Bayesian Nonparametric Single-Index Regression

George Karabatsos
College of Education
University of Illinois-Chicago
georgek@uic.edu

Isaac Newton Institute Workshop
CONSTRUCTION AND PROPERTIES OF BAYESIAN NONPARAMETRIC REGRESSION MODELS
6 August to 10 August 2007

Outline

I. Introduce basic framework of GLM and quasi-likelihood model.

II. Discuss positive aspects and features of the nonparametric quasi-likelihood model, which makes use of a *single index*.

III. Open problems with the model.

IV. A novel, Bayesian nonparametric approach to the model:
-- regression function is modeled by splines where the number of location of knots are unknown.
-- error variance and random effects are each modeled by an MDP

V. Methods for sampling the posterior.
(Mainly, slice sampling and adaptive Metropolis sampling).

VI. Application to school data, involving a high-dimensional covariate.

VII. Conclusions

This work was inspired by previous discussions with:
Tim Hanson, Alejandro Jara, Athanasios Kottas, Stephen Walker
(any errors are my own).

Introduction

- The Generalized Linear Model (GLM; McCullagh & Nelder, 1989; Nelder & Wedderburn, 1972) has seen many statistical applications.

- The GLM can be represented by:

$$y_i|\mathbf{x}_i, \boldsymbol{\omega} \sim \pi(\ g(\mathbf{x}_i^T\boldsymbol{\omega}),\ V(g(\mathbf{x}_i^T\boldsymbol{\omega}))\ ), \quad i = 1, \ldots, n$$

y_i observed univariate (continuous or discrete) outcome;

$\mathbf{x}_i$ p-dimensional vector of covariates (maybe high-dimensional);

$\boldsymbol{\omega}$ p-dimensional vector of regression coefficients;

$\mathbf{x}^T\boldsymbol{\omega}$ is the *single index*;

π expon family dist. with mean $g(\mathbf{x}^T\boldsymbol{\omega})$ and variance $V(\mathbf{x}^T\boldsymbol{\omega})$;

$g : \mathbb{R}^p \to R \subset \mathbb{R}$ a known mean function;

$V : \mathbb{R} \to [0, \infty)$ a known variance function.

Nonparametric Regression

$$y_i|\mathbf{x}_i, \boldsymbol{\omega} \sim \pi(\ g(\mathbf{x}_i^T\boldsymbol{\omega}),\ V(g(\mathbf{x}_i^T\boldsymbol{\omega}))\ ),$$
$$i = 1, \ldots, n$$

- In standard GLMs,
  the mean function $g(\cdot)$ and the variance function $V(\cdot)$
  are both assumed to be known.

- Chiou and Müller (1998, JASA) introduced
  a nonparametric (quasi-likelihood) regression model,
  where $g(\cdot)$, $V(\cdot)$ are estimated with smoothing techniques
  (e.g., local polynomial fitting by locally-weighted least squares),
  along with $\boldsymbol{\omega}$.
  .

Features of this NP Regression Model

$$y_i|\mathbf{x}_i, \boldsymbol{\omega} \sim \pi(\ g(\mathbf{x}_i^T\boldsymbol{\omega}),\ V(g(\mathbf{x}_i^T\boldsymbol{\omega}))\ ),$$
$$i = 1, \ldots, n$$

(1) Flexible modeling:

The mean function $g(\cdot)$ and the variance function $V(\cdot)$ can depend on the single index $\mathbf{x}^T\boldsymbol{\omega}$ in a flexible way.

Features of this NP Regression Model

$$y_i|\mathbf{x}_i, \boldsymbol{\omega} \sim \pi(\ g(\mathbf{x}_i^T\boldsymbol{\omega}),\ V(g(\mathbf{x}_i^T\boldsymbol{\omega}))\ ),$$
$$i = 1, \ldots, n$$

(2) Handles high-dimensional covariates:

Handles the "curse of dimensionality" (Bellman, 1961) by the projection of the (possibly) high-dimensional covariate **x** onto a univariate index $\mathbf{x}_i^T\boldsymbol{\omega}$, while capturing important features in the data.

- It has been argued that such a dimension reduction is a necessary step in data analysis (e.g., Li, 1991, JASA);

- Many data sets have a large number of covariates. This is especially true in educational research.

Application

Data

184 Students aged 15-16 from 10 U.S. high schools.

It is of interest to explore the factors (23 covariates)
that influence the students' self-efficacy with mathematics,
while allowing for school random effects.

It is known that this self-efficacy is key to a student's
success and interest in math-related fields.

Application

Outcome

MATHEFF	Mathematics self-efficacy

23 Covariates

RMHMWK	Relative time spent on Maths homework
COMPHOME	Computer facilities at home
HEDRES	Home educational resources
ATSCHL	Attitudes towards school
STUREL	Student-teacher relations at school
INTMAT	Interest in mathematics
INSTMOT	Instrumental motivation in mathematics
ANXMAT	Mathematics anxiety
SCMAT	Mathematics self-concept
CSTRAT	Control strategies
ELAB	Elaboration strategies
MEMOR	Memorisation strategies

Application

Outcome

MATHEFF Mathematics self-efficacy

23 Covariates (continued)

COMPLRN Competitive learning

COOPLRN Co-operative learning

TEACHSUP Teacher support in maths lessons

DISCLIM Disciplinary climate in maths lessons

INTUSE ICT: Internet/entertainment use

PRGUSE ICT: Programs/software use

ROUTCONF ICT: Confidence in routine tasks

INTCONF ICT: Confidence in internet tasks

HIGHCONF ICT: Confidence in high-level tasks

ATTCOMP ICT: Attitudes towards computers

ESCS Index of Socio-Economic and Cultural Status

Random effect: School membership (1-10)

Features of this NP Regression Model

$$y_i|\mathbf{x}_i, \boldsymbol{\omega} \sim \pi(\ g(\mathbf{x}_i^T\boldsymbol{\omega}),\ V(g(\mathbf{x}_i^T\boldsymbol{\omega}))\ ),$$
$$i = 1, \ldots, n$$

(2) Handles high-dimensional covariates:

Handles the "curse of dimensionality" (Bellman, 1961) by the projection of the (possibly) high-dimensional covariate $\mathbf{x}$ onto a univariate index $\mathbf{x}_i^T\boldsymbol{\omega}$, while capturing important features in the data.

- The nonparametric modeling of $g(\mathbf{x})$ (with $\mathbf{x}$ high-dimensional) can be challenging, due to the curse of dimensionality.
- It is easier to work with nonparametric modeling of $g(\mathbf{x}_i^T\boldsymbol{\omega})$. This approach always leads to one-dimensional nonparametric regression, regardless of the dimension of $\mathbf{x} = (x_1, \ldots, x_p)$.
- The projection of a high-dimensional $\mathbf{x}$ onto a single index $\mathbf{x}_i^T\boldsymbol{\omega}$ is the basic idea behind projection pursuit regression (Friedman & Steutzle, 1981, JASA).

Features of this NP Regression Model

(2) Handling high-dimensional covariates:

- We could use additive modeling and take:

$$g(\mathbf{x}) = \beta_0 + g_1(x_1) + g_2(x_2) + \cdots + g_p(x_p)$$

but then we would ignore interactions (among the p covariates).

Incorporating interactions in this framework would lead to an explosion of the number of regression terms.

- In binary regression, we could take:
$g(\mathbf{x}) = G(\mathbf{x}^T\boldsymbol{\omega})$ and $G \sim \text{DP}(m, G_0)$ (Newton, et al., 1996, *JASA*).

However, then we would not be able to account for any non-monotonicity in the true $g(\cdot)$.

Features of this NP Regression Model

$$y_i|\mathbf{x}_i, \boldsymbol{\omega} \sim \pi(\ g(\mathbf{x}_i^T\boldsymbol{\omega}),\ V(g(\mathbf{x}_i^T\boldsymbol{\omega}))\ ), \quad i = 1, \ldots, n$$

(3) Interactions:

The model automatically handles any interactions among

covariates $\mathbf{x} = (x_1, \ldots, x_p)$,

since a nonlinear link function $g(\cdot)$ is applied to

the single index $\mathbf{x}^T\boldsymbol{\omega}$.

Thus, the model provides an alternative to additive models, which reduce dimensionality, but does not easily handle interactions.

(4) Interpretability:

When $g(\cdot)$ is monotonic, $\boldsymbol{\omega}$ has the same interpretation as "effect" parameters as in ordinary linear models.

Features of this NP Regression Model

$$y_i|\mathbf{x}_i, \boldsymbol{\omega} \sim \pi(\ g(\mathbf{x}_i^T\boldsymbol{\omega}),\ V(g(\mathbf{x}_i^T\boldsymbol{\omega}))\ ),$$
$$i = 1, \ldots, n$$

(5) Handles Heteroscedasticity:

Since the variance function $V(\cdot)$ is allowed to depend on the single index $\mathbf{x}^T\boldsymbol{\omega}$ in a flexible way, we can account for heterogeneity in variance.

Incorrectly assuming homoscedasticity can lead to misleading statistical inferences.

Features of this NP Regression Model

$$y_i|\mathbf{x}_i, \boldsymbol{\omega} \sim \pi(\ g(\mathbf{x}_i^T\boldsymbol{\omega}),\ V(g(\mathbf{x}_i^T\boldsymbol{\omega}))\ ),$$
$$i = 1, \ldots, n$$

(6) Consistency, Efficiency:

The model addresses the fact that the consistency of the estimate of the regression coefficients ($\boldsymbol{\omega}$) depends on a correctly specified link function,
and that the efficiency of these estimates
depend on a correctly specified variance function.

Open Problems, NP regression model

$$y_i|\mathbf{x}_i, \boldsymbol{\omega} \sim \pi(\ g(\mathbf{x}_i^T\boldsymbol{\omega}),\ V(g(\mathbf{x}_i^T\boldsymbol{\omega}))\ ),$$
$$i = 1, \ldots, n$$

(1) Frequentist Only:

Available inference methods focus only on point estimation.

Fully Bayesian approaches to this model have yet to be devised.

- Point estimation only provides approximate inference.
- An *empirical* Bayes approach is considered by Antoniadis et al. (2003), assuming homoscedasticity.

Open Problems, NP regression model

$$y_i|\mathbf{x}_i, \boldsymbol{\omega} \sim \pi(\ g(\mathbf{x}_i^T\boldsymbol{\omega}),\ V(g(\mathbf{x}_i^T\boldsymbol{\omega}))\ ), \quad i = 1, \ldots, n$$

(2) No random effects:

- The model does not account for sources of variance that are not explained by the covariate vector **x**.
- More generally, we wish to be able to account for random effects to model correlation between different clusters of observations.
- The inclusion of random effects in this formulation renders the model applicable to various study designs, such as clustered, longitudinal, hierarchical, and spatial designs.
- "Off the shelf" LMMs and GLMMs (including "multilevel" models) accounts for such random effects. But, in such models, the mean function $g(\cdot)$ and the variance function $V(\cdot)$ are assumed to be known.

Bayesian Regression

$$y_i|\mathbf{x}_i, \boldsymbol{\omega} \sim \pi(\ g(\mathbf{x}_i^T\boldsymbol{\omega}),\ V(g(\mathbf{x}_i^T\boldsymbol{\omega}))\ ),$$
$$i = 1, \ldots, n$$

- $g(\mathbf{x}^T\boldsymbol{\omega})$ is modeled by splines, such that the **number** and **locations** of the knots are treated as unknown.
  - $g(\cdot)$ can take on any shape.
  - The set of candidate knots is defined by a large number $K_{\max}$ equidistant percentile points of the observed values of the single index: $\mathbf{x}_i^T\boldsymbol{\omega}$, $i = 1,\ldots,n$.
  - We constrain $\|\boldsymbol{\omega}\| = 1$ for identifiability.

- The random effects $\theta_1,\ldots,\theta_J$ are modeled by an MDP. (we avoid parametric assumptions about the distribution)
- The variance function $V(\cdot)$ is also modeled by a MDP The regression error variance (σ^2) changes nonparametrically with $g(\mathbf{x}'_i\boldsymbol{\omega}) + \theta_j$.

The most convenient way to describe the Dirichlet Process is through a representation based on a countably-infinite sampling strategy.

Let:

$\theta_1, \theta_2, \ldots, \theta_j, \ldots \overset{iid}{\sim} G_0$

$v_1, v_2, \ldots, v_j, \ldots \overset{iid}{\sim} \text{Beta}(1, \alpha) \;\; (\alpha > 0)$

Then a random distribution function F chosen from a Dirichlet Process prior with parameters (α, G_0) can be constructed via

$F(x) = \sum_{j=1}^{\infty} w_j \, \mathbf{1}(\theta_j \leq x),$

where $w_1 = v_1$ and for $j > 1$, we have $w_j = v_j \prod_{l<j} (1 - v_l)$.

MDP Mixture model:

$f_G(x) = \int K(x; \theta) \, G(\mathrm{d}\theta), \;\; \text{with } G \sim \text{DP}(\alpha, G_0).$

$x_j | \theta_j \overset{iid}{\sim} K(x; \theta_j), \;\; \text{and} \;\; \theta_j \sim G, \quad \text{with } G \sim \text{DP}(\alpha, G_0).$

$$y_i|\mu_i, \sigma_i^2 \sim \text{Normal}(g(\mathbf{x}_i^T\boldsymbol{\omega}) + \theta_{j(i)}, \sigma_i^2), \quad i = 1, \ldots, n,$$

$$g(\mathbf{x}_i^T\boldsymbol{\omega}) = \beta_0 + \sum_{k=1}^{K_{\max}} \beta_k (\mathbf{x}_i^T\boldsymbol{\omega} - t_k)_+^q$$

$$\boldsymbol{\omega} \sim \text{FisherVonMises}_p(\lambda, \boldsymbol{\xi})$$

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim \text{Uniform}(\boldsymbol{\beta}_{\gamma=1}) \times \delta_{\mathbf{0}}(\boldsymbol{\beta}_{\gamma=0})$$

$$\boldsymbol{\gamma}|\boldsymbol{\phi} \sim \text{Bernoulli}(\gamma_0|1) \prod_{k=1}^{K_{\max}} \text{Bernoulli}(\gamma_k|\phi_k)$$

$$\sigma_1^2, \ldots, \sigma_n^2|G_\sigma \overset{iid}{\sim} G_\sigma$$

$$G_\sigma|\alpha_\sigma, \rho_1, \rho_2 \sim \text{DirichletProcess}(\alpha_\sigma, \text{InverseGamma}(\rho_1, \rho_2))$$

$$\alpha_\sigma|a_\sigma, b_\sigma \sim \text{Gamma}(a_\sigma, b_\sigma)$$

$$\rho_1, \rho_2|b_\rho \overset{iid}{\sim} \text{Uniform}(0, b_\rho)$$

$$\theta_1, \ldots, \theta_J|G_\theta \overset{iid}{\sim} G_\theta$$

$$G_\theta|\alpha_\theta, \tau \sim \text{DirichletProcess}(\alpha_\theta, \text{Normal}(0, \tau))$$

$$\alpha_\theta|a_\theta, b_\theta \sim \text{Gamma}(a_\theta, b_\theta)$$

$$\tau|b_\tau \sim \text{Uniform}(0, b_\tau)$$

When the outcomes y_i are binary, ordinal, or counts,
that take on discrete values in $c = 0, \ldots, C$,
we can adopt a latent variable approach by taking:

$$\mathbf{1}(y_i = c) \sim \int_{z \in A_c} \text{Normal}(z \mid g(\mathbf{x}_i^T \boldsymbol{\omega}) + \theta_{j(i)}, \sigma_i^2)\text{dz},$$

$$i = 1, \ldots, n,$$

where $A_c = (c-1, c]$

(with $c - 1 = -1 \equiv -\infty$ and $c = C \equiv \infty$)

Posterior Sampling

Sampling ω: Adaptive Random Walk Metropolis (ARWM) step.

Spline coefficients: Use Neal's (2003, *Ann Stat*) slice sampler for the intercept β_0. Then use a novel, a bivariate slice sampler for each pair of spline parameters (γ_k, β_k) , k=1,…,$K_{\max}$ (with random scan).

Regression Variances: Use Damien-Walker's (1998) slice sampler for the regression variances.

Random Effects: Use Neal's (2000) algorithm to sample the cluster memberships of the random effects, and then use Neal's (2003) slice sampler for the random effects, given new clusters.

DP Hyperparameters: Use Neal's (2003) univariate slice sampler for τ (variance of random effects), and for ρ_1 and ρ_2 (inverse gamma parameters of error variances).

Sampling $\boldsymbol{\omega}$

To generate samples from the conditional posterior distribution:

$$p(\boldsymbol{\omega}|rest, Data) \propto \ell(Data|\boldsymbol{\omega}, rest) \times \text{FisherVonMises}_p(\lambda, \boldsymbol{\xi}^T\boldsymbol{\omega}),$$

we implement a special version of Atchadé and Rosenthal's (2005, *Bernoulli*) Adaptive Random-Walk Metropolis (ARWM) algorithm.

Sampling $\boldsymbol{\omega}$

Algorithm ARWM algorithm for sampling $p(\boldsymbol{\omega}|rest, Data)$

1. Generate a proposal $\boldsymbol{\omega}^*$ from the spherical distribution $H_p(\underline{\lambda}^{old}, \boldsymbol{\omega}^{old})$.

2. With probability $\alpha(\boldsymbol{\omega}^{old}, \boldsymbol{\omega}^{new}) = \min\left\{1, \dfrac{\ell(Data|\boldsymbol{\omega}^*, rest) \times \exp(\lambda \boldsymbol{\xi}^T \boldsymbol{\omega}^*)}{\ell(Data|\boldsymbol{\omega}^{old}, rest) \times \exp(\lambda \boldsymbol{\xi}^T \boldsymbol{\omega}^{old})}\right\}$,
   accept $\boldsymbol{\omega}^*$ as the new state ($\boldsymbol{\omega}^{new} = \boldsymbol{\omega}^*$), otherwise, set $\boldsymbol{\omega}^{new} = \boldsymbol{\omega}^{old}$.

3. Compute $\underline{\lambda}^{new} = q(\underline{\lambda}^{old} + \eta_t(.234 - \alpha(\boldsymbol{\omega}^{old}, \boldsymbol{\omega}^{new})))$.

Sampling $\boldsymbol{\omega}$

Algorithm ARWM algorithm for sampling $p(\boldsymbol{\omega}|rest, Data)$

1. Generate a proposal $\boldsymbol{\omega}^*$ from the spherical distribution $H_p(\underline{\lambda}^{old}, \boldsymbol{\omega}^{old})$.

2. With probability $\alpha(\boldsymbol{\omega}^{old}, \boldsymbol{\omega}^{new}) = \min\left\{1, \dfrac{\ell(Data|\boldsymbol{\omega}^*, rest) \times \exp(\lambda \boldsymbol{\xi}^T \boldsymbol{\omega}^*)}{\ell(Data|\boldsymbol{\omega}^{old}, rest) \times \exp(\lambda \boldsymbol{\xi}^T \boldsymbol{\omega}^{old})}\right\}$,
accept $\boldsymbol{\omega}^*$ as the new state ($\boldsymbol{\omega}^{new} = \boldsymbol{\omega}^*$), otherwise, set $\boldsymbol{\omega}^{new} = \boldsymbol{\omega}^{old}$.

3. Compute $\underline{\lambda}^{new} = q(\underline{\lambda}^{old} + \eta_t(.234 - \alpha(\boldsymbol{\omega}^{old}, \boldsymbol{\omega}^{new})))$.

Note:

(1) The spherical proposal distribution H_p is easy to sample (Saw, 1978).
If $y \sim \text{Normal}_p(\mu, \mathbf{I}_p)$,
then $\boldsymbol{\omega} = y/(y^T y)^{1/2} \sim H_p(\underline{\lambda} = (\tfrac{1}{2}\mu^T \mu)^{\frac{1}{2}}, \ \boldsymbol{\xi} = \mu/(\mu^T \mu)^{\frac{1}{2}})$.

(2) Since this spherical distribution is symmetric, the ratio of proposal densities cancel out in the acceptance ratio.

(3) The Fisher von Mises prior density is proportional to $\exp(\lambda \boldsymbol{\xi}^T \boldsymbol{\omega})$.

Sampling $\boldsymbol{\omega}$

Algorithm ARWM algorithm for sampling $p(\boldsymbol{\omega}|rest, Data)$

1. Generate a proposal $\boldsymbol{\omega}^*$ from the spherical distribution $H_p(\underline{\lambda}^{old}, \boldsymbol{\omega}^{old})$.

2. With probability $\alpha(\boldsymbol{\omega}^{old}, \boldsymbol{\omega}^{new}) = \min\left\{1, \frac{\ell(Data|\boldsymbol{\omega}^*, rest) \times \exp(\lambda \boldsymbol{\xi}^T \boldsymbol{\omega}^*)}{\ell(Data|\boldsymbol{\omega}^{old}, rest) \times \exp(\lambda \boldsymbol{\xi}^T \boldsymbol{\omega}^{old})}\right\}$,
accept $\boldsymbol{\omega}^*$ as the new state ($\boldsymbol{\omega}^{new} = \boldsymbol{\omega}^*$), otherwise, set $\boldsymbol{\omega}^{new} = \boldsymbol{\omega}^{old}$.

3. Compute $\underline{\lambda}^{new} = q(\underline{\lambda}^{old} + \eta_t(.234 - \alpha(\boldsymbol{\omega}^{old}, \boldsymbol{\omega}^{new})))$.

Note:

(4) In high-dimensional spaces (under general conditions), the efficiency of the Metropolis algorithm is optimized when the variance of the proposal distribution has an asymptotic acceptance rate of approximately .234. (Roberts & Rosenthal, 2001, *Statistical Science*)

Sampling $\boldsymbol{\omega}$

Algorithm ARWM algorithm for sampling $p(\boldsymbol{\omega}|rest, Data)$

1. Generate a proposal $\boldsymbol{\omega}^*$ from the spherical distribution $H_p(\underline{\lambda}^{old}, \boldsymbol{\omega}^{old})$.

2. With probability $\alpha(\boldsymbol{\omega}^{old}, \boldsymbol{\omega}^{new}) = \min\left\{1, \frac{\ell(Data|\boldsymbol{\omega}^*, rest) \times \exp(\lambda \boldsymbol{\xi}^T \boldsymbol{\omega}^*)}{\ell(Data|\boldsymbol{\omega}^{old}, rest) \times \exp(\lambda \boldsymbol{\xi}^T \boldsymbol{\omega}^{old})}\right\}$,
accept $\boldsymbol{\omega}^*$ as the new state ($\boldsymbol{\omega}^{new} = \boldsymbol{\omega}^*$), otherwise, set $\boldsymbol{\omega}^{new} = \boldsymbol{\omega}^{old}$.

3. Compute $\underline{\lambda}^{new} = q(\underline{\lambda}^{old} + \eta_t(.234 - \alpha(\boldsymbol{\omega}^{old}, \boldsymbol{\omega}^{new})))$.

Note:

(5) In Step 3, a Robbins-Monro (1951, *Ann Math Stat*) stochastic approximation algorithm updates $\underline{\lambda}$.
This increases $\underline{\lambda}$ (decreases the proposal variance) when $\alpha(\boldsymbol{\omega}^{old}, \boldsymbol{\omega}^{new})$ is below .234,
and decreases $\underline{\lambda}$ (increases the proposal variance) when $\alpha(\boldsymbol{\omega}^{old}, \boldsymbol{\omega}^{new})$ is above .234.

Sampling $\boldsymbol{\omega}$

Algorithm ARWM algorithm for sampling $p(\boldsymbol{\omega}|rest, Data)$

1. Generate a proposal $\boldsymbol{\omega}^*$ from the spherical distribution $H_p(\underline{\lambda}^{old}, \boldsymbol{\omega}^{old})$.

2. With probability $\alpha(\boldsymbol{\omega}^{old}, \boldsymbol{\omega}^{new}) = \min\left\{1, \frac{\ell(Data|\boldsymbol{\omega}^*, rest) \times \exp(\lambda \boldsymbol{\xi}^T \boldsymbol{\omega}^*)}{\ell(Data|\boldsymbol{\omega}^{old}, rest) \times \exp(\lambda \boldsymbol{\xi}^T \boldsymbol{\omega}^{old})}\right\}$,
accept $\boldsymbol{\omega}^*$ as the new state ($\boldsymbol{\omega}^{new} = \boldsymbol{\omega}^*$), otherwise, set $\boldsymbol{\omega}^{new} = \boldsymbol{\omega}^{old}$.

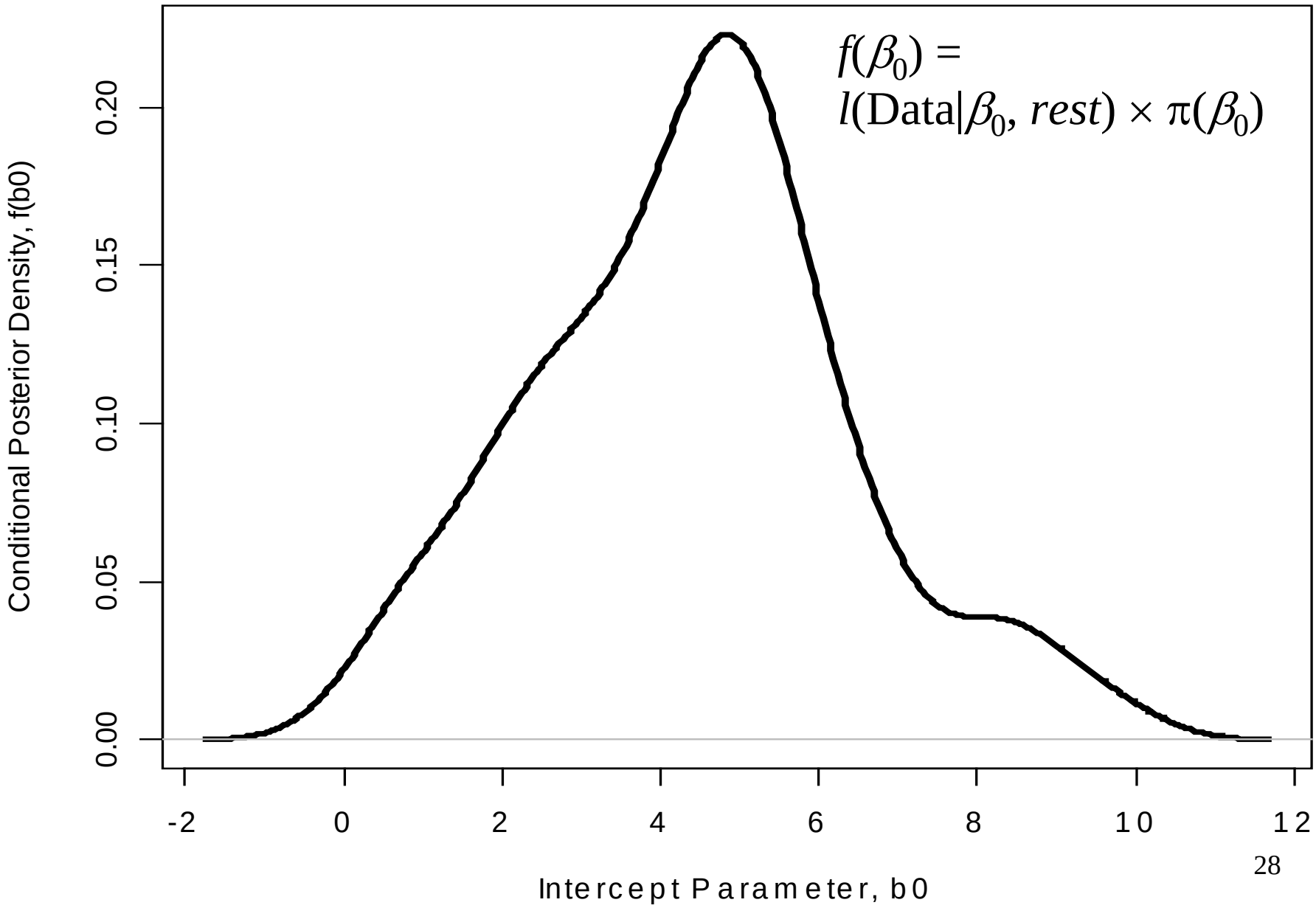
3. Compute $\underline{\lambda}^{new} = q(\underline{\lambda}^{old} + \eta_t(.234 - \alpha(\boldsymbol{\omega}^{old}, \boldsymbol{\omega}^{new})))$.

Note:

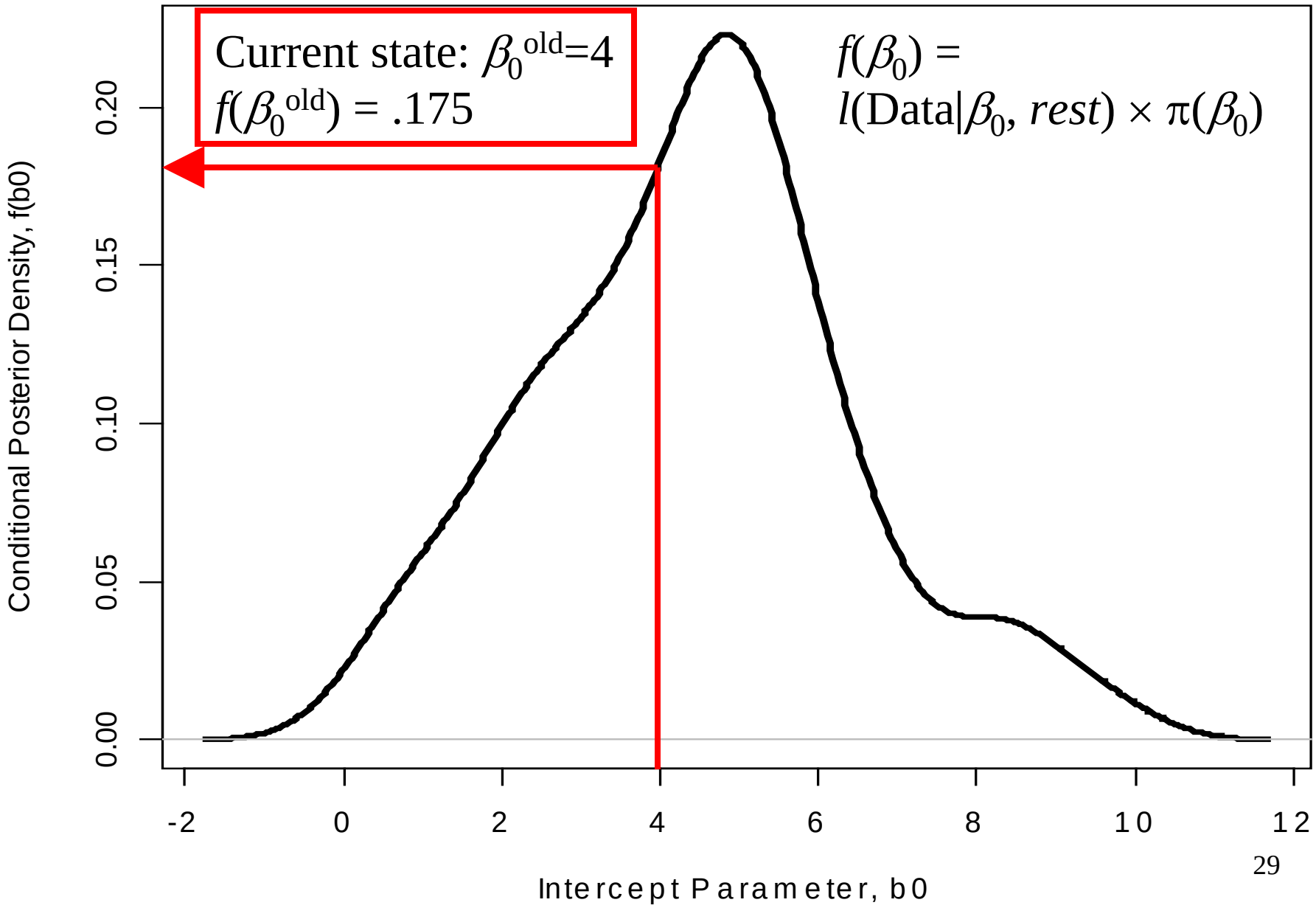
(6) We specify $\eta_t = 10/t$, for iterations $t = 1, \ldots, T$.

(7) Adaptive Metropolis algorithms preserve the ergodicity and stationarity of the target posterior distribution, provided that the size of the change of the proposal variance converges to zero as the number of sampling iterations increases (for details: Roberts & Rosenthal, *J Appl Prob*, to appear).

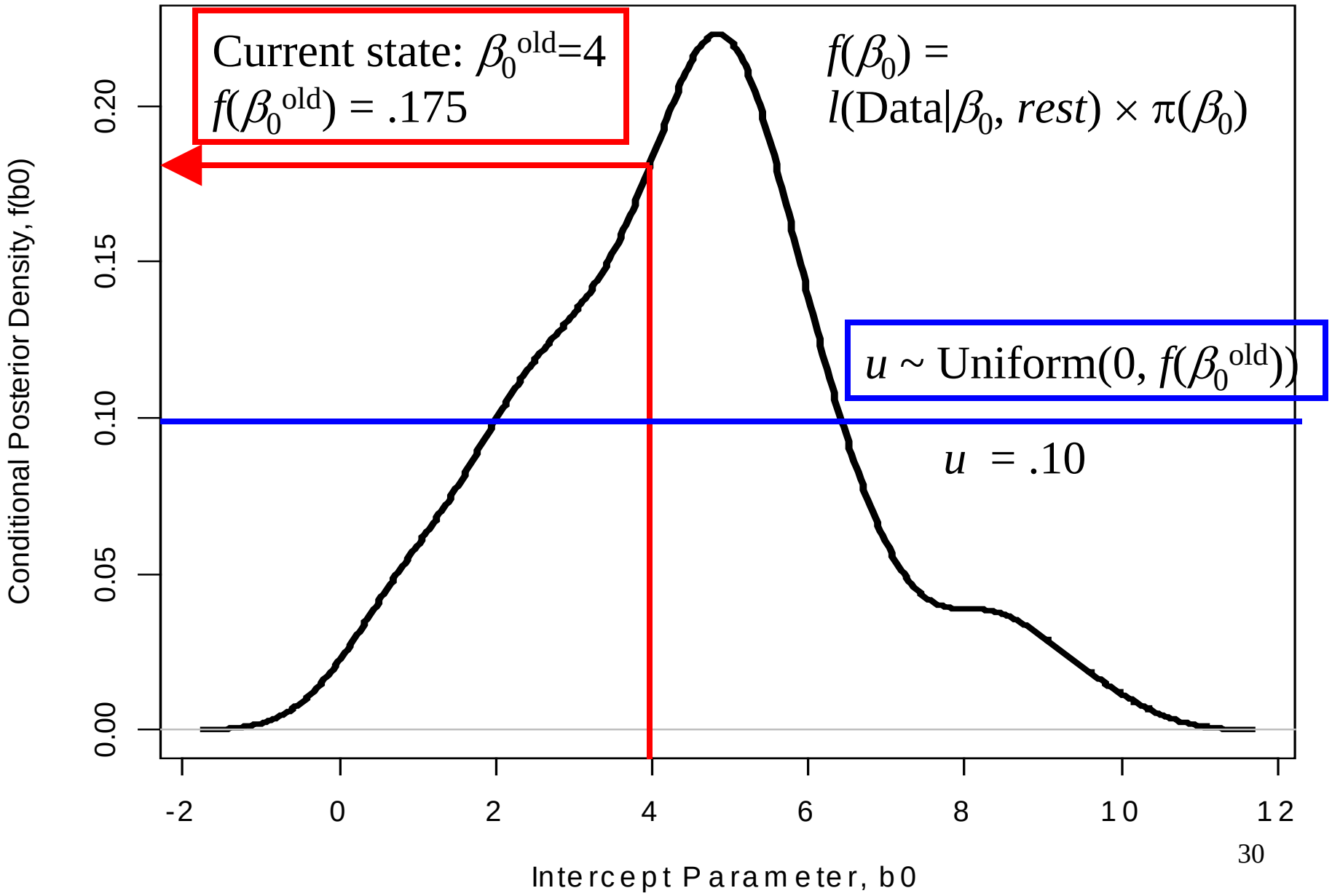
Slice Sampling the Intercept β_0



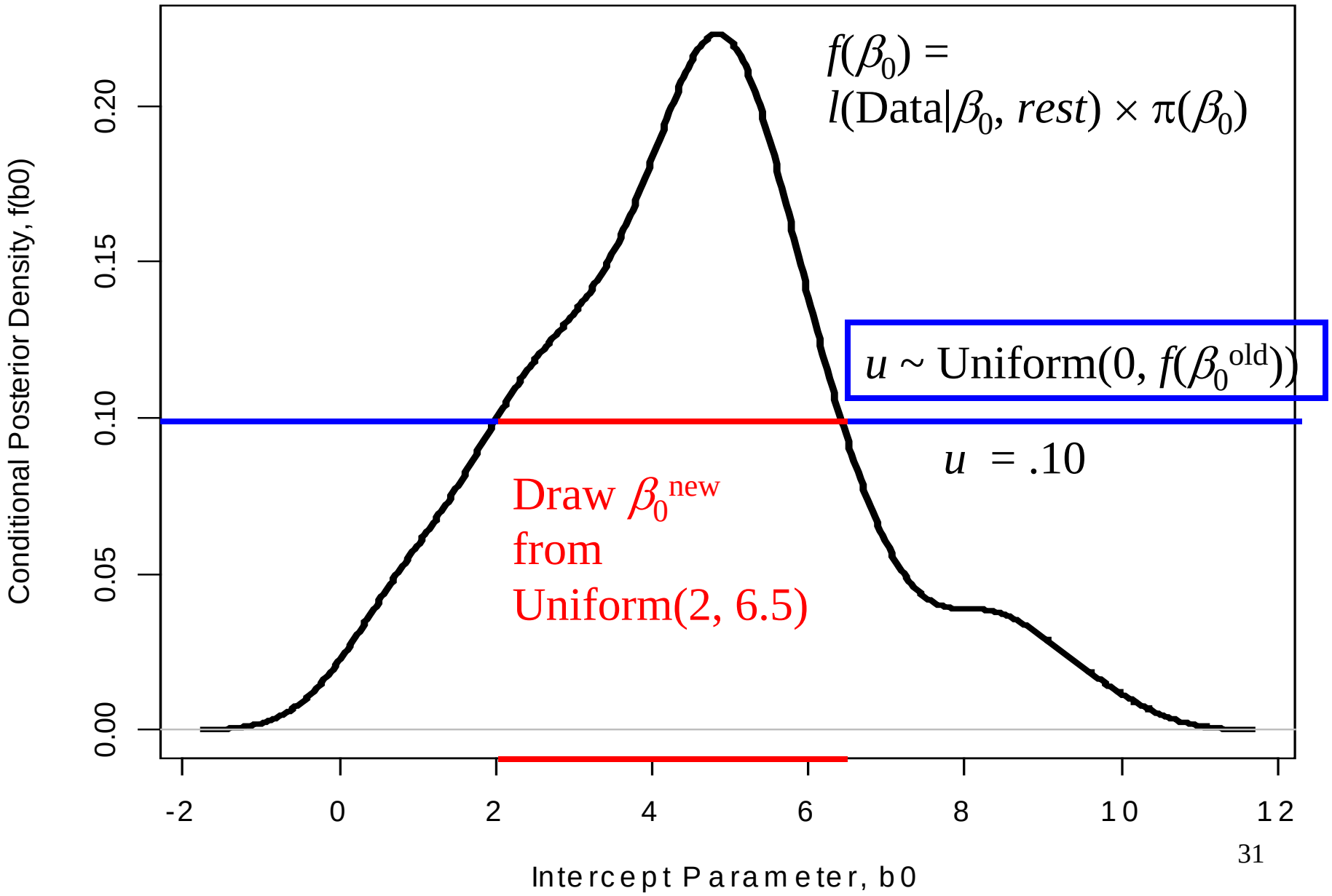
Slice Sampling the Intercept β_0



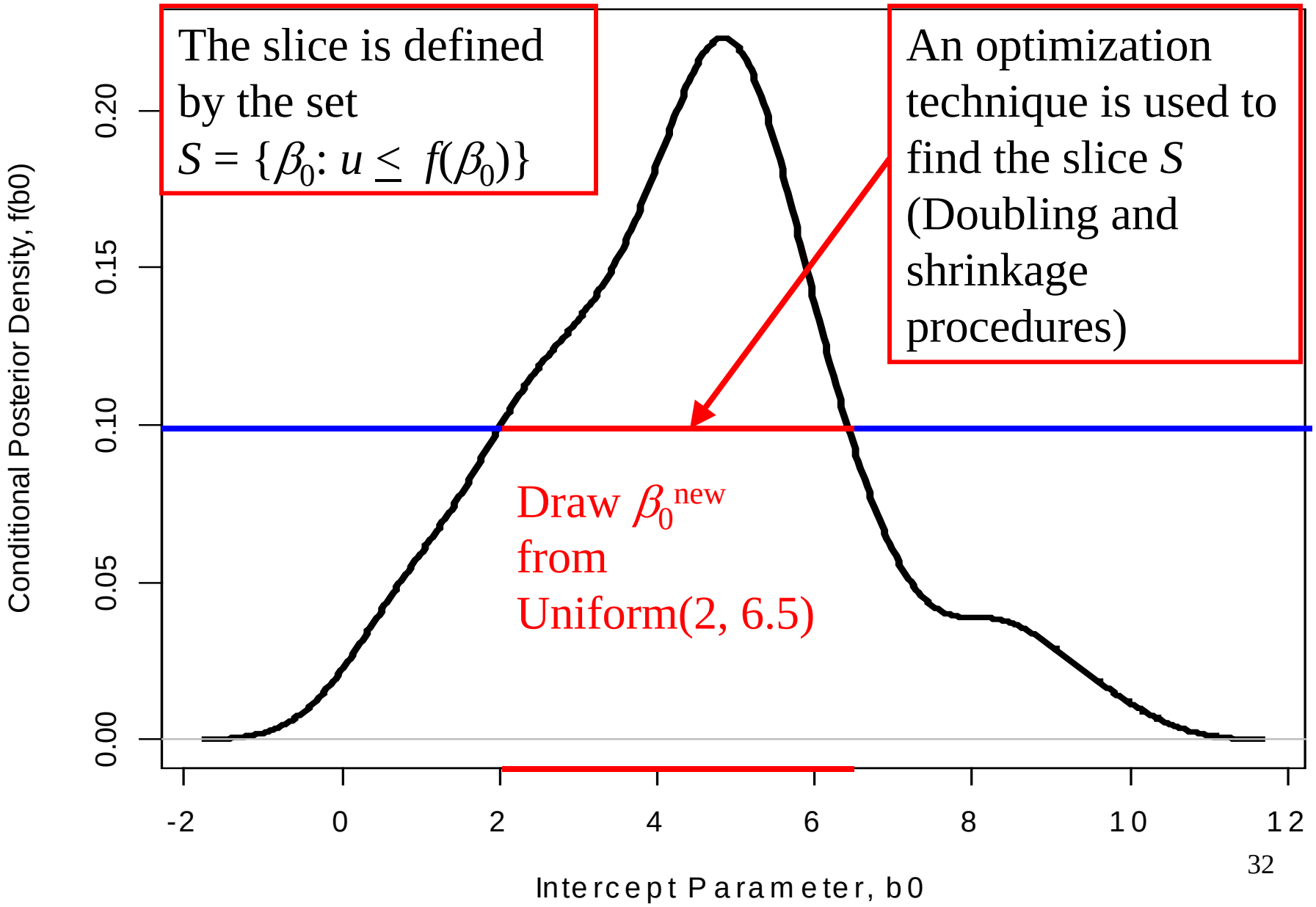
Slice Sampling the Intercept β_0



Slice Sampling the Intercept β_0



Slice Sampling the Intercept β_0



Bivariate Slice Sampling of (γ_k, β_k)

- Now, let: $f(\beta_k, \gamma_k) = l(\text{Data}|\beta_k, \gamma_k, \textit{rest}) \times \pi(\beta_0, \gamma_k)$.

- Recall $\beta_k \in (-\infty, \infty)$ and $\gamma_k \in \{0, 1\}$, for $k = 1,\ldots,K_{\max}$.

- Sampling:
  - Draw $u \sim \text{Uniform}(0, f(\beta_k^{\text{old}}, \gamma_k^{\text{old}}))$
  - Draw new state $(\beta_k^{\text{new}}, \gamma_k^{\text{new}}) \sim \text{Uniform}(S)$, with the slice: $S = \{(\beta_k, \gamma_k) : u \leq f(\beta_k, \gamma_k)\}$

- S is found by doubling and shrinkage procedures.
- This sampling is repeated for all candidate knots $k = 1,\ldots,K_{\max}$.

Slice Sampling the Variances

- We use Damien & Walker's (1998) slice sampling procedure (it is vary fast--important for sampling *n* parameters)
- Recall that the variances $\sigma_1^2,\ldots,\sigma_n^2$ are modeled by an MDP, with G_σ=InverseGamma(ρ_1, ρ_2) the baseline distribution.

- With $\sigma_i^{2(old)}$ the current state of σ_i^2,
  Draw $u_i \sim \text{Uniform}(0, \ell(y_i|\sigma_i^{2(old)}, rest))$,
  to define the slice $S_i = \{\sigma_i^2 : u_i \leq \ell(y_i|\sigma_i^2, rest)\}$
- The slice S_i (interval) can be found quickly (unimodality).

It can be shown:

$$f(\sigma_i^2|u_i, rest, data) = \frac{\alpha_\sigma G_\sigma(S_i)\text{IG}(\sigma_i^2, S_i|\rho_1, \rho_2) + \sum_{g\neq i}^{n-1} \mathbf{1}\left(\sigma_g^{2(old)} \in S_i\right)\delta_{\sigma_g^{2(old)}}(\sigma_i^2)}{\alpha_\sigma \text{IG}(S_i|\rho_1, \rho_2) + \sum_{g\neq i}^{n-1} \mathbf{1}\left(\sigma_g^{2(old)} \in S_i\right)}$$

$\text{IG}(\sigma_i^2, S_i|\rho_1, \rho_2)$ is proportional to $\text{IG}(\sigma_i^2|\rho_1, \rho_2)\mathbf{1}(\sigma_i^2 \in S_i)$.

Slice Sampling the Variances

- We use Damien & Walker's (1998) slice sampling procedure (it is vary fast--important for sampling *n* parameters)
- Recall that the variances $\sigma_1^2,\ldots,\sigma_n^2$ are modeled by an MDP, with G_σ=InverseGamma(ρ_1, ρ_2) the baseline distribution.

- With $\sigma_i^{2(old)}$ the current state of σ_i^2,
  Draw $u_i \sim \text{Uniform}(0, \ell(y_i|\sigma_i^{2(old)}, rest))$,
  to define the slice $S_i = \{\sigma_i^2 : u_i \leq \ell(y_i|\sigma_i^2, rest)\}$

- Then draw $\sigma_i^{2(new)}$ from $\text{IG}(\sigma_i^2|\rho_1,\rho_2)\mathbf{1}(\sigma_i^2 \in S_i)$
  wp $\dfrac{\alpha_\sigma \text{IG}(S_i|\rho_1,\rho_2)}{\alpha_\sigma \text{IG}(S_i|\rho_1,\rho_2) + \sum_{g\neq i}^{n-1} \mathbf{1}\left(\sigma_g^{2(old)} \in S_i\right)}$,
  otherwise,
  sample $\sigma_i^{2(new)}$ uniformly from $\{\sigma_g^{2(old)} : g \neq i, \sigma_g^{2(old)} \in S_i\}$.

Slice Sampling Random Effects

- Recall that the random effects $\theta_1,\ldots,\theta_J$ are modeled by an MDP, with the Normal(0, τ) baseline distribution.
- Sampling clusters:
  Let $\theta_1^{old},\ldots,\theta_c^{old},\ldots,\theta_C^{old}$ denote the distinct values of $\theta_1^{old},\ldots,\theta_J^{old}$.
  wp proportional to: $\frac{n_{-j,c}}{J-1+\alpha_\theta}\prod_{i\in j}\ell(y_i|\theta_c,rest)$

  each θ_j is assigned a cluster defined by θ_c,

  and wp proportional to: $\frac{\alpha_\theta/m}{J-1+\alpha_\theta}\prod_{i\in j}\ell(y_i|\theta_d,rest)$,

  each θ_j is assigned to a cluster defined by θ_d,

  where $\theta_1,\ldots,\theta_d,\ldots,\theta_m \sim_{iid}$ Normal$(0,\tau)$.

  (we choose $m = 300$).
- Then given the new clusters, the univariate slice sampler (Neal, 2003) is used to generate a new state θ_j^{new}, for all $j = 1,\ldots,J$.

Sampling DP hyperparameters

- For each of the DP precision parameters α_θ , α_σ, a Gibbs sample is obtained by generating a sample from a mixture of two gamma densities. This mixture distribution depends on the number of distinct values of the parameters. (Escobar & West, 1995, *JASA*).

- A sample of τ (variance of the random effects) and ρ_1 and ρ_2 (inverse gamma parameters of error variances) are each obtained with Neal's (2003) univariate slice sampler.

For τ, the slice is: $S_\tau = \{ \tau : u \leq \prod_j \text{Normal}(\theta_j \mid \tau) \times \text{Unif}(\tau \mid 0, b_\tau) \}$.

For ρ_1, the slice is: $S_{\rho 1} = \{\rho_1 : u \leq \prod_i \text{IG}(\sigma_i^2 \mid \rho_1, \rho_2) \times \text{Unif}(\rho_1 \mid 0, b_\rho) \}$

For ρ_2, the slice is: $S_{\rho 2} = \{\rho_2 : u \leq \prod_i \text{IG}(\sigma_i^2 \mid \rho_1, \rho_2) \times \text{Unif}(\rho_2 \mid 0, b_\rho) \}$

$$y_i|\mu_i,\sigma_i^2 \sim \text{Normal}(g(\mathbf{x}_i^T\boldsymbol{\omega}) + \theta_{j(i)}, \sigma_i^2), \quad i = 1,\ldots,n,$$

$K_{\max}$= 50 knots
order q = 1

$$g(\mathbf{x}_i^T\boldsymbol{\omega}) = \beta_0 + \sum_{k=1}^{50} \beta_k (\mathbf{x}_i^T\boldsymbol{\omega} - t_k)_+^1$$

$$\boldsymbol{\omega} \sim \text{FisherVonMises}_p\left(\lambda = 1, \boldsymbol{\xi} = (\sqrt{1/p},\ldots,\sqrt{1/p})^T\right)$$

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim \text{Uniform}(\boldsymbol{\beta}_{\gamma=1}) \times \delta_{\mathbf{0}}(\boldsymbol{\beta}_{\gamma=0})$$

$$\boldsymbol{\gamma}|\boldsymbol{\phi} \sim \text{Bernoulli}(\gamma_0|1) \prod_{k=1}^{50} \text{Bernoulli}(\gamma_k|\tfrac{1}{2})$$

$$\sigma_1^2,\ldots,\sigma_n^2|G_\sigma \overset{iid}{\sim} G_\sigma$$

$$G_\sigma|\alpha_\sigma,\rho_1,\rho_2 \sim \text{DirichletProcess}(\alpha_\sigma, \text{InverseGamma}(\rho_1,\rho_2))$$

$$\alpha_\sigma|a_\sigma,b_\sigma \sim \text{Gamma}(1,10)$$

$$\rho_1,\rho_2|b_\rho \overset{iid}{\sim} \text{Uniform}(0,10^5)$$

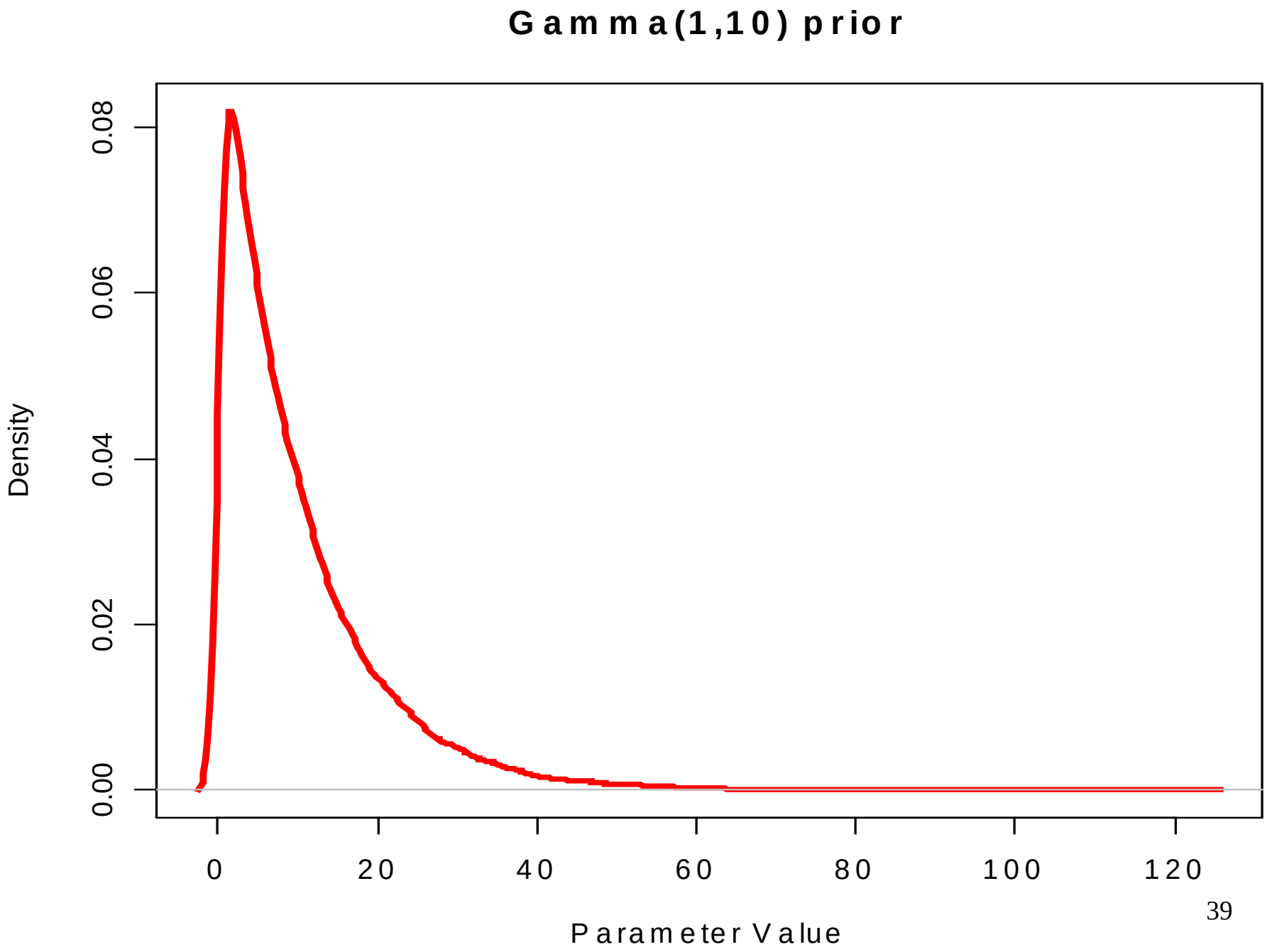
10 school Random effects

$$\theta_1,\ldots,\theta_{10}|G_\theta \overset{iid}{\sim} G_\theta$$

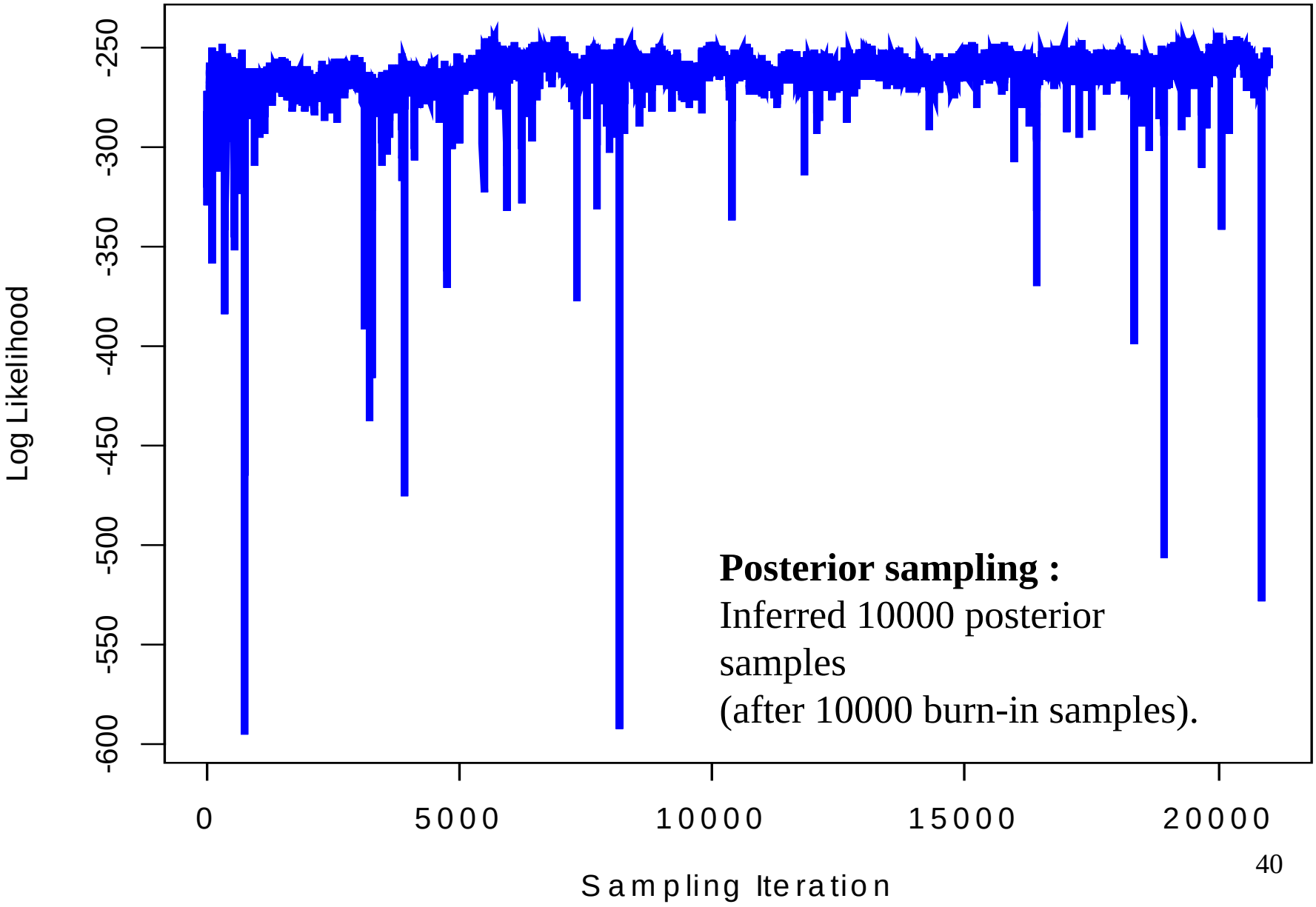
$$G_\theta|\alpha_\theta,\tau \sim \text{DirichletProcess}(\alpha_\theta, \text{Normal}(0,\tau))$$

$$\alpha_\theta|a_\theta,b_\theta \sim \text{Gamma}(1,10)$$

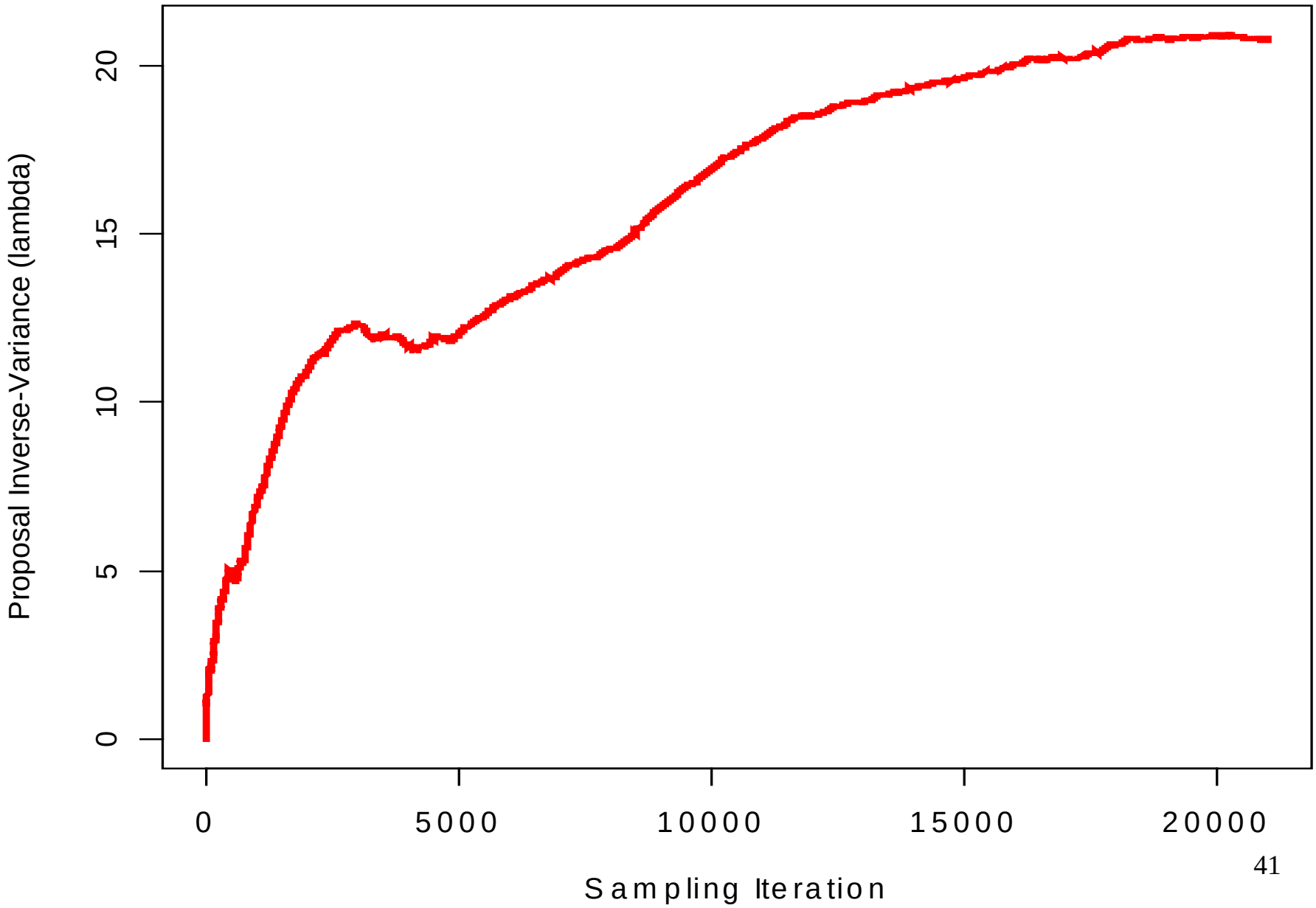
$$\tau|b_\tau \sim \text{Uniform}(0,10^5)$$


Gamma(1,10) prior
Density
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
100
120
Parameter Value

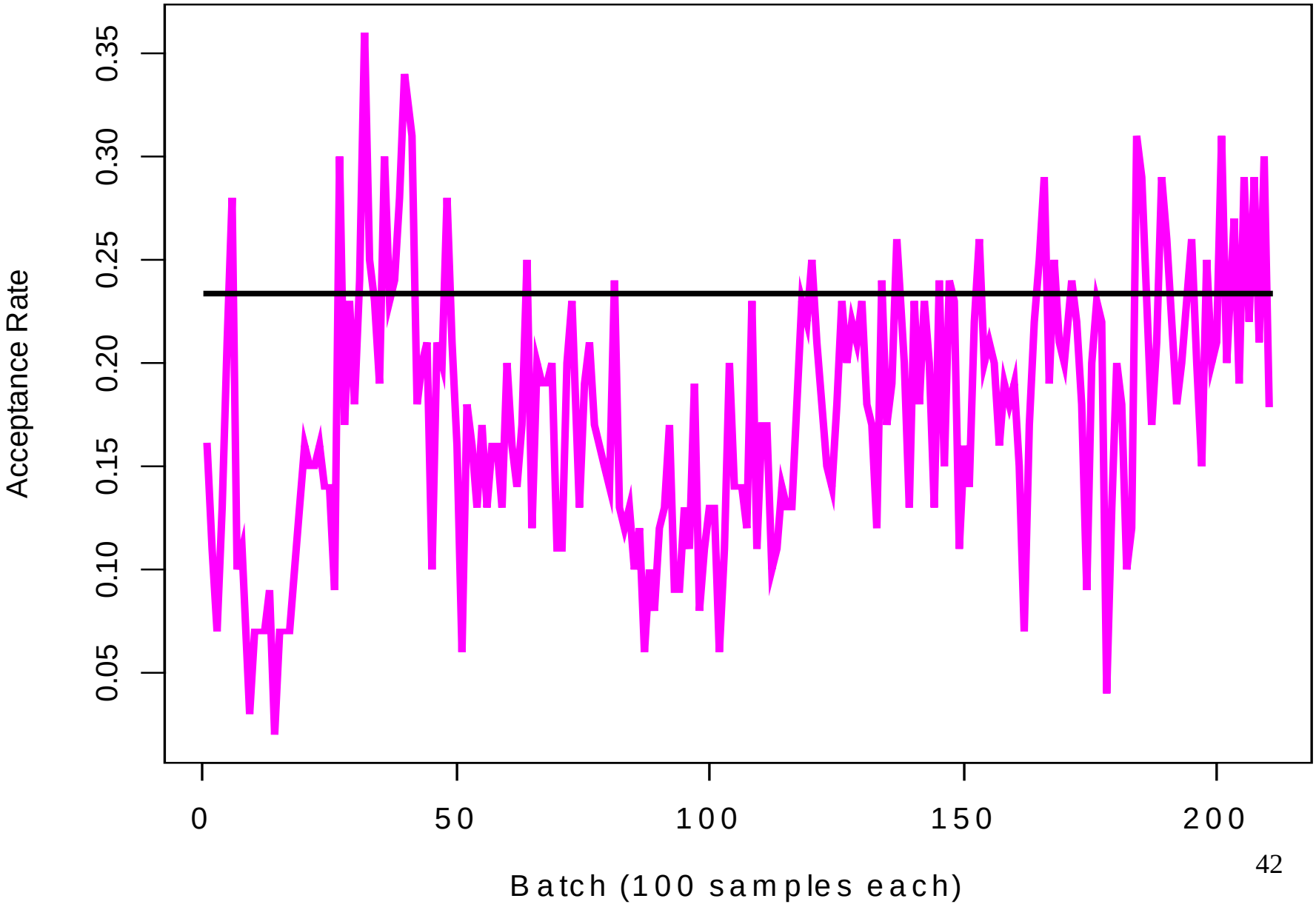
Log Likelihood

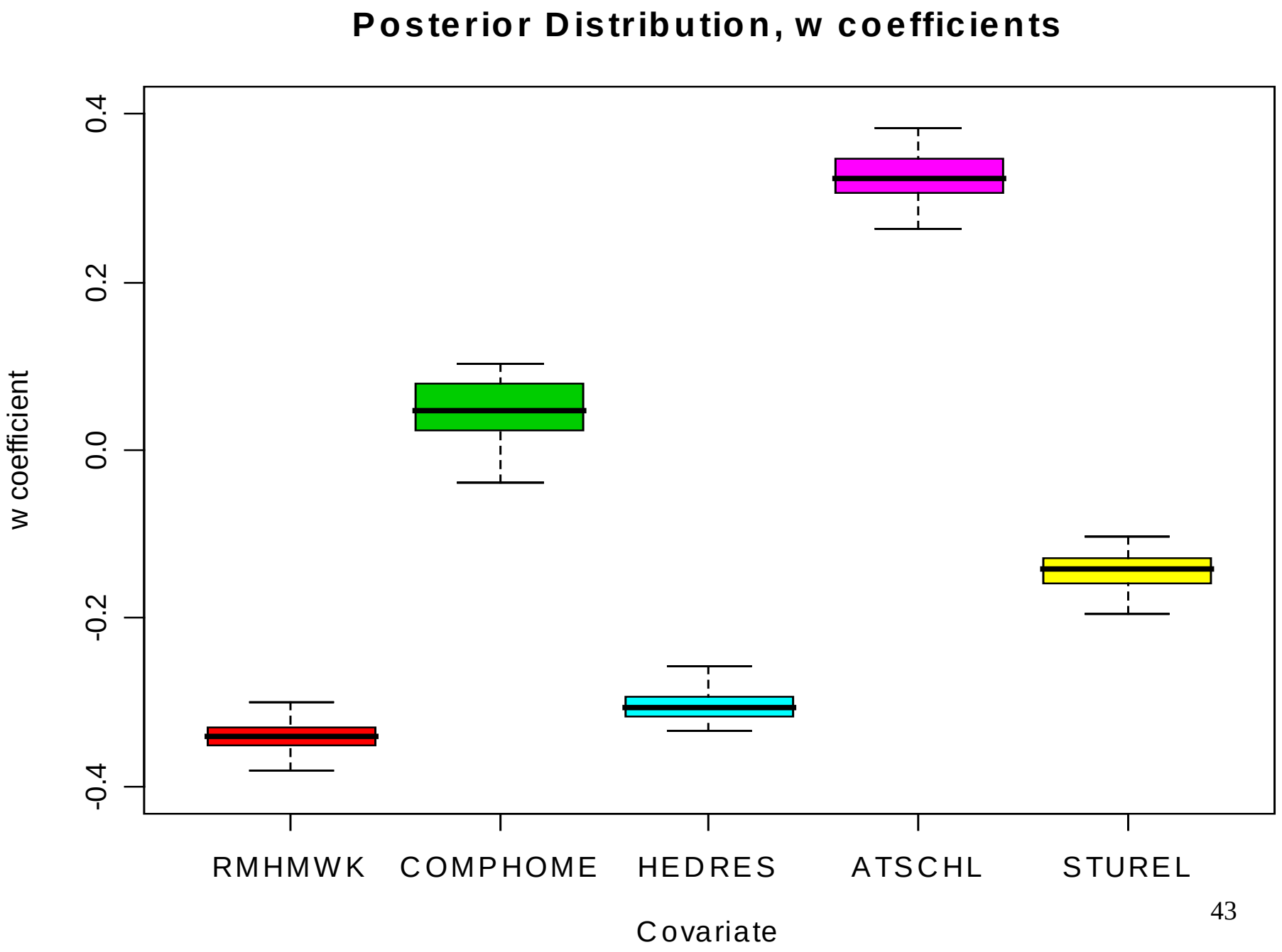


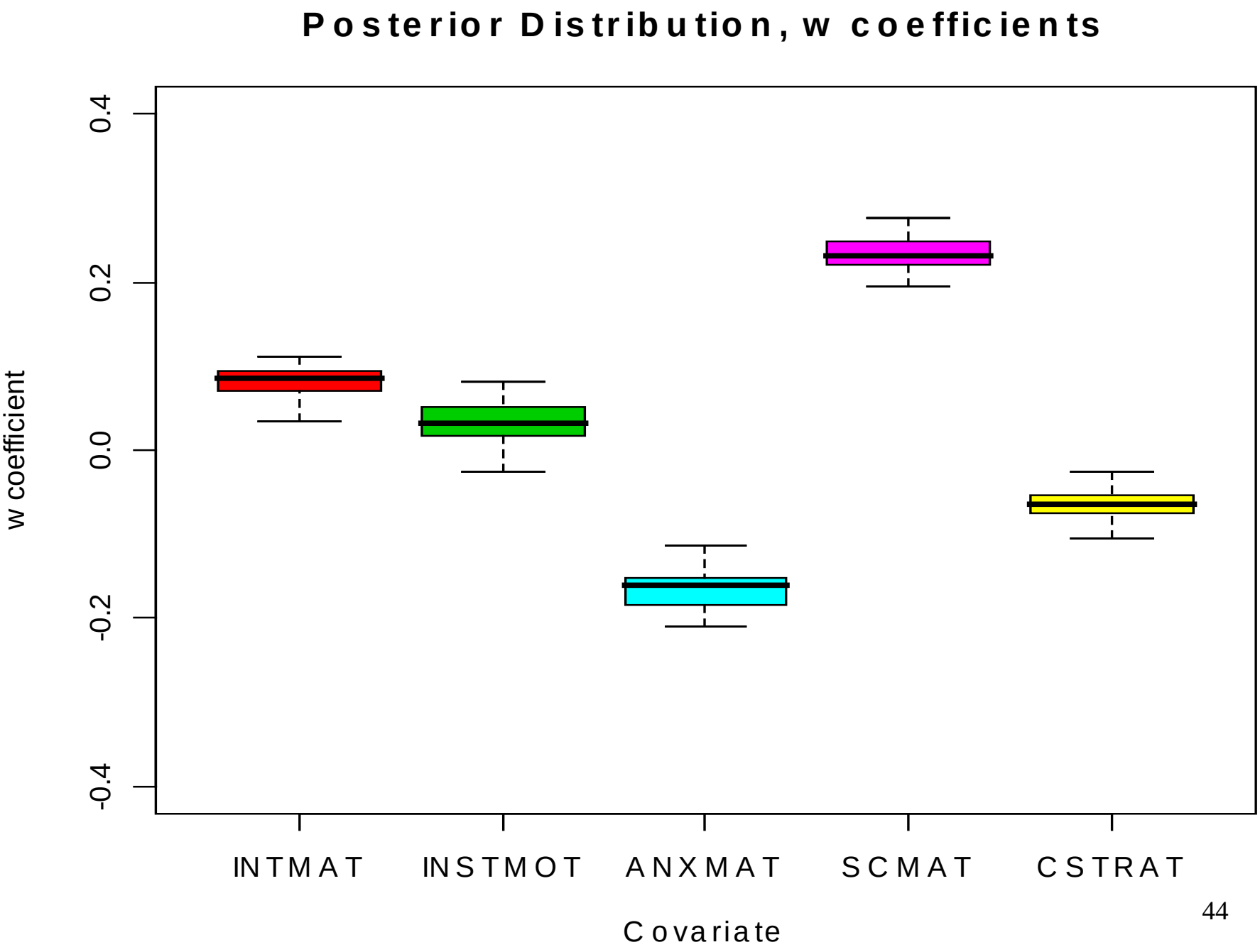
ARWM Convergence, ω

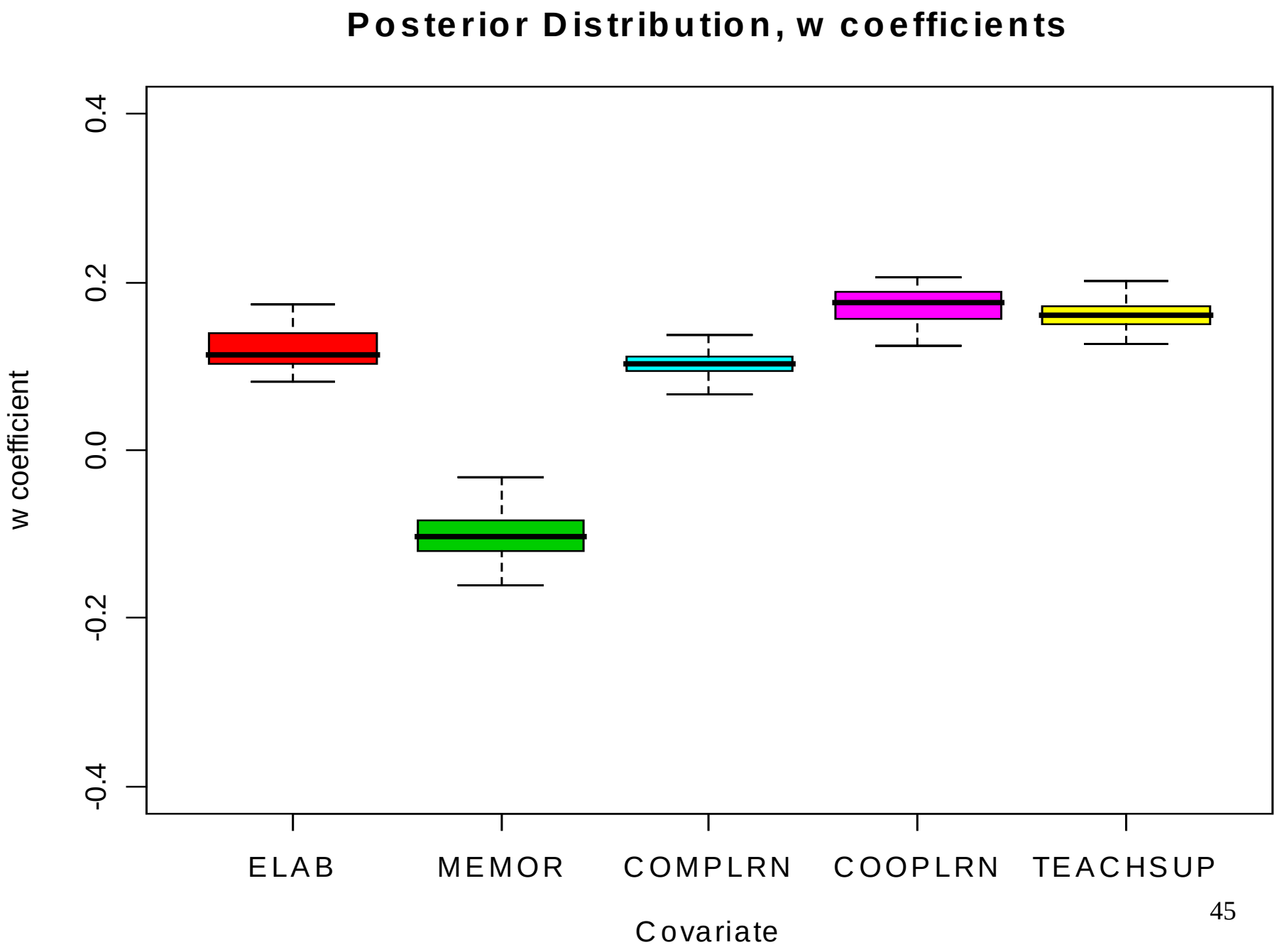


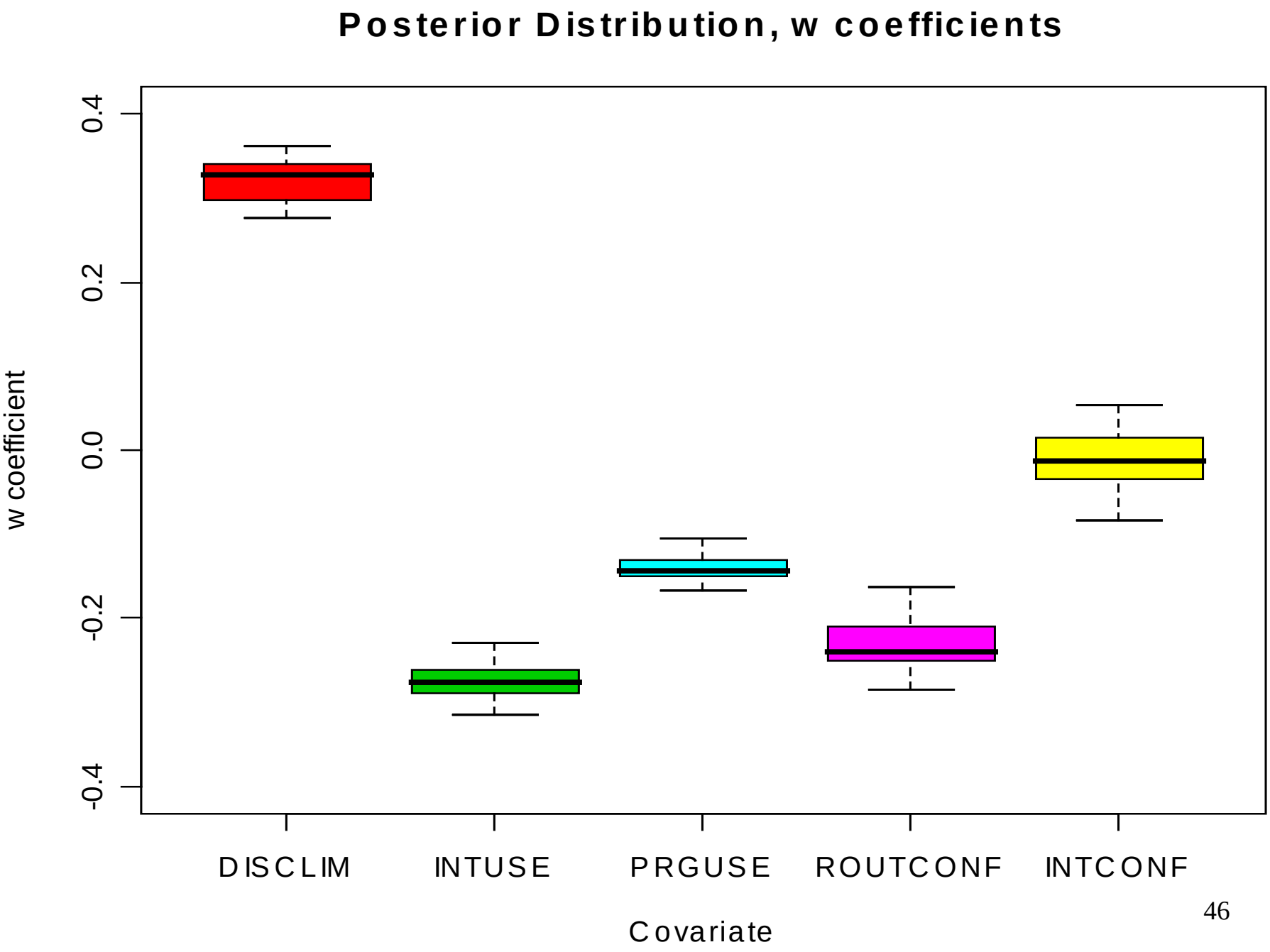
ARWM Convergence, ω

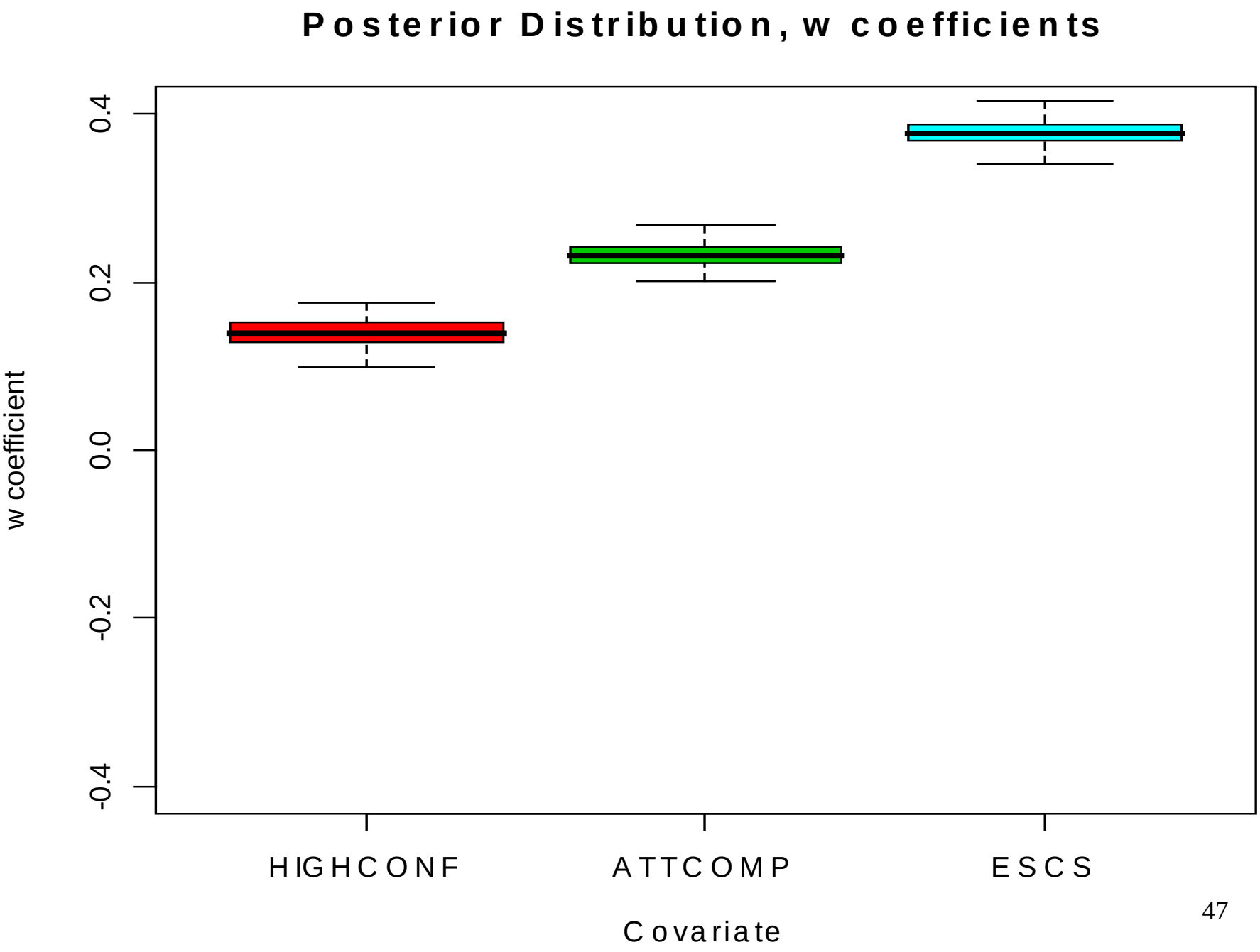



Posterior Distribution, w coefficients
w coefficient
0.4
0.2
0.0
-0.2
-0.4
RMHMWK
COMPHOME
HEDRES
ATSCHL
STUREL
Covariate

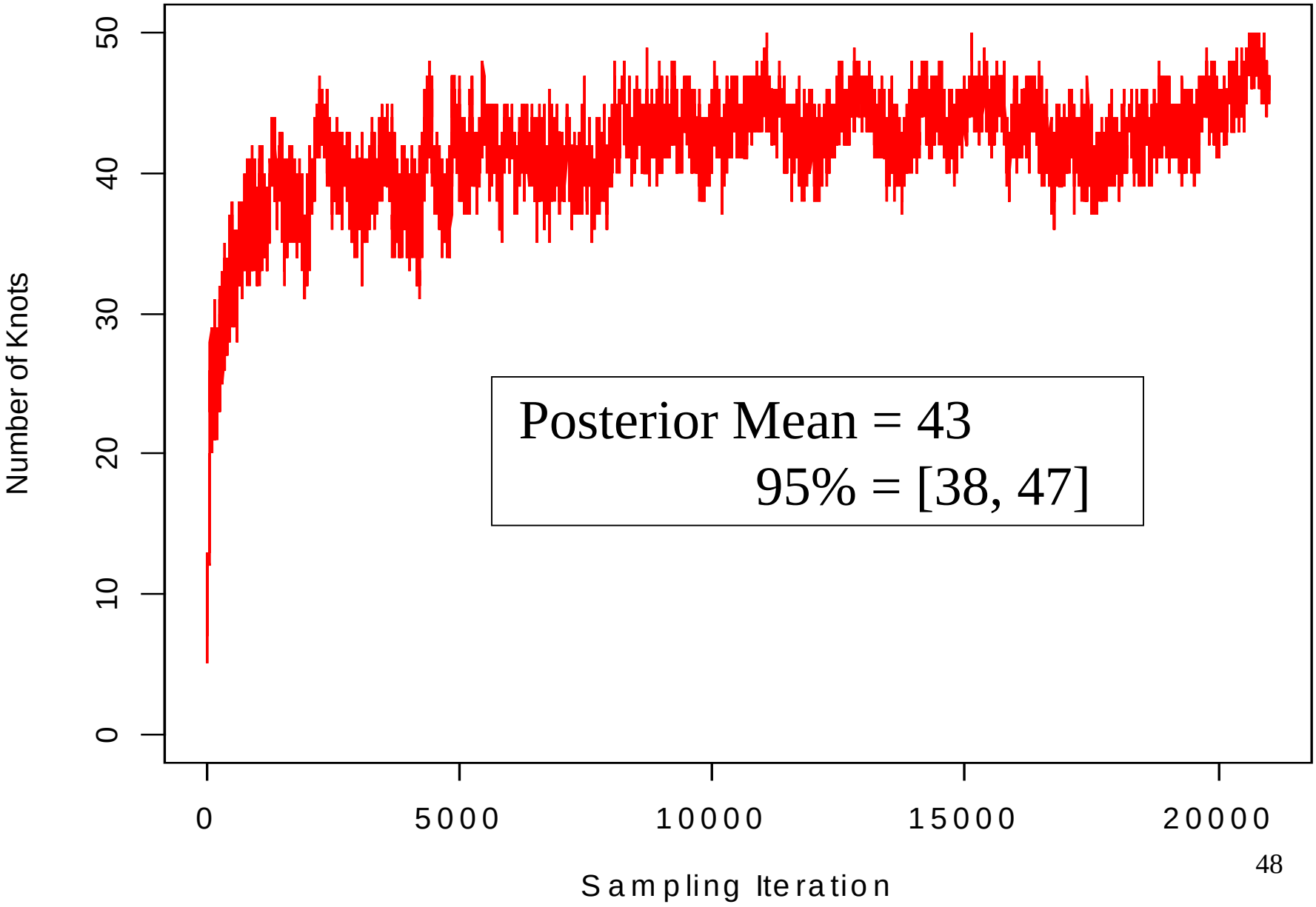

Posterior Distribution, w coefficients
w coefficient
0.4
0.2
0.0
-0.2
-0.4
INTMAT
INSTMOT
ANXMAT
SCMAT
CSTRAT
Covariate


Posterior Distribution, w coefficients
w coefficient
0.4
0.2
0.0
-0.2
-0.4
ELAB
MEMOR
COMPLRN
COOPLRN
TEACHSUP
Covariate

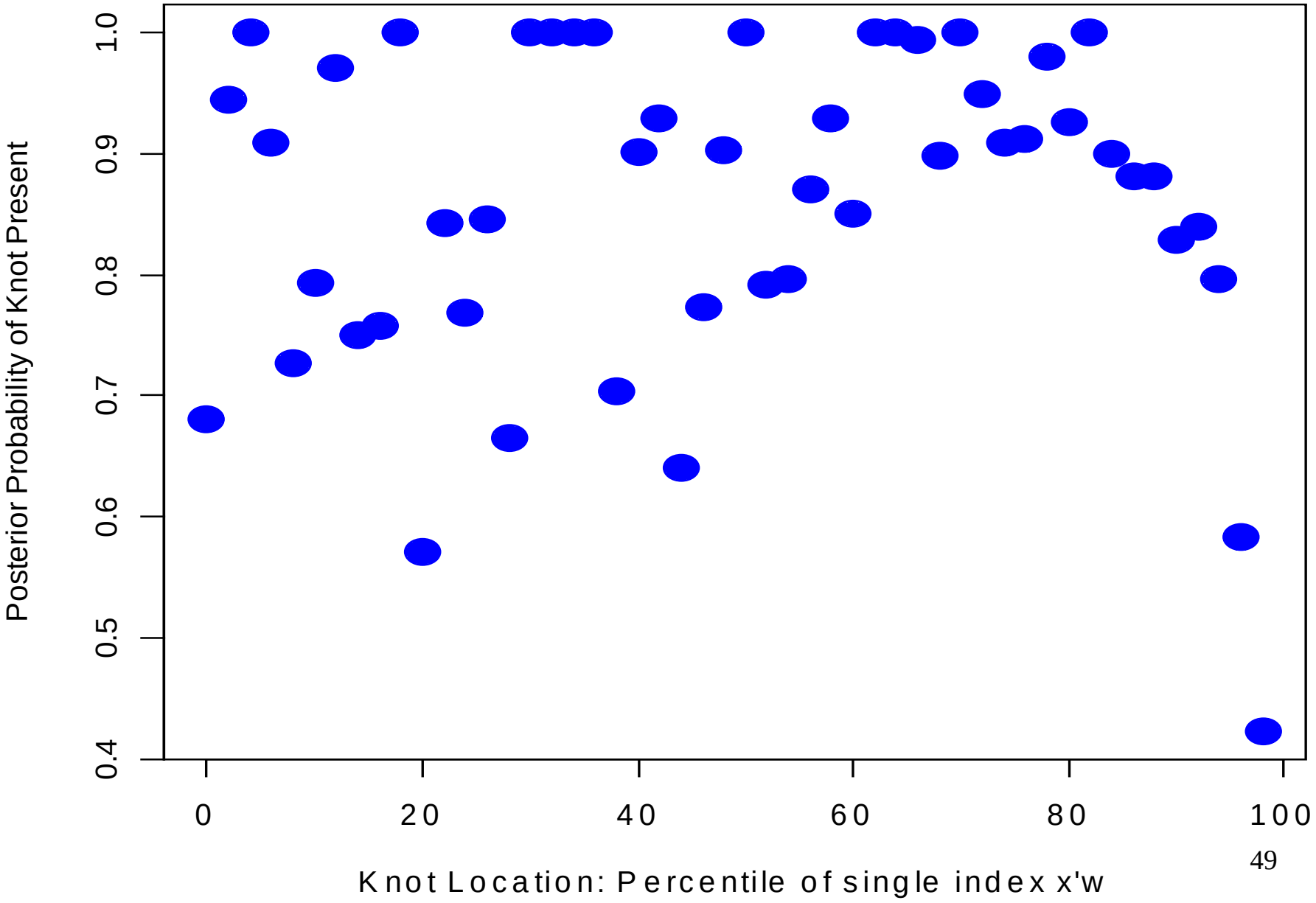

Posterior Distribution, w coefficients
w coefficient
0.4
0.2
0.0
-0.2
-0.4
DISCLIM
INTUSE
PRGUSE
ROUTCONF
INTCONF
Covariate


Posterior Distribution, w coefficients
w coefficient
0.4
0.2
0.0
-0.2
-0.4
HIGHCONF
ATTCOMP
ESCS
Covariate

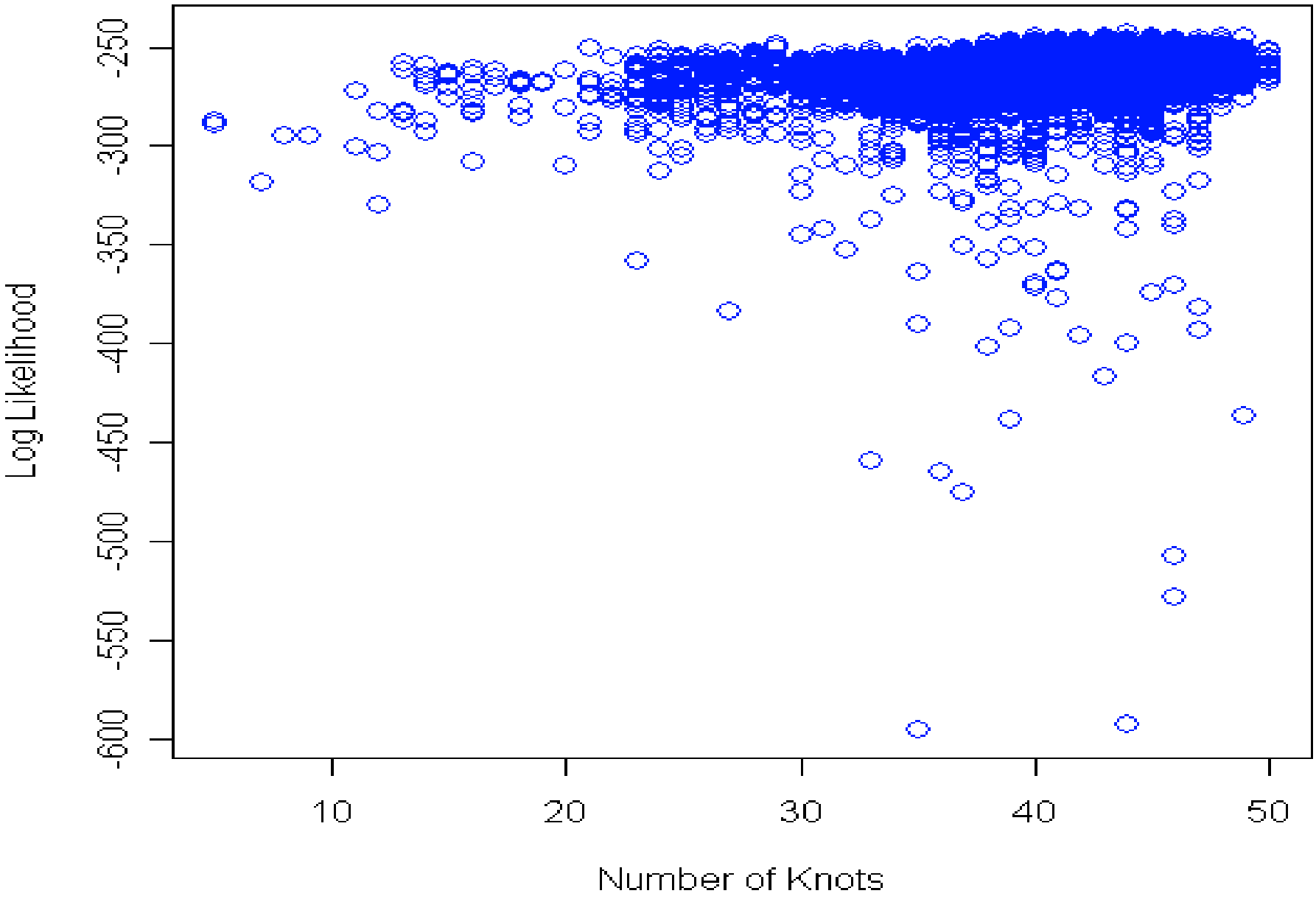
Number of Knots

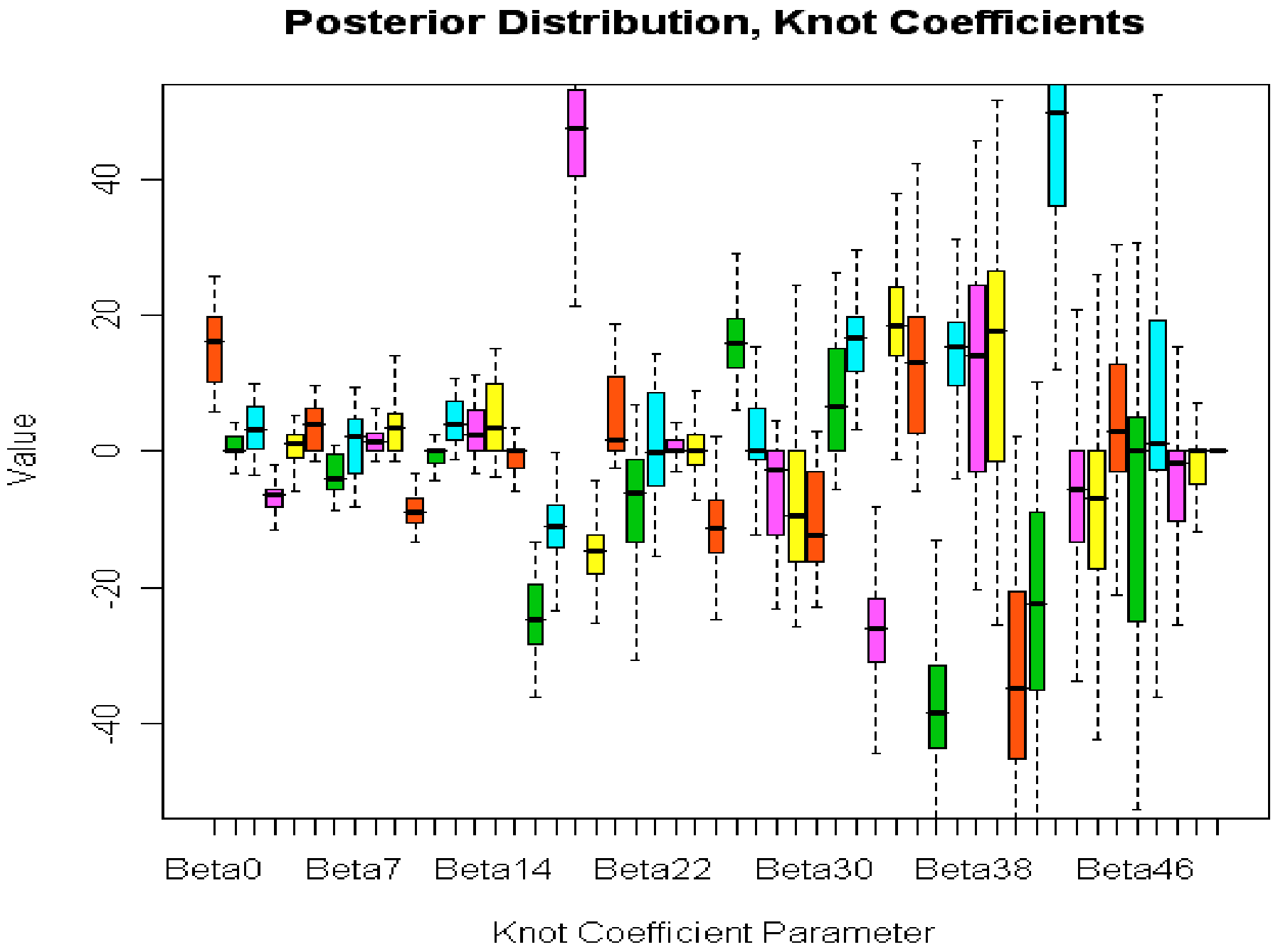


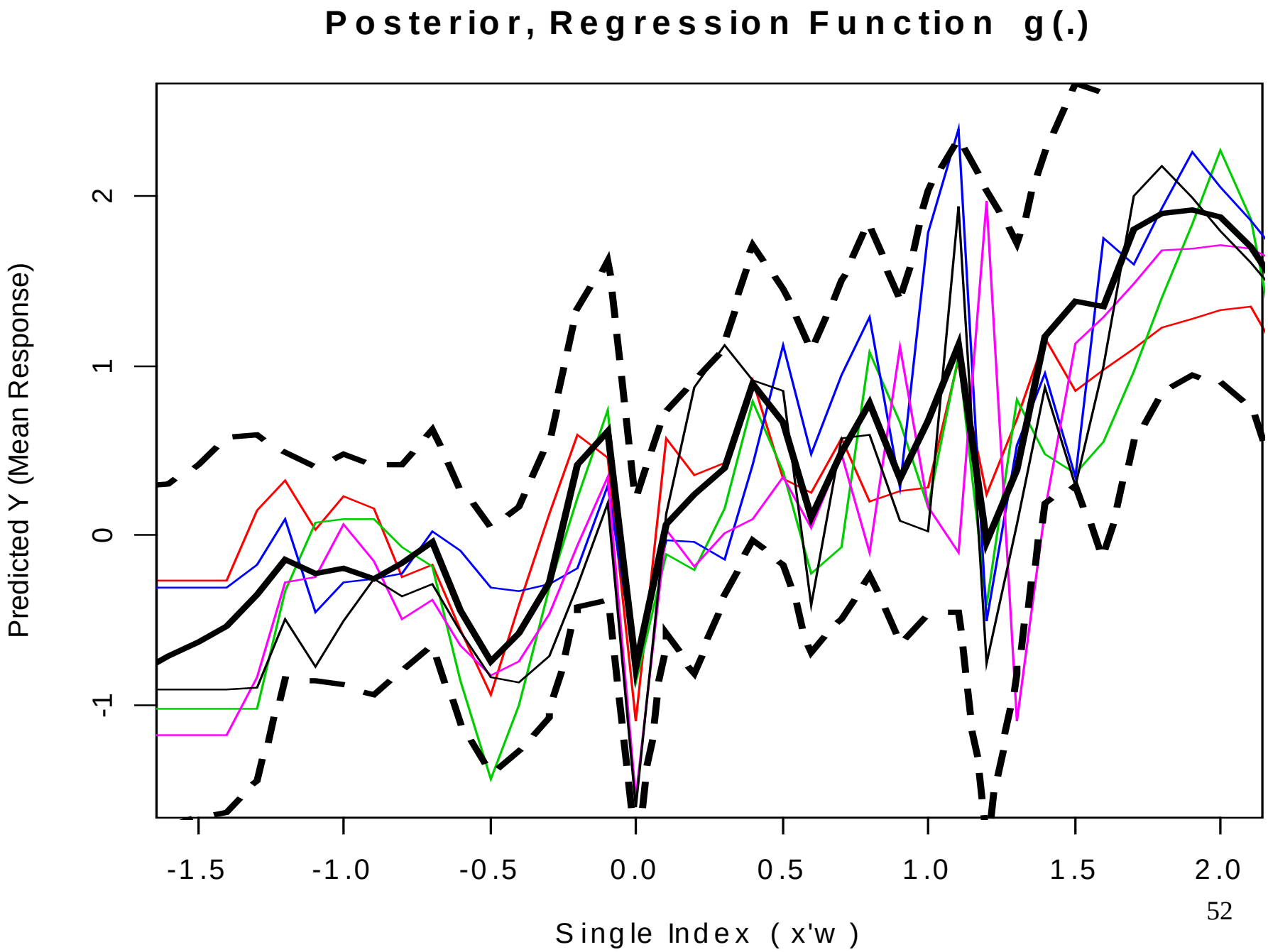
Knot Locations

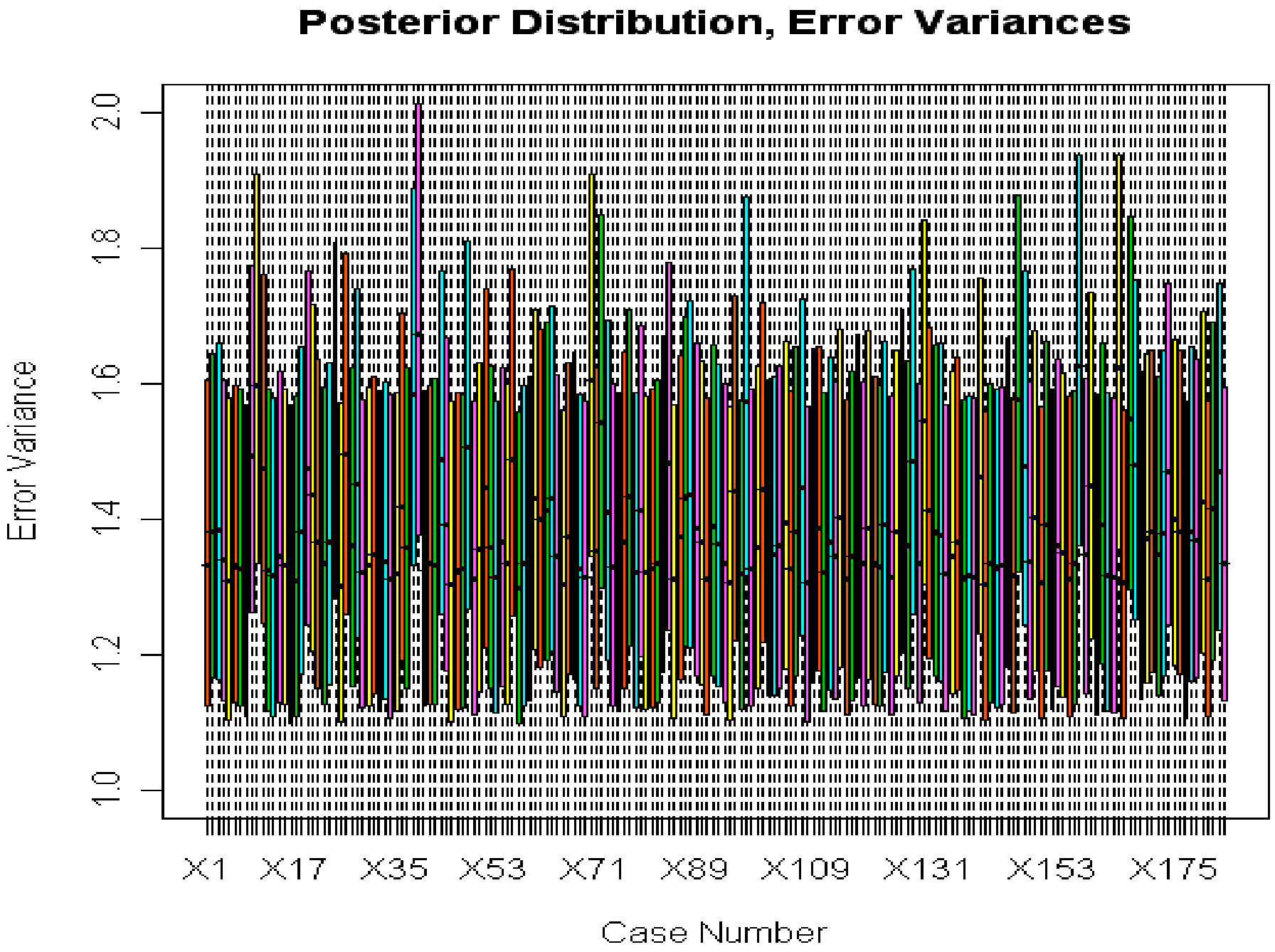


Number of Knots and Log Likelihood




Posterior Distribution, Knot Coefficients
Value
40
20
0
-20
-40
Beta0
Beta7
Beta14
Beta22
Beta30
Beta38
Beta46
Knot Coefficient Parameter


Posterior, Regression Function g(.)
Predicted Y (Mean Response)
2
1
0
-1
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
Single Index (x'w)


Posterior Distribution, Error Variances
Error Variance
1.0
1.2
1.4
1.6
1.8
2.0
X1
X17
X35
X53
X71
X89
X109
X131
X153
X175
Case Number

Posterior: DP Parameters

	Mean	2.5%	97.5%
Regression			
Error Variances:			
Nclus	182.92	180.00	184.00
Shape	16.48	8.54	27.31
Scale	22.21	11.73	36.28
alpha	1846.80	1585.10	2125.00
Random Effects:			
Nclus	1.07	1.00	2.00
Var	10752.23	11.28	79221.48
alpha	11.88	0.29	43.89

Conclusions

- The new model:
  - Easily handles a high-dimensional covariate **x**.
    - Number of splines is K_{max};
      Compare to $p \cdot K_{max}$ splines needed for additive models.
  - Provides an interpretable, and very flexible regression function for the single index $\mathbf{x}^T\boldsymbol{\omega}$,
    which describes deviations from monotonicity,
    and automatically describes interactions (to a degree).
  - Accounts for change in the regression error variance,
    over the mean response (heterogeneity of variance).
  - Describes random effects (the extra sources of variance not explained in the covariate vector **x**)
  - Discrete (binary, ordinal, or counts) outcomes are addressed with a latent variable approach.

Conclusions

- Perhaps consider nonparametric priors for the random effects and error variances, other than the MDP.
  For example, the MPT prior, the Bernstein polynomial prior.

- Posterior sampling is slow.
  - Gibbs sampler for ω ? (without component-wise sampling)
  - It would be nice to find a way to sample (β, γ) at once, rather than sampling β_0, (β_k, γ_k), $k=1,\ldots,K_{max}$, individually. In applications, this would place less restriction on the number of candidate knots, K_{max}.

Thank You!