Weakly informative priors

Andrew Gelman and Aleks Jakulin
Department of Statistics and Department of Political Science
Columbia University

3 Mar 2007

Themes

- Informative, noninformative, and weakly informative priors
- The sociology of shrinkage, or conservatism of Bayesian inference
- Collaborators

Themes

- Informative, noninformative, and weakly informative priors
- The sociology of shrinkage, or
  conservatism of Bayesian inference
- Collaborators

  Yu-Sung Su (Dept of Poli Sci, City Univ of New York)
  Masanao Yajima (Dept of Statistics, Columbia Univ)
  Maria Grazia Pittau (Dept of Economics, Univ of Rome)
  Gary King (Dept of Government, Harvard Univ)
  Samantha Cook (Statistics group, Google)
  Francis Tuerlinckx (Dept of Psychology, Univ of Leuven)

Themes

- Informative, noninformative, and weakly informative priors
- The sociology of shrinkage, or conservatism of Bayesian inference
- Collaborators
  - Yu-Sung Su (Dept of Poli Sci, City Univ of New York)
  - Masanao Yajima (Dept of Statistics, Columbia Univ)
  - Maria Grazia Pittau (Dept of Economics, Univ of Rome)
  - Gary King (Dept of Government, Harvard Univ)
  - Samantha Cook (Statistics group, Google)
  - Francis Tuerlinckx (Dept of Psychology, Univ of Leuven)

Themes

- Informative, noninformative, and weakly informative priors
- The sociology of shrinkage, or conservatism of Bayesian inference
- Collaborators
  - Yu-Sung Su (Dept of Poli Sci, City Univ of New York)
  - Masanao Yajima (Dept of Statistics, Columbia Univ)
  - Maria Grazia Pittau (Dept of Economics, Univ of Rome)
  - Gary King (Dept of Government, Harvard Univ)
  - Samantha Cook (Statistics group, Google)
  - Francis Tuerlinckx (Dept of Psychology, Univ of Leuven)

Themes

- Informative, noninformative, and weakly informative priors
- The sociology of shrinkage, or conservatism of Bayesian inference
- Collaborators
  - Yu-Sung Su (Dept of Poli Sci, City Univ of New York)
  - Masanao Yajima (Dept of Statistics, Columbia Univ)
  - Maria Grazia Pittau (Dept of Economics, Univ of Rome)
  - Gary King (Dept of Government, Harvard Univ)
  - Samantha Cook (Statistics group, Google)
  - Francis Tuerlinckx (Dept of Psychology, Univ of Leuven)

Themes

- Informative, noninformative, and weakly informative priors
- The sociology of shrinkage, or
  conservatism of Bayesian inference
- Collaborators
  - Yu-Sung Su (Dept of Poli Sci, City Univ of New York)
  - Masanao Yajima (Dept of Statistics, Columbia Univ)
  - Maria Grazia Pittau (Dept of Economics, Univ of Rome)
  - Gary King (Dept of Government, Harvard Univ)
  - Samantha Cook (Statistics group, Google)
  - Francis Tuerlinckx (Dept of Psychology, Univ of Leuven)

Themes

- Informative, noninformative, and weakly informative priors
- The sociology of shrinkage, or
  conservatism of Bayesian inference
- Collaborators
  - Yu-Sung Su (Dept of Poli Sci, City Univ of New York)
  - Masanao Yajima (Dept of Statistics, Columbia Univ)
  - Maria Grazia Pittau (Dept of Economics, Univ of Rome)
  - Gary King (Dept of Government, Harvard Univ)
  - Samantha Cook (Statistics group, Google)
  - Francis Tuerlinckx (Dept of Psychology, Univ of Leuven)

Themes

- Informative, noninformative, and weakly informative priors
- The sociology of shrinkage, or conservatism of Bayesian inference
- Collaborators
  - Yu-Sung Su (Dept of Poli Sci, City Univ of New York)
  - Masanao Yajima (Dept of Statistics, Columbia Univ)
  - Maria Grazia Pittau (Dept of Economics, Univ of Rome)
  - Gary King (Dept of Government, Harvard Univ)
  - Samantha Cook (Statistics group, Google)
  - Francis Tuerlinckx (Dept of Psychology, Univ of Leuven)

Themes

- Informative, noninformative, and weakly informative priors
- The sociology of shrinkage, or
  conservatism of Bayesian inference
- Collaborators
  - Yu-Sung Su (Dept of Poli Sci, City Univ of New York)
  - Masanao Yajima (Dept of Statistics, Columbia Univ)
  - Maria Grazia Pittau (Dept of Economics, Univ of Rome)
  - Gary King (Dept of Government, Harvard Univ)
  - Samantha Cook (Statistics group, Google)
  - Francis Tuerlinckx (Dept of Psychology, Univ of Leuven)

Themes

- Informative, noninformative, and weakly informative priors
- The sociology of shrinkage, or
  conservatism of Bayesian inference
- Collaborators
  - Yu-Sung Su (Dept of Poli Sci, City Univ of New York)
  - Masanao Yajima (Dept of Statistics, Columbia Univ)
  - Maria Grazia Pittau (Dept of Economics, Univ of Rome)
  - Gary King (Dept of Government, Harvard Univ)
  - Samantha Cook (Statistics group, Google)
  - Francis Tuerlinckx (Dept of Psychology, Univ of Leuven)

Themes

- Informative, noninformative, and weakly informative priors
- The sociology of shrinkage, or conservatism of Bayesian inference
- Collaborators
  - Yu-Sung Su (Dept of Poli Sci, City Univ of New York)
  - Masanao Yajima (Dept of Statistics, Columbia Univ)
  - Maria Grazia Pittau (Dept of Economics, Univ of Rome)
  - Gary King (Dept of Government, Harvard Univ)
  - Samantha Cook (Statistics group, Google)
  - Francis Tuerlinckx (Dept of Psychology, Univ of Leuven)

What does this have to do with MCMC?

- I'm speaking at Jun Liu's MCMC conference
- We don't have to be trapped by decades-old models
- The folk theorem about computation and modeling
- The example of BUGS

What does this have to do with MCMC?

- I'm speaking at Jun Liu's MCMC conference
- We don't have to be trapped by decades-old models
- The folk theorem about computation and modeling
- The example of BUGS

What does this have to do with MCMC?

- I'm speaking at Jun Liu's MCMC conference
- We don't have to be trapped by decades-old models
- The folk theorem about computation and modeling
- The example of BUGS

What does this have to do with MCMC?

- I'm speaking at Jun Liu's MCMC conference
- We don't have to be trapped by decades-old models
- The folk theorem about computation and modeling
- The example of BUGS

What does this have to do with MCMC?

- I'm speaking at Jun Liu's MCMC conference
- We don't have to be trapped by decades-old models
- The folk theorem about computation and modeling
- The example of BUGS

Information in prior distributions

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

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 θ
- Weakly informative prior dist

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 θ
- Weakly informative prior dist

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 θ
- Weakly informative prior dist

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

Weakly informative priors: some examples

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

Weakly informative priors: some examples

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

Weakly informative priors: some examples

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

Weakly informative priors: some examples

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

Weakly informative priors: some examples

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

Weakly informative priors: some examples

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

Weakly informative priors: some examples

- Variance parameters
- Covariance matrices
- Logistic regression coefficients
- 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 τ

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 τ

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 τ

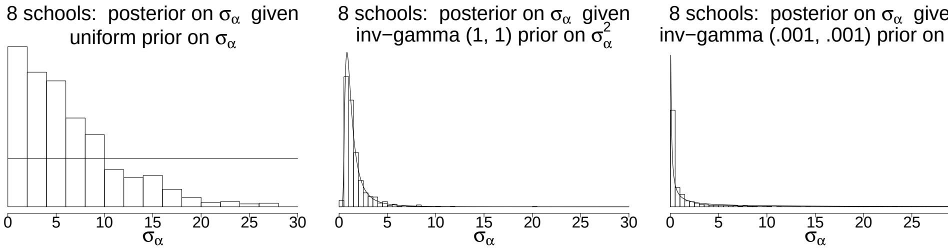
Weakly informative priors for variance parameter

- 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 τ

Weakly informative priors for variance parameter

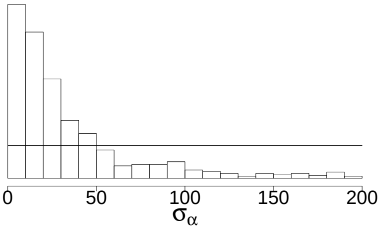
- Basic hierarchical model
- Traditional inverse-gamma$(0.001, 0.001)$ prior can be highly informative (in a bad way)!
- Noninformative uniform prior works better
- But if #groups is small ($J = 2$, 3, even 5), a weakly informative prior helps by shutting down huge values of τ

Priors for variance parameter: $J = 8$ goups

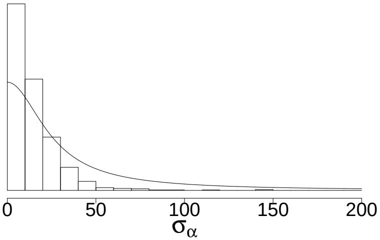


Priors for variance parameter: $J = 3$ groups

3 schools: posterior on σ_α given uniform prior on σ_α



3 schools: posterior on σ_α given half−Cauchy (25) prior on σ_α



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

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

Separation in logistic regression

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

1960                              1968
            coef.est coef.se                  coef.est coef.se
(Intercept) -0.14     0.23        (Intercept)  0.47     0.24
female       0.24     0.14        female      -0.01     0.15
black       -1.03     0.36        black       -3.64     0.59
income       0.03     0.06        income      -0.03     0.07

1964                              1972
            coef.est coef.se                  coef.est coef.se
(Intercept)  -1.15     0.22       (Intercept)  0.67     0.18
female       -0.09     0.14       female      -0.25     0.12
black       -16.83   420.40       black       -2.63     0.27
income        0.19     0.06       income       0.09     0.05
```

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

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

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`

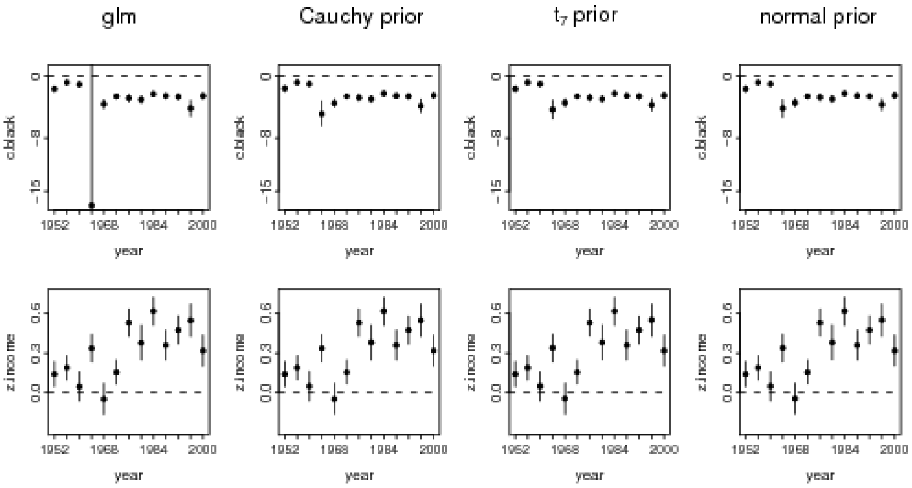
Weakly informative priors for logistic regression coefficients

- 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`

Regularization in action!



Weakly informative priors for population variation in a physiological model

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

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

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 θ_j
  - Group-level model $\theta_j \sim \mathrm{N}(\mu, \tau^2)$
  - No-pooling estimate $\hat{\theta}_j = y_j$
  - Bayesian partial-pooling estimate $E(\theta_j|y)$
- Weak Bayes estimate: same as Bayes, but replacing τ with 2τ
- An example of the "inconsistent Gibbs" algorithm
- Why would we do this??

Intentional underpooling in hierarchical models

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

Intentional underpooling in hierarchical models

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

Intentional underpooling in hierarchical models

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

Intentional underpooling in hierarchical models

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

Intentional underpooling in hierarchical models

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

Intentional underpooling in hierarchical models

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

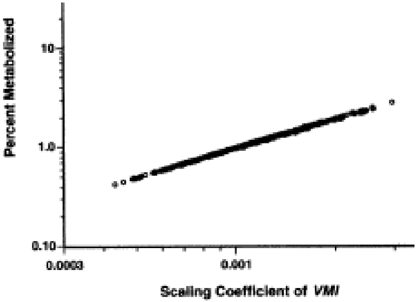
Intentional underpooling in hierarchical models

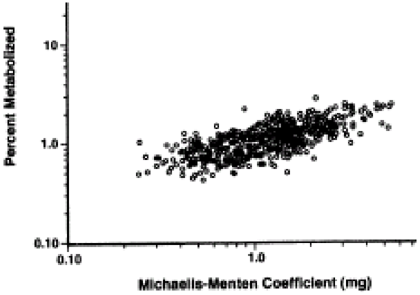
- Basic hierarchical model:
  - Data y_j on parameters θ_j
  - Group-level model $\theta_j \sim \mathrm{N}(\mu, \tau^2)$
  - No-pooling estimate $\hat{\theta}_j = y_j$
  - Bayesian partial-pooling estimate $\mathrm{E}(\theta_j|y)$
- Weak Bayes estimate: same as Bayes, but replacing τ with 2τ
- An example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

Intentional underpooling in hierarchical models

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

Static sensitivity analysis: what happens if we add prior information?





Conservatism of Bayesian inference

- Consider the logistic regression example
- 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

- Consider the logistic regression example
- 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

- Consider the logistic regression example
- 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

- Consider the logistic regression example
- 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

- Consider the logistic regression example
- 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

- Consider the logistic regression example
- 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

- Consider the logistic regression example
- 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

Another example

Dose	#deaths/#animals
−0.86	0/5
−0.30	1/5
−0.05	3/5
0.73	5/5

- Slope of a logistic regression of Pr(death) on dose:
  - Maximum likelihood est is 7.8 ± 4.9
  - With weakly-informative prior: Bayes est is 4.4 ± 1.9
- Which is truly conservative?
- The sociology of shrinkage

Another example

Dose	#deaths/#animals
−0.86	0/5
−0.30	1/5
−0.05	3/5
0.73	5/5

- Slope of a logistic regression of Pr(death) on dose:
  - Maximum likelihood est is 7.8 ± 4.9
  - With weakly-informative prior: Bayes est is 4.4 ± 1.9
- Which is truly conservative?
- The sociology of shrinkage

Another example

Dose	#deaths/#animals
-0.86	$0/5$
-0.30	$1/5$
-0.05	$3/5$
0.73	$5/5$

- Slope of a logistic regression of Pr(death) on dose:
  - Maximum likelihood est is 7.8 ± 4.9
  - With weakly-informative prior: Bayes est is 4.4 ± 1.9
- Which is truly conservative?
- The sociology of shrinkage

Another example

Dose	#deaths/#animals
−0.86	0/5
−0.30	1/5
−0.05	3/5
0.73	5/5

- Slope of a logistic regression of Pr(death) on dose:
  - Maximum likelihood est is 7.8 ± 4.9
  - With weakly-informative prior: Bayes est is 4.4 ± 1.9
- Which is truly conservative?
- The sociology of shrinkage

Another example

Dose	#deaths/#animals
−0.86	0/5
−0.30	1/5
−0.05	3/5
0.73	5/5

- Slope of a logistic regression of Pr(death) on dose:
  - Maximum likelihood est is 7.8 ± 4.9
  - With weakly-informative prior: Bayes est is 4.4 ± 1.9
- Which is truly conservative?
- The sociology of shrinkage

Another example

Dose	#deaths/#animals
−0.86	0/5
−0.30	1/5
−0.05	3/5
0.73	5/5

- Slope of a logistic regression of Pr(death) on dose:
  - Maximum likelihood est is 7.8 ± 4.9
  - With weakly-informative prior: Bayes est is 4.4 ± 1.9
- Which is truly conservative?
- The sociology of shrinkage

A hierarchical framework

- Consider many possible datasets
- The "true prior" is the distribution of β'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?

A hierarchical framework

- Consider many possible datasets
- The "true prior" is the distribution of β'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?

A hierarchical framework

- Consider many possible datasets
- The "true prior" is the distribution of β'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?

A hierarchical framework

- Consider many possible datasets
- The "true prior" is the distribution of β'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?

A hierarchical framework

- Consider many possible datasets
- The "true prior" is the distribution of β'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?

A hierarchical framework

- Consider many possible datasets
- The "true prior" is the distribution of β'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 cross-validation and average predictive error
- The optimal prior distribution for β'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 cross-validation and average predictive error
- The optimal prior distribution for β'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 cross-validation and average predictive error
- The optimal prior distribution for β's is (approx) Cauchy $(0, 1)$
- Our Cauchy $(0, 2.5)$ prior distribution is weakly informative!

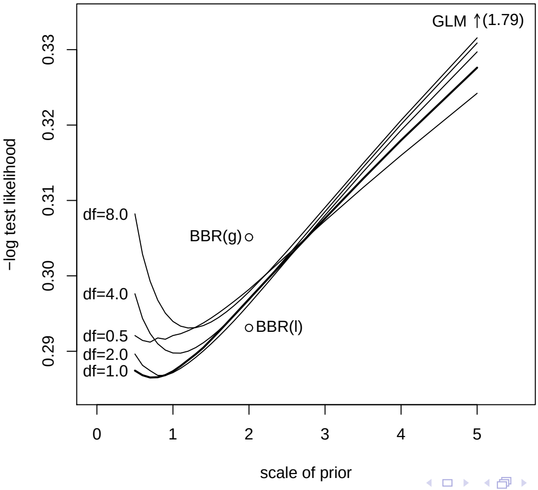
Evaluation using a corpus of datasets

- Compare classical glm to Bayesian estimates using various prior distributions
- Evaluate using cross-validation and average predictive error
- The optimal prior distribution for β'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 cross-validation and average predictive error
- The optimal prior distribution for β'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



Conclusion

- "Noninformative priors" are really weakly informative
- "Weakly informative" is a more general and useful concept
- Regularization

Conclusion

- “Noninformative priors” are really weakly informative
- “Weakly informative” is a more general and useful concept
- Regularization
  - Better inferences
  - Stability of computation

Conclusion

- "Noninformative priors" are really weakly informative
- "Weakly informative" is a more general and useful concept
- Regularization
  - Better inferences
  - Stability of computation

Conclusion

- "Noninformative priors" are really weakly informative
- "Weakly informative" is a more general and useful concept
- Regularization
  - Better inferences
  - Stability of computation

Conclusion

- "Noninformative priors" are really weakly informative
- "Weakly informative" is a more general and useful concept
- Regularization
  - Better inferences
  - Stability of computation

Conclusion

- “Noninformative priors” are really weakly informative
- “Weakly informative” is a more general and useful concept
- Regularization
  - Better inferences
  - Stability of computation

References: our work

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515–533.

Gelman, A., Bois, F., and Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association* **91**, 1400–1412.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. London: CRC Press.

Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., and Jakulin, A. (2007). Bayes: radical, liberal, or conservative? *Statistica Sinica* **17**, 422–426.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y. S. (2007). A default prior distribution for logistic and other regression models. Technical report, Department of Statistics, Columbia University.

Gelman, A., and King, G. (1990). Estimating the electoral consequences of legislative redistricting. *Journal of the American Statistical Association* **85**, 274–282.

Gelman, A., and Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* **15**, 373–390.

References: work of others

Cook, S. R., and Rubin, D. B. (2006). Constructing vague but proper prior distributions in complex Bayesian models. Technical report, Department of Statistics, Harvard University.

Greenland, S. (2007). Bayesian perspectives for epidemiological research. II. Regression analysis. *International Journal of Epidemiology* **36**, 1–8.

MacLehose, R. F., Dunson, D. B., Herring, A. H., and Hoppin, J. A. (2007). Bayesian methods for highly correlated exposure data. *Epidemiology* **18**, 199–207.

O'Malley, A. J., and Zaslavsky, A. M. (2005). Cluster-level covariance analysis for survey data with structured nonresponse. Technical report, Department of Health Care Policy, Harvard Medical School.

Thomas, D. C., Witte, J. S., and Greenland, S. (2007). Dissecting effects of complex mixtures: who's afraid of informative priors? *Epidemiology* **18**, 186–190.