$(n_i, 0)$. Therefore, we're done:

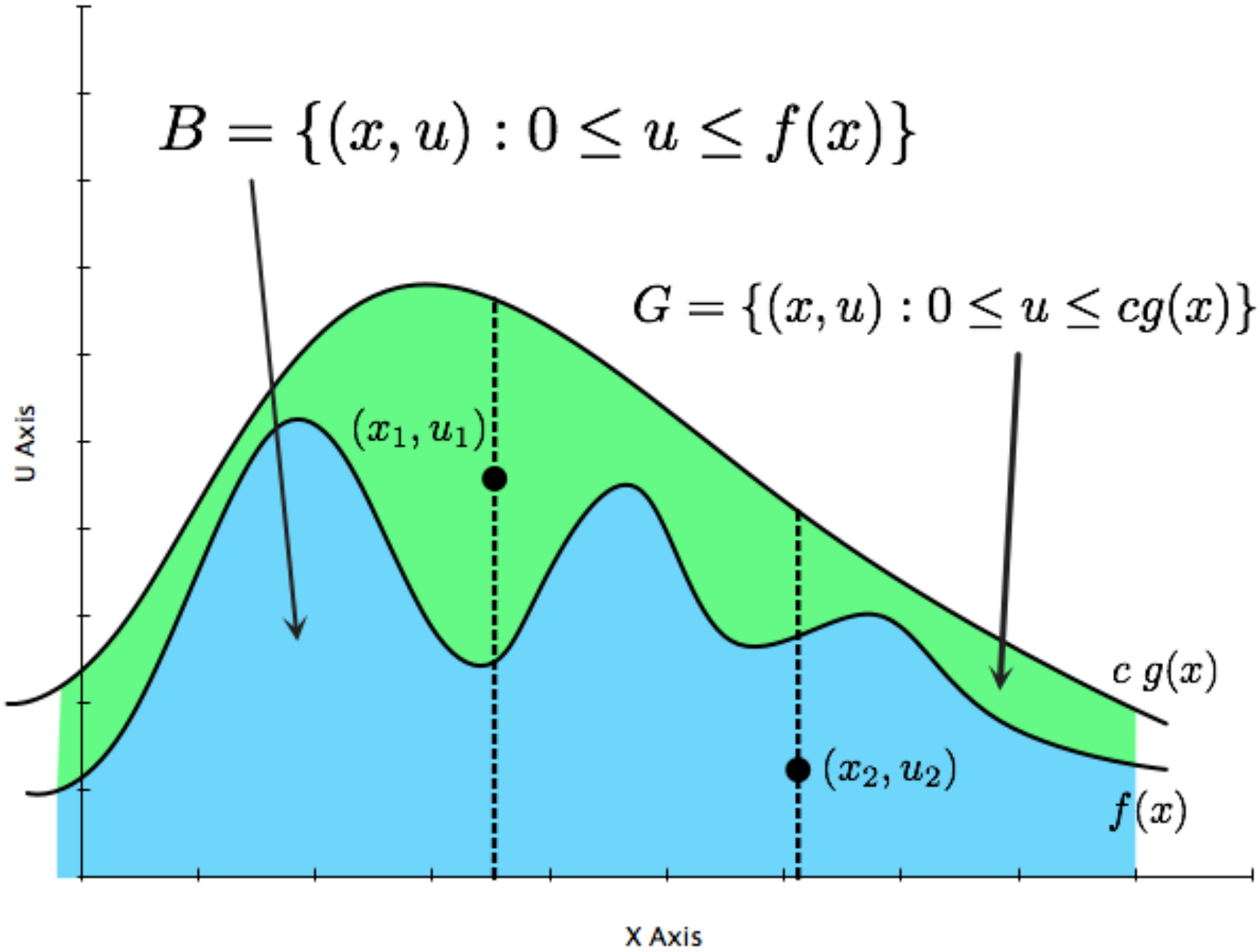
$$(\omega \mid \psi_i) \sim \text{PG}(n_i, \psi_i) \, .$$

The key step in applying these results is the existence of a fast, efficient routine for simulating PG(1,z) random variates.

The previous expression will obviously not work!

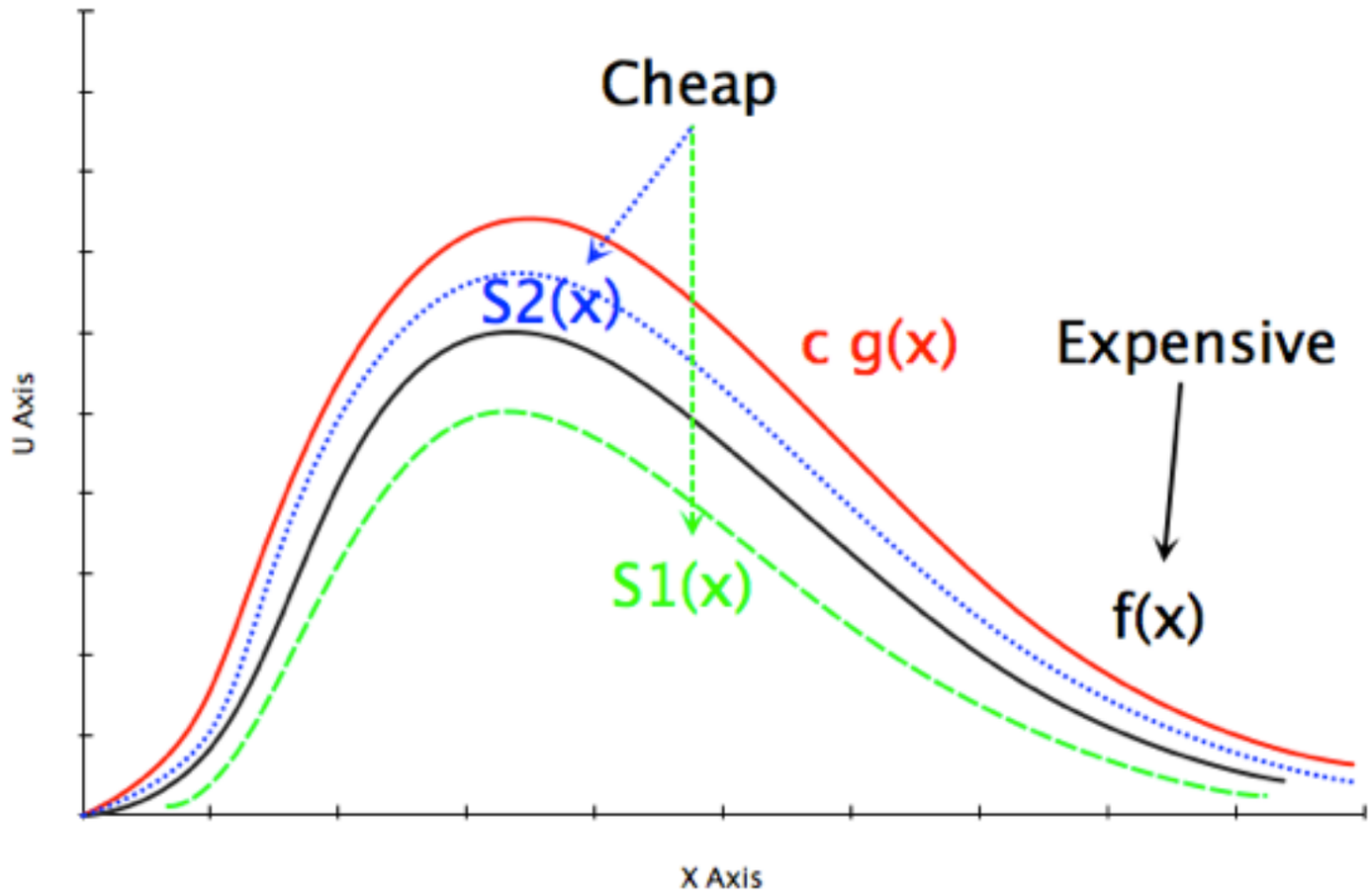
$$X \stackrel{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2 + c^2/(4\pi^2)} \,.$$

Remember accept/reject sampling:

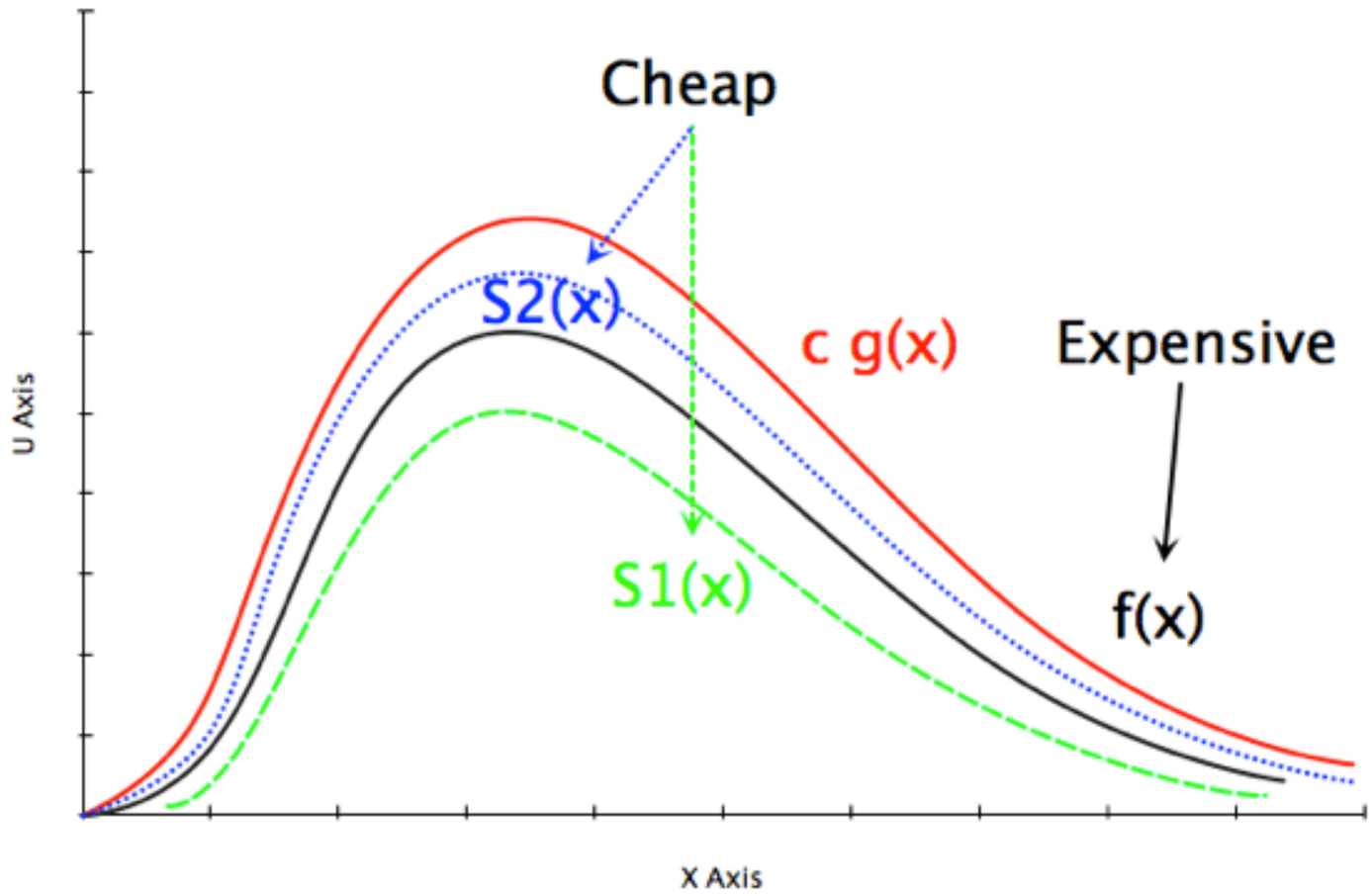


Draw $x \sim g; u \sim U(0, cg(x))$ until $u \leq f(x)$.

The squeeze principle:



The squeeze principle:



Suppose that, for all x, $S_1(x) \leq f(x)$ and $S_2(x) > f(x)$. Then to sample from $f(x)$:

1. $x \sim g$ and $u \sim U(0, cg(x))$ as before.
2. $u \leq S_1(x)$: accept.
3. $u > S_2(x)$: reject.
4. $u \leq f(x)$: accept. If not, reject.

If the bounds are good, we rarely have to evaluate $f(x)$.

The Polya-Gamma density can be expressed as an infinite alternating sum:

$$f(x) = \lim_{n\to\infty} S_n(x), \;\; S_n(x) = \sum_{i=0}^{n} (-1)^i a_i(x)$$

for which the partial sums S_i satisfy

$$\forall x, \;\; S_0(x) > S_2(x) > \ldots > f(x) > \ldots > S_3(x) > S_1(x) \; .$$

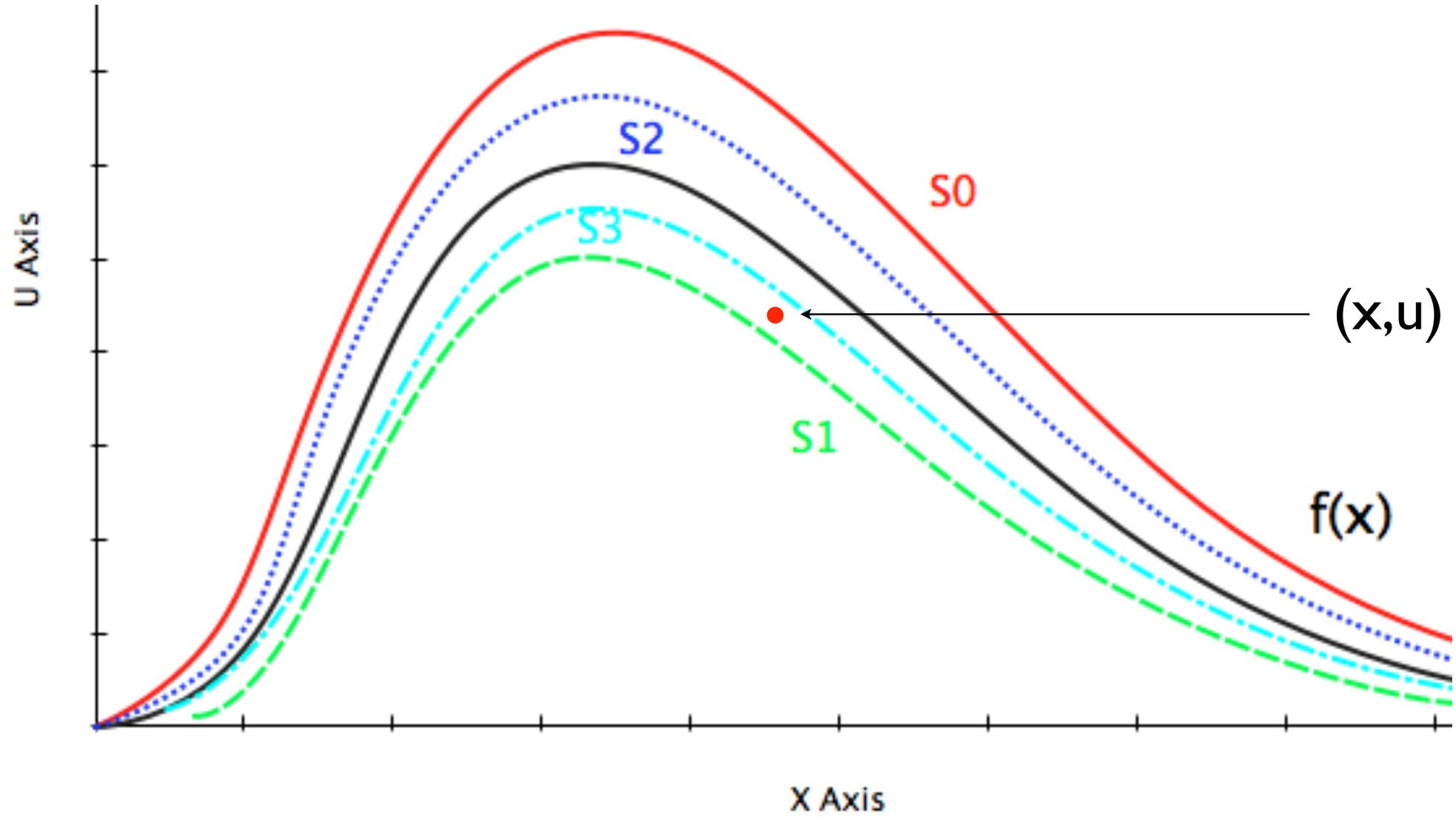
This is equivalent to $\{a_i(x)\}_{i=1}^{\infty}$ decreasing in i for all x.

In that case:

- $u < f(x)$ if and only if $u \leq S_i(x)$ for some odd i;
- $u > f(x)$ if and only if $u \geq S_i(x)$ for some even i.

We only need as many terms as necessary to find that $u \leq S_i(x)$ for odd i or $u \geq S_i(x)$ for even i.

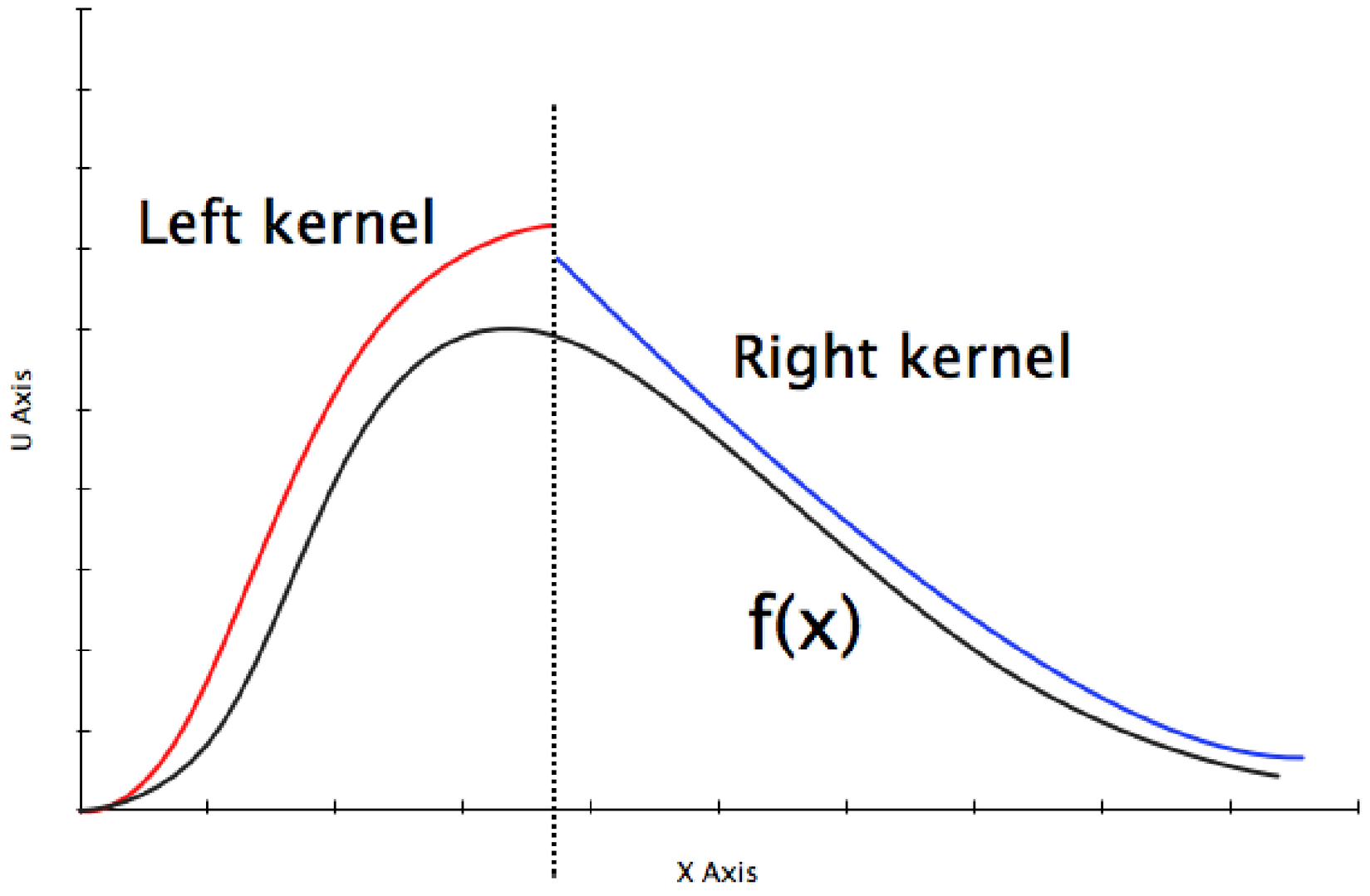
Just like the basic "squeeze" picture, with an infinite family of bounds that get progressively tighter with more terms.

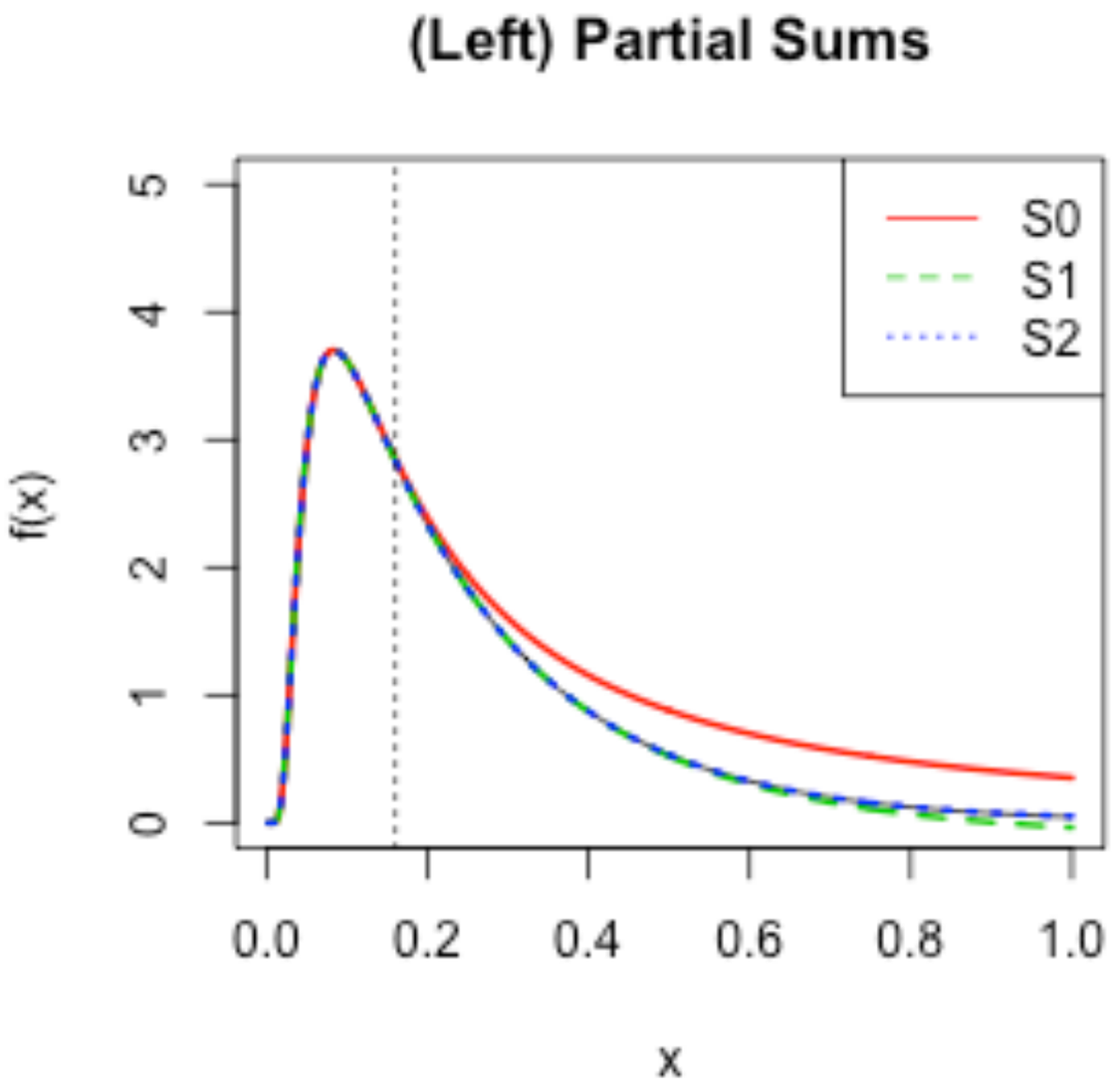


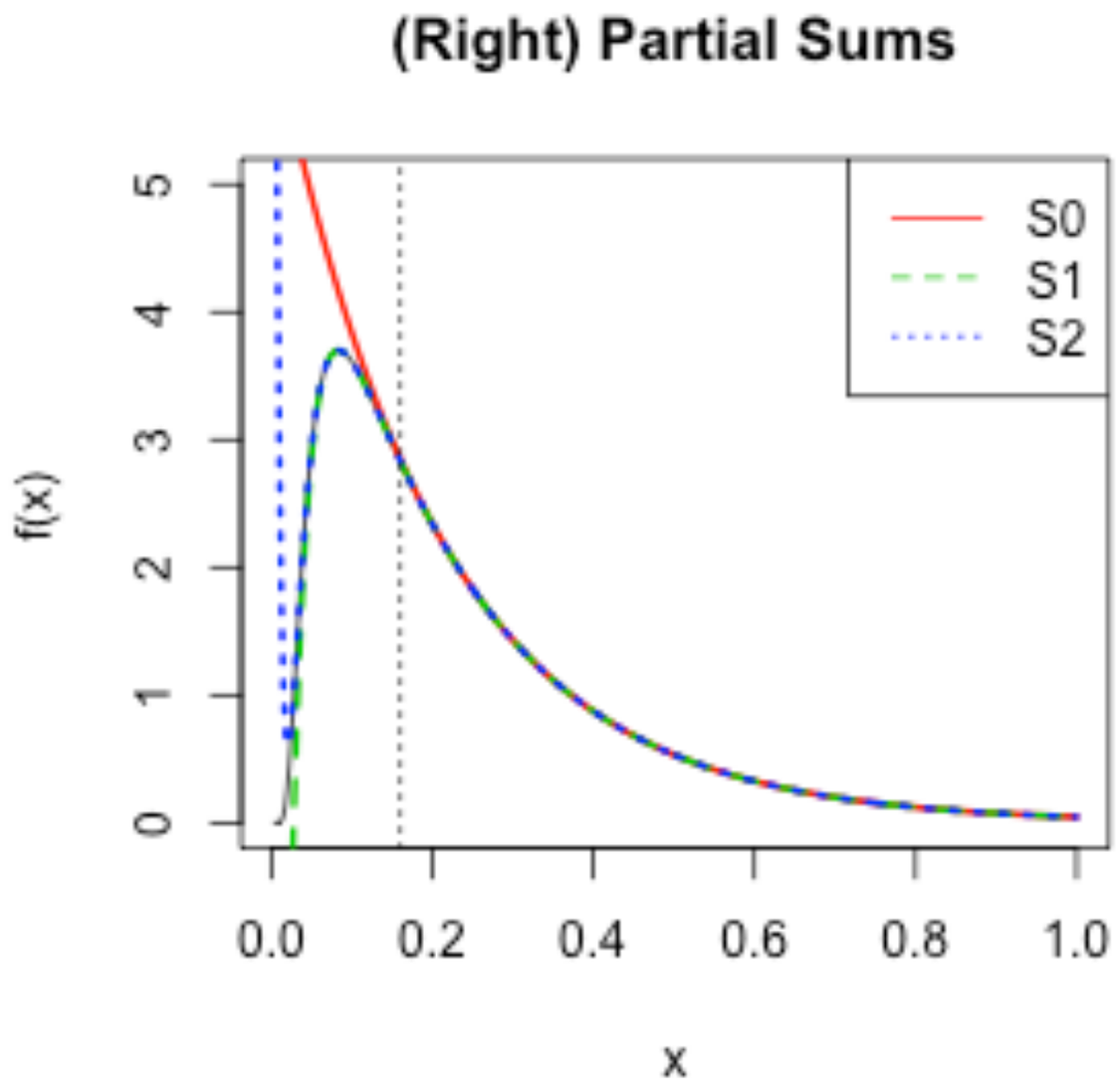
The next bound is computed recursively from the last bound.

A complication: we must ``glue together" two different infinite series representations in order to satisfy the even/odd squeezing criterion where $a_i(x)$ in decreasing in i for all x:

$$f(x) = \sum_{i=0}^{\infty} (-1)^i a_i^L(x) = \sum_{i=0}^{\infty} (-1)^i a_i^R(x) \; .$$




(Left) Partial Sums
S0
S1
S2
f(x)
x
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0


(Right) Partial Sums
S0
S1
S2
f(x)
x
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0

Two things could go wrong:

1) We reject many proposals.

2) We must calculate many partial sums before deciding whether to accept the proposal.

~~Two things could go wrong:~~

~~1) We reject many proposals.~~

~~2) We must calculate many partial sums before deciding whether to accept the proposal.~~

For the best choice of truncation point ($t = 1/[2\pi]$),

$$\inf_{z \geq 0} P(u \leq f(x \mid 1, z)) = \inf_{z \geq 0} \frac{1}{c(z)} \approx \frac{1}{1.0008} = 0.9992\,.$$

That is, we reject no more than 8 out of every 10000 draws.

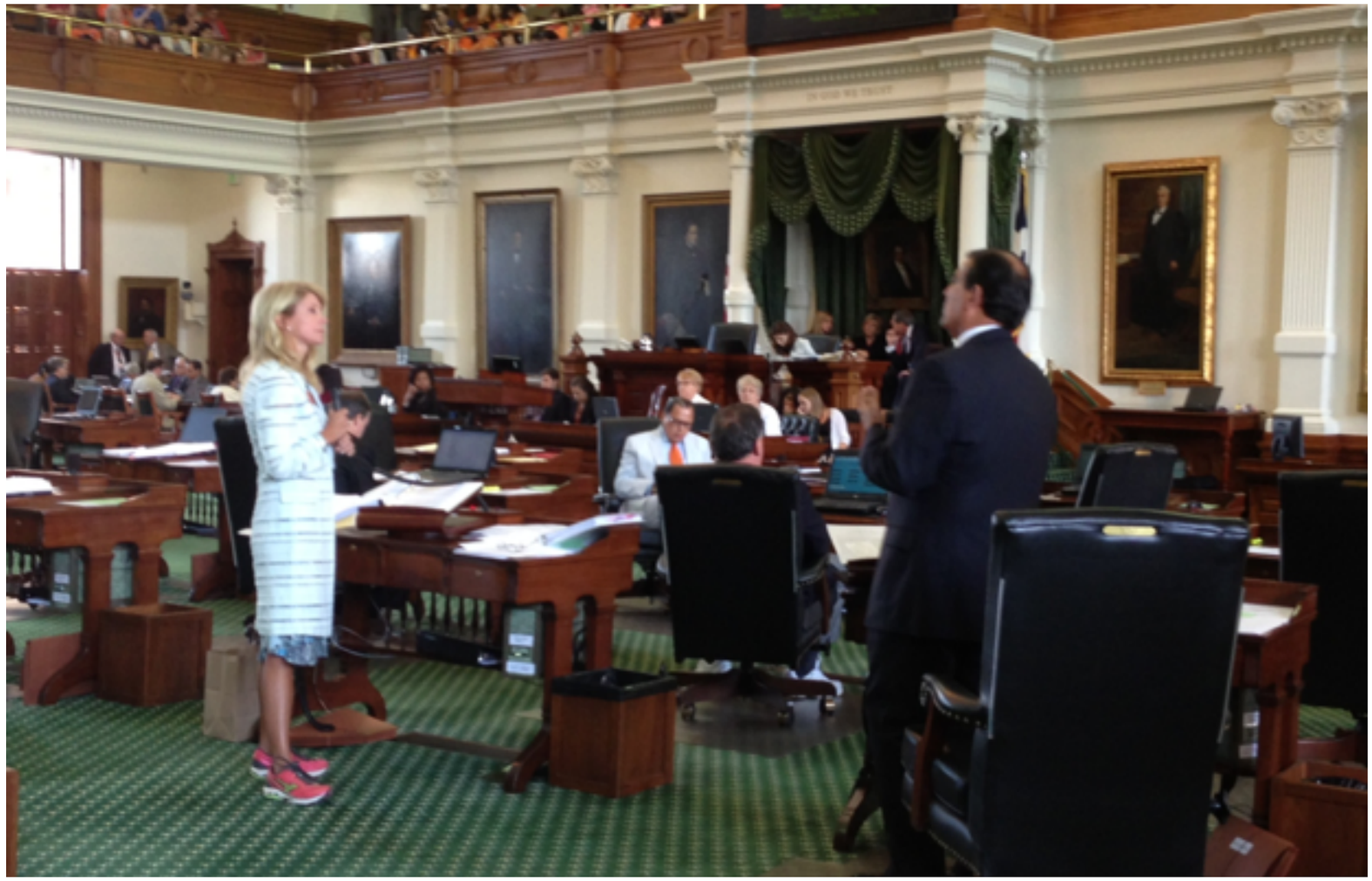
And for the worst case z $(= 1.378)$, the distribution of exit times is:

n	1	2	3	4
$P(L = n)$	0.9991977	8.023e-04	1.728e-09	8.213e-18

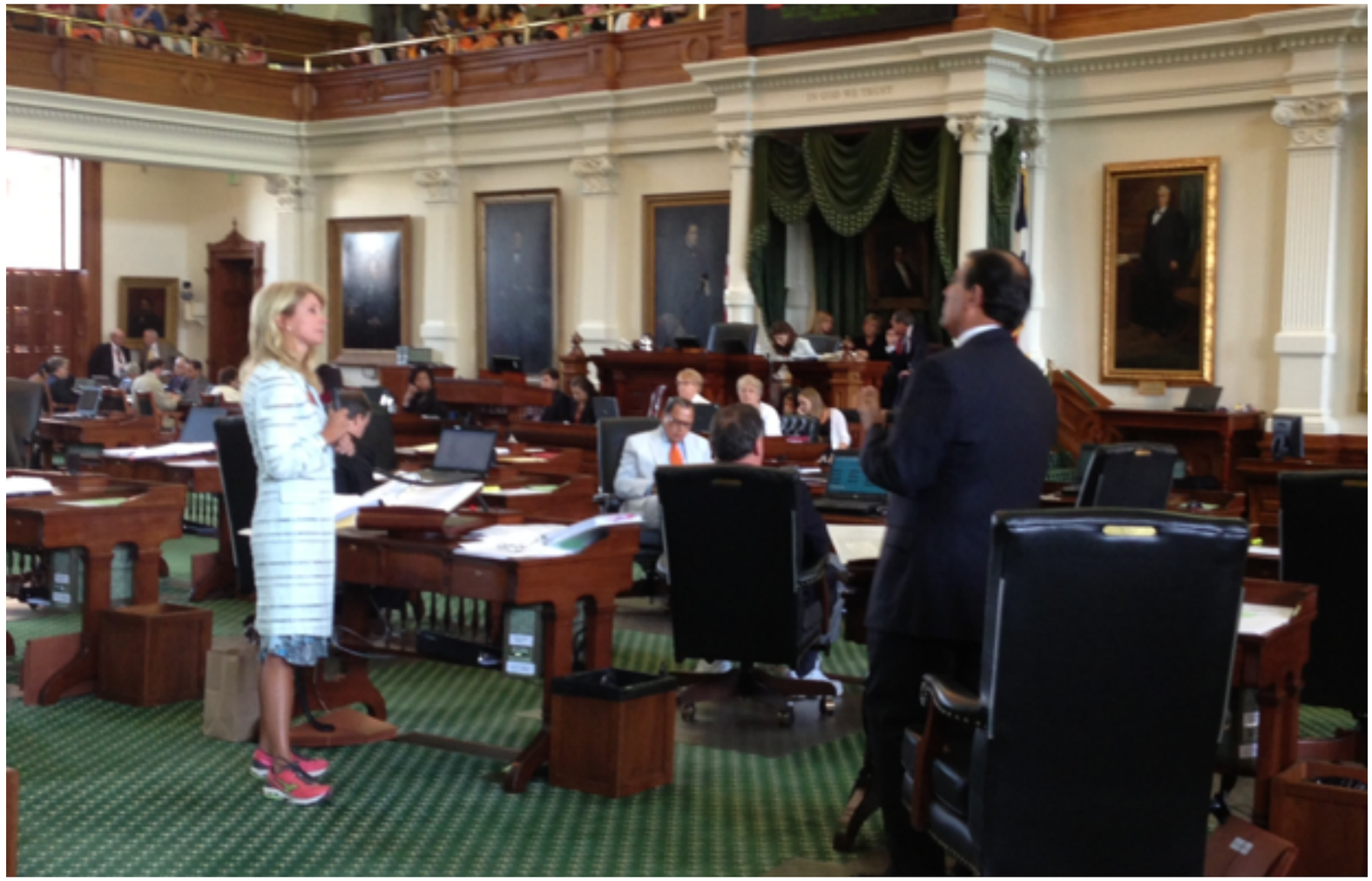
.

Some models we've fit using this trick:

- multinomial logistic regression
- variable selection in logistic regression
- mixed-effects logit models
- contingency tables, with and without fixed margins
- negative-binomial regression
- logit hidden-Markov and general state-space models
- online EM for logit and negative-binomial models
- supervised correlated topic models
- spatial models for binary and count outcomes
- negative-binomial and logistic factor models
- multiple testing in the presence of covariates



Example 1: Wendy Davis's pink sneakers

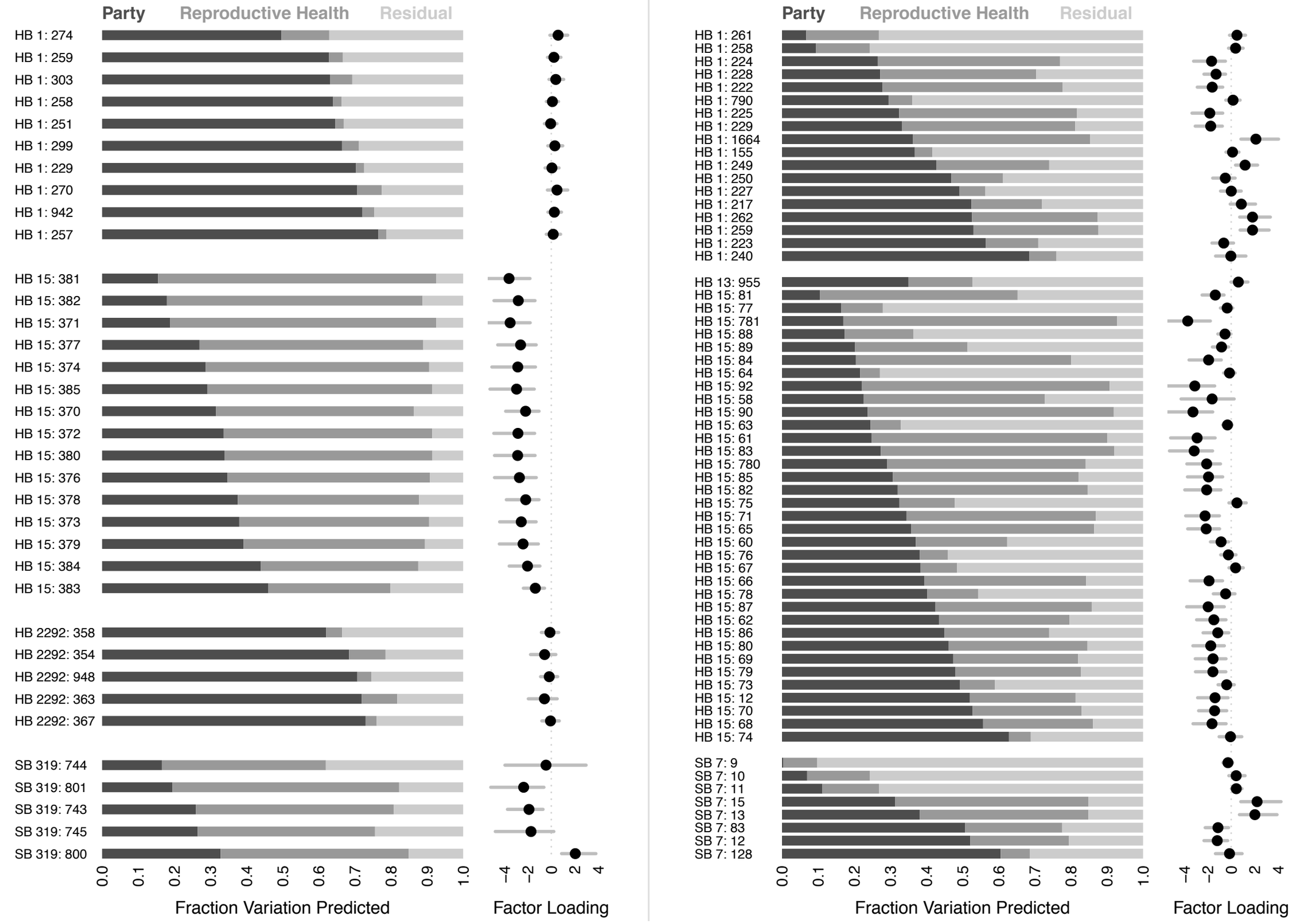


$$y_{ij} \sim \text{Bern}(w_{ij}) = \text{Vote by Rep } i \text{ on bill } j$$

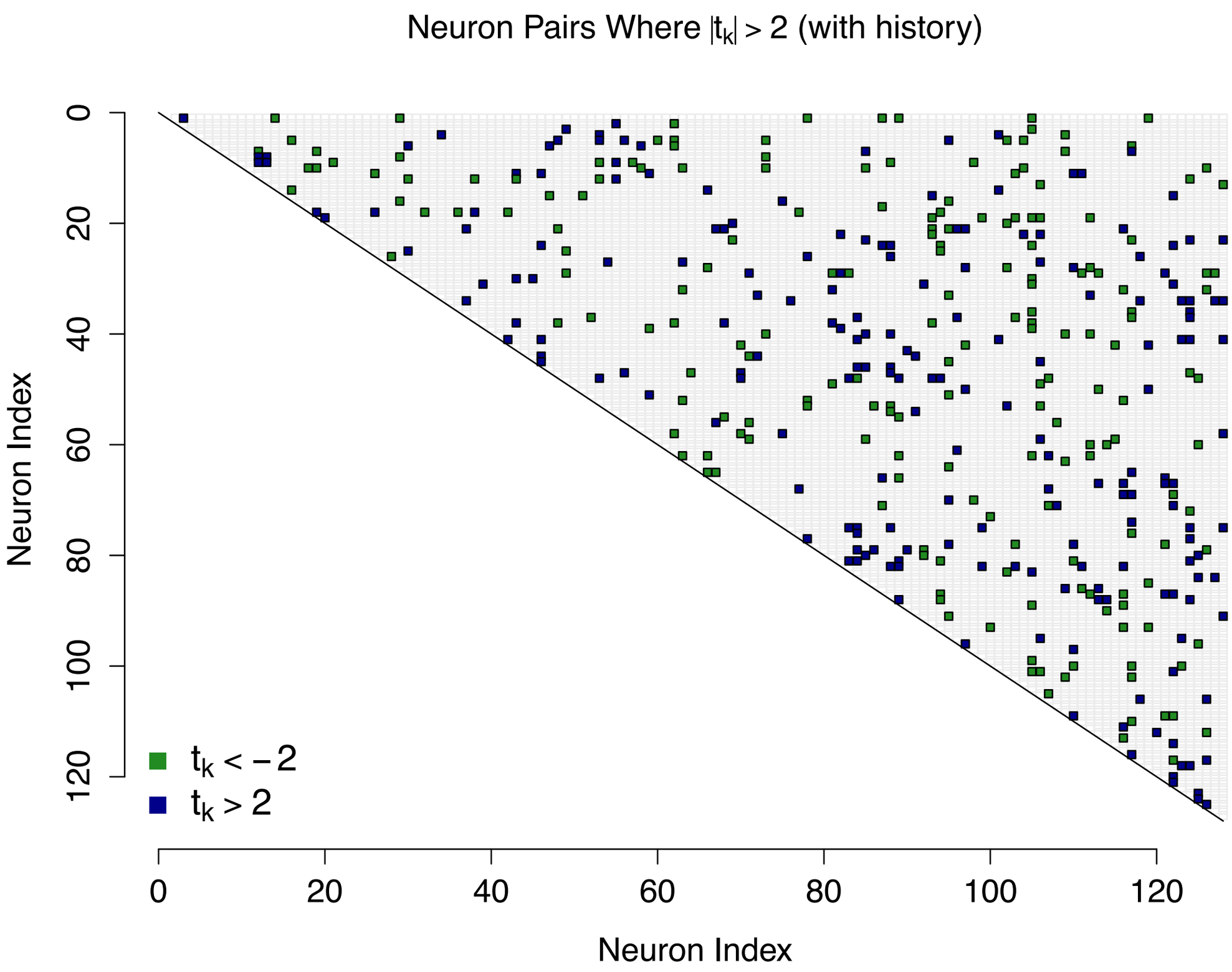
$$w_{ij} = \frac{1}{1 + \exp\{-(b_j^T f_i + \alpha_j)\}}$$

$$f_i = \text{(Partisanship factor, Family-planning factor)}$$

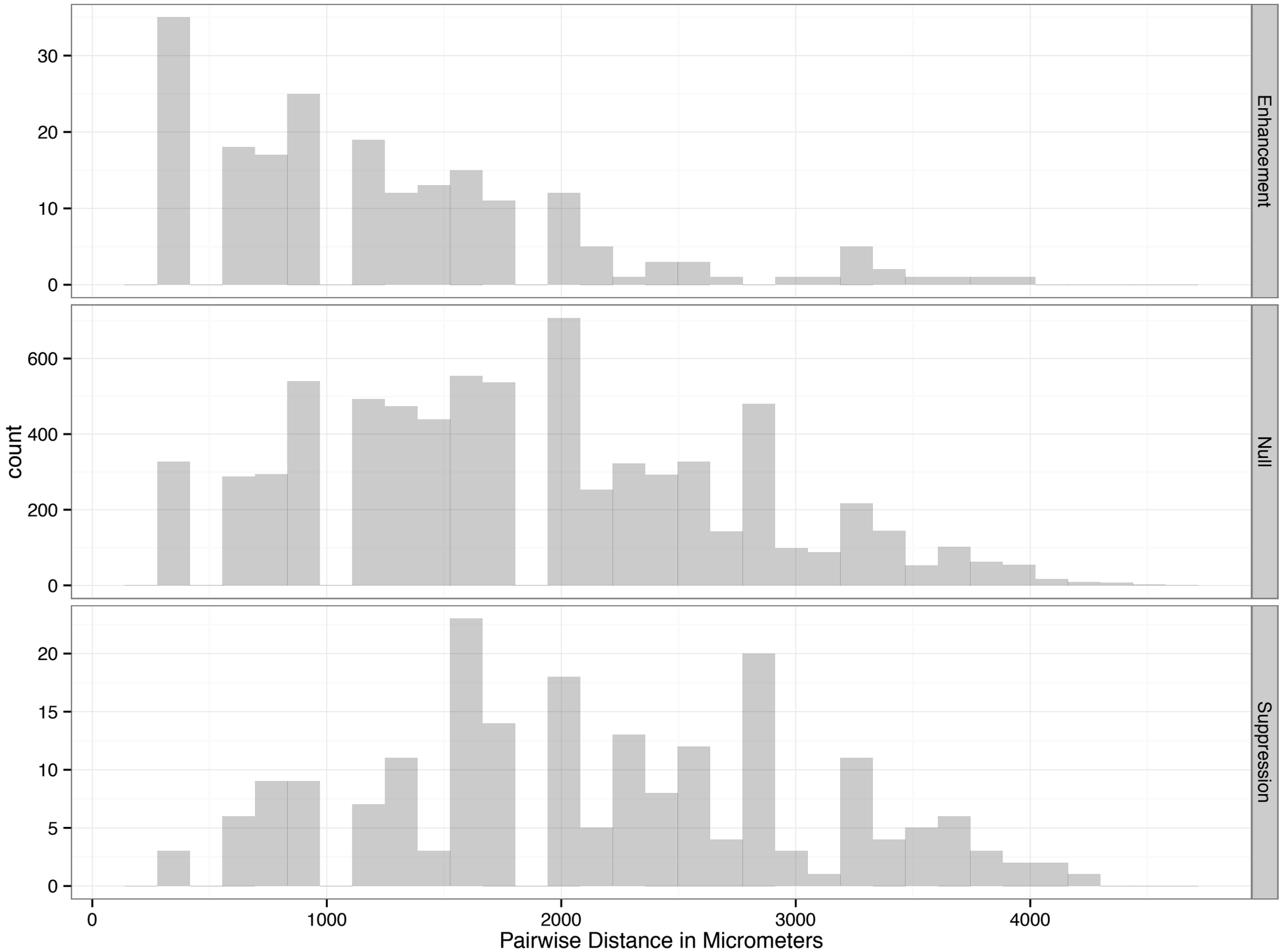
Example 1: Wendy Davis's pink sneakers



A Bayesian ANOVA for the reproductive-health bills in two sessions



Example 2: detection of excess synchrony in neuron firing rates
Scott, Kelly, Smith, and Kass (2013)



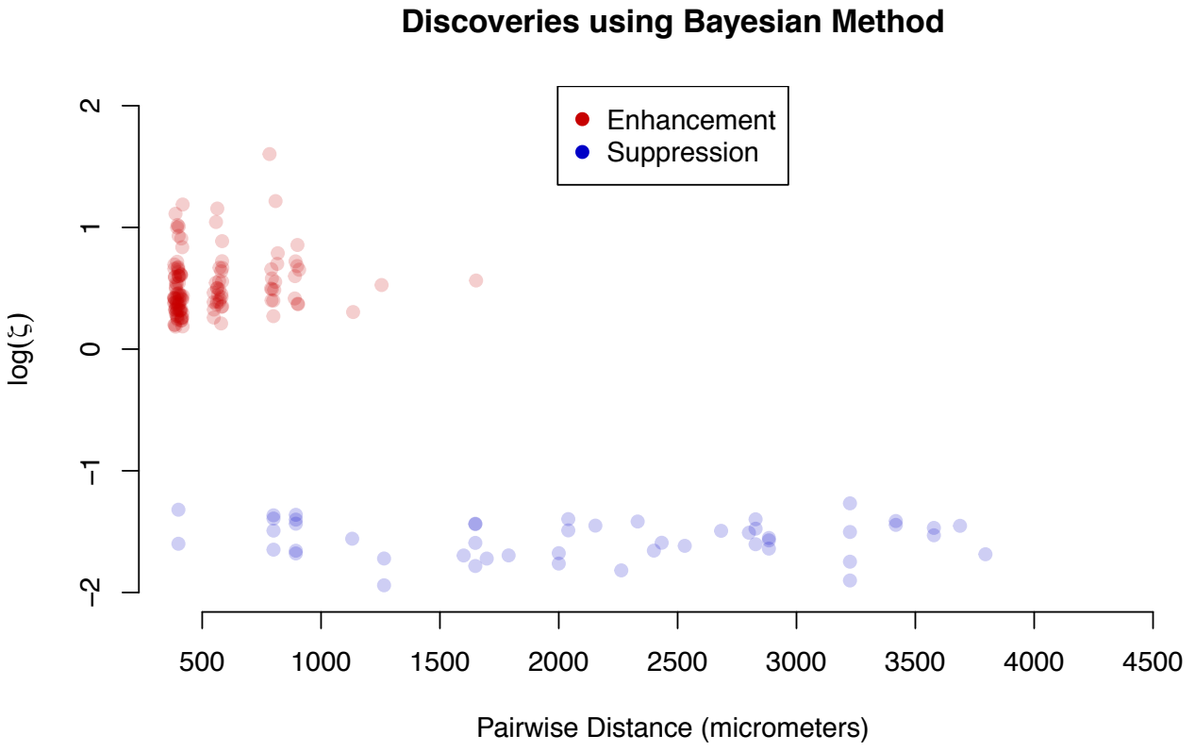
Distance stratified by t statistics.

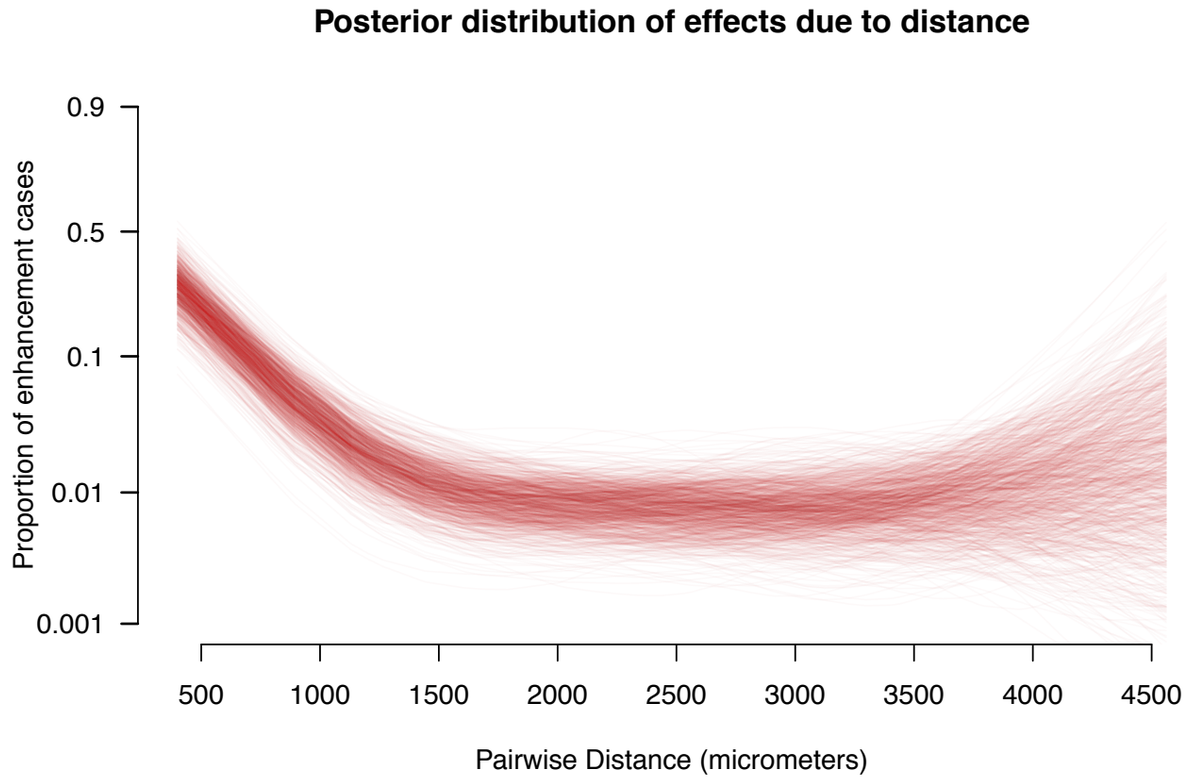
The classic approach (popularized by Efron):

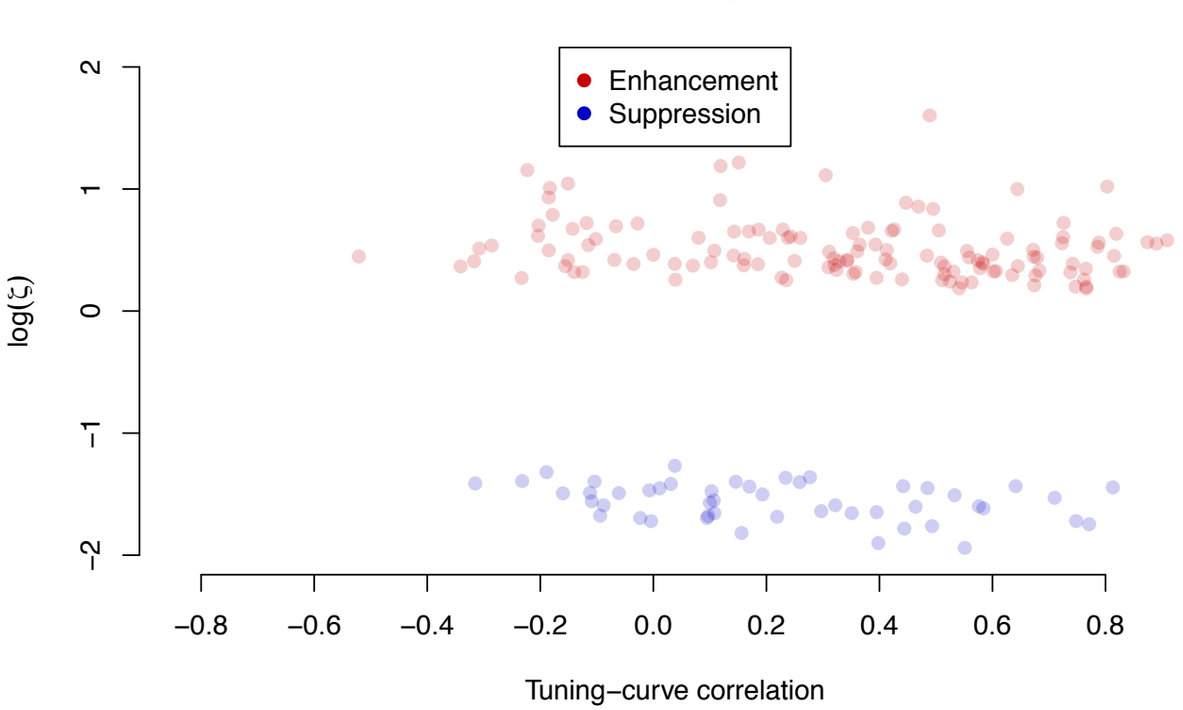
$$
\begin{aligned}
z_i &\sim N(\theta_i, \sigma_i^2) \\
\theta_i &\sim \begin{cases} \delta_0\,, & \gamma_i = 0 \\ \Pi\,, & \gamma_i \neq 0 \end{cases} \\
\gamma_i &\sim \mathrm{Bern}(w)
\end{aligned}
$$

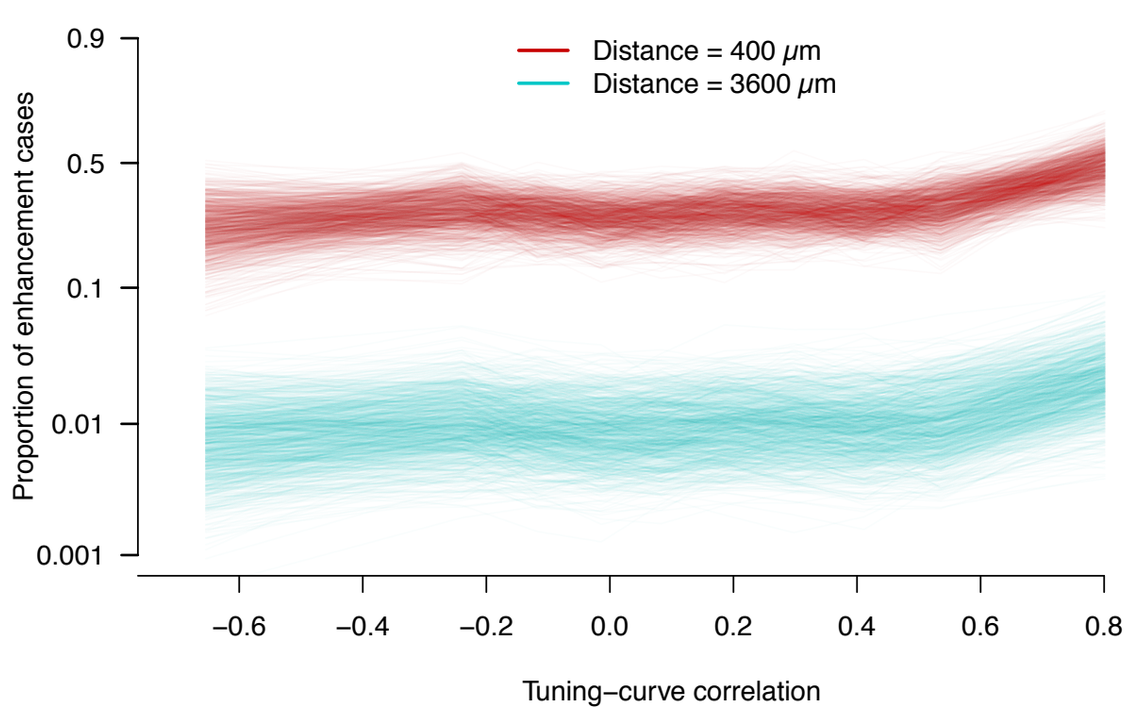
A natural extension involving covariates:

$$
\begin{aligned}
z_i &\sim N(\theta_i, \sigma_i^2) \\
\theta_i &\sim \begin{cases} \delta_0\,, & \gamma_i = 0 \\ \Pi\,, & \gamma_i \neq 0 \end{cases} \\
p(\gamma_i = 1) &= \frac{1}{1 + \exp(-x_i^T \beta)}
\end{aligned}
$$

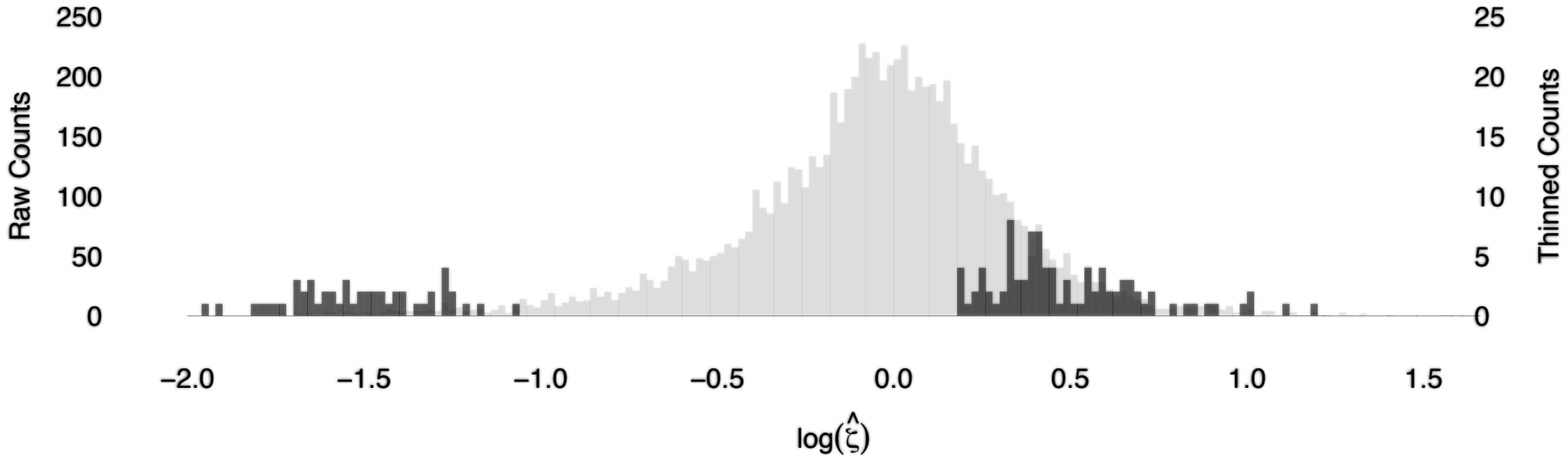


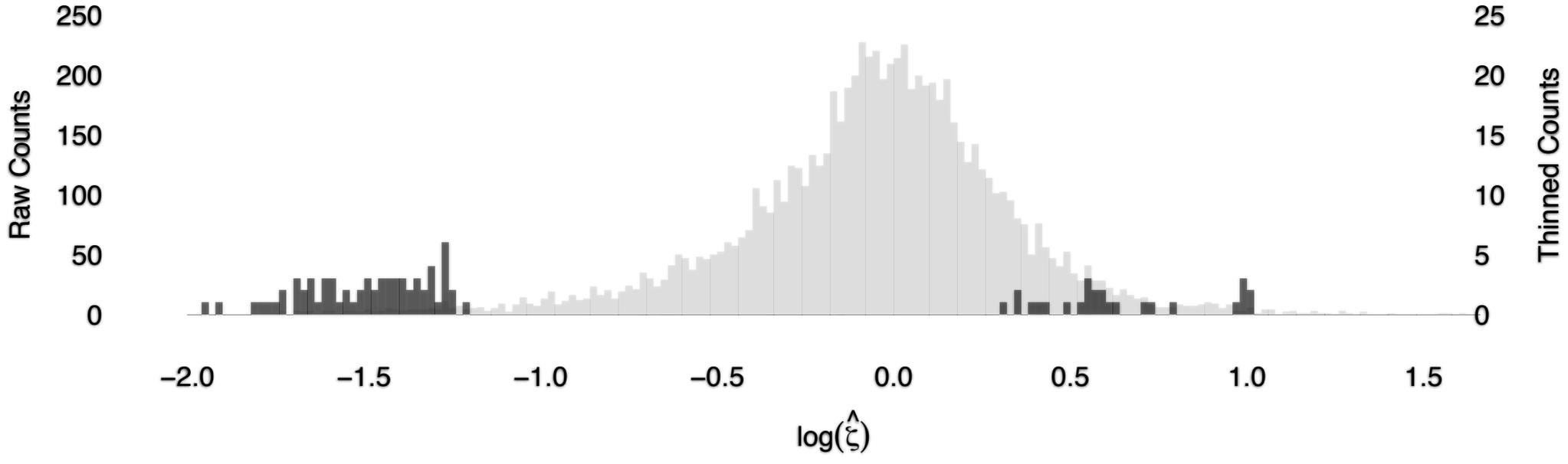


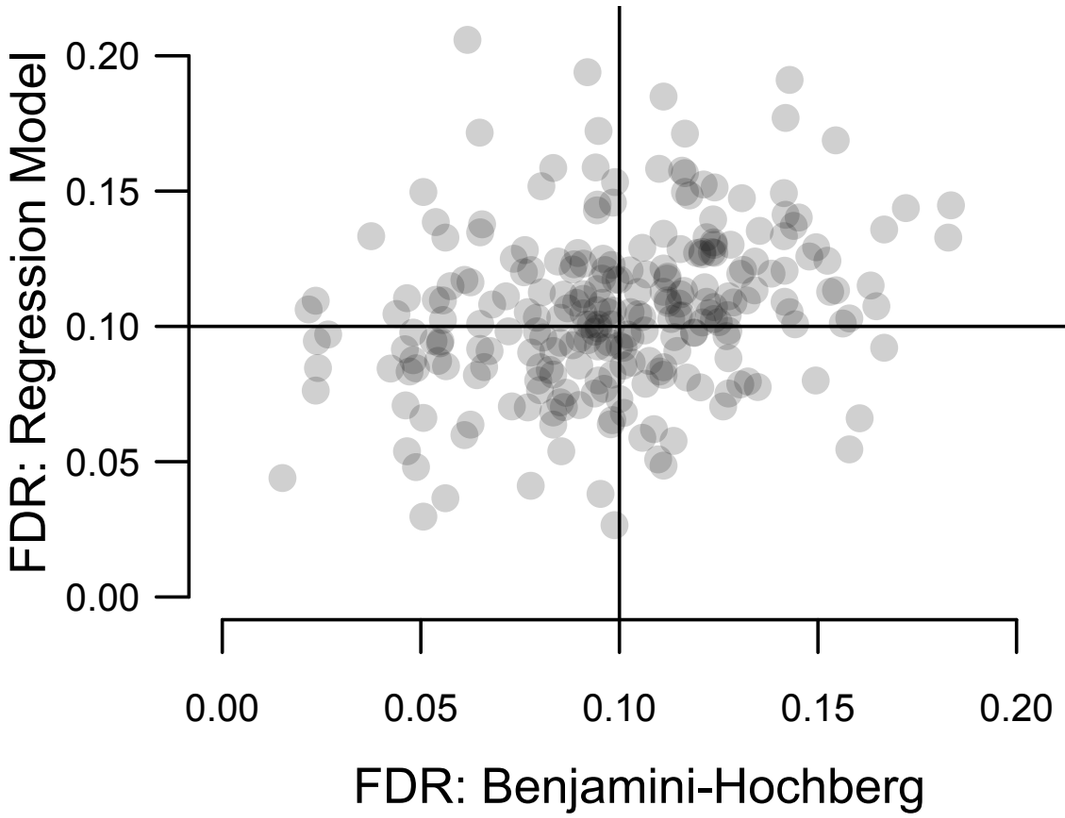


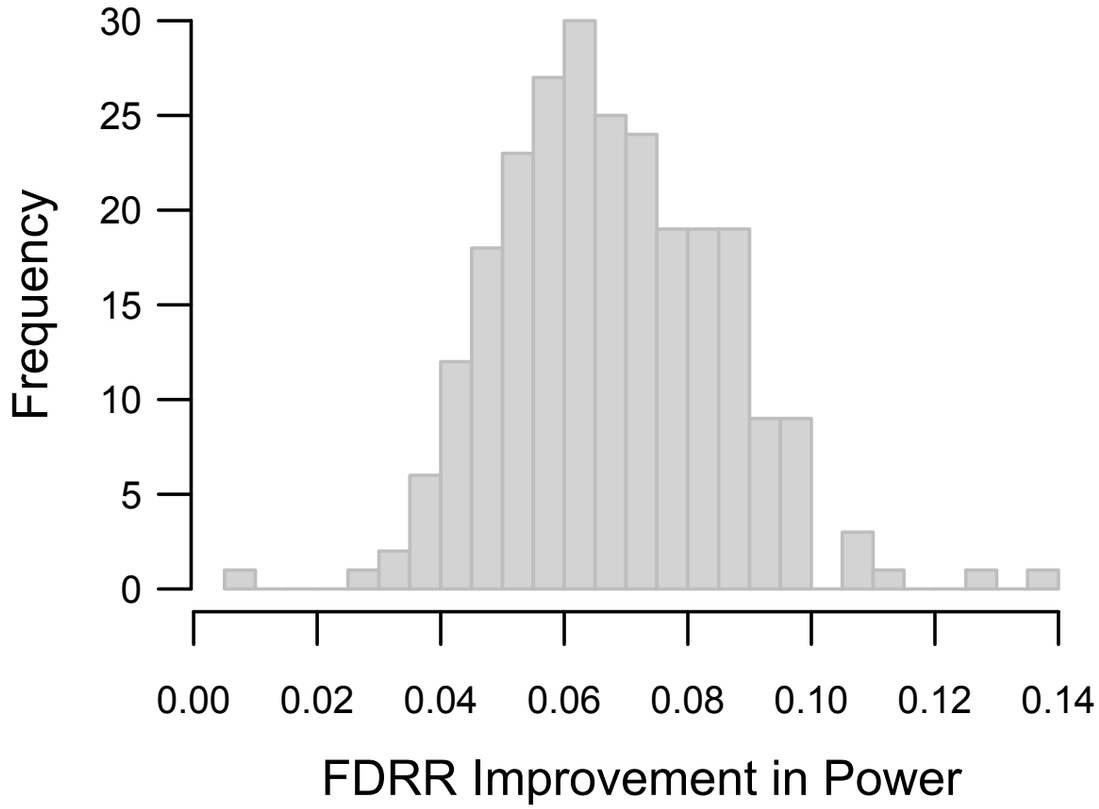


With a Gaussian-process prior over the covariate effects


Thinned counts: Bayesian method
Raw Counts
250
200
150
100
50
0
Thinned Counts
25
20
15
10
5
0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
log(ξ̂)

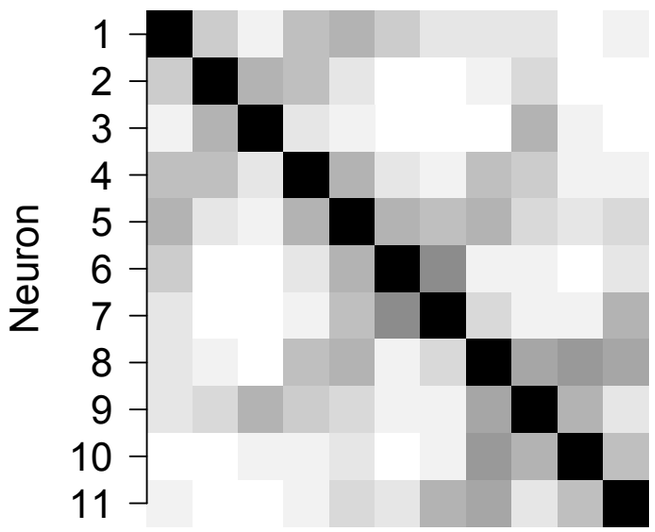

Thinned counts: Benjamini–Hochberg
Raw Counts
250
200
150
100
50
0
Thinned Counts
25
20
15
10
5
0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
log(ξ̂)


FDR: Regression Model
0.20
0.15
0.10
0.05
0.00
0.00
0.05
0.10
0.15
0.20
FDR: Benjamini-Hochberg

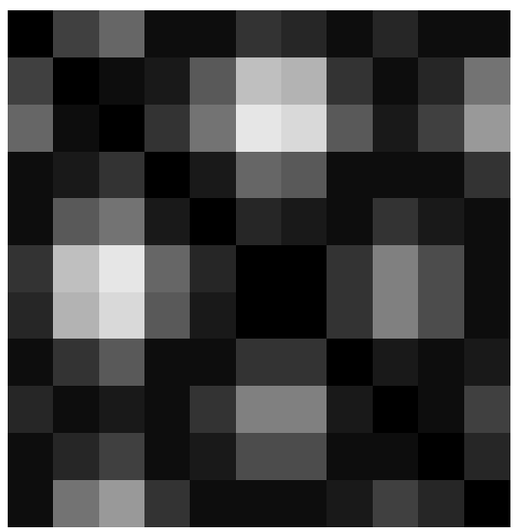

Frequency
30
25
20
15
10
5
0
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
FDRR Improvement in Power

Example 3: a dynamic factor model fit to neural recordings in the primate visual cortex.

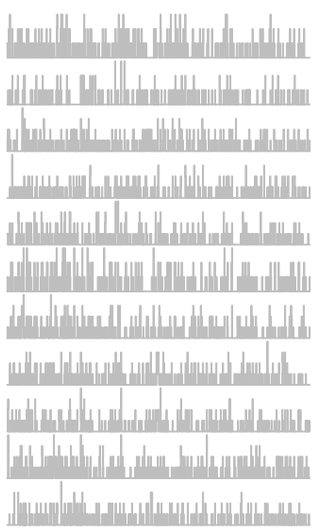
$$
\begin{aligned}
y_{tk} &\sim \mathrm{NB}(\xi, e^{\psi_{tk}}) \quad \text{for} \quad k = 1, \ldots K \\
\psi_t &= \alpha + B f_t \\
f_t &= \Gamma f_{t-1} + \epsilon_t \,, \quad \epsilon_t \sim N(0, \tau^2 I) \,.
\end{aligned}
$$



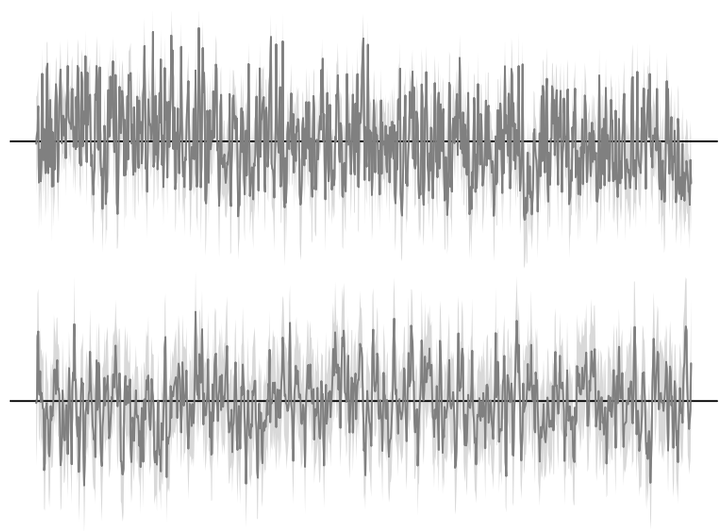
Correlation among spike counts



Estimated correlation of latent states



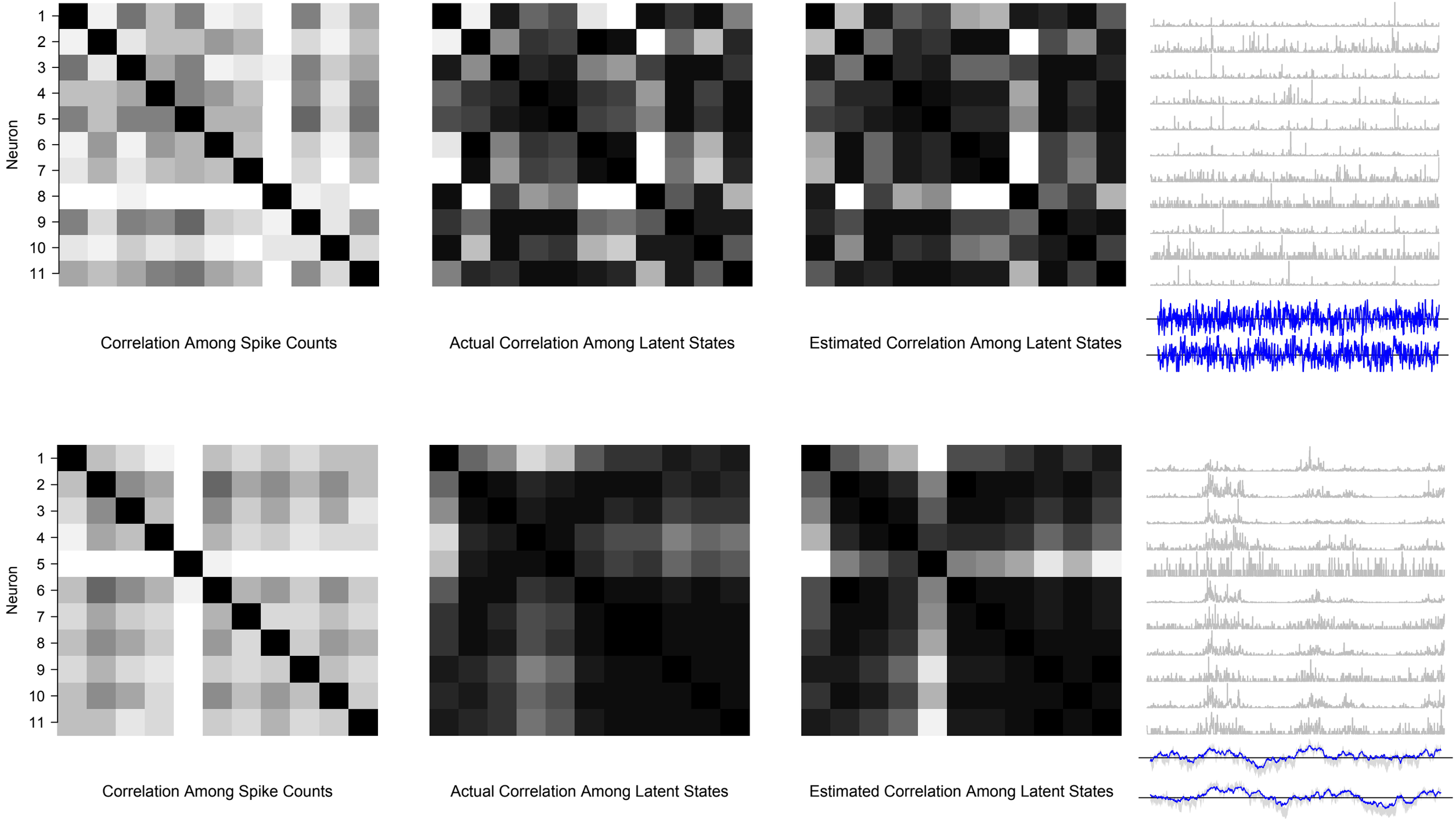
Spike counts



Posterior mean factor scores

Pillow and Scott (2012)

In simulations, we can distinguish low-autocorrelation from high-autocorrelation latent features.



A connection with variational methods:

Let $l_i(\beta)$ be the ith contribution to the log-likelihood. Using basic properties of concave functions, we have

$$\begin{aligned} l_i(\beta) &= y_i\psi_i - n_i \log\{1+\exp(\psi_i)\} \\ &= (y_i - n_i/2)\psi_i - n_i \log\{\exp(\psi_i/2) + \exp(-\psi_i/2)\} \\ &= (y_i - n_i/2)\psi_i - n_i \inf_{\lambda_i} \left\{\lambda_i \psi_i^2 - \phi^\star(\lambda_i)\right\} \end{aligned}$$

A connection with variational methods:

Let $l_i(\beta)$ be the ith contribution to the log-likelihood. Using basic properties of concave functions, we have

$$\begin{aligned} l_i(\beta) &= y_i\psi_i - n_i \log\{1+\exp(\psi_i)\} \\ &= (y_i - n_i/2)\psi_i - n_i \log\{\exp(\psi_i/2) + \exp(-\psi_i/2)\} \\ &= (y_i - n_i/2)\psi_i - n_i \inf_{\lambda_i} \left\{\lambda_i \psi_i^2 - \phi^\star(\lambda_i)\right\} \end{aligned}$$

By a complete different route, the Pólya-Gamma argument gives

$$\begin{aligned} l_i(\beta) &= (y_i - n_i/2)\psi_i - n_i \log\{\exp(\psi_i/2) + \exp(-\psi_i/2)\} \\ &= (y_i - n_i/2)\psi_i - n_i \mathbb{E}_{\lambda_i} \left\{\lambda_i \psi_i^2 - \log p(\lambda_i)\right\} \end{aligned}$$

A connection with variational methods:

Let $l_i(\beta)$ be the ith contribution to the log-likelihood. Using basic properties of concave functions, we have

$$\begin{aligned} l_i(\beta) &= y_i\psi_i - n_i \log\{1+\exp(\psi_i)\} \\ &= (y_i - n_i/2)\psi_i - n_i \log\{\exp(\psi_i/2) + \exp(-\psi_i/2)\} \\ &= (y_i - n_i/2)\psi_i - n_i \inf_{\lambda_i} \left\{\lambda_i\psi_i^2 - \phi^\star(\lambda_i)\right\} \end{aligned}$$

By a complete different route, the Pólya-Gamma argument gives

$$\begin{aligned} l_i(\beta) &= (y_i - n_i/2)\psi_i - n_i \log\{\exp(\psi_i/2) + \exp(-\psi_i/2)\} \\ &= (y_i - n_i/2)\psi_i - n_i \mathbb{E}_{\lambda_i} \left\{\lambda_i\psi_i^2 - \log p(\lambda_i)\right\} \end{aligned}$$

Both give $\hat{\lambda}(\psi) = \mathbb{E}(\lambda \mid \psi) = \frac{1}{4\psi}\tanh(\psi/2)$.

The main lesson

Whatever you can do with a Gaussian model, you can do with a logit or negative-binomial model, using Polya-Gamma data augmentation.

The paper

Bayesian inference for logistic models using Polya-Gamma latent variables (with Nick Polson and Jesse Windle).
Journal of the American Statistical Association (Theory and Methods), 2013.

The R package

BayesLogit (available on CRAN; includes a Polya-Gamma sampler as a separate routine)

Thank you!

Handling the logistic "error" term:

Holmes and Held (2006) exploit the fact that the logistic distribution is a mixture of normals.

$$
\begin{aligned}
z_i &= x_i^T \beta + \epsilon_i \\
(\epsilon_i \mid \phi_i) &\sim \mathsf{N}(0, \phi_i) \\
\phi_i &= (2\lambda_i^2) \\
\lambda_i &\sim \mathsf{KS}(1) \, .
\end{aligned}
$$

Now the logit model is a mixture of probits.

The second-layer latent variables involve non-conjugate distributions, and are updated via adaptive rejection sampling.

Frühwirth-Schnatter and Frühwirth (2010) use a discrete approximation to the logistic distribution.

$$
\begin{aligned}
z_i &= x_i^T \beta + \epsilon_i \\
(\epsilon_i \mid \phi_i) &\sim \mathsf{N}(0, \phi_i) \\
\phi_i &\sim \sum_{k=1}^{K} w_k \delta_{\phi^{(k)}}
\end{aligned}
$$

Again, the logit model is (almost) a mixture of probits.

This method (and its variations) yields slowly mixing samplers. Also, a good approximation to the link function does not imply a good approximation to the posterior distribution.