Sampling methods for generalized linear models

Martyn Plummer

International Agency for
Research on Cancer

AppliBUGS, 26 June 2012

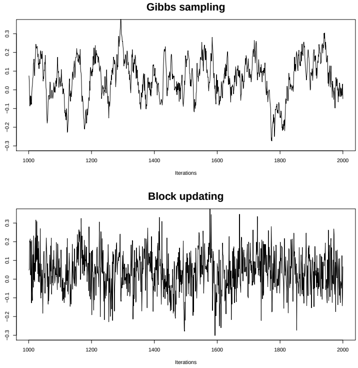
JAGS: Just Another Gibbs Sampler

- A clone of BUGS
- Written in C++
  - Cross-platform (Linux, Windows, Mac OS X, ...)
- Object-based R interface (rjags)
  - Other interfaces: R2jags, runjags

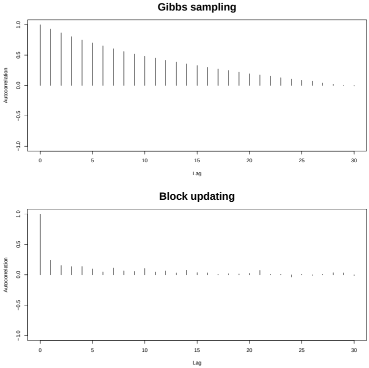
JAGS aims (more or less) for compatibility with BUGS. In particular, models are described in the same way using “the BUGS language”

Good and bad "mixing"

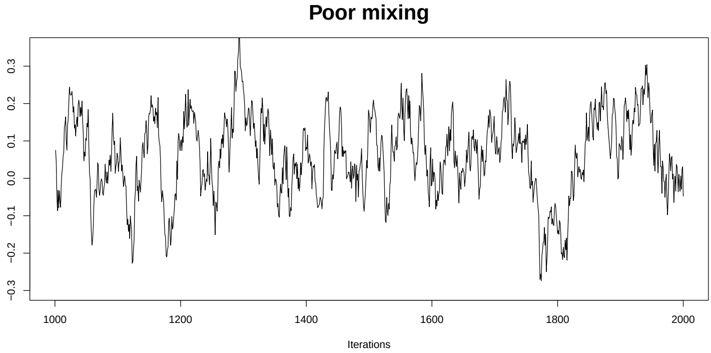
Trace plots show the evolution of the sampled values by the number of iterations.



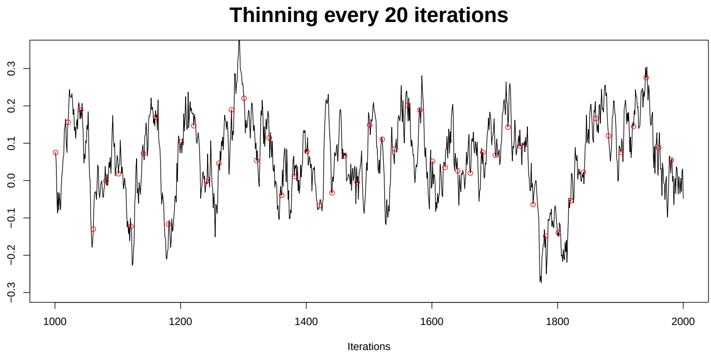
Autocorrelation plots show the extent to which current value depends on previous ones.



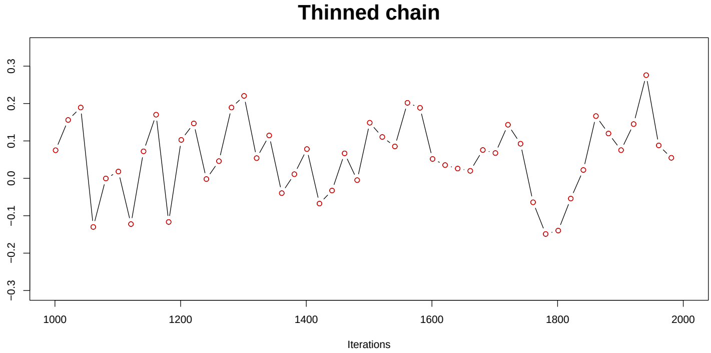
“Solving” autocorrelation by thinning



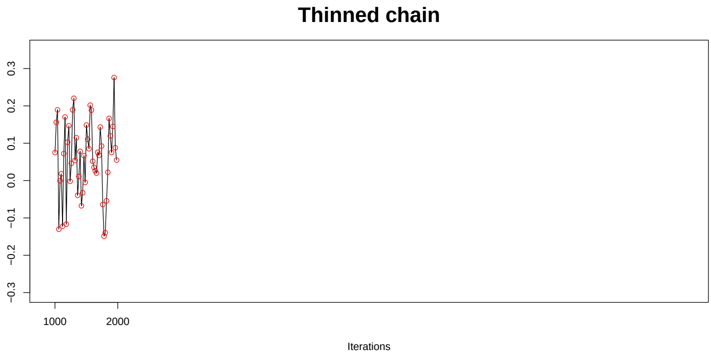
“Solving” autocorrelation by thinning



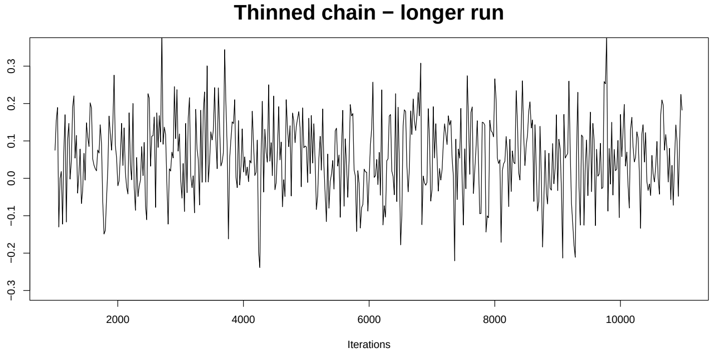
"Solving" autocorrelation by thinning



"Solving" autocorrelation by thinning



“Solving” autocorrelation by thinning



The search for better sampling algorithms

- We reject idea of “universal” sampling algorithm.
- Focus on methods that can sample particular model classes efficiently.
- Block updating of highly correlated parameters key.
- Parameters in the linear predictor of a generalized linear model (GLMs) are suitable candidates for block updating

Correlation in GLMs

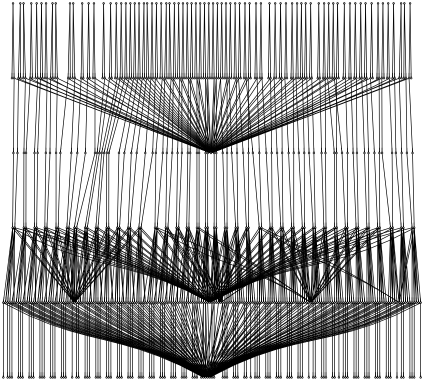
When sampling individual nodes, centering of covariates can reduce cross-correlations between parameters, and so improve mixing.

```
y[i] ~ dnorm(mu[i], tau)
mu[i] <- alpha + beta * (x[i] - mean(x))
```

Tricks like this are not necessary if we update α, β together.

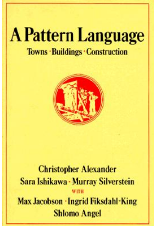
An abstract view of the KIDNEY model

This is what a right-censored survival analysis problem looks like as a graphical model.



JAGS needs to inspect the graph and identify blocks of stochastic nodes that can be updated together.

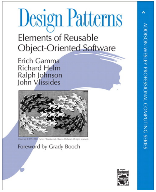
Design Patterns



A Pattern Language

Towns · Buildings · Construction

Christopher Alexander
Sara Ishikawa · Murray Silverstein
WITH
Max Jacobson · Ingrid Fiksdahl-King
Shlomo Angel



Design Patterns

Elements of Reusable Object-Oriented Software

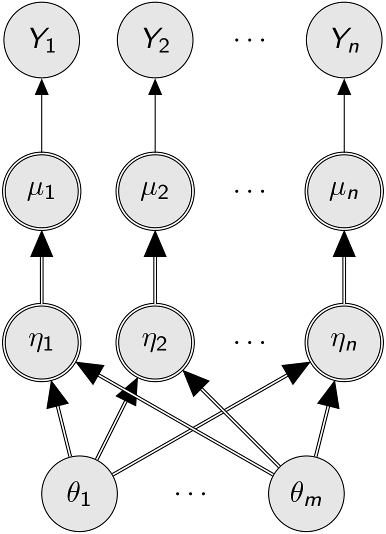
Erich Gamma
Richard Helm
Ralph Johnson
John Vlissides

Foreword by Grady Booch

ADDISON-WESLEY PROFESSIONAL COMPUTING SERIES

- Design patterns are reusable solutions to commonly recurring design problems
- Originally developed in architecture, the patterns concept has been translated to software development.
- This may be a useful way of thinking about efficient sampling of graphical models, which often have a rich structure.
- First we need to look for recurring design motifs

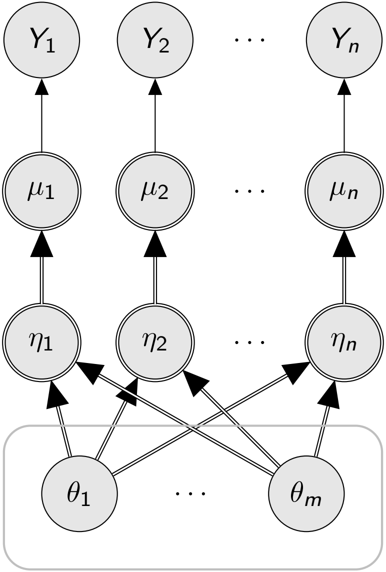
GLM as a design motif



A GLM is a sub-graph with the following elements

- **parameters** θ with prior normal distribution
- **linear predictors** η are linear functions of the parameters (intermediate nodes omitted).
- **link functions** transform linear predictor η to mean value μ
- **Outcome variables Y** depend on parameters θ via the mean μ

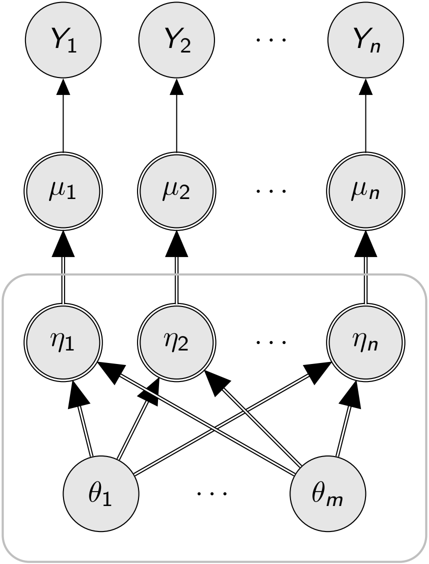
GLM as a design motif



A GLM is a sub-graph with the following elements

- **parameters** θ with prior normal distribution
- **linear predictors** η are linear functions of the parameters (intermediate nodes omitted).
- **link functions** transform linear predictor η to mean value μ
- **Outcome variables Y** depend on parameters θ via the mean μ

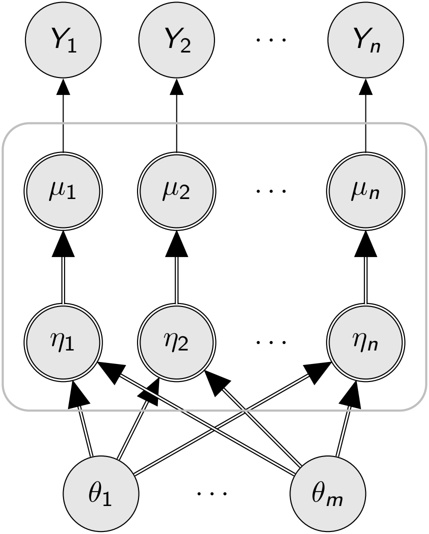
GLM as a design motif



A GLM is a sub-graph with the following elements

- **parameters** θ with prior normal distribution
- **linear predictors** η are linear functions of the parameters (intermediate nodes omitted).
- **link functions** transform linear predictor η to mean value μ
- **Outcome variables Y** depend on parameters θ via the mean μ

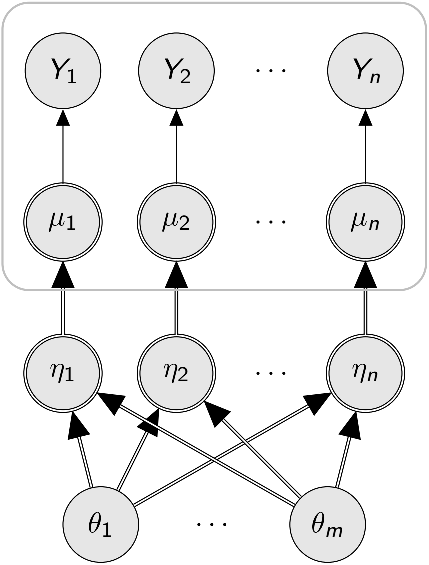
GLM as a design motif



A GLM is a sub-graph with the following elements

- **parameters** θ with prior normal distribution
- **linear predictors** η are linear functions of the parameters (intermediate nodes omitted).
- **link functions** transform linear predictor η to mean value μ
- **Outcome variables Y** depend on parameters θ via the mean μ

GLM as a design motif



A GLM is a sub-graph with the following elements

- **parameters** $\boldsymbol{\theta}$ with prior normal distribution
- **linear predictors** $\boldsymbol{\eta}$ are linear functions of the parameters (intermediate nodes omitted).
- **link functions** transform linear predictor $\boldsymbol{\eta}$ to mean value $\boldsymbol{\mu}$
- **Outcome variables Y** depend on parameters $\boldsymbol{\theta}$ via the mean $\boldsymbol{\mu}$

Mixed models

- Mixed models are usually described using notation due to Laird and Ware (1982)

$$\begin{aligned} \boldsymbol{\eta} &= \mathbf{X}\beta + \mathbf{Z}\mathbf{b} \\ \mathbf{b} &\sim N(0, \Psi) \end{aligned}$$

- In software the distinction between fixed effects (β) and random effects ($\mathbf{b}$) is usually implicit in the model syntax.
- If JAGS finds a GLM motif in the model, can it partition the nodes $\theta_1 \ldots \theta_m$ into fixed and random effects?

Distinguishing between fixed and random effects

Classical	Bayesian
Maximum likelihood	MCMC
$\boldsymbol{\beta}$ unknown parameter $\mathbf{b}$ random variable	$\boldsymbol{\beta}, \mathbf{b}$ both random variables
Marginal likelihood $p(\mathbf{Y} \mid \boldsymbol{\beta}, \Psi)$	Local likelihood $p(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{b})$
Need to estimate Ψ	Ψ fixed when updating $\boldsymbol{\beta}, \mathbf{b}$
$\mathbf{X}$ dense $\mathbf{Z}$ sparse	$(\mathbf{Z}, \mathbf{X})$ form a single design matrix with both dense and sparse components

There are no mixed models, only models with sparse design matrices.

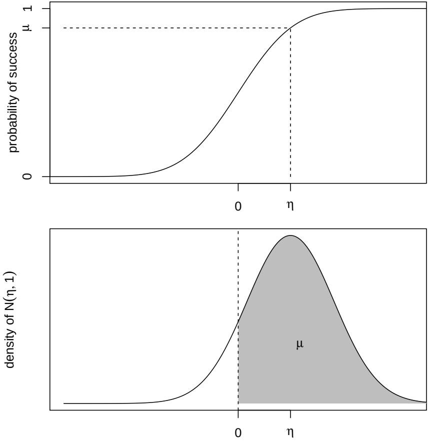
Sparse matrix algebra in JAGS

- JAGS uses the sparse matrix libraries CSparse and CHOLMOD developed by Timothy A Davis.
- The same sampling engine handles both fixed-effect (dense) and mixed (sparse) linear models.
- The set of parameters of the linear model $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{b})$ has multivariate normal posterior distribution, and so can be efficiently sampled.
- Programming is relatively easy due to the fundamental connection between graph theory and sparse matrix algebra

Extending the linear model sampler

- It seems natural to preserve the benefits of the linear sampler by extending its scope.
- This also has the benefit of code reuse as a single sampling "engine" can address multiple models
- Some GLMs can be reduced to linear form by data augmentation (adding additional nodes to the graph)
- Methods have been proposed for Poisson regression and logistic regression, which are coincidentally the most common models in epidemiology

Albert and Chib (1993) approach to binary probit models



$$\mu \equiv P(Y = 1 \mid \eta) = \Phi(\eta)$$

Albert and Chib (1993) introduce a latent variable

$$Z \sim N(\eta, 1)$$

and make the outcome Y a deterministic function of Z

$$Y = I\{Z \geq 0\}$$

Graphical representation of Albert-Chib (1993)

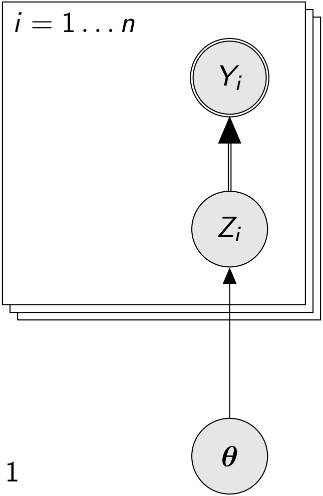
Nodes $Z_1 \ldots Z_n$ are intermediate between the parameters $\boldsymbol{\theta}$ and the outcomes $Y_1 \ldots Y_n$ so that

$$\boldsymbol{\theta} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$$

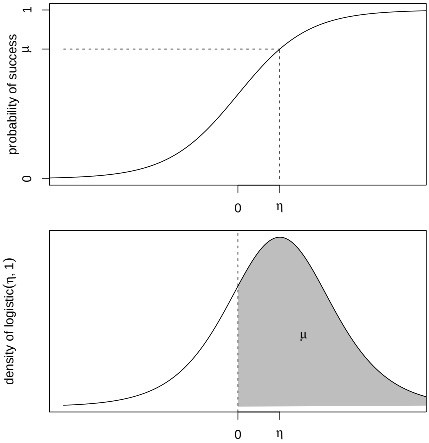
When sampling

- $\boldsymbol{\theta}$ is updated given $\mathbf{Z}$ ignoring $\mathbf{Y}$, thereby reducing the problem to a linear model.
- $Z_1 \ldots Z_n$ are updated individually from the truncated normal

$$\begin{array}{lcll} Z_i & \sim & N(\eta_i, 1) I\{Z_i \geq 0\} & \text{if} \quad Y_i = 1 \\ Z_i & \sim & N(\eta_i, 1) I\{Z_i < 0\} & \text{if} \quad Y_i = 0 \end{array}$$

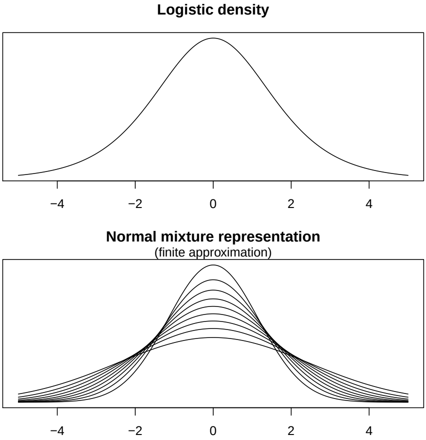


Holmes and Held (2006) approach to binary logit models



Logistic regression models with a binary outcome also have a latent variable representation, where the latent Z has a logistic distribution.

Mixture representation of logistic distribution



The logistic distribution is a scale mixture of normals, where the scale parameter has a Kolmogorov-Smirnov distribution

$$\begin{aligned} Z \mid \psi &\sim N(0,(2\psi)^2) \\ \psi &\sim KS \end{aligned}$$

Andrews and Mallows (1974).

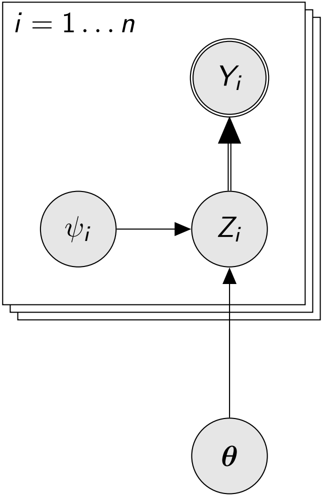
Graphical representation of Holmes and Held logistic model

The logistic-mixture model adds further auxiliary nodes $\psi_1 \ldots \psi_n$, such that

$$\psi_i \perp\!\!\!\perp \mathbf{Y} \mid Z_i$$

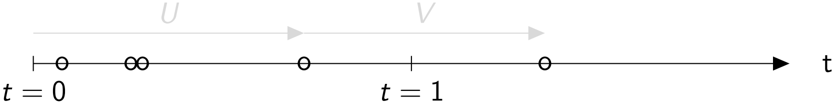
Hence ψ_i is updated from $[\psi_i \mid Z_i, \boldsymbol{\theta}]$

- The Kolmogorov-Smirnov density has no closed-form expression.
- But Devroye (1986) provides two alternating series expansions that can be used for rejection sampling of ψ_i.
- Z_i is updated from the truncated logistic distribution $Z_i \mid Y_i, \boldsymbol{\theta}$



Frühwirth-Schnatter et al (2009) Poisson regression

In a Poisson regression model, $Y \sim \text{Po}(\lambda)$ where $\lambda = \exp(\eta)$. Frühwirth-Schnatter et al (2009) model Y in terms of an underlying Poisson process of rate λ, as the number of events before time 1.

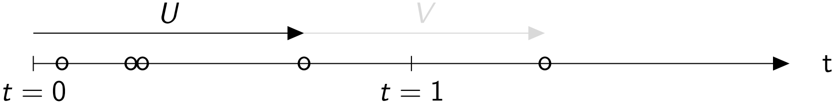


- Sufficient statistics for η are
  - U, the last arrival before $t = 1$
  - V, the next inter-arrival time.
- On a log scale, the model is linear

$$
\begin{aligned}
-\log(U) &= \eta - \log(\epsilon) \quad \text{where} \quad \epsilon \sim \Gamma(y-1, 1) \\
-\log(V) &= \eta - log(\xi) \quad \text{where} \quad \xi \sim \exp(1)
\end{aligned}
$$

Frühwirth-Schnatter et al (2009) Poisson regression

In a Poisson regression model, $Y \sim \text{Po}(\lambda)$ where $\lambda = \exp(\eta)$. Frühwirth-Schnatter et al (2009) model Y in terms of an underlying Poisson process of rate λ, as the number of events before time 1.

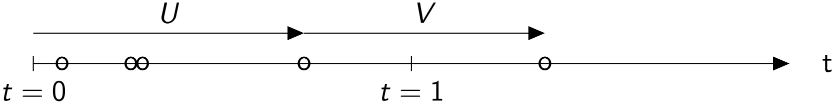


- Sufficient statistics for η are
  - U, the last arrival before $t = 1$
  - V, the next inter-arrival time.
- On a log scale, the model is linear

$$
\begin{array}{rclcccl}
-\log(U) & = & \eta - \log(\epsilon) & \text{where} & \epsilon & \sim & \Gamma(y-1, 1) \\
-\log(V) & = & \eta - log(\xi) & \text{where} & \xi & \sim & \exp(1)
\end{array}
$$

Frühwirth-Schnatter et al (2009) Poisson regression

In a Poisson regression model, $Y \sim \text{Po}(\lambda)$ where $\lambda = \exp(\eta)$. Frühwirth-Schnatter et al (2009) model Y in terms of an underlying Poisson process of rate λ, as the number of events before time 1.

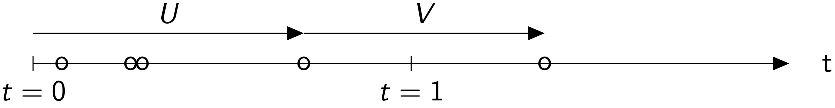


- Sufficient statistics for η are
  - U, the last arrival before $t = 1$
  - V, the next inter-arrival time.
- On a log scale, the model is linear

$$\begin{aligned} -\log(U) &= \eta - \log(\epsilon) \quad \text{where} \quad \epsilon \sim \Gamma(y-1,1) \\ -\log(V) &= \eta - log(\xi) \quad \text{where} \quad \xi \sim \exp(1) \end{aligned}$$

Frühwirth-Schnatter et al (2009) Poisson regression

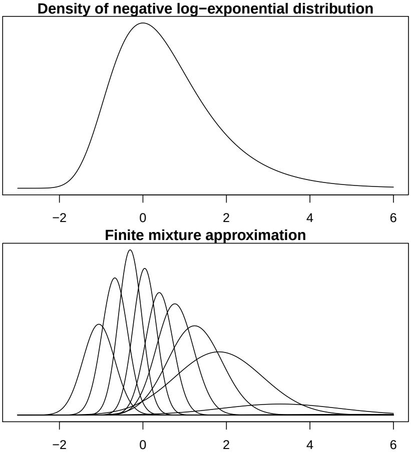
In a Poisson regression model, $Y \sim \mathrm{Po}(\lambda)$ where $\lambda = \exp(\eta)$. Frühwirth-Schnatter et al (2009) model Y in terms of an underlying Poisson process of rate λ, as the number of events before time 1.



- Sufficient statistics for η are
  - U, the last arrival before $t = 1$
  - V, the next inter-arrival time.
- On a log scale, the model is linear

$$\begin{array}{rclcccl} -\log(U) & = & \eta - \log(\epsilon) & \text{where} & \epsilon & \sim & \Gamma(y-1,1) \\ -\log(V) & = & \eta - log(\xi) & \text{where} & \xi & \sim & \exp(1) \end{array}$$

Mixture representation of negative log-gamma


Density of negative log−exponential distribution

Finite mixture approximation

The GMRFlib library contains code for approximating the negative log-gamma distribution as a finite mixture of normals. This code is borrowed by JAGS, under the GPL.

Using this mixture approximation reduced the model to a normal linear model.

Frühwirth-Schnatter et al (2009) logistic regression

In a logistic regression model $Y \sim \text{Bin}(\pi, n)$ where $\pi = \lambda/(1 + \lambda)$ and $\lambda = \exp(\eta)$. Frühwirth-Schnatter et al (2009) model Y as the sum of n Bernoulli trials with probability π.

$$Y = \sum_{i=1}^{n} Y_j$$

In each trial, Y_j is a function of two latent variables U_j, V_j

$$\begin{aligned} U_j &\sim \exp(\lambda) \\ V_j &\sim \exp(1) \\ Y_j &= I\{U_j < V_j\} \end{aligned}$$

The sufficient statistic for η is $U = \sum_j U_j$ and

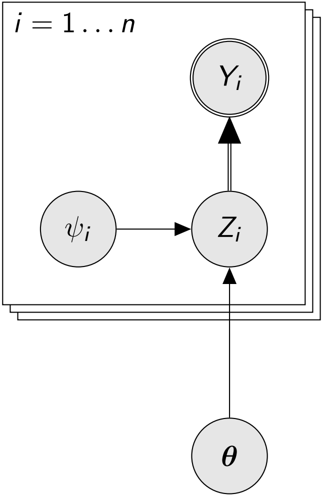
$$-\log(U) = \eta - \log(\epsilon) \text{ where } \epsilon \sim \Gamma(n, 1)$$

The same mixture representation for $-\log(\epsilon)$ reduces the model to a normal linear model

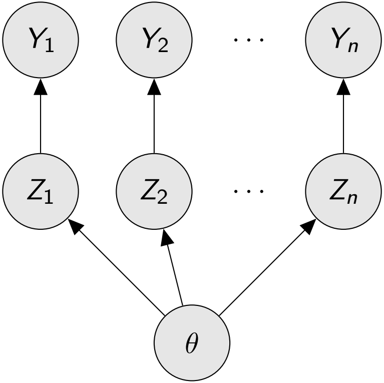
Graphical representation

These Poisson and logistic regression models have the same graphical representation as the Holmes and Held logistic model.

- ψ_i is an integer value that determines which mixture component is in use.
- $Z_i = (U_i, V_i)$ can be efficiently sampled given $\boldsymbol{\theta}$, Y_i and marginalizing over ψ_i.
- Sampling of ψ_i given $Z_i, \boldsymbol{\theta}$ is trivial.



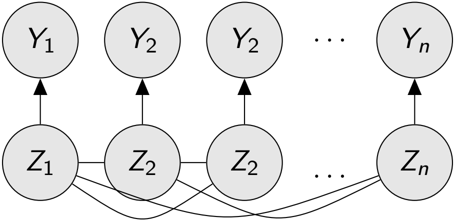
Patterns again



- All of these graphical contain the same distinctive "fantail" motif.
- Whenever this motif is seen in a graph, sampling individual nodes produces poor mixing (if the prior variance of Z is large).
- When $\mathbf{Z}$ is sampled, it depends strongly on the current value of $\boldsymbol{\theta}$
- Conversely, when $\boldsymbol{\theta}$ is sampled, it depends strongly on the current value of $\mathbf{Z}$.

Improved binary probit sampler

Holmes and Held (2006) suggested an improved sampler for the probit model that removes the dependency of Z on $\boldsymbol{\theta}$. When updating Z, we integrate out $\boldsymbol{\theta}$.



In this marginal model, $\mathbf{Z}$ has a joint multivariate normal prior.

- $\mathbf{Z}$ is updated element-wise
- $[Z_i \mid Y_i, \mathbf{Z}_{-i}]$ has a truncated normal distribution. Its mean and variance can be calculated from the posterior variance of $\boldsymbol{\theta} \mid Z$, which is already calculated by the update method for $\boldsymbol{\theta}$.
- This produces substantial increase in efficiency of the sampler

Improved binary-logit sampler?

Holmes and Held (2006) extended the marginal sampling method to the binary logistic model, but found little improvement

- To preserve multivariate normality of $\mathbf{Z}$, we need to condition on the current values of the mixture parameters $\boldsymbol{\psi}$
- But the coverage of the normal mixture component $Z_i \mid \psi_i$ can be much smaller than the full distribution of Z_i.
- This reduces the step-size for the update of each element Z_i and so reduces mixing.

ASIS

Yu and Meng (2011) propose a general strategy for improving mixing of data augmentation models by interleaving two different parameterizations

- A sufficient (centered) parameterization (Z)
- An ancillary (non-centred) parameterization ($\tilde{Z} = Z - \mathrm{E}(Z)$)

They call this Ancillary-Sufficiency Interweaving Strategy (ASIS). Motivated by Basu's theorem on independence of complete sufficient and ancillary statistics.

Variance components

- Poor mixing of variance components can also affect generalized linear mixed models.
- Andrew Gelman has proposed a conjugate prior for the *standard deviation* of a normal random effect – the half-t_d distribution.
- Gelman's proposal uses a *redundant parameterization* of the t distribution (normal numerator, chi-squared denominator) which effectively captures the ancillary/sufficient parameterization of Yu and Meng.
- In BUGS, this is equivalent to an $F_{d,1}$ distribution on the precision parameter. An experimental sampler in JAGS handles this case

Conclusions

MCMC performance in GLMs can be improved with a combination of strategies that transform one model into aother.

- Data augmentation (adding nodes) simplifies update methods for GLMs but may lead to poor sampling performance.
- Marginalization (removing nodes) removes unwanted dependencies between nodes, improving sampling performance.
- Redundant parameterization (splitting an identifiable node into non-identifiable components) combines features of both.