Dynamic Generalized Linear Models

Jesse Windle

Oct. 24, 2012

Contents

1 Introduction

- Useful when the response is categorical or count data.
- Used in ecology, finance, economics, political science, neuroscience, epidemiology.
- Ralph Reed is data mining your Kindle [Becker, 2012].
  - Response: Vote Republican

- Predictors: hunting license, read the Bible, has read “Going Rogue” by Sarah Palin, drive a pickup, married, income.

- Sometimes such models evolve in time.
  - Spike train data in neuroscience.
- Outline
  - Posterior inference via data augmentation in static case.
  - Then posterior inference for dynamic case.

2 Binary Data (Static Case)

- $y_i \in \{0, 1\}$ response, x_i row-vector of predictors.
- $P(y_i = 1) = p_i$
- ~~$p_i = x_i\beta$~~
- Transform to some other coordinate system:
  - $\psi_i = g(p_i) \iff p_i = \sigma(\psi_i)$, σ is a sigmoidal function.
  - $\psi_i = x_i\beta \leftarrow$ linear model in new coordinate system.
  - A term in LH:

$$p_i^{y_i}(1-p_i)^{1-y_i} \iff \sigma(y_i)^{y_i}(1-\sigma(y_i))^{1-y_i}.$$

    LH:

$$\prod_{i=1}^{n} \sigma(y_i)^{y_i}(1-\sigma(y_i))^{1-y_i}.$$

- Problem: we don’t know this distribution.
- Possible Solutions:
  - Metropolis-Hastings (requires picking a proposal, tuning).
  - Data augmentation + Gibbs sampling (don’t even have to think... once you have the augmentation).

3 Data Augmentation (de-marginalization) by 4 examples

- Idea:
  - $p(\beta)$ is hard to simulate.
  - Find joint distribution so that

    $$p(\beta) = \int p(\beta, \omega) d\omega$$

    and

    $$p(\beta|\omega) \text{ and } p(\omega|\beta)$$

    are easy(er) to simulate.
  - Gibbs sample.

3.1 Example 1: CDF method

- Albert and Chib [1993]
- $\sigma(\psi_i) = \Phi(\psi_i)$, Gaussian CDF.
- So

$$\begin{cases} P(y_i = 1) = p_i \\ p_i = \Phi(\psi_i), \qquad \psi_i = x_i \beta. \end{cases}$$

- Data generating "density:"

$$p(y_i|\psi_i) = \delta_1(y_i)\Phi(\psi_i) + \delta_0(y_i)(1 - \Phi(\psi_i)).$$

- De-marginalize:

$$\Phi(\psi_i) = \int_{-\infty}^{\infty} N(z_i; \psi_i, 1)\mathbf{1}_{(0,\infty)}(z_i).$$

  So

$$p(y_i|\psi_i) = \int_{-\infty}^{\infty} \left[\delta_1(y_i)\mathbf{1}_{(0,\infty)}(z_i) + \delta_0(y_i) \underbrace{\mathbf{1}_{(-\infty,0)}(z_i)}_{1-\mathbf{1}_{(0,\infty)}(z_i)}\right] N(z_i|\psi_i, 1) dz_i.$$

Thus

$$\begin{aligned} p(y_i, z_i|\psi_i) &= p(y_i|z_i)p(z_i|\psi_i) \\ &= \Big[\delta_1(y_i)\mathbf{1}_{(0,\infty)}(z_i) + \delta_0(y_i)\mathbf{1}_{(-\infty,0)}(z_i)\Big] N(z_i|\psi_i, 1). \end{aligned}$$

- Augmented model:

$$\begin{cases} y_i = \mathbf{1}\{z_i > 0\} \\ z_i = x_i\beta + \varepsilon_i, \qquad \varepsilon_i \sim N(0,1). \end{cases}$$

- Gibbs sample:
  - $p(\beta|z, y) \to$ Normal given normal prior.
  - $p(z|y, \beta) = \prod_i p(z_i|y_i, \beta)$ where

    $$p(z_i|y_i, \beta) \text{ is truncated normal.}$$

    Check it out for yourself.

3.2 Example 2: Mixture of Normals

- Holmes and Held [2006]
- Goal: take intractable distribution and represent it as a mixture of normals.
- Previously no interpretation of ψ_i.
- Logistic regression:
  - Same: $P(y_i = 1) = p_i$.
  - Different:

    $$\psi_i = \log \frac{p_i}{1 - p_i} \quad \text{(log odds scale)}$$

    iff

    $$p_i = \frac{e^{\psi_i}}{1 + e^{\psi_i}}.$$

- Play the same game...

$$\begin{cases} y_i = \mathbf{1}\{z_i > 0\} \\ z_i = x_i\beta + \varepsilon_i, \qquad \varepsilon_i \sim \text{Lo}(0,1). \end{cases}$$

Lo is the logistic distribution.

$p(\beta|z_i, y_i)$?

- Andrews and Mallows [1974]

$$\varepsilon \sim \text{Lo}(0,1) \iff \begin{cases} \varepsilon \sim N(0,\xi) \\ \xi = 4\lambda^2 \\ \lambda \sim \text{KS} = \text{Kolmogorov-Smirnov.} \end{cases}$$

So

$$p(\varepsilon) = \int_0^\infty p(\varepsilon|\lambda)p(\lambda)d\lambda.$$

- Augmented Model

$$\begin{cases} y_i = \mathbf{1}\{z_i > 0\} \\ z_i = x_i\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0,\xi_i) \\ \xi_i = 4\lambda_i^2 \\ \lambda_i \sim \text{KS}. \end{cases}$$

- Gibbs Sample:

$$\begin{cases} p(z_i|y_i,\beta,\xi) & \sim \text{ trunc. normal} \\ p(\beta|y,z,\xi) & \sim \text{ normal} \\ p(\xi_i|\beta,z,y) & \sim \text{ something tractable.} \end{cases}$$

3.3 Example 3: Discrete Mixture of Normals

- Frühwirth-Schnatter and Frühwirth [2007, 2010].
- Logistic continued...

$$\begin{cases} y_i = \mathbf{1}\{z_i > 0\} \\ z_i = x_i\beta + \varepsilon_i, \quad \varepsilon_i \sim \text{Lo}(0,1). \end{cases}$$

- Approximate Lo using a mixture of normals:

$$\begin{cases} \varepsilon_i \sim N(0,\sigma_{r_i}^2) \\ r_i \sim \text{MN}(1,w). \end{cases}$$

Then

$$\begin{cases} y_i = \mathbf{1}\{z_i > 0\} \\ z_i = x_i\beta + \varepsilon_i, \qquad \varepsilon_i N(0, \sigma_{r_i}^2) \\ r_i = \text{MN}(1, w). \end{cases}$$

- The only thing that changes is that

$$p(r_i|z_i, \beta, y) = \text{MN}.$$

3.4 Example 4: "Laplace" Method

- Logistic continued...
-

$$\begin{aligned} p(y_i|p_i) &\propto p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \Big(\frac{e^{\psi_i}}{1+e^{\psi_i}}\Big)^{y_i}\Big(\frac{1}{1+e^{\psi_i}}\Big)^{1-y_i} \\ &= \frac{e^{\psi_i y_i}}{1+e^{\psi_i}} \\ &= e^{\psi_i(y_i-1/2)}2^{-1/2}\cosh^{-1}(\psi_i/2). \end{aligned}$$

So

$$p(y_i|\psi_i) \propto e^{\psi_i\kappa_i}\cosh^{-1}(\psi_i/2)$$

where $\kappa_i = y_i - 1/2$.

- De-marginalize (via Laplace transform):

$$\cosh^{-1}(\psi_i/2) = \int_0^\infty e^{-\xi_i\psi_i^2/2}p(\xi_i)d\xi_i$$

where $\xi_i \sim \text{PG}(1,0)$. Then

$$p(y_i|\psi_i) \propto e^{\psi_i\kappa_i}\int_0^\infty e^{-\xi_i\psi_i^2/2}p(\xi_i)d\xi_i$$

so

$$p(y_i, \xi_i|\psi_i) \propto e^{\psi_i\kappa_i - \xi_i\psi_i^2/2}p(\xi_i).$$

So

$$p(y, \xi|\beta) \propto e^{\kappa' X\beta - \frac{1}{2}\beta' X\Xi X\beta}p(\xi)$$

where $p(\xi) = \prod p(\xi_i)$.

- Gibbs Sample:

$$p(\beta|\xi, y) \propto p(\beta)e^{\kappa' X\beta - \frac{1}{2}\beta' X \Xi X \beta} \leftarrow \text{ normal kernel.}$$

$$p(\xi_i|y,\beta) \propto e^{-\xi_i \psi_i^2/2} p(\xi_i) \leftarrow \text{PG}(1, \psi_i).$$

ONLY ONE LAYER OF LATENTS!

- Data generating process:

$$\begin{cases} \xi_i \sim \text{PG}(1, \psi_i) \\ y_i \sim \text{Binom}(1, p_i) \\ p_i = \frac{e^{\psi_i}}{1+e^{\psi_i}} \\ \psi_i = x_i\beta. \end{cases}$$

4 Aside: Expectation Maximization

Tangential, but a great opportunity to discuss EM, aka, a way to find the posterior mode. Following Gelman's red book.

We want $\underset{\beta}{\text{argmax}}\, p(\beta|y)$. We will assume everything is conditioned on y.

We have some sort of augmentation variable ξ. Then

$$p(\beta) = \frac{p(\beta, \xi)}{p(\xi|\beta)}.$$

Then

$$\log p(\beta) = \log p(\beta, \xi) - \log p(\xi|\beta).$$

Think of these as functions of β and ξ.

Take the expectation over $(\xi|\beta^{old})$:

$$\log p(\beta) = \mathbb{E}_{\xi|\beta^{old}}[\log p(\beta, \xi)] - \mathbb{E}_{\xi|\beta^{old}}[\log p(\xi|\beta)].$$

Fact:

$$\mathbb{E}_{\xi|\beta^{old}}[\log p(\xi|\beta)] \geq \mathbb{E}_{\xi|\beta^{old}}[\log p(\xi|\beta^{old})] \text{ for all } \beta.$$

Thus if β satisfies

$$\mathbb{E}_{\xi|\beta^{old}}[\log p(\beta, \xi)] \geq \mathbb{E}_{\xi|\beta^{old}}[\log p(\beta^{old}, \xi)]$$

then

$$\log p(\beta) \geq \log p(\beta^{old}).$$

In the Laplace transform method:

$$p(\beta, \xi) = c(y) p(\beta) e^{\kappa X' \beta - \frac{1}{2}\beta' X' \Xi X \beta} p(\xi).$$

So

$$\begin{aligned} \mathbb{E}_{\xi|\beta^{old}}[\log p(\beta, \xi)] &= \cancel{\log c(y)} \\ \text{quadratic form} &\begin{cases} + \log p(\beta) \\ + \kappa X' \beta - \frac{1}{2} \beta' X' \mathbb{E}_{\xi|\beta^{old}}(\Xi) X' \beta \end{cases} \\ &\cancel{\mathbb{E}_{\xi|\beta^{old}}[\log p(\xi)]}. \end{aligned}$$

We can calculate $\mathbb{E}[\xi_i]$ using the MGF, i.e. the Laplace transform.

Then $\beta^{t-1} \to \mathbb{E}_{\xi|\beta^{t-1}}(\xi) \to \beta^t$.

5 Dynamic Generalized Linear Models

Now our index is time.

5.1 The Old School Way

- West et al. [1985]
- Still: $\psi_t = \log \frac{p_t}{1-p_t}$
- We will evolve filtered distribution of ψ_t by linear Bayes.
- Observation equation: $P(y_t = 1) = p_t$.
- Evolution equation:

$$\begin{cases} \psi_t = x_t \beta_t \\ \beta_t = G_t \beta_{t-1} + \omega_t, \quad \omega_t \sim [0, W]. \end{cases}$$

- We only specify the first two moments!
- Just like DLM: once you understand how to do one update you are done.

- Prior: $\beta_{t-1}|D_{t-1} \sim [m_{t-1}, C_{t-1}]$.
- Evolution: $\beta_t|D_{t-1} \sim [a_t, R_t]$ where $a_t = G_{t-1}m_{t-1}$ and $R_t = G_{t-1}C_{t-1}G'_{t-1}$.
- Forecast: $\psi_t|D_{t-1} \sim [f_t, q_t]$ where $f_t = x_t a_t$ and $q_t = x'_t R_t x_t$.
- Update:

  Now we get to specify a distribution. $\psi_t|D_{t-1}$ is the "prior" for the observation of $(y_t|p_t)$. The conjugate prior for p_t is a Beta distribution. There is a one-to-one relationship between (f_t, q_t) and (r_t, s_t) so that

  $$\psi_t|D_{t-1} \sim [f_t, q_t]$$

  iff

  $$p_t|D_{t-1} \sim Beta(r_t, s_t).$$

  Use that to pick r_t and s_t.

  Hit the prior with the new observation and we have

  $$p_t|D_t \sim Beta(r_t^*, s_t^*).$$

  Now go backwards to get

  $$\psi_t|D_t \sim [f_t^*, q_t^*].$$

  This is exact in that once you specify that $p_t|D_{t-1}$ is Beta, then everything follows without any approximation.

  However, we only have an approximate update to β_t. We can use the distributions

  $$\psi_t|D_t$$

  and

  $$\begin{pmatrix} \psi_t \\ \beta_t \end{pmatrix} |D_{t-1} \sim \left[\begin{pmatrix} f_t \\ a_t \end{pmatrix}, \begin{pmatrix} q_t & x_t R_t \\ R_t x'_t & R_t \end{pmatrix} \right]$$

  to update β_t by linear Bayes to get an approximation of the two moments of $\beta_t|D_t$.

5.2 Using data augmentation

- Instead of having a normal prior on β we now have a stochastic process "prior" for β.
- Just change $\psi_i = x_i\beta$ to $\psi_t = x_t\beta_t$.

- Probit: CDF.
  - Augmented model:

$$\begin{cases} y_t = \mathbf{1}\{z_t > 0\} \\ z_t = x_t\beta_t + \varepsilon_t, & \varepsilon_t \sim N(0,1) \\ \beta_t = G_t\beta_{t-1} + \omega_t, & \omega_t \sim N(0,W). \end{cases}$$

  - Gibbs:
    * $p(z_t|\beta_t, y_t) \leftarrow$ same.
    * $p(\{\beta_t\}|z) \leftarrow$ FFBS.
- Logistic: mixture of normals.
  - Augmented model:

$$\begin{cases} y_t = \mathbf{1}\{z_t > 0\} \\ z_t = x_t\beta_t + \varepsilon_t, & \varepsilon_i \sim N(0,\xi_t) \\ \beta_t = G_{t-1}\beta_{t-1} + \omega_t, & \omega_t \sim N(0,W) \\ \xi_t = 4\lambda_t^2 \\ \lambda_t \sim \text{KS}. \end{cases}$$

  - Gibbs: all the same, but
    * $p(\{\beta_t\}|z,\xi) \leftarrow$ FFBS.
- Logistic: discrete mixture of normals.
  - Augmented model:

$$\begin{cases} y_t = \mathbf{1}\{z_t > 0\} \\ z_t = x_t\beta_t + \varepsilon_t, & \varepsilon_t N(0,\sigma_{r_t}^2) \\ \beta_t = G_{t-1}\beta_{t-1} + \omega_t, & \omega_t \sim N(0,W) \\ r_t = \text{MN}(1,w). \end{cases}$$

  - Gibbs: all the same, but
    * $p(\{\beta_t\}|z,r) \leftarrow$ FFBS.
- Logistic: Polya-Gamma:

– Look at posterior for β:

$$\begin{aligned} p(\{\beta_t\}|\xi, y) &\propto e^{\kappa' X\beta - \frac{1}{2}\beta' X' \Xi X' \beta} p(\beta) \\ &\propto e^{-\frac{1}{2}(z - X\beta)^2} p(\beta), \quad \text{where } \Xi z = \kappa \\ &\propto p(z|\beta) p(\beta) \end{aligned}$$

where

$$\begin{cases} z_t = x_t \beta_t + \varepsilon_t, & \varepsilon_t \sim N(0, 1/\xi_t) \\ \beta_t = G_{t-1}\beta_{t-1} + \eta_t, & \eta_t \sim N(0, W). \end{cases}$$

So, again, we can just FFBS for β.

References

J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, June 1993.

D. F. Andrews and C. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(36):99–102, 1974.

J. Becker. An Evangelical is back from exile, lifting Romney. *The New York Times*, September 22 2012.

S. Frühwirth-Schnatter and R. Frühwirth. Auxiliary mixture sampling with applications to logistic models. *Computational Statistics and Data Analysis*, 51:3509–3528, 2007.

S. Frühwirth-Schnatter and R. Frühwirth. Data augmentation and mcmc for binary and multinomial logit models. In *Statistical Modelling and Regression Structures*, pages 111–132. Springer-Verlag, 2010. Available from UT library online.

C. C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168, 2006.

M. West, J. Harrison, and H. S. Migon. Dynamics generalized linear models and bayesian forecasting. *Journal of the American Statistical Association*, 80(389):73–83, March 1985.