Session 3

Bayesian Methods

Yuri Balasanov

University of Chicago, MScA

For any inquiries contact the author, Yuri Balasanov, at ybalasan@uchicago.edu or yuri.balasanov@iLykei.com .

Outline of the Session

- Conjugate priors for Gaussian, Poisson, exponential data
- Predicting distribution for data
- Shrinkage and Stein Paradox
- Random effects
- Hierarchical model as an explanation of random effects
- Motivation for hierarchical models
- Role of exchangeability in hierarchical models

Gaussian Conjugate Distributions

- Consider the case when the model for the data is Gaussian: $Y \sim \mathbb{N}\left(\theta, \sigma^2\right)$, where σ^2 is the known variance and θ is the unknown mean value.
  For such model conjugate family of prior distributions for the parameter is also Gaussian, $\theta \sim \mathbb{N}\left(\mu_0, \tau_0^2\right)$ with hyperparameters μ_0, τ_0^2
- Then posterior distribution is obtained analytically:

$$p\left(\theta | \, y_1, \ldots, y_n\right) = p\left(\theta | \, \bar{y}\right) = \phi\left(\mu_n, \tau_n^2\right),$$

where

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

- Variables $\frac{1}{\tau_0^2}$ and $\frac{n}{\sigma^2}$ are called prior precision and data precision, correspondingly. If $\tau_0^2 = \sigma^2$ then prior distribution has approximately same weight as 1 observation equal to μ_0

Gaussian Model with Unknown Variance

Let the likelihood function for unknown variance of Gaussian distribution be

$$p\left(y|\,\sigma^2\right) = \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{n}{2\sigma^2}v\right)$$

with sufficient statistic

$$v = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mu\right)^2.$$

Then the conjugate prior is the inverse-gamma distribution

$$p\left(\sigma^2\right) \propto \left(\sigma^2\right)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right)$$

with hyperparameters (α, β).

Poisson Model

Likelihood function for Poisson model is

$$p\left(y_1,\ldots,y_n|\,\theta\right)=p\left(y|\,\theta\right)\propto\exp\left(-n\theta\right)\exp\left(t\left(y\right)\log\left(\theta\right)\right),$$

where $t\left(y\right)=\sum y_i$ is the sufficient statistic.
Then the natural conjugate prior is of the form

$$p\left(\theta\right)\propto\left(\exp\left(-\theta\right)\right)^{\eta}\exp\left(\nu\log\left(\theta\right)\right)\propto\exp\left(-\beta\theta\right)\theta^{\alpha-1}.$$

The last expression is gamma distribution density function $Gamma\left(\alpha,\beta\right)$ with hyperparameters $\left(\alpha,\beta\right)$.
If prior is

$$p\left(\theta\right)=Gamma\left(\alpha,\beta\right)$$

then posterior is

$$p\left(\theta|\,y\right)=Gamma\left(\alpha+n\bar{y},\beta+n\right)$$

Exponential Model

Exponential model is a particular case of gamma model with parameters $\alpha = 1$, $\beta = \theta$, and is used to describe waiting times
It has likelihood function for single observation y

$$p\left(y|\,\theta\right) = \theta \exp\left(-y\theta\right), y > 0,$$

where $\theta = \left(\mathbb{E}\left[y|\,\theta\right]\right)^{-1}$ is the intensity parameter.
The conjugate prior for such model is gamma distribution

$$p\left(\theta|\,\alpha, \beta\right) = Gamma\left(\alpha, \beta\right)$$

and the corresponding posterior distribution is

$$p\left(\theta|\,y\right) = Gamma\left(\alpha + 1, \beta + yk\right)$$

Marginal Distribution of Data

For known forms of prior and posterior densities for conjugate families one can find marginal distribution of data

$$p(y) = \frac{p(y|\theta)\, p(\theta)}{p(\theta|y)}.$$

For example Poisson model for single observation y gives prior predictive distribution

$$\begin{aligned} p(y) &= \frac{Pois(y|\theta)\, Gamma(\theta|\alpha,\beta)}{Gamma(\theta|\alpha+y,\beta+1)} = \frac{\Gamma(\alpha+y)\,\beta^{\alpha}}{\Gamma(\alpha)\, y!\,(\beta+1)^{(\alpha+y)}} \\ &= \binom{\alpha+y-1}{y} \left(\frac{\beta}{\beta+1}\right)^{\alpha} \left(\frac{1}{\beta+1}\right)^{y} = NBin(\alpha,\beta), \end{aligned}$$

which is the negative binomial distribution.
This shows that negative binomial distribution is a mixture of Poisson distributions with intensities θ that follow gamma distribution

Predictive Distribution For Binomial Data

In general prior predictive distribution is defined as

$$p(y) = \int p(y, \theta)\, d\theta = \int p(y|\theta)\, p(\theta)\, d\theta$$

For binomial distribution it is

$$\mathbb{P}\{y = 1\} = \int_0^1 Beta(\theta|\alpha, \beta)\, \theta d\theta = E[\theta]$$

The expected value is taken with respect to the prior distribution
Predictive posterior distribution is

$$\begin{aligned} p(\tilde{y}|D) &= \int p(\tilde{y}, \theta|D)\, d\theta = \int p(\tilde{y}|\theta, D)\, p(\theta|D)\, d\theta \\ &= \int p(\tilde{y}|\theta)\, p(\theta|D)\, d\theta = \int \theta p(\theta|D)\, d\theta = \mathbb{E}[\theta|D] \end{aligned}$$

In case of binomial distribution posterior predictive probability of success is the expectation of the posterior distribution

Concept of Shrinkage I

Stein Paradox

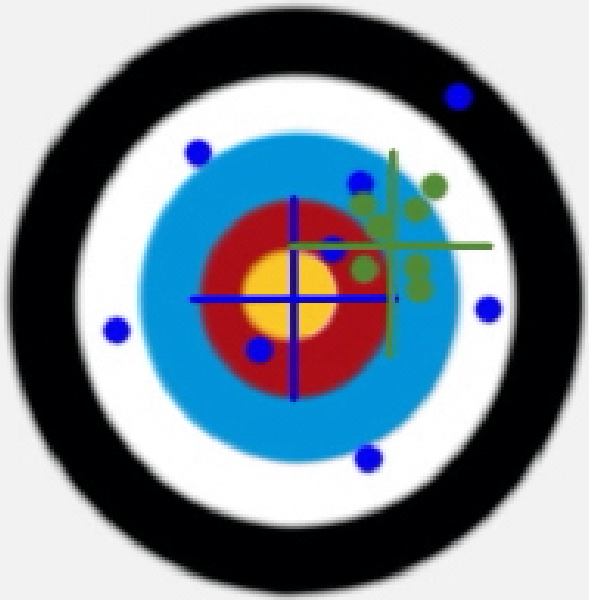
- A famous result by Charles Stein from Stanford (1955), shows that MLE $\hat{\mu}_j$ can be improved in terms of mean squared error if instead of group averages the estimates are mixes of group averages with the grand total:

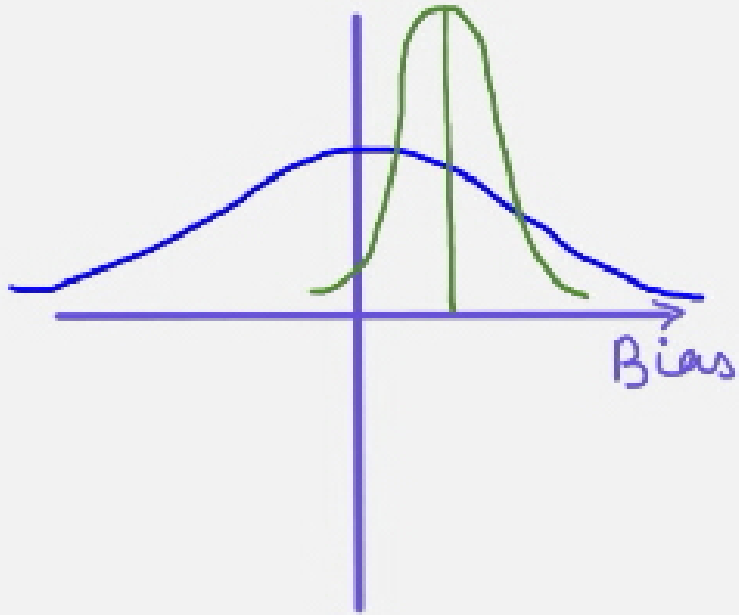
$$\hat{\mu}_j^S = \bar{y}_{..} + \left(1 - \frac{(J-3)\,\sigma^2}{n\sum_{j=1}^{n}\left(\bar{y}_{..} - \bar{y}_{\cdot j}\right)^2}\right)^{+}\left(\bar{y}_{\cdot j} - \bar{y}_{..}\right), j = 1, \ldots, J; J \geq 3$$

- Stein estimate shows that mean squared errors improve if unbiased MLE $\hat{\mu}_j$ are replaced by biased estimates "pulled" towards the grand mean
- The effect of pulling estimates towards the grand mean is called **shrinkage**. The factor by which Stein estimate is closer to the grand mean relative to MLE is called **shrinkage factor**.

Concept of Shrinkage II

Bias vs. Mean-Square Error.





Allowing some bias may improve mean-square error.

Concept of Shrinkage III

Example

In a multiple studies of medical treatment in different hospitals data y_{ij} show effect of treatment in study case i in hospital j, i.e. y_{ij} is the difference of proportions of success in treated group and in control group in study i, $i = 1, \ldots, n$, in hospital j, $j = 1, \ldots, J$. It is reasonable to assume $y_{ij} \sim \mathbb{N}\left(\mu_j, \sigma_j\right)$. Let σ_j be known for simplicity. What are the best estimates for μ_j?

- One method of solving the problem: organize y_{ij} in a matrix with n rows and a column for each hospital. Then calculate column means as $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$. We know that estimates $\hat{\mu}_j$ have all the good properties of MLE.
- But such estimation does not use information that treatment in all hospitals was the same.
- How can we use this information to improve the estimate?

Concept of Shrinkage IV

Fixed Effects

- Assume fixed effects model

$$y_{ij} = \mu + \beta_1 x_{i1} + \ldots + \beta_{J-1} x_{iJ-1} + \epsilon_{ij}, \epsilon_{ij} \sim \mathbb{N}(0, \sigma),$$

  where x_{ij} is a dummy variable identifying hospital j. This model estimates the grand mean μ and specific hospital means as offsets to the grand mean: $\mu_j = \mu + \beta_{j-1}, \beta_0 = 0$.
- Classical solution of such problem is based on ANOVA framework and F-test, which compares the between-units variance $\tau^2 = \mathbb{V}\left[\mu_j\right]$ and within-units variance $\sigma^2 = \mathbb{V}\left[\epsilon_{ij}\right]$.
- In this framework the grand mean is playing a role, but only as a general level around which the group levels are offsets.
- Weakness of the F-test approach is that σ^2 is the within-group variance for all hospitals.

Concept of Shrinkage V

Mixed Effects

- Mixed effects model assumes that group means are randomly drawn from one distribution
- Mixed effects estimates are, like Stein's estimate, weighted averages of grand mean and group means mixed by shrinkage factor

$$y_{ij} = \mu + \alpha_1 I_{\{j=1\}} + \ldots + \alpha_J I_{\{j=J\}} + \epsilon_{ij},$$

where α_j are random effects, $\alpha_j \sim \mathbb{N}\left(\mu_j, \sigma\right)$.

- Look at random effects from the point of view of Bayesian approach:

$$y_{ij} \sim \mathbb{N}\left(\mu + \mu_j, \sigma\right), \mu_j \sim \mathbb{N}\left(\mu_0, \tau_0\right)$$

- For conjugate normal distribution $\mu_j \Big| y \sim \mathbb{N}\left(\mu_j, \tau_j\right), \frac{1}{\tau_j^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$

$$\mu_j = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}_{\cdot j}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = B_j\mu_0 + (1 - B_j)\,\bar{y}_{\cdot j}; B_j = \frac{\frac{1}{\tau_0^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

Comparison of Fixed and Random Effects

Fixed effects

- Analyzes clusters within given dataset
- Efficient when clusters are big
- Requires many dummy variables and slopes for them
- Provides unbiased estimates, variance may be large
- Estimates cluster means, tests if they are the same

Random effects

- Analyzes population of clusters
- Efficient when there are many clusters
- Instead of slopes for dummy variables estimates variance
- Provides lower MSE; better for prediction
- Estimates variation of cluster means, tests if it is zero

Motivation for Hierarchical Models

- Realization of Stein paradox creates an argument in favor of a hierarchical model instead of ANOVA framework
- When hierarchical Bayesian model may be more useful?
  When you expect that there is some common information in multiple groups, but group means are not fixed, but distributed around some common center: general effectiveness of drug that may vary depending on conditions of treatments in different hospitals, individual responses of patients, etc.
- How do you continue Bayesian analysis with new data after group means with common prior have different posteriors?

Example

Simplest hierarchical model. Data y_{ij} are generated by binomial model with parameter θ_j: patient i treated in hospital j with binary success. Parameters θ_j are populated from prior distribution depending on parameter ϕ. So, the hierarchy is $\phi \longrightarrow \theta \longrightarrow y$.

Further Motivation

Example

The hierarchy of the previous example is the simplest setup of Bayesian analysis. To take it to the next level assume that hospitals are grouped by state, so that parameter ϕ is itself a realization from a population of states according to distribution $p(\phi|\lambda)$ with super-hyperparameter λ. Then the hierarchy is $\lambda \longrightarrow \phi \longrightarrow \theta \longrightarrow y$.

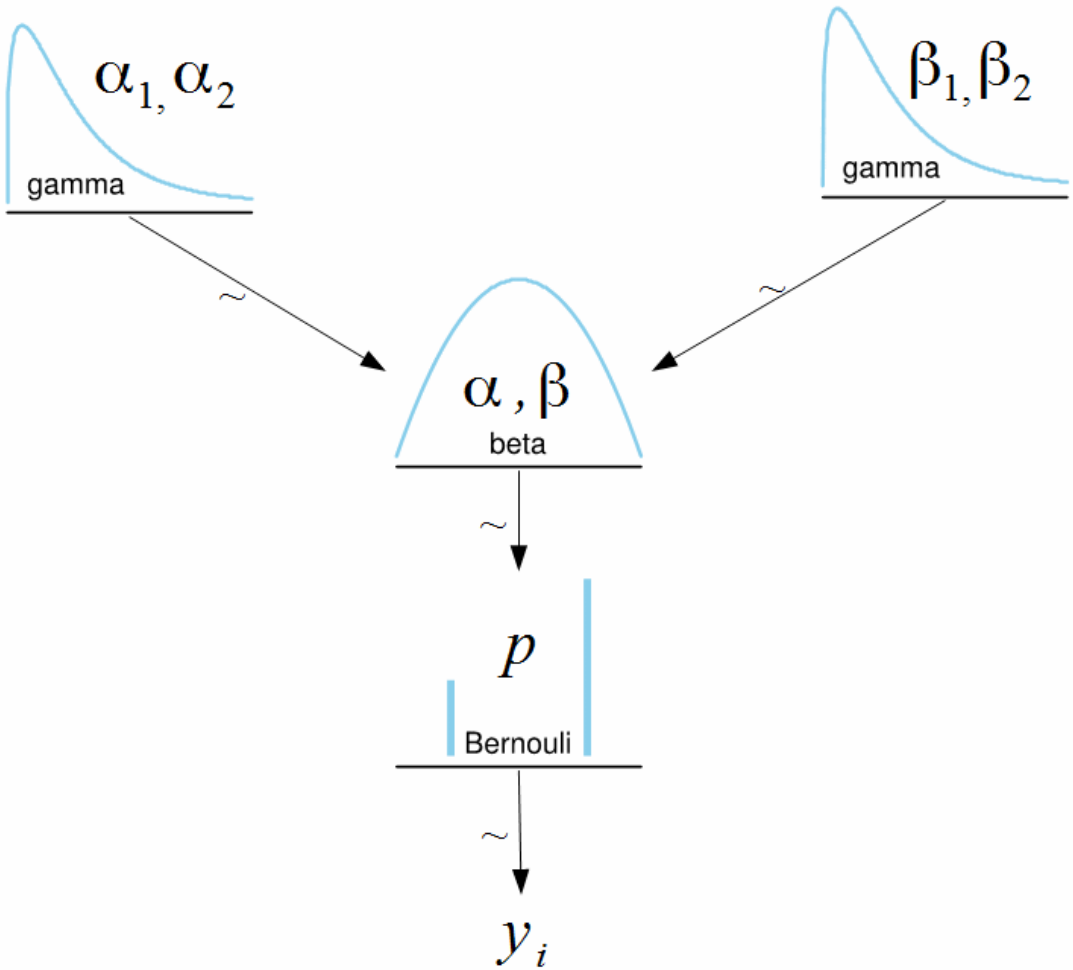
All parameters of a hierarchical model exist in the same probability space and have joint distribution.
Bayes rule applies to the joint parameter distribution:

$$\begin{aligned} p(\theta,\phi,\lambda|y) &\propto p(y|\theta,\phi,\lambda)\,p(\theta,\phi,\lambda) = p(y|\theta)\,p(\theta|\phi,\lambda)\,p(\phi,\lambda) \\ &= p(y|\theta)\,p(\theta|\phi)\,p(\phi|\lambda)\,p(\lambda) \end{aligned}$$

Advantages of hierarchical parameterization:

- Dependencies are often meaningful for given application
- Model can be efficiently coded and run by using MCMC sampling

Diagram of Hierarchical Model



Posterior Predictive Distributions

- Hierarchical models are characterized by their parameters (θ) and their hyperparameters (ϕ) and super-hyperparameters (λ)
- There are 2 different predictive posterior distributions of interest:
  1. Distribution of future $\tilde{y}$ corresponding to a currently known superpopulation of the parameter θ. In this case future observations are drawn from posterior for θ based on the current experiment
  2. Distribution of future $\tilde{y}$ corresponding to a future $\tilde{\theta}$ drawn from the same superpopulation. In this case posterior distribution comes from future experiment. First draw $\tilde{\theta}$ of future experiment using posterior draws of ϕ, then draw $\tilde{y}$ given the simulated $\tilde{\theta}$.
- Posterior predictive distribution is defined as

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y)\, d\theta = \int p(\tilde{y}|\theta, y)\, p(\theta|y)\, d\theta \\ &= \int p(\tilde{y}|\theta)\, p(\theta|y)\, d\theta \end{aligned}$$

Role of Exchangeability

Consider experiments $1, \ldots, J$ resulting in data samples $y_j = \{y_{1j}, \ldots, y_{n_j j}\}$ informing about parameter $\theta_j, j = 1, \ldots, J$.
If there is no information besides the data y, that can help ordering or grouping θ_j one must assume probabilistic exchangeability of parameters.

Definition

Random variables $\{\theta_1, \ldots, \theta_J\}$ are exchangeable in their joint distribution if $p(\theta_1, \ldots, \theta_J)$ is invariant to permutations of indices $\{1, \ldots, J\}$.

Example

Same coin flipped 3 times with 1 head and 2 tails.
$P\{0,0,1\} = P\{0,1,0\} = P\{1,0,0\}$. Joint distribution of the results of the experiment is exchangeable.

Example

A time series of GDP for number of years is unlikely to be exchangeable

Further Notes on Exchangeability I

- See example about exchangeability and sampling in the workshop
- In some cases when data are not exchangeable they may be partially or conditionally exchangeable:
  - If observations can be grouped then a hierarchical model may be used allowing a submodel for each group, but assuming the group properties exchangeable
  - If there is an information x_i about y_i that makes y_i not exchangeable then joint model for (x_i, y_i) or conditional model for $y_i \,|\, x_i$ may work

Example

See tumor in rats study in the workshop. If there was an additional information that experiments were held in different laboratories we could assume partial exchangeability and use 2-level hierarchical model for variation within each lab and between them

Further Notes on Exchangeability II

Example

See divorce rates example in the workshop. If previous year divorce rates x_i were available, but again without naming states they correspond to then y_i are not exchangeable, but (x_i, y_i) are.

Example

In the tumor in rats example there is an additional variable that may allow to distinguish observations: number n_i of animals in each experiment. A hypothesis about the role of n_i to check would be if a model for pairs $(y, n)_j$ can work.
The first step in that direction would be just plotting $\frac{y_i}{n_i}$ against n_i.
This would be a plausible hypothesis because in case of rare disease larger series may show significantly different results.
In this particular study, though, no dependence on n_i was detected as the author indicates.