Introduction to Bayesian Modeling of Epidemiologic Data

David Dunson[1], Amy Herring[2] & Rich MacLehose[1]

[1]National Institute of Environmental Health Sciences, NIH & [2]Department of Biostatistics, UNC at Chapel Hill

June 19, 2007

Outline of Workshop

1. Introduction to Bayesian modeling (*David Dunson*)
2. Bayesian modeling in SAS (*Amy Herring*)
3. Hierarchical models (*Rich MacLehose*)

Background on Perchlorate

- Contaminant found in groundwater, drinking water & soils - mainly in southwest US

Background on Perchlorate

- Contaminant found in groundwater, drinking water & soils - mainly in southwest US
- Primary source industrial & military - perchlorate used as oxidizing agent (e.g., in rocket fuel)

Background on Perchlorate

- Contaminant found in groundwater, drinking water & soils - mainly in southwest US
- Primary source industrial & military - perchlorate used as oxidizing agent (e.g., in rocket fuel)
- Concern about effects of perchlorate on the thyroid (*known to inhibit thyroid's ability to absorb iodine from the blood*)

Background on Perchlorate

- Contaminant found in groundwater, drinking water & soils - mainly in southwest US
- Primary source industrial & military - perchlorate used as oxidizing agent (e.g., in rocket fuel)
- Concern about effects of perchlorate on the thyroid (*known to inhibit thyroid's ability to absorb iodine from the blood*)
- EPA conducted extensive risk assessment → NAS review of health effects (*recommended new reference dose*)

Two Generation Rodent Study (Argus, 1999)

- Male rats were exposed to ammonium perchlorate through drinking water

Two Generation Rodent Study (Argus, 1999)

- Male rats were exposed to ammonium perchlorate through drinking water
- 30 rats/group with doses of 0, 0.01, 0.1, 1.0 & 30 mg/kg/day

Two Generation Rodent Study (Argus, 1999)

- Male rats were exposed to ammonium perchlorate through drinking water
- 30 rats/group with doses of 0, 0.01, 0.1, 1.0 & 30 mg/kg/day
- Male rats in P1 generation exposed 70+ days before mating, through mating period & until sacrifice at 21-22 weeks

Two Generation Rodent Study (Argus, 1999)

- Male rats were exposed to ammonium perchlorate through drinking water
- 30 rats/group with doses of 0, 0.01, 0.1, 1.0 & 30 mg/kg/day
- Male rats in P1 generation exposed 70+ days before mating, through mating period & until sacrifice at 21-22 weeks
- F1 generation treated similarly, with additional exposure during gestation & lactation

Two Generation Rodent Study (Argus, 1999)

- Male rats were exposed to ammonium perchlorate through drinking water
- 30 rats/group with doses of 0, 0.01, 0.1, 1.0 & 30 mg/kg/day
- Male rats in P1 generation exposed 70+ days before mating, through mating period & until sacrifice at 21-22 weeks
- F1 generation treated similarly, with additional exposure during gestation & lactation
- At 19 weeks for F1 rats, thyroid tissues examined histologically

Two Generation Rodent Study (Argus, 1999)

- Male rats were exposed to ammonium perchlorate through drinking water
- 30 rats/group with doses of 0, 0.01, 0.1, 1.0 & 30 mg/kg/day
- Male rats in P1 generation exposed 70+ days before mating, through mating period & until sacrifice at 21-22 weeks
- F1 generation treated similarly, with additional exposure during gestation & lactation
- At 19 weeks for F1 rats, thyroid tissues examined histologically
- 2/30 male rats in 30 mg/kg/day dose group had thyroid follicular cell adenomas, with one of these rats having two adenomas.

Analyzing the Perchlorate data

- Frequentist analysis: comparing 0/30 tumors in control rats with 2/30 tumors in the high dose group → non-significant (*Fisher's exact test p-value=0.49*)

Analyzing the Perchlorate data

- Frequentist analysis: comparing 0/30 tumors in control rats with 2/30 tumors in the high dose group → non-significant (*Fisher's exact test p-value=0.49*)
- Ignores the prior knowledge that thyroid follicular cell adenomas are very rare in 19 week rats

Analyzing the Perchlorate data

- Frequentist analysis: comparing 0/30 tumors in control rats with 2/30 tumors in the high dose group → non-significant (*Fisher's exact test p-value=0.49*)
- Ignores the prior knowledge that thyroid follicular cell adenomas are very rare in 19 week rats
- The National Toxicology Program (NTP) routinely collects tumor incidence data for control rats in two year studies.

Analyzing the Perchlorate data

- Frequentist analysis: comparing 0/30 tumors in control rats with 2/30 tumors in the high dose group $\longrightarrow$ non-significant (*Fisher's exact test p-value=0.49*)
- Ignores the prior knowledge that thyroid follicular cell adenomas are very rare in 19 week rats
- The National Toxicology Program (NTP) routinely collects tumor incidence data for control rats in two year studies.
- Would our conclusion change if we included information from the NTP data base?

Some prior information

- In 67 recent NTP studies, $38/3419 = 1.1\%$ of male rats developed thyroid follicular cell adenomas by death in a two year study.

Some prior information

- In 67 recent NTP studies, $38/3419 = 1.1\%$ of male rats developed thyroid follicular cell adenomas by death in a two year study.
- Results from Portier et al. (1986) suggest probability of developing thyroid follicular cell adenoma increases in proportion to $age^{4.78}$

Some prior information

- In 67 recent NTP studies, $38/3419 = 1.1\%$ of male rats developed thyroid follicular cell adenomas by death in a two year study.
- Results from Portier et al. (1986) suggest probability of developing thyroid follicular cell adenoma increases in proportion to $age^{4.78}$
- Average survival time in NTP study for control male rat is 95.2 weeks

Some prior information

- In 67 recent NTP studies, $38/3419 = 1.1\%$ of male rats developed thyroid follicular cell adenomas by death in a two year study.
- Results from Portier et al. (1986) suggest probability of developing thyroid follicular cell adenoma increases in proportion to $age^{4.78}$
- Average survival time in NTP study for control male rat is 95.2 weeks
- Suggests that the ratio of probability of thyroid follicular cell adenomas at 19 weeks to the lifetime probability in a 2-year study is $(19/95.2)^{4.78} = 0.0005$

Some prior information

- In 67 recent NTP studies, $38/3419 = 1.1\%$ of male rats developed thyroid follicular cell adenomas by death in a two year study.
- Results from Portier et al. (1986) suggest probability of developing thyroid follicular cell adenoma increases in proportion to $age^{4.78}$
- Average survival time in NTP study for control male rat is 95.2 weeks
- Suggests that the ratio of probability of thyroid follicular cell adenomas at 19 weeks to the lifetime probability in a 2-year study is $(19/95.2)^{4.78} = 0.0005$
- Question: How do we incorporate this information in analysis?

Frequentist vs Bayes

- Suppose we are interested in a parameter θ (*e.g., probability of thyroid FCA by 19 weeks in control rats*)

Frequentist vs Bayes

- Suppose we are interested in a parameter θ (*e.g., probability of thyroid FCA by 19 weeks in control rats*)
- Frequentists would typically rely on the MLE, which would be $\widehat{\theta} = 0/30 = 0$ in the perchlorate example

Frequentist vs Bayes

- Suppose we are interested in a parameter θ (*e.g., probability of thyroid FCA by 19 weeks in control rats*)
- Frequentists would typically rely on the MLE, which would be $\widehat{\theta} = 0/30 = 0$ in the perchlorate example
- Bayesians instead rely on the *posterior distribution* of θ

Frequentist vs Bayes

- Suppose we are interested in a parameter θ (*e.g., probability of thyroid FCA by 19 weeks in control rats*)
- Frequentists would typically rely on the MLE, which would be $\widehat{\theta} = 0/30 = 0$ in the perchlorate example
- Bayesians instead rely on the *posterior distribution* of θ
- Obtained in updating one's *prior distribution* with the likelihood for the data.

Bayes' Rule

- Let $\pi(\theta) =$ prior distribution of parameter θ

Bayes' Rule

- Let $\pi(\theta)$ = prior distribution of parameter θ
- Let $L(\mathbf{y}\,|\theta)$ = likelihood of data $\mathbf{y}$ given parameter θ

Bayes' Rule

- Let $\pi(\theta) =$ prior distribution of parameter θ
- Let $L(\mathbf{y}\,|\theta) =$ likelihood of data $\mathbf{y}$ given parameter θ
- Then, the posterior is defined as:

$$\pi(\theta \mid \mathbf{y}) = \frac{\pi(\theta)\, L(\mathbf{y}\,|\theta)}{\int \pi(\theta)\, L(\mathbf{y} \mid \theta) d\theta},$$

which is the prior $\times$ the likelihood divided by a normalizing constant

Bayes' Rule

- Let $\pi(\theta)$ = prior distribution of parameter θ
- Let $L(\mathbf{y}\,|\theta)$ = likelihood of data $\mathbf{y}$ given parameter θ
- Then, the posterior is defined as:

$$\pi(\theta \mid \mathbf{y}) = \frac{\pi(\theta)\, L(\mathbf{y}\,|\theta)}{\int \pi(\theta)\, L(\mathbf{y} \mid \theta) d\theta},$$

  which is the prior $\times$ the likelihood divided by a normalizing constant

- The posterior, $\pi(\theta \mid \mathbf{y})$, represents the state of knowledge about θ after *updating* the prior, $\pi(\theta)$, with the information in the data, $\mathbf{y}$.

Bayesian Updating

- As an example of *Bayesian updating*, let θ=probability of preterm birth (PTB)

Bayesian Updating

- As an example of *Bayesian updating*, let θ=probability of preterm birth (PTB)
- Typical choice of prior for θ is the beta(a, b) distribution

Bayesian Updating

- As an example of *Bayesian updating*, let θ=probability of preterm birth (PTB)
- Typical choice of prior for θ is the beta(a, b) distribution
- a, b=*hyperparameters* characterizing uncertainty in θ before incorporating information in data from current study

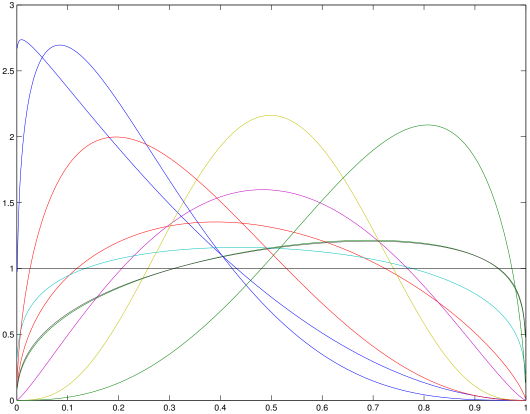
Bayesian Updating

- As an example of *Bayesian updating*, let θ=probability of preterm birth (PTB)
- Typical choice of prior for θ is the beta(a, b) distribution
- a, b=*hyperparameters* characterizing uncertainty in θ before incorporating information in data from current study
- $a/(a+b)$=prior expectation for θ & $a+b$=prior sample size

Bayesian Updating

- As an example of *Bayesian updating*, let θ=probability of preterm birth (PTB)
- Typical choice of prior for θ is the beta(a, b) distribution
- a, b=*hyperparameters* characterizing uncertainty in θ before incorporating information in data from current study
- $a/(a+b)$=prior expectation for θ & $a+b$=prior sample size
- beta$(1, 1)$ corresponds to uniform distribution $\rightarrow$ has as much information as two subjects (*one with PTB & one without*)

Beta prior distributions for different hyperparameters



Preterm Birth Example

- θ=probability of preterm birth

Preterm Birth Example

- θ=probability of preterm birth
- Consider two different priors: (1) a uniform prior expressing ignorance; and (2) a beta(10,90) prior.

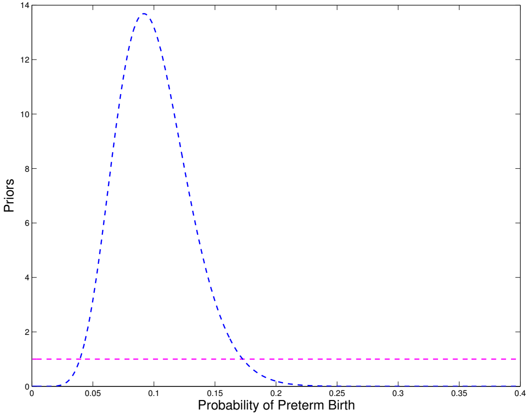
Preterm Birth Example

- θ=probability of preterm birth
- Consider two different priors: (1) a uniform prior expressing ignorance; and (2) a beta(10,90) prior.
- The beta(10,90) prior implies a 95% prior probability of $\theta \in [0.05, 0.17]$ (*wide range of plausible values for probability preterm birth*)

Preterm Birth Example

- θ=probability of preterm birth
- Consider two different priors: (1) a uniform prior expressing ignorance; and (2) a beta(10,90) prior.
- The beta(10,90) prior implies a 95% prior probability of $\theta \in [0.05, 0.17]$ (*wide range of plausible values for probability preterm birth*)
- We collect data for 100 women & observe 7/100 preterm births.

Different Priors



Updating the beta prior

- The beta prior is *conjugate* to the binomial likelihood

Updating the beta prior

- The beta prior is *conjugate* to the binomial likelihood
- For conjugate priors, the posterior $\pi(\theta \mid \mathbf{y})$ is available analytically and has the same form as the prior

Updating the beta prior

- The beta prior is *conjugate* to the binomial likelihood
- For conjugate priors, the posterior $\pi(\theta \mid \mathbf{y})$ is available analytically and has the same form as the prior
- Let $y_i = 1$ if woman i has a preterm birth and $y_i = 0$ otherwise, with $\Pr(y_i = 1) = \theta$

Updating the beta prior

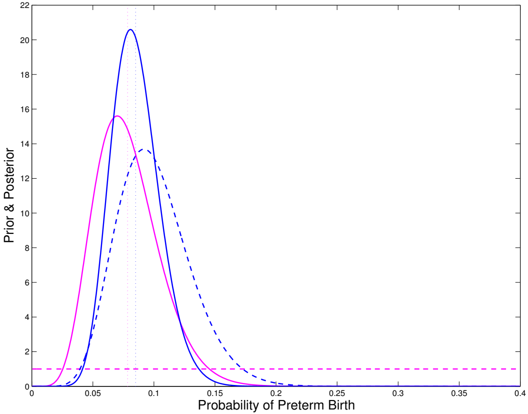
- The beta prior is *conjugate* to the binomial likelihood
- For conjugate priors, the posterior $\pi(\theta \mid \mathbf{y})$ is available analytically and has the same form as the prior
- Let $y_i = 1$ if woman i has a preterm birth and $y_i = 0$ otherwise, with $\Pr(y_i = 1) = \theta$
- Likelihood is Bernoulli: $L(\mathbf{y} \mid \theta) = \prod_i \theta^{y_i}(1-\theta)^{1-y_i}$

Updating the beta prior

- The beta prior is *conjugate* to the binomial likelihood
- For conjugate priors, the posterior $\pi(\theta \mid \mathbf{y})$ is available analytically and has the same form as the prior
- Let $y_i = 1$ if woman i has a preterm birth and $y_i = 0$ otherwise, with $\Pr(y_i = 1) = \theta$
- Likelihood is Bernoulli: $L(\mathbf{y} \mid \theta) = \prod_i \theta^{y_i}(1-\theta)^{1-y_i}$
- The posterior distribution of θ is then

$$\pi(\theta \mid \mathbf{y}) = \text{beta}\left(a + \sum_i y_i, b + \sum_i (1 - y_i)\right).$$

Prior and Posteriors



Returning to the Perchlorate Example

- Let $\theta = \gamma \times \rho$, θ=prob tumor in 19 weeks, γ=prob tumor in lifetime & ρ=proportion of tumors developing by 19 weeks

Returning to the Perchlorate Example

- Let $\theta = \gamma \times \rho$, θ=prob tumor in 19 weeks, γ=prob tumor in lifetime & ρ=proportion of tumors developing by 19 weeks
- We choose beta$(38, 3381)$ prior for probability of developing thyroid FCA for a control male rat in a two-year study (γ)

Returning to the Perchlorate Example

- Let $\theta = \gamma \times \rho$, θ=prob tumor in 19 weeks, γ=prob tumor in lifetime & ρ=proportion of tumors developing by 19 weeks
- We choose beta$(38, 3381)$ prior for probability of developing thyroid FCA for a control male rat in a two-year study (γ)
- Based on the $38/(38+3381)$ rats observed with these tumors in NTP studies

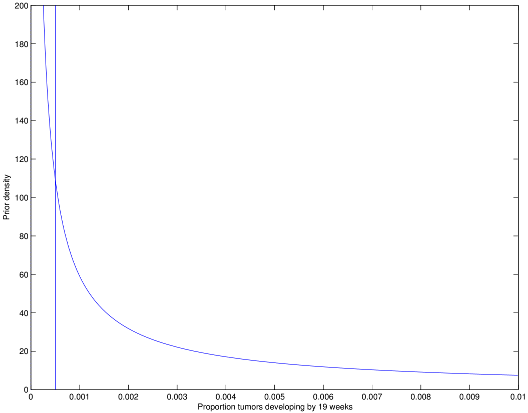
Returning to the Perchlorate Example

- Let $\theta = \gamma \times \rho$, θ=prob tumor in 19 weeks, γ=prob tumor in lifetime & ρ=proportion of tumors developing by 19 weeks
- We choose beta$(38, 3381)$ prior for probability of developing thyroid FCA for a control male rat in a two-year study (γ)
- Based on the $38/(38 + 3381)$ rats observed with these tumors in NTP studies
- We choose beta$(0.11, 2.6)$ prior for ratio:

$$\rho = \frac{\text{probability of developing tumor by 19 weeks}}{\text{probability of developing tumor in two year study}}.$$

Centered on 0.0005 with 95% probability of falling within [0.0000,0.379]

Prior for proportion of thyroid FCA by 19 weeks (ρ)



Conclusions from Perchlorate Example

- $\theta = \gamma \times \rho$=probability of developing thyroid FCA by 19 weeks for control male rat

Conclusions from Perchlorate Example

- $\theta = \gamma \times \rho$=probability of developing thyroid FCA by 19 weeks for control male rat
- We update priors for γ and ρ with data from the Argus (1999) study to obtain posterior distribution for θ.

Conclusions from Perchlorate Example

- $\theta = \gamma \times \rho$=probability of developing thyroid FCA by 19 weeks for control male rat
- We update priors for γ and ρ with data from the Argus (1999) study to obtain posterior distribution for θ.
- The posterior mean of θ is $1/100,000$

Conclusions from Perchlorate Example

- $\theta = \gamma \times \rho$=probability of developing thyroid FCA by 19 weeks for control male rat
- We update priors for γ and ρ with data from the Argus (1999) study to obtain posterior distribution for θ.
- The posterior mean of θ is $1/100,000$
- How likely it is to observe 2 or more rats out of 30 with tumors under the null hypothesis of no effect of perchlorate?

Conclusions from Perchlorate Example

- $\theta = \gamma \times \rho$=probability of developing thyroid FCA by 19 weeks for control male rat
- We update priors for γ and ρ with data from the Argus (1999) study to obtain posterior distribution for θ.
- The posterior mean of θ is $1/100,000$
- How likely it is to observe 2 or more rats out of 30 with tumors under the null hypothesis of no effect of perchlorate?
- This probability is $< 1/100,000 \rightarrow$ data support causal effect of perchlorate on increased thyroid tumor incidence

More Complex Models

- Posterior calculation for preterm birth example relied on conjugate prior

More Complex Models

- Posterior calculation for preterm birth example relied on conjugate prior
- Posterior calculation for perchlorate example relied on numeric integration - easy for two parameters

More Complex Models

- Posterior calculation for preterm birth example relied on conjugate prior
- Posterior calculation for perchlorate example relied on numeric integration - easy for two parameters
- For epidemiologic analyses (*e.g.*, *logistic regression*, *survival analysis*), conjugate priors not available & dimension high

More Complex Models

- Posterior calculation for preterm birth example relied on conjugate prior
- Posterior calculation for perchlorate example relied on numeric integration - easy for two parameters
- For epidemiologic analyses (*e.g.*, *logistic regression*, *survival analysis*), conjugate priors not available & dimension high
- In such settings, there are multiple parameters in θ and one needs to compute the joint posterior:

$$\pi(\theta \,|\, \mathbf{y}) = \frac{\pi(\theta)\, L(\mathbf{y} \,|\theta)}{\int \pi(\theta)\, L(\mathbf{y} \,|\, \theta) d\theta}.$$

Example: Bayesian Logistic Regression

- Logistic regression model:

$$\operatorname{logit} \Pr(y_i = 1 \mid \mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i' \boldsymbol{\beta},$$

with $\mathbf{x}_i = (1, x_{i2}, \ldots, x_{ip})'$ a vector of predictors & $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ coefficients for these predictors

Example: Bayesian Logistic Regression

- Logistic regression model:

$$\operatorname{logit} \Pr(y_i = 1 \mid \mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i' \boldsymbol{\beta},$$

  with $\mathbf{x}_i = (1, x_{i2}, \ldots, x_{ip})'$ a vector of predictors & $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ coefficients for these predictors

- A Bayesian specification of the model is completed with a prior for the coefficients, $\pi(\boldsymbol{\beta}) = \mathrm{N}_p(\boldsymbol{\beta}_0, \boldsymbol{\Sigma})$.

Example: Bayesian Logistic Regression

- Logistic regression model:

$$\text{logit}\,\Pr(y_i = 1 \,|\, \mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta},$$

  with $\mathbf{x}_i = (1, x_{i2}, \ldots, x_{ip})'$ a vector of predictors & $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ coefficients for these predictors
- A Bayesian specification of the model is completed with a prior for the coefficients, $\pi(\boldsymbol{\beta}) = \mathrm{N}_p(\boldsymbol{\beta}_0, \boldsymbol{\Sigma})$.
- Here, $\boldsymbol{\beta}_0$ is one's best *guess* at the coefficient values prior to observing the data from the current study

Example: Bayesian Logistic Regression

- Logistic regression model:

$$\text{logit} \Pr(y_i = 1 \mid \mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i' \boldsymbol{\beta},$$

  with $\mathbf{x}_i = (1, x_{i2}, \ldots, x_{ip})'$ a vector of predictors & $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ coefficients for these predictors
- A Bayesian specification of the model is completed with a prior for the coefficients, $\pi(\boldsymbol{\beta}) = \text{N}_p(\boldsymbol{\beta}_0, \boldsymbol{\Sigma})$.
- Here, $\boldsymbol{\beta}_0$ is one's best *guess* at the coefficient values prior to observing the data from the current study
- $\boldsymbol{\Sigma}$=covariance matrix quantifying uncertainty in this guess

Some Different Possibilities for the Prior

I. Informative Prior

- Review literature & choose a prior to be centered on previous estimates of coefficients.

Some Different Possibilities for the Prior

I. Informative Prior

- Review literature & choose a prior to be centered on previous estimates of coefficients.
- In the absence of previous estimates, choose a subjective value synthesizing knowledge of the literature

Some Different Possibilities for the Prior

I. Informative Prior

- Review literature & choose a prior to be centered on previous estimates of coefficients.
- In the absence of previous estimates, choose a subjective value synthesizing knowledge of the literature
- Prior variance chosen so that a 90 or 95% prior interval contains a wide range of plausible values

Some Different Possibilities for the Prior

I. Informative Prior

- Review literature & choose a prior to be centered on previous estimates of coefficients.
- In the absence of previous estimates, choose a subjective value synthesizing knowledge of the literature
- Prior variance chosen so that a 90 or 95% prior interval contains a wide range of plausible values
- Useful to choose informative priors for intercept and confounding coefficients, as there is typically substantial information about these coefficients

Some Possible Priors (continued)

II. Shrinkage Priors

- Choose a prior centered on zero with modest variance

Some Possible Priors (continued)

II. Shrinkage Priors

- Choose a prior centered on zero with modest variance
- When little information is available about a parameter, results in *shrinkage* towards zero

Some Possible Priors (continued)

II. Shrinkage Priors

- Choose a prior centered on zero with modest variance
- When little information is available about a parameter, results in *shrinkage* towards zero
- Avoids unstable estimates - particularly problematic in high dimensions & for correlated predictors.

Some Possible Priors (continued)

II. Shrinkage Priors

- Choose a prior centered on zero with modest variance
- When little information is available about a parameter, results in *shrinkage* towards zero
- Avoids unstable estimates - particularly problematic in high dimensions & for correlated predictors.
- As more information becomes available that the parameter (*e.g., the exposure odds ratio*) is non-zero, the likelihood will dominate.

Some Possible Priors (continued)

III. **Non-Informative Priors**

- Choose a prior that has high variance or is *flat* in some sense to express ignorance about the parameter value

Some Possible Priors (continued)

III. **Non-Informative Priors**

- Choose a prior that has high variance or is *flat* in some sense to express ignorance about the parameter value
- Often yields similar results to maximum likelihood - what's the point?

Some Possible Priors (continued)

III. **Non-Informative Priors**

- Choose a prior that has high variance or is *flat* in some sense to express ignorance about the parameter value
- Often yields similar results to maximum likelihood - what's the point?
- No prior is truly non-informative - flat or high variance priors assign most of their probability outside a plausible range for the parameter values.

Some Possible Priors (continued)

III. **Non-Informative Priors**

- Choose a prior that has high variance or is *flat* in some sense to express ignorance about the parameter value
- Often yields similar results to maximum likelihood - what's the point?
- No prior is truly non-informative - flat or high variance priors assign most of their probability outside a plausible range for the parameter values.
- Can lead to poor results when insufficient information available about a given parameter in the current data set - typically, the case when many predictors are collected.

Bayes Logistic Regression (continued)

- Posterior distribution:

$$\pi(\boldsymbol{\beta} \mid \mathbf{y}) = \frac{\mathrm{N}_p(\boldsymbol{\beta}; \boldsymbol{\beta}_0, \boldsymbol{\Sigma}) \prod_{l=1}^{n} L(y_i; \mathbf{x}_i, \boldsymbol{\beta})}{\int \mathrm{N}_p(\boldsymbol{\beta}; \boldsymbol{\beta}_0, \boldsymbol{\Sigma}) \prod_{l=1}^{n} L(y_i; \mathbf{x}_i, \boldsymbol{\beta}) d\boldsymbol{\beta}},$$

  where $L(y_i; \mathbf{x}_i, \boldsymbol{\beta})$ is the likelihood contribution for individual i

Bayes Logistic Regression (continued)

- Posterior distribution:

$$\pi(\boldsymbol{\beta} \mid \mathbf{y}) = \frac{\mathrm{N}_p(\boldsymbol{\beta}; \boldsymbol{\beta}_0, \boldsymbol{\Sigma}) \prod_{l=1}^{n} L(y_i; \mathbf{x}_i, \boldsymbol{\beta})}{\int \mathrm{N}_p(\boldsymbol{\beta}; \boldsymbol{\beta}_0, \boldsymbol{\Sigma}) \prod_{l=1}^{n} L(y_i; \mathbf{x}_i, \boldsymbol{\beta}) d\boldsymbol{\beta}},$$

  where $L(y_i; \mathbf{x}_i, \boldsymbol{\beta})$ is the likelihood contribution for individual i

- Note that we can write the numerator in this expression in closed form

Bayes Logistic Regression (continued)

- Posterior distribution:

$$\pi(\boldsymbol{\beta} \mid \mathbf{y}) = \frac{\mathrm{N}_p(\boldsymbol{\beta}; \boldsymbol{\beta}_0, \boldsymbol{\Sigma}) \prod_{l=1}^{n} L(y_i; \mathbf{x}_i, \boldsymbol{\beta})}{\int \mathrm{N}_p(\boldsymbol{\beta}; \boldsymbol{\beta}_0, \boldsymbol{\Sigma}) \prod_{l=1}^{n} L(y_i; \mathbf{x}_i, \boldsymbol{\beta}) d\boldsymbol{\beta}},$$

  where $L(y_i; \mathbf{x}_i, \boldsymbol{\beta})$ is the likelihood contribution for individual i
- Note that we can write the numerator in this expression in closed form
- However, the denominator involves a nasty high-dimensional integral that has no analytic solution.

Calculating the Posterior Distribution

- To calculate the posterior, one can potentially rely on a large sample approximation

Calculating the Posterior Distribution

- To calculate the posterior, one can potentially rely on a large sample approximation
- As $n \to \infty$, the posterior is normally distributed centered on the maximum likelihood estimate

Calculating the Posterior Distribution

- To calculate the posterior, one can potentially rely on a large sample approximation
- As $n \to \infty$, the posterior is normally distributed centered on the maximum likelihood estimate
- Impact of the prior decreases as the sample size increases in general

Calculating the Posterior Distribution

- To calculate the posterior, one can potentially rely on a large sample approximation
- As $n \to \infty$, the posterior is normally distributed centered on the maximum likelihood estimate
- Impact of the prior decreases as the sample size increases in general
- However, even for moderate to large samples, asymptotic normal approximation may be inaccurate

Calculating the Posterior Distribution

- To calculate the posterior, one can potentially rely on a large sample approximation
- As $n \to \infty$, the posterior is normally distributed centered on the maximum likelihood estimate
- Impact of the prior decreases as the sample size increases in general
- However, even for moderate to large samples, asymptotic normal approximation may be inaccurate
- In logistic regression for rare outcomes or rare exposure categories, posterior can be highly skewed

MCMC - Basic Idea

- Markov chain Monte Carlo (MCMC) provides an approach for generating samples from the posterior distribution

MCMC - Basic Idea

- Markov chain Monte Carlo (MCMC) provides an approach for generating samples from the posterior distribution
- This does not give us an approximation to $\pi(\theta \mid \mathbf{y})$ directly

MCMC - Basic Idea

- Markov chain Monte Carlo (MCMC) provides an approach for generating samples from the posterior distribution
- This does not give us an approximation to $\pi(\theta \mid \mathbf{y})$ directly
- However, from these samples we can obtain summaries of the posterior distribution for θ

MCMC - Basic Idea

- Markov chain Monte Carlo (MCMC) provides an approach for generating samples from the posterior distribution
- This does not give us an approximation to $\pi(\theta \mid \mathbf{y})$ directly
- However, from these samples we can obtain summaries of the posterior distribution for θ
- Summaries of exact posterior distributions of $g(\theta)$, for any functional $g(\cdot)$, can also be obtained.

MCMC - Basic Idea

- Markov chain Monte Carlo (MCMC) provides an approach for generating samples from the posterior distribution
- This does not give us an approximation to $\pi(\theta \,|\, \mathbf{y})$ directly
- However, from these samples we can obtain summaries of the posterior distribution for θ
- Summaries of exact posterior distributions of $g(\theta)$, for any functional $g(\cdot)$, can also be obtained.
- For example, if θ is the log-odds ratio, then we could choose $g(\theta) = \exp(\theta)$ to obtain the odds ratio

How does MCMC work?

- Let $\theta^t = (\theta_1^t, \ldots, \theta_p^t)$ denote the value of the $p \times 1$ vector of parameters at iteration t.

How does MCMC work?

- Let $\theta^t = (\theta_1^t, \ldots, \theta_p^t)$ denote the value of the $p \times 1$ vector of parameters at iteration t.
- $\theta^0 =$ initial value used to start the chain (*shouldn't be sensitive*)

How does MCMC work?

- Let $\theta^t = (\theta_1^t, \ldots, \theta_p^t)$ denote the value of the $p \times 1$ vector of parameters at iteration t.
- $\theta^0 =$ initial value used to start the chain (*shouldn't be sensitive*)
- MCMC generates θ^t from a distribution that depends on the data & potentially on θ^{t-1}, but not on $\theta^1, \ldots, \theta^{t-2}$.

How does MCMC work?

- Let $\theta^t = (\theta_1^t, \ldots, \theta_p^t)$ denote the value of the $p \times 1$ vector of parameters at iteration t.
- $\theta^0 =$ initial value used to start the chain (*shouldn't be sensitive*)
- MCMC generates θ^t from a distribution that depends on the data & potentially on θ^{t-1}, but not on $\theta^1, \ldots, \theta^{t-2}$.
- This results in a Markov chain with stationary distribution $\pi(\theta \mid \mathbf{y})$ under some conditions on the sampling distribution

Different flavors of MCMC

- The most commonly used MCMC algorithms are:

Different flavors of MCMC

- The most commonly used MCMC algorithms are:
  - Metropolis sampling (*Metropolis et al., 1953*)

Different flavors of MCMC

- The most commonly used MCMC algorithms are:
  - Metropolis sampling (*Metropolis et al., 1953*)
  - Metropolis-Hastings (MH) (*Hastings, 1970*)

Different flavors of MCMC

- The most commonly used MCMC algorithms are:
  - Metropolis sampling (*Metropolis et al., 1953*)
  - Metropolis-Hastings (MH) (*Hastings, 1970*)
  - Gibbs sampling (*Geman & Geman, 1984; Gelfand & Smith, 1990*)

Different flavors of MCMC

- The most commonly used MCMC algorithms are:
  - Metropolis sampling (*Metropolis et al., 1953*)
  - Metropolis-Hastings (MH) (*Hastings, 1970*)
  - Gibbs sampling (*Geman & Geman, 1984; Gelfand & Smith, 1990*)
- Easy overview of Gibbs - Casella & George (1992, *The American Statistician*, 46, 167-174)

Different flavors of MCMC

- The most commonly used MCMC algorithms are:
  - Metropolis sampling (*Metropolis et al., 1953*)
  - Metropolis-Hastings (MH) (*Hastings, 1970*)
  - Gibbs sampling (*Geman & Geman, 1984; Gelfand & Smith, 1990*)
- Easy overview of Gibbs - Casella & George (1992, *The American Statistician*, 46, 167-174)
- Easy overview of MH - Chib & Greenberg (1995, *The American Statistician*)

Gibbs Sampling

- Start with initial value $\theta^0 = (\theta_1^0, \ldots, \theta_p^0)$

Gibbs Sampling

- Start with initial value $\theta^0 = (\theta_1^0, \ldots, \theta_p^0)$
- For iterations $t = 1, \ldots, T$,

Gibbs Sampling

- Start with initial value $\theta^0 = (\theta_1^0, \ldots, \theta_p^0)$
- For iterations $t = 1, \ldots, T$,
  1. Sample θ_1^t from the conditional posterior distribution

$$\pi(\theta_1 \mid \theta_2 = \theta_2^{t-1}, \ldots, \theta_p = \theta_p^{t-1}, \mathbf{y})$$

Gibbs Sampling

- Start with initial value $\theta^0 = (\theta_1^0, \ldots, \theta_p^0)$
- For iterations $t = 1, \ldots, T$,
  1. Sample θ_1^t from the conditional posterior distribution

$$\pi(\theta_1 \mid \theta_2 = \theta_2^{t-1}, \ldots, \theta_p = \theta_p^{t-1}, \mathbf{y})$$

  2. Sample θ_2^t from the conditional posterior distribution

$$\pi(\theta_2 \mid \theta_1 = \theta_1^t, \theta_3 = \theta_3^{t-1}, \ldots, \theta_p = \theta_p^{t-1}, \mathbf{y})$$

Gibbs Sampling

- Start with initial value $\theta^0 = (\theta_1^0, \dots, \theta_p^0)$
- For iterations $t = 1, \dots, T$,
  1. Sample θ_1^t from the conditional posterior distribution
  $$\pi(\theta_1 \mid \theta_2 = \theta_2^{t-1}, \dots, \theta_p = \theta_p^{t-1}, \mathbf{y})$$
  2. Sample θ_2^t from the conditional posterior distribution
  $$\pi(\theta_2 \mid \theta_1 = \theta_1^t, \theta_3 = \theta_3^{t-1}, \dots, \theta_p = \theta_p^{t-1}, \mathbf{y})$$
  3. Similarly, sample $\theta_3^t, \dots, \theta_p^t$ from the conditional posterior distributions given current values of other parameters.

Gibbs Sampling (continued)

- Under mild regularity conditions, samples converge to stationary distribution $\pi(\theta \,|\, \mathbf{y})$

Gibbs Sampling (continued)

- Under mild regularity conditions, samples converge to stationary distribution $\pi(\theta \mid \mathbf{y})$
- At the start of the sampling, the samples are <u>not</u> from the posterior distribution $\pi(\theta \mid \mathbf{y})$.

Gibbs Sampling (continued)

- Under mild regularity conditions, samples converge to stationary distribution $\pi(\theta \mid \mathbf{y})$
- At the start of the sampling, the samples are not from the posterior distribution $\pi(\theta \mid \mathbf{y})$.
- It is necessary to discard the initial samples as a *burn-in* to allow convergence

Gibbs Sampling (continued)

- Under mild regularity conditions, samples converge to stationary distribution $\pi(\theta \mid \mathbf{y})$
- At the start of the sampling, the samples are not from the posterior distribution $\pi(\theta \mid \mathbf{y})$.
- It is necessary to discard the initial samples as a *burn-in* to allow convergence
- In simple models such as logistic regression, convergence typically occurs quickly & burn-in of 100 iterations should be sufficient (*to be conservative SAS uses 2,000 as default*)

Example - DDE & Preterm Birth

- Scientific interest: Association between DDE exposure & preterm birth adjusting for possible confounding variables

Example - DDE & Preterm Birth

- Scientific interest: Association between DDE exposure & preterm birth adjusting for possible confounding variables
- Data from US Collaborative Perinatal Project (CPP) - n = 2380 children out of which 361 were born preterm

Example - DDE & Preterm Birth

- Scientific interest: Association between DDE exposure & preterm birth adjusting for possible confounding variables
- Data from US Collaborative Perinatal Project (CPP) - n = 2380 children out of which 361 were born preterm
- Analysis: Bayesian analysis using a probit model:

$$\Pr(y_i = 1 \,|\, \mathbf{x}_i, \beta) = \Phi(\beta_1 + \beta_2 dde_i + \beta_3 z_{i1} + \cdots + \beta_7 z_{i5}).$$

Example - DDE & Preterm Birth

- Scientific interest: Association between DDE exposure & preterm birth adjusting for possible confounding variables
- Data from US Collaborative Perinatal Project (CPP) - n = 2380 children out of which 361 were born preterm
- Analysis: Bayesian analysis using a probit model:

$$\Pr(y_i = 1 \mid \mathbf{x}_i, \beta) = \Phi(\beta_1 + \beta_2 dde_i + \beta_3 z_{i1} + \cdots + \beta_7 z_{i5}).$$

- Chose normal prior with mean 0 and variance 4.

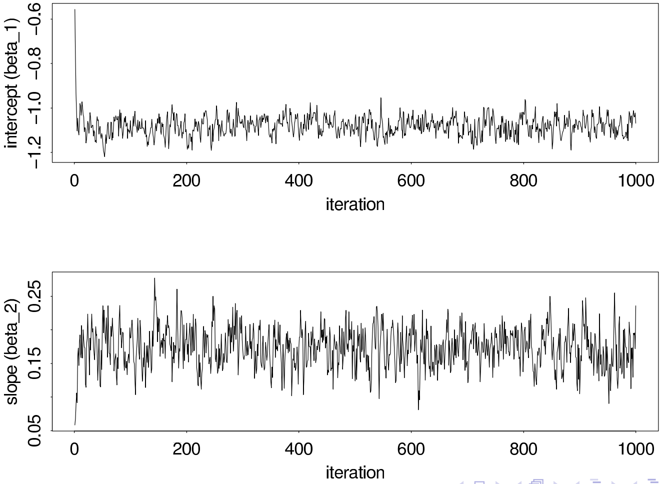
Example - DDE & Preterm Birth

- Scientific interest: Association between DDE exposure & preterm birth adjusting for possible confounding variables
- Data from US Collaborative Perinatal Project (CPP) - n = 2380 children out of which 361 were born preterm
- Analysis: Bayesian analysis using a probit model:

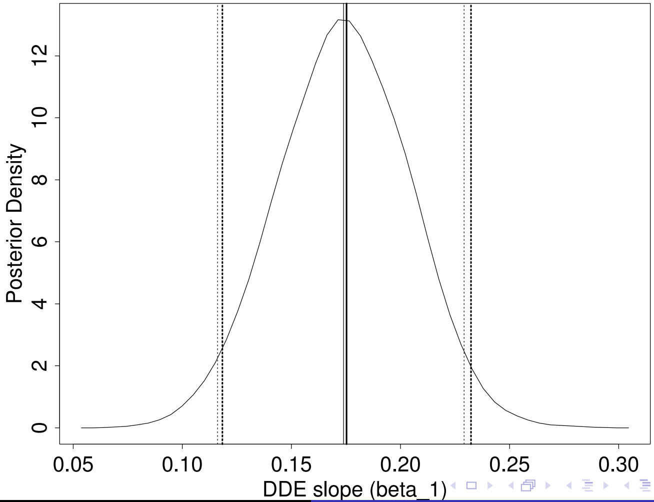
$$\Pr(y_i = 1 \,|\, \mathbf{x}_i, \beta) = \Phi(\beta_1 + \beta_2 dde_i + \beta_3 z_{i1} + \cdots + \beta_7 z_{i5}).$$

- Chose normal prior with mean 0 and variance 4.
- Probit model is similar to logistic regression, but with different link

Gibbs Sampling output for preterm birth example



Estimated Posterior Density



Some MCMC Terminology

- Convergence: initial drift in the samples towards a stationary distribution

Some MCMC Terminology

- Convergence: initial drift in the samples towards a stationary distribution
- Burn-in: samples at start of the chain that are discarded to allow convergence

Some MCMC Terminology

- Convergence: initial drift in the samples towards a stationary distribution
- Burn-in: samples at start of the chain that are discarded to allow convergence
- Slow mixing: tendency for high autocorrelation in the samples.

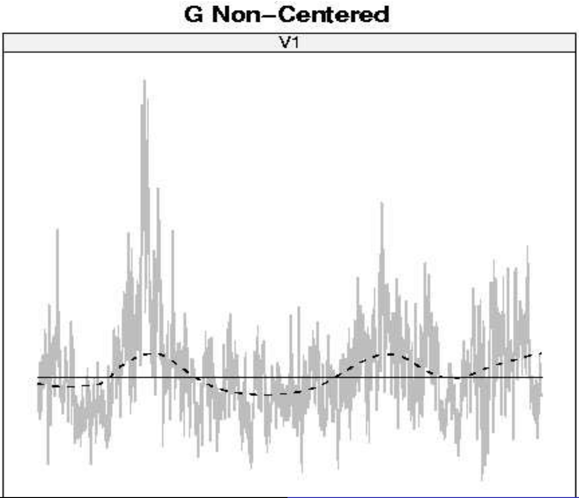
Some MCMC Terminology

- Convergence: initial drift in the samples towards a stationary distribution
- Burn-in: samples at start of the chain that are discarded to allow convergence
- Slow mixing: tendency for high autocorrelation in the samples.
- Thinning: practice of collecting every kth iteration to reduce autocorrelation

Some MCMC Terminology

- Convergence: initial drift in the samples towards a stationary distribution
- Burn-in: samples at start of the chain that are discarded to allow convergence
- Slow mixing: tendency for high autocorrelation in the samples.
- Thinning: practice of collecting every kth iteration to reduce autocorrelation
- Trace plot: plot of sampled values of a parameter vs iteration #

Example - trace plot with poor mixing



Poor mixing Gibbs sampler

- Exhibits “snaking” behavior in trace plot with cyclic local trends in the mean

Poor mixing Gibbs sampler

- Exhibits “snaking” behavior in trace plot with cyclic local trends in the mean
- Poor mixing in the Gibbs sampler caused by high posterior correlation in the parameters

Poor mixing Gibbs sampler

- Exhibits "snaking" behavior in trace plot with cyclic local trends in the mean
- Poor mixing in the Gibbs sampler caused by high posterior correlation in the parameters
- Decreases efficiency & <u>many</u> more samples need to be collected to maintain low Monte Carlo error in posterior summaries

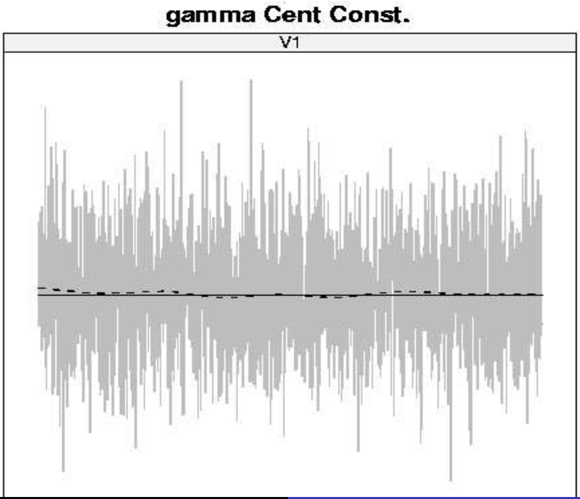
Poor mixing Gibbs sampler

- Exhibits "snaking" behavior in trace plot with cyclic local trends in the mean
- Poor mixing in the Gibbs sampler caused by high posterior correlation in the parameters
- Decreases efficiency & <u>many</u> more samples need to be collected to maintain low Monte Carlo error in posterior summaries
- For very poor mixing chain, may even need millions of iterations.

Poor mixing Gibbs sampler

- Exhibits "snaking" behavior in trace plot with cyclic local trends in the mean
- Poor mixing in the Gibbs sampler caused by high posterior correlation in the parameters
- Decreases efficiency & many more samples need to be collected to maintain low Monte Carlo error in posterior summaries
- For very poor mixing chain, may even need millions of iterations.
- Routinely examine trace plots!

Example - trace plot with good mixing



How to summarize results from the MCMC chain?

- Posterior mean: estimated by average of samples collected after discarding burn-in

How to summarize results from the MCMC chain?

- Posterior mean: estimated by average of samples collected after discarding burn-in
- Posterior mean provides alternative to maximum likelihood estimate as a single summary.

How to summarize results from the MCMC chain?

- Posterior mean: estimated by average of samples collected after discarding burn-in
- Posterior mean provides alternative to maximum likelihood estimate as a single summary.
- As a Bayesian alternative to the confidence interval, one can use a credible interval

How to summarize results from the MCMC chain?

- Posterior mean: estimated by average of samples collected after discarding burn-in
- Posterior mean provides alternative to maximum likelihood estimate as a single summary.
- As a Bayesian alternative to the confidence interval, one can use a credible interval
- The $100(1-\alpha)\%$ credible interval ranges from the $\alpha/2$ to $1-\alpha/2$ empirical percentiles of the collected samples

How to summarize results from the MCMC chain?

- Posterior mean: estimated by average of samples collected after discarding burn-in
- Posterior mean provides alternative to maximum likelihood estimate as a single summary.
- As a Bayesian alternative to the confidence interval, one can use a credible interval
- The $100(1-\alpha)\%$ credible interval ranges from the $\alpha/2$ to $1-\alpha/2$ empirical percentiles of the collected samples
- Credible intervals can be calculated for functionals (e.g., odds ratios) by first applying the function to each MCMC sample

Posterior probabilities

- Often interest focuses on the weight of evidence of $H_1 : \theta_j > 0$

Posterior probabilities

- Often interest focuses on the weight of evidence of $H_1 : \theta_j > 0$
- The posterior probability of H_1 can be calculated easily from the MCMC output as simply the proportion of collected samples having $\theta_j > 0$.

Posterior probabilities

- Often interest focuses on the weight of evidence of $H_1 : \theta_j > 0$
- The posterior probability of H_1 can be calculated easily from the MCMC output as simply the proportion of collected samples having $\theta_j > 0$.
- A high value (e.g., greater than 0.95) suggests strong evidence in favor of H_1

Marginal posterior density estimation

- Summary statistics such as the mean, median, standard deviation, etc provide an incomplete picture

Marginal posterior density estimation

- Summary statistics such as the mean, median, standard deviation, etc provide an incomplete picture
- Since we have many samples from the posterior, we can accurately estimate the *exact* posterior density

Marginal posterior density estimation

- Summary statistics such as the mean, median, standard deviation, etc provide an incomplete picture
- Since we have many samples from the posterior, we can accurately estimate the *exact* posterior density
- This can be done using a kernel-smoothed density estimation procedure applied to the samples

How to get started?

- It is not necessary to understand MCMC theory to implement Bayesian analyses

How to get started?

- It is not necessary to understand MCMC theory to implement Bayesian analyses
- WinBUGS is a general software package for implementing MCMC in a very broad variety of models

How to get started?

- It is not necessary to understand MCMC theory to implement Bayesian analyses
- WinBUGS is a general software package for implementing MCMC in a very broad variety of models
- WinBUGS can accommodate hierarchical models, missing data, spatial correlation, etc (*Rich will illustrate*)

How to get started?

- It is not necessary to understand MCMC theory to implement Bayesian analyses
- WinBUGS is a general software package for implementing MCMC in a very broad variety of models
- WinBUGS can accommodate hierarchical models, missing data, spatial correlation, etc (*Rich will illustrate*)
- SAS also has several new Bayes Procs available (*Amy will illustrate*)