Bayesian generalized linear models and an appropriate default prior

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su
Columbia University

14 August 2008
Logistic regression

Classical logistic regression
The problem of separation
Bayesian solution

\[ y = \logit^{-1}(x) \]

slope = 1/4
A clean example

estimated $\Pr(y=1) = \text{logit}^{-1}(-1.40 + 0.33 \, x)$

slope = 0.33/4
The problem of separation

slope = infinity?

Gelman, Jakulin, Pittau, Su
Bayesian generalized linear models and an appropriate default prior
Separation is no joke!

```r
glm (vote ~ female + black + income, family=binomial(link="logit"))
```

<table>
<thead>
<tr>
<th></th>
<th>1960 coef.est</th>
<th>1968 coef.est</th>
<th>1964 coef.est</th>
<th>1972 coef.est</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.14</td>
<td>0.47</td>
<td>-1.15</td>
<td>0.67</td>
</tr>
<tr>
<td>female</td>
<td>0.24</td>
<td>-0.01</td>
<td>-0.09</td>
<td>-0.25</td>
</tr>
<tr>
<td>black</td>
<td>-1.03</td>
<td>-3.64</td>
<td>-16.83</td>
<td>-2.63</td>
</tr>
<tr>
<td>income</td>
<td>0.03</td>
<td>-0.03</td>
<td>0.19</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
Bayesian logistic regression

*In the arm (Applied Regression and Multilevel modeling) package*

- Replaces `glm()`, estimates are more numerically and computationally stable
- Use EM-like algorithm
- We went inside `glm.fit` to augment the iteratively weighted least squares step
- Default choices for tuning parameters (we'll get back to this!)

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
Bayesian logistic regression

- In the \texttt{arm} (Applied Regression and Multilevel modeling) package
- Replaces \texttt{glm()}, estimates are more numerically and computationally stable
- Student-\textit{t} prior distributions for regression coefs
- Use EM-like algorithm
- We went inside \texttt{glm.fit} to augment the iteratively weighted least squares step
- Default choices for tuning parameters (we'll get back to this!)

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default
Bayesian logistic regression

- In the `arm` (Applied Regression and Multilevel modeling) package
- Replaces `glm()`, estimates are more numerically and computationally stable
- Student-$t$ prior distributions for regression coefficients
- Use EM-like algorithm
- We went inside `glm.fit` to augment the iteratively weighted least squares step
- Default choices for tuning parameters (we'll get back to this!)
Bayesian logistic regression

- In the `arm` (Applied Regression and Multilevel modeling) package
- Replaces `glm()`, estimates are more numerically and computationally stable
- Student-t prior distributions for regression coefs
- Use EM-like algorithm
- We went inside `glm.fit` to augment the iteratively weighted least squares step
- Default choices for tuning parameters (we'll get back to this!)

Gelman, Jakulin, Pittau, Su
Bayesian generalized linear models and an appropriate default
Bayesian logistic regression

- In the **arm** (Applied Regression and Multilevel modeling) package
- Replaces `glm()`, estimates are more numerically and computationally stable
- Student-\(t\) prior distributions for regression coefs
- Use EM-like algorithm
- We went inside `glm.fit` to augment the iteratively weighted least squares step
- Default choices for tuning parameters (we’ll get back to this!)

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default
Bayesian logistic regression

In the arm (Applied Regression and Multilevel modeling) package

Replaces glm(), estimates are more numerically and computationally stable

Student-t prior distributions for regression coefs

Use EM-like algorithm

We went inside glm.fit to augment the iteratively weighted least squares step

Default choices for tuning parameters (we’ll get back to this!)
Bayesian logistic regression

- **In the arm (Applied Regression and Multilevel modeling) package**
- **Replaces glm(), estimates are more numerically and computationally stable**
- **Student-t prior distributions for regression coefs**
- **Use EM-like algorithm**
- **We went inside glm.fit to augment the iteratively weighted least squares step**
- **Default choices for tuning parameters (we’ll get back to this!)**

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default
Bayesian logistic regression

- In the arm (Applied Regression and Multilevel modeling) package
- Replaces glm(), estimates are more numerically and computationally stable
- Student-t prior distributions for regression coefficients
- Use EM-like algorithm
- We went inside glm.fit to augment the iteratively weighted least squares step
- Default choices for tuning parameters (we’ll get back to this!)
Regularization in action!

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
What else is out there?

- glm (maximum likelihood): fails under separation, gives noisy answers for sparse data
- Augment with prior “successes” and “failures”: doesn’t work well for multiple predictors
- brlr (Jeffreys-like prior distribution): computationally unstable
- brglm (improvement on brlr): doesn’t do enough smoothing
- BBR (Laplace prior distribution): OK, not quite as good as bayesglm
- Non-Bayesian machine learning algorithms: understate uncertainty in predictions
What else is out there?

- **glm** (maximum likelihood): fails under separation, gives noisy answers for sparse data
- Augment with prior “successes” and “failures”: doesn’t work well for multiple predictors
- **brlr** (Jeffreys-like prior distribution): computationally unstable
- **brglm** (improvement on brlr): doesn’t do enough smoothing
- **BBR** (Laplace prior distribution): OK, not quite as good as bayesglm
- Non-Bayesian machine learning algorithms: understate uncertainty in predictions
What else is out there?

- **glm** (maximum likelihood): fails under separation, gives noisy answers for sparse data
- Augment with prior “successes” and “failures”: doesn’t work well for multiple predictors
- **brlr** (Jeffreys-like prior distribution): computationally unstable
- **brglm** (improvement on brlr): doesn’t do enough smoothing
- **BBR** (Laplace prior distribution): OK, not quite as good as bayesglm
- Non-Bayesian machine learning algorithms: understate uncertainty in predictions

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
What else is out there?

- **glm** (maximum likelihood): fails under separation, gives noisy answers for sparse data
- Augment with prior “successes” and “failures”: doesn’t work well for multiple predictors
- **brlr** (Jeffreys-like prior distribution): computationally unstable
- **brglm** (improvement on **brlr**): doesn’t do enough smoothing
- **BBR** (Laplace prior distribution): OK, not quite as good as **bayesglm**
- Non-Bayesian machine learning algorithms: understate uncertainty in predictions

Gelman, Jakulin, Pittau, Su: Bayesian generalized linear models and an appropriate default prior
What else is out there?

- glm (maximum likelihood): fails under separation, gives noisy answers for sparse data
- Augment with prior “successes” and “failures”: doesn’t work well for multiple predictors
- brlr (Jeffreys-like prior distribution): computationally unstable
- brglm (improvement on brlr): doesn’t do enough smoothing
- BBR (Laplace prior distribution): OK, not quite as good as bayesglm
- Non-Bayesian machine learning algorithms: understate uncertainty in predictions
What else is out there?

- glm (maximum likelihood): fails under separation, gives noisy answers for sparse data
- Augment with prior “successes” and “failures”: doesn’t work well for multiple predictors
- brlr (Jeffreys-like prior distribution): computationally unstable
- brglm (improvement on brlr): doesn’t do enough smoothing
- BBR (Laplace prior distribution): OK, not quite as good as bayesglm

- Non-Bayesian machine learning algorithms: understate uncertainty in predictions
What else is out there?

- `glm` (maximum likelihood): fails under separation, gives noisy answers for sparse data
- Augment with prior “successes” and “failures”: doesn’t work well for multiple predictors
- `brlr` (Jeffreys-like prior distribution): computationally unstable
- `brglm` (improvement on `brlr`): doesn’t do enough smoothing
- `BBR` (Laplace prior distribution): OK, not quite as good as `bayesglm`
- Non-Bayesian machine learning algorithms: understate uncertainty in predictions
Information in prior distributions

- Informative prior dist
  - A full generative model for the data
- Noninformative prior dist
- Weakly informative prior dist

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
Information in prior distributions

- **Informative prior dist**
  - A full generative model for the data
- **Noninformative prior dist**
  - Let the data speak
  - Goal: valid inference for any $\theta$
- **Weakly informative prior dist**
  - Purposely include less information than we actually have
  - Goal: regularization, stabilization

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default
Information in prior distributions

- Informative prior dist
  - A full generative model for the data
- Noninformative prior dist
  - Let the data speak
  - Goal: valid inference for any $\theta$
- Weakly informative prior dist
  - Purposely include less information than we actually have
  - Goal: regularization, stabilization
Information in prior distributions

- Informative prior dist
  - A full generative model for the data
- Noninformative prior dist
  - Let the data speak
  - Goal: valid inference for any $\theta$
- Weakly informative prior dist
  - Purposely include less information than we actually have
  - Goal: regularization, stabilization

Gelman, Jakulin, Pittau, Su
Bayesian generalized linear models and an appropriate default prior
Information in prior distributions

- Informative prior dist
  - A full generative model for the data

- Noninformative prior dist
  - Let the data speak
    - Goal: valid inference for any $\theta$

- Weakly informative prior dist
  - Purposely include less information than we actually have
  - Goal: regularization, stabilization
Information in prior distributions

- Informative prior dist
  - A full generative model for the data
- Noninformative prior dist
  - Let the data speak
  - Goal: valid inference for any $\theta$
- Weakly informative prior dist
  - Purposely include less information than we actually have
  - Goal: regularization, stabilization
Information in prior distributions

- Informative prior dist
  - A full generative model for the data

- Noninformative prior dist
  - Let the data speak
  - Goal: valid inference for any $\theta$

- Weakly informative prior dist
  - Purposely include less information than we actually have
  - Goal: regularization, stabilization
Information in prior distributions

- Informative prior dist
  - A full generative model for the data

- Noninformative prior dist
  - Let the data speak
  - Goal: valid inference for any $\theta$

- Weakly informative prior dist
  - Purposely include less information than we actually have
  - Goal: regularization, stabilization
Information in prior distributions

- Informative prior dist
  - A full generative model for the data

- Noninformative prior dist
  - Let the data speak
  - Goal: valid inference for any $\theta$

- Weakly informative prior dist
  - Purposely include less information than we actually have
  - Goal: regularization, stabilization

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coefs are almost always between $-5$ and $5$:
  - $5$ on the logit scale takes you from $0.01$ to $0.50$ or from $0.50$ to $0.99$
- Smoking and lung cancer
- Independent Cauchy prior dists with center 0 and scale $2.5$
- Rescale each predictor to have mean 0 and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of glm

Gelman, Jakulin, Pittau, Su
Bayesian generalized linear models and an appropriate default
Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coeffs are almost always between $-5$ and $5$:
  - $5$ on the logit scale takes you from $0.01$ to $0.50$ or from $0.50$ to $0.99$
  - Smoking and lung cancer
- Independent Cauchy prior dists with center 0 and scale 2.5
- Rescale each predictor to have mean 0 and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of glm

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coefs are almost always between $-5$ and $5$:
  - 5 on the logit scale takes you from 0.01 to 0.50 or from 0.50 to 0.99
  - Smoking and lung cancer
- Independent Cauchy prior dists with center 0 and scale 2.5
- Rescale each predictor to have mean 0 and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of glm
Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coefs are almost always between $-5$ and $5$:
  - $5$ on the logit scale takes you from 0.01 to 0.50 or from 0.50 to 0.99
- Smoking and lung cancer
- Independent Cauchy prior dists with center 0 and scale 2.5
- Rescale each predictor to have mean 0 and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of glm

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coeffs are almost always between $-5$ and $5$:
  - $5$ on the logit scale takes you from $0.01$ to $0.50$ or from $0.50$ to $0.99$
  - Smoking and lung cancer
- Independent Cauchy prior dists with center $0$ and scale $2.5$
- Rescale each predictor to have mean $0$ and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of glm

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coefs are almost always between $-5$ and $5$:
  - $5$ on the logit scale takes you from 0.01 to 0.50 or from 0.50 to 0.99
  - Smoking and lung cancer
- Independent Cauchy prior dists with center 0 and scale 2.5
- Rescale each predictor to have mean 0 and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of glm
Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coefs are almost always between $-5$ and $5$:
  - 5 on the logit scale takes you from 0.01 to 0.50 or from 0.50 to 0.99
  - Smoking and lung cancer
- Independent Cauchy prior dists with center 0 and scale 2.5
- Rescale each predictor to have mean 0 and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of glm

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default
Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coefs are almost always between $-5$ and $5$:
  - $5$ on the logit scale takes you from $0.01$ to $0.50$ or from $0.50$ to $0.99$
  - Smoking and lung cancer
- Independent Cauchy prior dists with center $0$ and scale $2.5$
- Rescale each predictor to have mean $0$ and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of glm
Prior distributions

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
Another example

<table>
<thead>
<tr>
<th>Dose</th>
<th>#deaths/#animals</th>
</tr>
</thead>
<tbody>
<tr>
<td>−0.86</td>
<td>0/5</td>
</tr>
<tr>
<td>−0.30</td>
<td>1/5</td>
</tr>
<tr>
<td>−0.05</td>
<td>3/5</td>
</tr>
<tr>
<td>0.73</td>
<td>5/5</td>
</tr>
</tbody>
</table>

- Slope of a logistic regression of $\text{Pr}(\text{death})$ on dose:
  - Maximum likelihood est is $7.8 \pm 4.9$
  - With weakly-informative prior: Bayes est is $4.4 \pm 1.9$

- Which is truly conservative?
- The sociology of shrinkage
Another example

<table>
<thead>
<tr>
<th>Dose</th>
<th>#deaths/#animals</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.86</td>
<td>0/5</td>
</tr>
<tr>
<td>-0.30</td>
<td>1/5</td>
</tr>
<tr>
<td>-0.05</td>
<td>3/5</td>
</tr>
<tr>
<td>0.73</td>
<td>5/5</td>
</tr>
</tbody>
</table>

- Slope of a logistic regression of Pr(death) on dose:
  - Maximum likelihood est is $7.8 \pm 4.9$
  - With weakly-informative prior: Bayes est is $4.4 \pm 1.9$

- Which is truly conservative?
- The sociology of shrinkage

Gelman, Jakulin, Pittau, Su
Another example

<table>
<thead>
<tr>
<th>Dose</th>
<th>#deaths/#animals</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.86</td>
<td>0/5</td>
</tr>
<tr>
<td>-0.30</td>
<td>1/5</td>
</tr>
<tr>
<td>-0.05</td>
<td>3/5</td>
</tr>
<tr>
<td>0.73</td>
<td>5/5</td>
</tr>
</tbody>
</table>

- Slope of a logistic regression of $\Pr(\text{death})$ on dose:
  - Maximum likelihood est is $7.8 \pm 4.9$
  - With weakly-informative prior: Bayes est is $4.4 \pm 1.9$

- Which is truly conservative?
- The sociology of shrinkage
Another example

<table>
<thead>
<tr>
<th>Dose</th>
<th>#deaths/#animals</th>
</tr>
</thead>
<tbody>
<tr>
<td>−0.86</td>
<td>0/5</td>
</tr>
<tr>
<td>−0.30</td>
<td>1/5</td>
</tr>
<tr>
<td>−0.05</td>
<td>3/5</td>
</tr>
<tr>
<td>0.73</td>
<td>5/5</td>
</tr>
</tbody>
</table>

- Slope of a logistic regression of \( \Pr(\text{death}) \) on dose:
  - Maximum likelihood est is \( 7.8 \pm 4.9 \)
  - With weakly-informative prior: Bayes est is \( 4.4 \pm 1.9 \)

- Which is truly conservative?
- The sociology of shrinkage
Another example

<table>
<thead>
<tr>
<th>Dose</th>
<th>#deaths/#animals</th>
</tr>
</thead>
<tbody>
<tr>
<td>−0.86</td>
<td>0/5</td>
</tr>
<tr>
<td>−0.30</td>
<td>1/5</td>
</tr>
<tr>
<td>−0.05</td>
<td>3/5</td>
</tr>
<tr>
<td>0.73</td>
<td>5/5</td>
</tr>
</tbody>
</table>

- Slope of a logistic regression of Pr(death) on dose:
  - Maximum likelihood est is $7.8 \pm 4.9$
  - With weakly-informative prior: Bayes est is $4.4 \pm 1.9$

- Which is truly conservative?

- The sociology of shrinkage
Another example

<table>
<thead>
<tr>
<th>Dose</th>
<th>#deaths/#animals</th>
</tr>
</thead>
<tbody>
<tr>
<td>−0.86</td>
<td>0/5</td>
</tr>
<tr>
<td>−0.30</td>
<td>1/5</td>
</tr>
<tr>
<td>−0.05</td>
<td>3/5</td>
</tr>
<tr>
<td>0.73</td>
<td>5/5</td>
</tr>
</tbody>
</table>

▶ Slope of a logistic regression of Pr(death) on dose:
  - Maximum likelihood est is $7.8 \pm 4.9$
  - With weakly-informative prior: Bayes est is $4.4 \pm 1.9$

▶ Which is truly conservative?
▶ The sociology of shrinkage
Maximum likelihood and Bayesian estimates

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
Conservatism of Bayesian inference

- Problems with maximum likelihood when data show separation:
  - Coefficient estimate of $-\infty$
  - Estimated predictive probability of 0 for new cases
- Is this conservative?
- Not if evaluated by log score or predictive log-likelihood
Conservatism of Bayesian inference

- Problems with maximum likelihood when data show separation:
  - Coefficient estimate of $-\infty$
  - Estimated predictive probability of 0 for new cases
- Is this conservative?
- Not if evaluated by log score or predictive log-likelihood
Conservatism of Bayesian inference

- Problems with maximum likelihood when data show separation:
  - Coefficient estimate of $-\infty$
  - Estimated predictive probability of 0 for new cases
- Is this conservative?
- Not if evaluated by log score or predictive log-likelihood
Conservatism of Bayesian inference

- Problems with maximum likelihood when data show separation:
  - Coefficient estimate of $-\infty$
  - Estimated predictive probability of 0 for new cases
- Is this conservative?
- Not if evaluated by log score or predictive log-likelihood
Conservatism of Bayesian inference

- Problems with maximum likelihood when data show separation:
  - Coefficient estimate of $-\infty$
  - Estimated predictive probability of 0 for new cases
- Is this conservative?
  - Not if evaluated by log score or predictive log-likelihood
Conservatism of Bayesian inference

- Problems with maximum likelihood when data show separation:
  - Coefficient estimate of $-\infty$
  - Estimated predictive probability of 0 for new cases
- Is this conservative?
- Not if evaluated by log score or predictive log-likelihood
Which one is conservative?

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
Prior as population distribution

- Consider many possible datasets
- The “true prior” is the distribution of $\beta$’s across these datasets
- Fit one dataset at a time
- A “weakly informative prior” has less information (wider variance) than the true prior
- Open question: How to formalize the tradeoffs from using different priors?
Prior as population distribution

- Consider many possible datasets
- The “true prior” is the distribution of $\beta$’s across these datasets
- Fit one dataset at a time
- A “weakly informative prior” has less information (wider variance) than the true prior
- Open question: How to formalize the tradeoffs from using different priors?
Prior as population distribution

- Consider many possible datasets
- The “true prior” is the distribution of $\beta$’s across these datasets
- Fit one dataset at a time
- A “weakly informative prior” has less information (wider variance) than the true prior
- Open question: How to formalize the tradeoffs from using different priors?
Prior as population distribution

- Consider many possible datasets
- The “true prior” is the distribution of $\beta$’s across these datasets
- Fit one dataset at a time
- A “weakly informative prior” has less information (wider variance) than the true prior
- Open question: How to formalize the tradeoffs from using different priors?
Prior as population distribution

- Consider many possible datasets
- The “true prior” is the distribution of $\beta$'s across these datasets
- Fit one dataset at a time
- A “weakly informative prior” has less information (wider variance) than the true prior

Open question: How to formalize the tradeoffs from using different priors?
Prior as population distribution

- Consider many possible datasets
- The “true prior” is the distribution of $\beta$’s across these datasets
- Fit one dataset at a time
- A “weakly informative prior” has less information (wider variance) than the true prior
- Open question: How to formalize the tradeoffs from using different priors?
Evaluation using a corpus of datasets

- Compare classical glm to Bayesian estimates using various prior distributions
- Evaluate using 5-fold cross-validation and average predictive error
- The optimal prior distribution for $\beta$’s is (approx) Cauchy $(0, 1)$
- Our Cauchy $(0, 2.5)$ prior distribution is weakly informative!
Evaluation using a corpus of datasets

- Compare classical glm to Bayesian estimates using various prior distributions
- Evaluate using 5-fold cross-validation and average predictive error
- The optimal prior distribution for $\beta$’s is (approx) Cauchy $(0, 1)$
- Our Cauchy $(0, 2.5)$ prior distribution is weakly informative
Evaluation using a corpus of datasets

- Compare classical glm to Bayesian estimates using various prior distributions
- Evaluate using 5-fold cross-validation and average predictive error
- The optimal prior distribution for $\beta$’s is (approx) Cauchy (0, 1)
- Our Cauchy (0, 2.5) prior distribution is weakly informative!
Evaluation using a corpus of datasets

- Compare classical glm to Bayesian estimates using various prior distributions
- Evaluate using 5-fold cross-validation and average predictive error
- The optimal prior distribution for $\beta$’s is (approx) Cauchy $(0, 1)$
- Our Cauchy $(0, 2.5)$ prior distribution is weakly informative!
Evaluation using a corpus of datasets

- Compare classical glm to Bayesian estimates using various prior distributions
- Evaluate using 5-fold cross-validation and average predictive error
- The optimal prior distribution for $\beta$’s is (approx) Cauchy (0, 1)
- Our Cauchy (0, 2.5) prior distribution is weakly informative!
Expected predictive loss, avg over a corpus of datasets

Gelman, Jakulin, Pittau, Su
Bayesian generalized linear models and an appropriate default prior
Priors for other regression models

- Probit
- Ordered logit/probit
- Poisson
- Linear regression with normal errors
Priors for other regression models

- Probit
- Ordered logit/probit
- Poisson
- Linear regression with normal errors
Priors for other regression models

- Probit
- Ordered logit/probit
- Poisson
- Linear regression with normal errors
Priors for other regression models

- Probit
- Ordered logit/probit
- Poisson
- Linear regression with normal errors
Priors for other regression models

- Probit
- Ordered logit/probit
- Poisson
- Linear regression with normal errors
Other examples of weakly informative priors

- Variance parameters
- Covariance matrices
- Population variation in a physiological model
- Mixture models
- Intentional underpooling in hierarchical models
Other examples of weakly informative priors

- Variance parameters
- Covariance matrices
- Population variation in a physiological model
- Mixture models
- Intentional underpooling in hierarchical models
Other examples of weakly informative priors

- Variance parameters
- Covariance matrices
- Population variation in a physiological model
- Mixture models
- Intentional underpooling in hierarchical models
Other examples of weakly informative priors

- Variance parameters
- Covariance matrices
- Population variation in a physiological model
- Mixture models
- Intentional underpooling in hierarchical models
Other examples of weakly informative priors

- Variance parameters
- Covariance matrices
- Population variation in a physiological model
- Mixture models
- Intentional underpooling in hierarchical models
Other examples of weakly informative priors

- Variance parameters
- Covariance matrices
- Population variation in a physiological model
- Mixture models
- Intentional underpooling in hierarchical models
Conclusions

- “Noninformative priors” are actually weakly informative
- “Weakly informative” is a more general and useful concept
- Regularization

- Why use weakly informative priors rather than informative priors?

Gelman, Jakulin, Pittau, Su
Bayesian generalized linear models and an appropriate default
Conclusions

- “Noninformative priors” are actually weakly informative
- “Weakly informative” is a more general and useful concept
- Regularization
  - Better inferences
  - Stability of computation (bayesglm)
- Why use weakly informative priors rather than informative priors?
“Noninformative priors” are actually weakly informative

“Weakly informative” is a more general and useful concept

Regularization

- Better inferences
- Stability of computation (bayesglm)

Why use weakly informative priors rather than informative priors?
Conclusions

- “Noninformative priors” are actually weakly informative
- “Weakly informative” is a more general and useful concept
- Regularization
  - Better inferences
  - Stability of computation (bayesglm)
- Why use weakly informative priors rather than informative priors?
Conclusions

- "Noninformative priors" are actually weakly informative
- "Weakly informative" is a more general and useful concept
- Regularization
  - Better inferences
  - Stability of computation (bayesglm)
- Why use weakly informative priors rather than informative priors?
  - Conformity with statistical culture ("conservatism")
  - Labor-saving device
  - Robustness
Conclusions

- “Noninformative priors” are actually weakly informative
- “Weakly informative” is a more general and useful concept
- Regularization
  - Better inferences
  - Stability of computation (bayesglm)
- Why use weakly informative priors rather than informative priors?
  - Conformity with statistical culture (“conservatism”)
  - Labor-saving device
  - Robustness
Conclusions

- “Noninformative priors” are actually weakly informative
- “Weakly informative” is a more general and useful concept
- Regularization
  - Better inferences
  - Stability of computation (bayesglm)
- Why use weakly informative priors rather than informative priors?
  - Conformity with statistical culture (“conservatism”)
  - Labor-saving device
  - Robustness
Conclusions

- “Noninformative priors” are actually weakly informative
- “Weakly informative” is a more general and useful concept
- Regularization
  - Better inferences
  - Stability of computation (\texttt{bayesglm})
- Why use weakly informative priors rather than informative priors?
  - Conformity with statistical culture (“conservatism”)
  - Labor-saving device
  - Robustness

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
Conclusions

- “Noninformative priors” are actually weakly informative
- “Weakly informative” is a more general and useful concept
- Regularization
  - Better inferences
  - Stability of computation (bayesglm)
- Why use weakly informative priors rather than informative priors?
  - Conformity with statistical culture ("conservatism")
  - Labor-saving device
  - Robustness
Conclusions

- “Noninformative priors” are actually weakly informative
- “Weakly informative” is a more general and useful concept
- Regularization
  - Better inferences
  - Stability of computation (bayesglm)
- Why use weakly informative priors rather than informative priors?
  - Conformity with statistical culture ("conservatism")
  - Labor-saving device
  - Robustness
Other examples of weakly informative priors

- Variance parameters
- Covariance matrices
- Population variation in a physiological model
- Mixture models
- Intentional underpooling in hierarchical models
Other examples of weakly informative priors

- Variance parameters
- Covariance matrices
- Population variation in a physiological model
- Mixture models
- Intentional underpooling in hierarchical models
Other examples of weakly informative priors

- Variance parameters
- Covariance matrices
- Population variation in a physiological model
- Mixture models
- Intentional underpooling in hierarchical models
Other examples of weakly informative priors

- Variance parameters
- Covariance matrices
- Population variation in a physiological model
- Mixture models
- Intentional underpooling in hierarchical models
Other examples of weakly informative priors

- Variance parameters
- Covariance matrices
- Population variation in a physiological model
- Mixture models
- Intentional underpooling in hierarchical models
Other examples of weakly informative priors

- Variance parameters
- Covariance matrices
- Population variation in a physiological model
- Mixture models
- Intentional underpooling in hierarchical models
Weakly informative priors for variance parameter

- Basic hierarchical model
- Traditional inverse-gamma(0.001, 0.001) prior can be highly informative (in a bad way)!
- Noninformative uniform prior works better
- But if $\#$groups is small ($J = 2, 3$, even 5), a weakly informative prior helps by shutting down huge values of $\tau$
Weakly informative priors for variance parameter

- Basic hierarchical model
- Traditional inverse-gamma(0.001, 0.001) prior can be highly informative (in a bad way)!
- Noninformative uniform prior works better
- But if #groups is small ($J = 2, 3, \text{ even } 5$), a weakly informative prior helps by shutting down huge values of $\tau$
Weakly informative priors for variance parameter

- Basic hierarchical model
- Traditional inverse-gamma(0.001, 0.001) prior can be highly informative (in a bad way)!
- Noninformative uniform prior works better
- But if \#groups is small (J = 2, 3, even 5), a weakly informative prior helps by shutting down huge values of \( \tau \)
Weakly informative priors for variance parameter

- Basic hierarchical model
- Traditional inverse-gamma(0.001, 0.001) prior can be highly informative (in a bad way)!
- Noninformative uniform prior works better
- But if #groups is small ($J = 2, 3, \text{ or even } 5$), a weakly informative prior helps by shutting down huge values of $\tau$
Weakly informative priors for variance parameter

- Basic hierarchical model
- Traditional inverse-gamma(0.001, 0.001) prior can be highly informative (in a bad way)!
- Noninformative uniform prior works better
- But if \#groups is small (\(J = 2, 3,\) even 5), a weakly informative prior helps by shutting down huge values of \(\tau\)
Priors for variance parameter: \( J = 8 \) groups

8 schools: posterior on \( \sigma_\alpha \) given uniform prior on \( \sigma_\alpha \)

8 schools: posterior on \( \sigma_\alpha \) given inv–gamma (1, 1) prior on \( \sigma_\alpha^2 \)

8 schools: posterior on \( \sigma_\alpha \) given inv–gamma (.001, .001) prior on \( \sigma_\alpha^2 \)
Priors for variance parameter: \( J = 3 \) groups

3 schools: posterior on \( \sigma_\alpha \) given uniform prior on \( \sigma_\alpha \)

3 schools: posterior on \( \sigma_\alpha \) given half-Cauchy (25) prior on \( \sigma_\alpha \)

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
Weakly informative priors for covariance matrices

- Inverse-Wishart has problems
- Correlations can be between 0 and 1
- Set up models so prior expectation of correlations is 0
- Goal: to be weakly informative about correlations and variances
- Scaled inverse-Wishart model uses redundant parameterization
Weakly informative priors for covariance matrices

- Inverse-Wishart has problems
- Correlations can be between 0 and 1
- Set up models so prior expectation of correlations is 0
- Goal: to be weakly informative about correlations and variances
- Scaled inverse-Wishart model uses redundant parameterization
Weakly informative priors for covariance matrices

- Inverse-Wishart has problems
- Correlations can be between 0 and 1
- Set up models so prior expectation of correlations is 0
- Goal: to be weakly informative about correlations and variances
- Scaled inverse-Wishart model uses redundant parameterization
Weakly informative priors for covariance matrices

- Inverse-Wishart has problems
- Correlations can be between 0 and 1
- Set up models so prior expectation of correlations is 0
- Goal: to be weakly informative about correlations and variances
- Scaled inverse-Wishart model uses redundant parameterization

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
Weakly informative priors for covariance matrices

- Inverse-Wishart has problems
- Correlations can be between 0 and 1
- Set up models so prior expectation of correlations is 0
- Goal: to be weakly informative about correlations and variances
- Scaled inverse-Wishart model uses redundant parameterization
Weakly informative priors for covariance matrices

- Inverse-Wishart has problems
- Correlations can be between 0 and 1
- Set up models so prior expectation of correlations is 0
- Goal: to be weakly informative about correlations and variances
- Scaled inverse-Wishart model uses redundant parameterization
Weakly informative priors for population variation in a physiological model

- Pharmacokinetic parameters such as the “Michaelis-Menten coefficient”
- Wide uncertainty: prior guess for $\theta$ is 15 with a factor of 100 of uncertainty, $\log \theta \sim N(\log(15), \log(10)^2)$
- Population model: data on several people $j$, $\log \theta_j \sim N(\log(15), \log(10)^2)$
- Hierarchical prior distribution:
  - $\log \theta_j \sim N(\mu, \sigma^2)$, $\sigma \approx \log(2)$
  - $\mu \sim N(\log(15), \log(10)^2)$

- Weakly informative
Weakly informative priors for population variation in a physiological model

- Pharamcokinetic parameters such as the “Michaelis-Menten coefficient”
- Wide uncertainty: prior guess for $\theta$ is 15 with a factor of 100 of uncertainty, $\log \theta \sim N(\log(15), \log(10)^2)$
- Population model: data on several people $j$, $\log \theta_j \sim N(\log(15), \log(10)^2)$
- Hierarchical prior distribution:
  - $\log \theta_j \sim N(\mu, \sigma^2)$, $\sigma \approx \log(2)$
  - $\mu \sim N(\log(15), \log(10)^2)$
Weakly informative priors for population variation in a physiological model

- Pharamcokineti...“Michaelis-Menten coefficient”
- Wide uncertainty: prior guess for $\theta$ is 15 with a factor of 100 of uncertainty, $\log \theta \sim N(\log(15), \log(10)^2)$
- Population model: data on several people $j$, $\log \theta_j \sim N(\log(15), \log(10)^2)$
- Hierarchical prior distribution: $\log \theta_j \sim N(\mu, \sigma^2)$, $\sigma \approx \log(2), \mu \sim N(\log(15), \log(10)^2)$
- Weakly informative
Weakly informative priors for population variation in a physiological model

- Pharmacokinetic parameters such as the “Michaelis-Menten coefficient”
- Wide uncertainty: prior guess for $\theta$ is 15 with a factor of 100 of uncertainty, $\log \theta \sim N(\log(15), \log(10)^2)$
- Population model: data on several people $j$, $\log \theta_j \sim N(\log(15), \log(10)^2)$
- Hierarchical prior distribution:
  - $\log \theta_j \sim N(\mu, \sigma^2)$, $\sigma \approx \log(2)$
  - $\mu \sim N(\log(15), \log(10)^2)$
- Weakly informative
Pharamcokinetic parameters such as the “Michaelis-Menten coefficient”

Wide uncertainty: prior guess for $\theta$ is 15 with a factor of 100 of uncertainty, $\log \theta \sim N(\log(15), \log(10)^2)$

Population model: data on several people $j$, $\log \theta_j \sim N(\log(15), \log(10)^2)$

Hierarchical prior distribution:
- $\log \theta_j \sim N(\mu, \sigma^2)$, $\sigma \approx \log(2)$
- $\mu \sim N(\log(15), \log(10)^2)$

Weakly informative priors for population variation in a physiological model
Pharmacokinetic parameters such as the “Michaelis-Menten coefficient”

- Wide uncertainty: prior guess for $\theta$ is 15 with a factor of 100 of uncertainty, $\log(\theta) \sim N(\log(15), \log(10)^2)$
- Population model: data on several people $j$, $\log(\theta_j) \sim N(\log(15), \log(10)^2)$

Hierarchical prior distribution:
- $\log(\theta_j) \sim N(\mu, \sigma^2)$, $\sigma \approx \log(2)$
- $\mu \sim N(\log(15), \log(10)^2)$

Weakly informative priors for population variation in a physiological model
Weakly informative priors for population variation in a physiological model

- Pharamcokinetic parameters such as the “Michaelis-Menten coefficient”
- Wide uncertainty: prior guess for $\theta$ is 15 with a factor of 100 of uncertainty, $\log \theta \sim N(\log(15), \log(10)^2)$
- Population model: data on several people $j$, $\log \theta_j \sim N(\log(15), \log(10)^2)$
- Hierarchical prior distribution:
  - $\log \theta_j \sim N(\mu, \sigma^2)$, $\sigma \approx \log(2)$
  - $\mu \sim N(\log(15), \log(10)^2)$

Weakly informative

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default
Weakly informative priors for population variation in a physiological model

- Pharmacokinetic parameters such as the “Michaelis-Menten coefficient”
- Wide uncertainty: prior guess for $\theta$ is 15 with a factor of 100 of uncertainty, $\log \theta \sim N(\log(15), \log(10)^2)$
- Population model: data on several people $j$, $\log \theta_j \sim N(\log(15), \log(10)^2)$
- Hierarchical prior distribution:
  - $\log \theta_j \sim N(\mu, \sigma^2)$, $\sigma \approx \log(2)$
  - $\mu \sim N(\log(15), \log(10)^2)$
- Weakly informative
Weakly informative priors for mixture models

- Well-known problem of fitting the mixture model likelihood
- The maximum likelihood fits are weird, with a single point taking half the mixture
- Bayes with flat prior is just as bad
- These solutions don't "look" like mixtures
- There must be additional prior information—or, to put it another way, regularization
- Simple constraints, for example, a prior dist on the variance ratio
- Weakly informative

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default
Well-known problem of fitting the mixture model likelihood

The maximum likelihood fits are weird, with a single point taking half the mixture

Bayes with flat prior is just as bad

These solutions don’t “look” like mixtures

There must be additional prior information—or, to put it another way, regularization

Simple constraints, for example, a prior dist on the variance ratio

Weakly informative
Weakly informative priors for mixture models

- Well-known problem of fitting the mixture model likelihood
- The maximum likelihood fits are weird, with a single point taking half the mixture
- Bayes with flat prior is just as bad
- These solutions don’t “look” like mixtures
- There must be additional prior information—or, to put it another way, regularization
- Simple constraints, for example, a prior dist on the variance ratio
- Weakly informative
Weakly informative priors for mixture models

- Well-known problem of fitting the mixture model likelihood
- The maximum likelihood fits are weird, with a single point taking half the mixture
- Bayes with flat prior is just as bad
- These solutions don’t “look” like mixtures
- There must be additional prior information—or, to put it another way, regularization
- Simple constraints, for example, a prior dist on the variance ratio
- Weakly informative

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default
Weakly informative priors for mixture models

- Well-known problem of fitting the mixture model likelihood
- The maximum likelihood fits are weird, with a single point taking half the mixture
- Bayes with flat prior is just as bad
- These solutions don’t “look” like mixtures
- There must be additional prior information—or, to put it another way, regularization
- Simple constraints, for example, a prior dist on the variance ratio
- Weakly informative
Weakly informative priors for mixture models

- Well-known problem of fitting the mixture model likelihood
- The maximum likelihood fits are weird, with a single point taking half the mixture
- Bayes with flat prior is just as bad
- These solutions don’t “look” like mixtures
- There must be additional prior information—or, to put it another way, regularization
- Simple constraints, for example, a prior dist on the variance ratio
- Weakly informative
Weakly informative priors for mixture models

- Well-known problem of fitting the mixture model likelihood
- The maximum likelihood fits are weird, with a single point taking half the mixture
- Bayes with flat prior is just as bad
- These solutions don’t “look” like mixtures
- There must be additional prior information—or, to put it another way, regularization
- Simple constraints, for example, a prior dist on the variance ratio

Weakly informative
Weakly informative priors for mixture models

- Well-known problem of fitting the mixture model likelihood
- The maximum likelihood fits are weird, with a single point taking half the mixture
- Bayes with flat prior is just as bad
- These solutions don’t “look” like mixtures
- There must be additional prior information—or, to put it another way, regularization
- Simple constraints, for example, a prior dist on the variance ratio
- Weakly informative
Intentional underpooling in hierarchical models

- Basic hierarchical model:
  - Data $y_j$ on parameters $\theta_j$
  - Group-level model $\theta_j \sim N(\mu, \tau^2)$
  - No-pooling estimate $\hat{\theta}_j = y_j$
  - Bayesian partial-pooling estimate $E(\theta_j|y)$
  - Weak Bayes estimate: same as Bayes, but replacing $\tau$ with $2\tau$

- An example of the "incompatible Gibbs" algorithm
- Why would we do this??
Intentional underpooling in hierarchical models

Basic hierarchical model:
- Data $y_j$ on parameters $\theta_j$
- Group-level model $\theta_j \sim N(\mu, \tau^2)$
- No-pooling estimate $\hat{\theta}_j = y_j$
- Bayesian partial-pooling estimate $E(\theta_j | y)$
- Weak Bayes estimate: same as Bayes, but replacing $\tau$ with $2\tau$
- An example of the “incompatible Gibbs” algorithm
- Why would we do this??

Gelman, Jakulin, Pittau, Su
Bayesian generalized linear models and an appropriate default
Intentional underpooling in hierarchical models

- Basic hierarchical model:
  - Data $y_j$ on parameters $\theta_j$
  - Group-level model $\theta_j \sim \mathcal{N}(\mu, \tau^2)$
  - No-pooling estimate $\hat{\theta}_j = y_j$
  - Bayesian partial-pooling estimate $E(\theta_j|y)$
- Weak Bayes estimate: same as Bayes, but replacing $\tau$ with $2\tau$
- An example of the “incompatible Gibbs” algorithm
- Why would we do this??
Intentional underpooling in hierarchical models

- Basic hierarchical model:
  - Data $y_j$ on parameters $\theta_j$
  - Group-level model $\theta_j \sim \mathcal{N}(\mu, \tau^2)$
  - No-pooling estimate $\hat{\theta}_j = y_j$
  - Bayesian partial-pooling estimate $E(\theta_j | y)$

- Weak Bayes estimate: same as Bayes, but replacing $\tau$ with $2\tau$
- An example of the “incompatible Gibbs” algorithm
- Why would we do this??

Gelman, Jakulin, Pittau, Su
Bayesian generalized linear models and an appropriate default prior
Intentional underpooling in hierarchical models

- Basic hierarchical model:
  - Data $y_j$ on parameters $\theta_j$
  - Group-level model $\theta_j \sim N(\mu, \tau^2)$
  - No-pooling estimate $\hat{\theta}_j = y_j$
  - Bayesian partial-pooling estimate $E(\theta_j | y)$

- Weak Bayes estimate: same as Bayes, but replacing $\tau$ with $2\tau$

- An example of the “incompatible Gibbs” algorithm

- Why would we do this??

Gelman, Jakulin, Pittau, Su

Bayesian generalized linear models and an appropriate default prior
Intentional underpooling in hierarchical models

- Basic hierarchical model:
  - Data $y_j$ on parameters $\theta_j$
  - Group-level model $\theta_j \sim \text{N}(\mu, \tau^2)$
  - No-pooling estimate $\hat{\theta}_j = y_j$
  - Bayesian partial-pooling estimate $\mathbb{E}(\theta_j|y)$

- Weak Bayes estimate: same as Bayes, but replacing $\tau$ with $2\tau$
- An example of the “incompatible Gibbs” algorithm
- Why would we do this??
Intentional underpooling in hierarchical models

- Basic hierarchical model:
  - Data $y_j$ on parameters $\theta_j$
  - Group-level model $\theta_j \sim N(\mu, \tau^2)$
  - No-pooling estimate $\hat{\theta}_j = y_j$
  - Bayesian partial-pooling estimate $E(\theta_j | y)$

- Weak Bayes estimate: same as Bayes, but replacing $\tau$ with $2\tau$

- An example of the “incompatible Gibbs” algorithm

- Why would we do this??
Intentional underpooling in hierarchical models

- Basic hierarchical model:
  - Data $y_j$ on parameters $\theta_j$
  - Group-level model $\theta_j \sim N(\mu, \tau^2)$
  - No-pooling estimate $\hat{\theta}_j = y_j$
  - Bayesian partial-pooling estimate $E(\theta_j | y)$
- Weak Bayes estimate: same as Bayes, but replacing $\tau$ with $2\tau$
- An example of the “incompatible Gibbs” algorithm
- Why would we do this??

Gelman, Jakulin, Pittau, Su
Bayesian generalized linear models and an appropriate default prior
Intentional underpooling in hierarchical models

- Basic hierarchical model:
  - Data $y_j$ on parameters $\theta_j$
  - Group-level model $\theta_j \sim N(\mu, \tau^2)$
  - No-pooling estimate $\hat{\theta}_j = y_j$
  - Bayesian partial-pooling estimate $E(\theta_j | y)$

- Weak Bayes estimate: same as Bayes, but replacing $\tau$ with $2\tau$

- An example of the “incompatible Gibbs” algorithm

- Why would we do this??