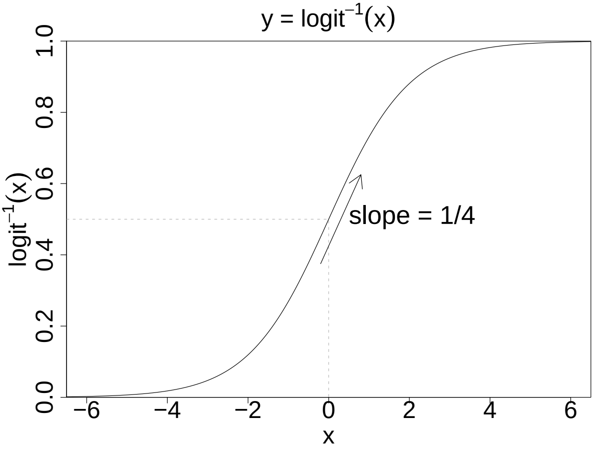
Bayesian generalized linear models and an appropriate default prior

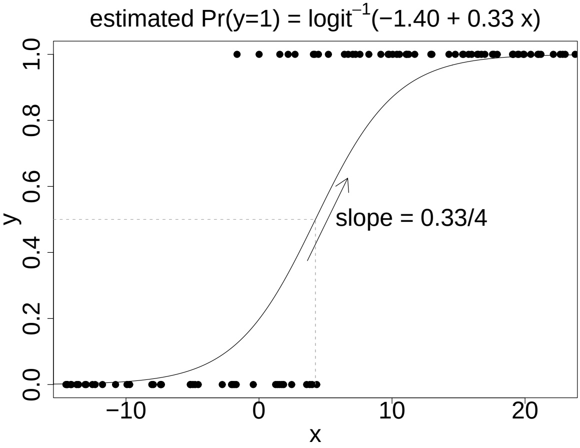
Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su
Columbia University

14 August 2008

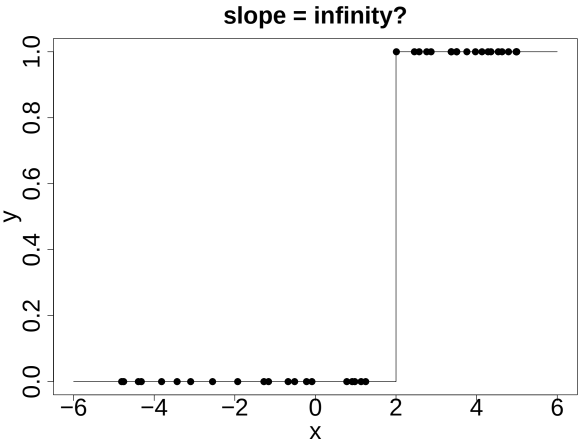
Logistic regression



A clean example



The problem of separation



Separation is no joke!

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

1960

	coef.est	coef.se
(Intercept)	-0.14	0.23
female	0.24	0.14
black	-1.03	0.36
income	0.03	0.06

1968

	coef.est	coef.se
(Intercept)	0.47	0.24
female	-0.01	0.15
black	-3.64	0.59
income	-0.03	0.07

1964

	coef.est	coef.se
(Intercept)	-1.15	0.22
female	-0.09	0.14
black	-16.83	420.40
income	0.19	0.06

1972

	coef.est	coef.se
(Intercept)	0.67	0.18
female	-0.25	0.12
black	-2.63	0.27
income	0.09	0.05

bayesglm()

- 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!)

`bayesglm()`

- 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!)

bayesglm()

- 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!)

bayesglm()

- 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!)

bayesglm()

- 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!)

bayesglm()

- 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!)

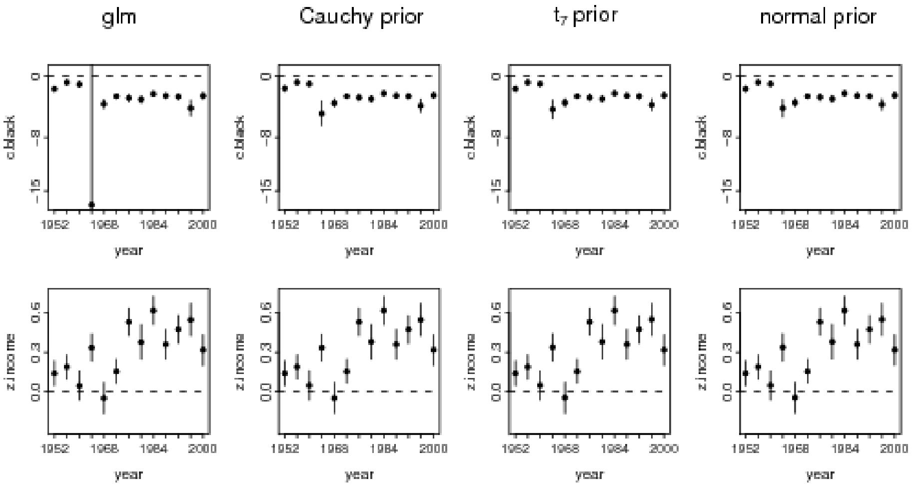
`bayesglm()`

- 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!)

`bayesglm()`

- 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!)

Regularization in action!



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

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

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

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: 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: 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: 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: 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: regularization, stabilization

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`

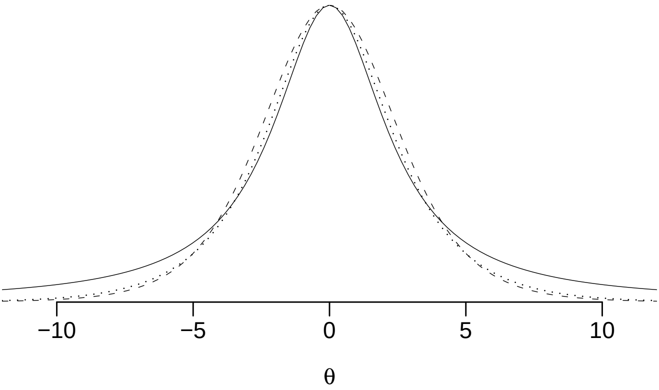
Weakly informative priors for logistic regression coefficients

- 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`

Prior distributions



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	$5/5$

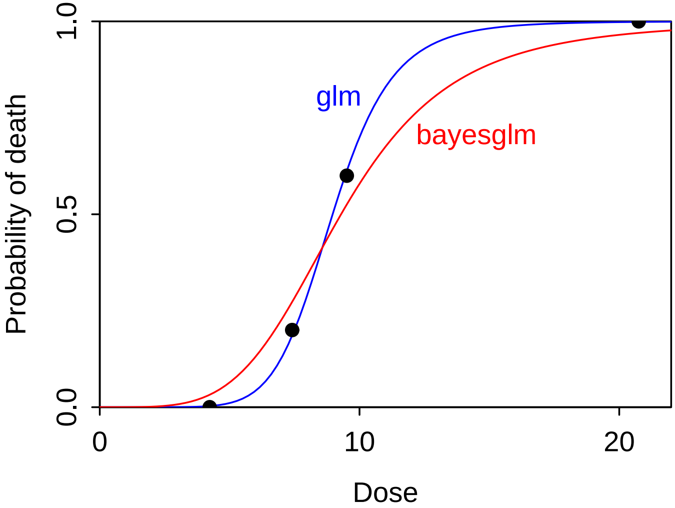
- Slope of a logistic regression of Pr(death) on dose:
  - Maximum likelihood est is 7.8 ± 4.9
  - With weakly-informative prior: Bayes est is 4.4 ± 1.9
- Which is truly conservative?
- The sociology of shrinkage

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

Maximum likelihood and Bayesian estimates



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

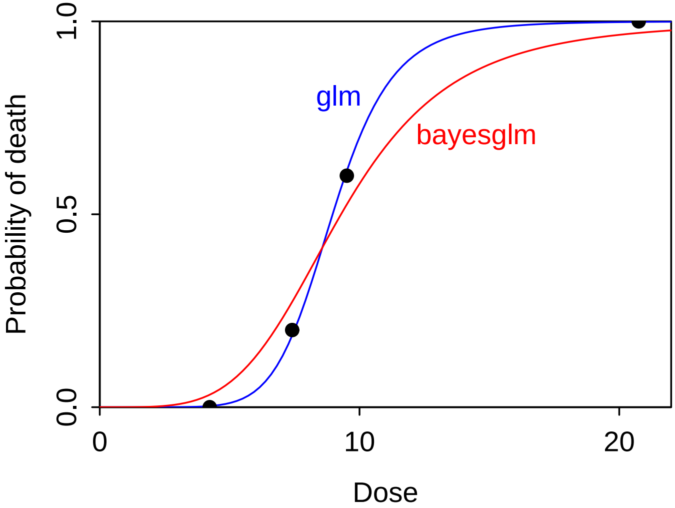
Conservatism of Bayesian inference

- 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?



Prior as population distribution

- 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?

Prior as population distribution

- 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?

Prior as population distribution

- 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?

Prior as population distribution

- 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?

Prior as population distribution

- 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?

Prior as population distribution

- 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 5-fold 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 5-fold 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 5-fold 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!

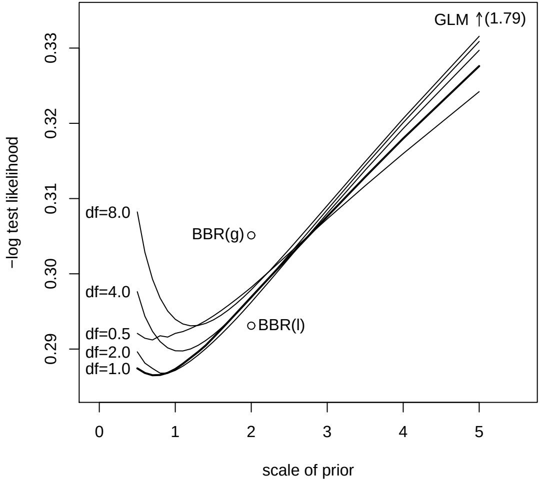
Evaluation using a corpus of datasets

- Compare classical glm to Bayesian estimates using various prior distributions
- Evaluate using 5-fold 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 5-fold 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



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?

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?

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 (`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

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 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

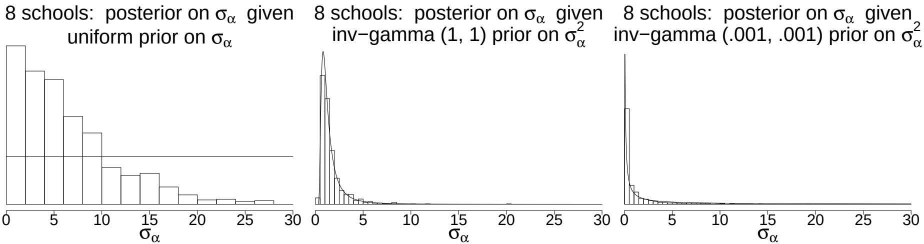
Weakly informative priors for variance parameter

- 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

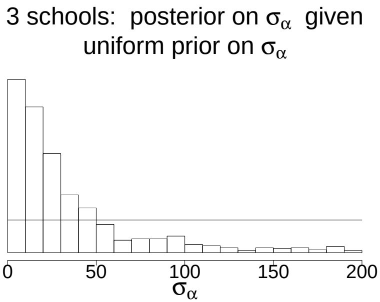
Weakly informative priors for variance parameter

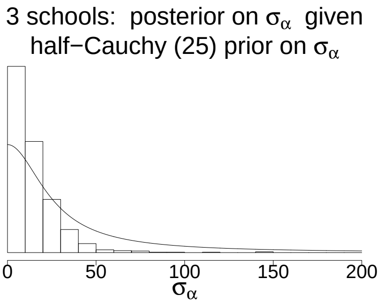
- Basic hierarchical model
- Traditional inverse-gamma$(0.001, 0.001)$ prior can be highly informative (in a bad way)!
- Noninformative uniform prior works better
- But if #groups is small ($J = 2$, 3, even 5), a weakly informative prior helps by shutting down huge values of τ

Priors for variance parameter: $J = 8$ groups



Priors for variance parameter: $J = 3$ groups





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

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 $\mathrm{E}(\theta_j|y)$
- Weak Bayes estimate: same as Bayes, but replacing τ with 2τ
- 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 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 "incompatible 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 "incompatible 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 "incompatible 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 "incompatible 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 "incompatible 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 "incompatible 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 "incompatible 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 "incompatible Gibbs" algorithm
- Why would we do this??