CS 5135/6035 Learning Probabilistic Models

Lecture 17: Multiparameter Models

Gowtham Atluri

October 28, 2018

Reading Material

- Gelman et al. Bayesian Data Analysis
  - Chapter 3. Introduction to multiparameter models
- Kevin Murphy, Conjugate Bayesian analysis of the Gaussian distribution
  https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf

Topics

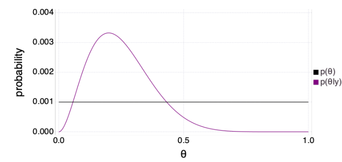
- Multiparameter models
- Joint posterior density
- Nuisance parameters
- Conditional posterior density
- Marginal posterior density
- Univariate Gaussian parameter estimation
  - Marginal posteriors for μ and σ^2
- Multivariate Gaussian parameter estimation

Bayesian Estimation: Single-Parameter models

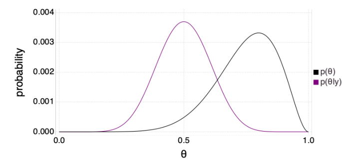
Scenario: Coin toss experiment (where 2 heads and 8 tails are observed)

- Goal is to estimate θ

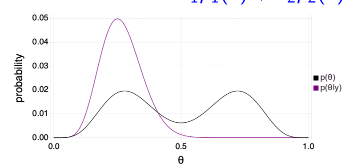
Flat prior: $p(\theta) = k$



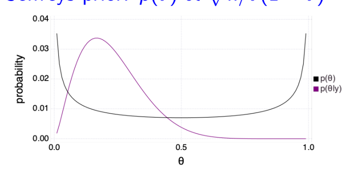
Conjugate prior: $p(\theta) = Beta(a, b)$



Mixture of Priors: $\pi_1 p_1(\theta) + \pi_2 p_2(\theta)$



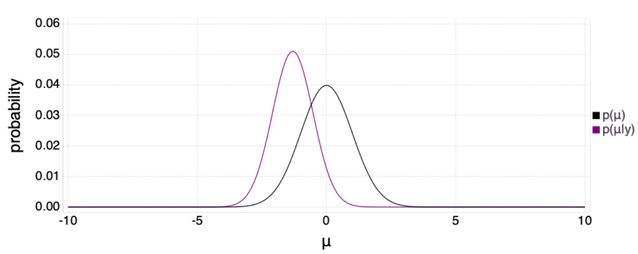
Jeffreys prior: $p(\theta) \propto \sqrt{n/\theta(1-\theta)}$



Bayesian Estimation: Single-Parameter models

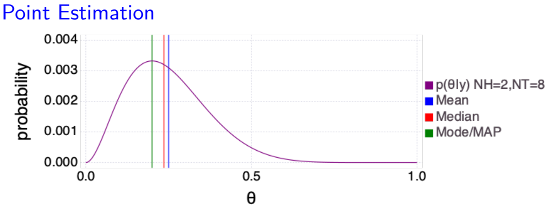
Scenario: The temperatures, in Celsius, in Minneapolis during the first week of March 2018 are observed as $\{-2.5, -9.9, -12.1, -8.9, -6.0, -4.8, 2.4\}$

- Goal is to estimate μ, assuming σ^2 is known.
- Natural Conjugate Gaussian Prior $p(\mu) = \mathcal{N}(0, 1)$
- Posterior is also Gaussian $p(\mu|y) = \mathcal{N}(\mu_p, \sigma_p^2)$

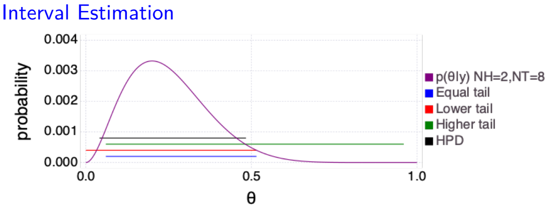


Point and Interval estimation

Point Estimation



Interval Estimation



Multiparameter Models

- Virtually every practical problem in statistics involves more than one parameter
- $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$
  - For example, $\boldsymbol{\theta} = (\mu, \sigma^2)$ when data is Gaussian $\mathcal{N}(\mu, \sigma^2)$
- Bayesian approach for estimating the joint posterior $p(\boldsymbol{\theta}|y)$
  - For example, $p(\mu, \sigma^2|y)$ in the case of the Gaussian
- Although a problem may include several parameters of interest
  - conclusions are often drawn about one, or a few, parameters at a time
  - e.g., one may be interested in a point-estimate of the mean μ.
- The parameters in the joint posterior about which one is not interested in making inferences are referred to as nuisance parameters

Averaging over 'nuisance parameters'

- A joint posterior for $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is expressed as $p(\theta_1, \theta_2|y)$
  - $p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)p(\theta_1, \theta_2)$
- Let us say, we are only interested in inference for θ_1
  - then θ_2 is a nuisance parameter.
- E.g., $y|\mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu (= \theta_1)$ and $\sigma^2 (= \theta_2)$ are unknown
  - interest commonly centers on μ.
- We seek conditional distribution of the parameter of interest, given the observed data: $p(\theta_1|y)$
- This is derived from the joint posterior density by marginalizing θ_2

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)d\theta_2$$

Averaging over 'nuisance parameters'

- Alternatively, joint posterior density can be factored to yield

$$p(\theta_1|y) = \int p(\theta_1|\theta_2, y)p(\theta_2|y)d\theta_2$$

- This can be treated as a mixture of conditional posterior $p(\theta_1|\theta_2, y)$
  - Where mixing weights are $p(\theta_2|y)$
  - Note that $p(\theta_2|y)$ depends on both data and prior model
- Rarely this integral is evaluated explicity, generally simulation is used
  - *Step 1:* First θ_2 is drawn from the marginal posterior $p(\theta_2|y)$
  - *Step 2:* Then θ_1 is drawn from its conditional posterior $p(\theta_1|\theta_2, y)$
  - Thus integration is performed indirectly

Normal data with a noninformative prior

Scenario: The temperatures, in Celsius, in Minneapolis during the first week of March 2018 are observed as $\{-2.5, -9.9, -12.1, -8.9, -6.0, -4.8, 2.4\}$

- Goal is to estimate μ and σ^2
- We will use the noninformative prior $p(\mu, \sigma^2) \propto 1/\sigma^2$
  - assumes that μ and σ^2 are independent
- Data points y follow a Gaussian distribution

$$\mathcal{N}(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big(-\frac{1}{2\sigma^2}(y-\mu)^2\Big)$$

- Likelihood

$$p(y|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\Big)$$

Joint posterior for Gaussian

$$p(y|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\Big) \qquad p(\mu, \sigma^2) \propto 1/\sigma^2$$

$$\begin{aligned}
p(\mu, \sigma^2|y) &\propto (\sigma^2)^{-n/2}\exp\left(-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^{n}(y_i \qquad -\mu)^2\right) \times \tfrac{1}{\sigma^2} \\
&= (\sigma^2)^{-1-n/2}\exp\left(-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - \mu)^2\right) \\
&\vdots \\
&= (\sigma^2)^{-(n+2)/2}\exp\left(-\tfrac{1}{2\sigma^2}\left[\textstyle\sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2\right]\right) \\
&= (\sigma^2)^{-(n+2)/2}\exp\left(-\tfrac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y}-\mu)^2\right]\right)
\end{aligned}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the sample variance

Determining the marginal posterior

$$p(\mu, \sigma^2|y) = (\sigma^2)^{-(n+2)/2}\exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y}-\mu)^2\right]\right)$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the sample variance

To determine the marginal posterior for μ, we need to do marginalization

$$p(\mu|y) = \int p(\mu, \sigma^2|y)d\sigma^2 = \int p(\mu|\sigma^2, y)p(\sigma^2|y)d\sigma^2$$

For this we need to determine $\underbrace{p(\mu|\sigma^2, y)}_{\text{conditional posterior}}$ and $\underbrace{p(\sigma^2|y)}_{\text{marginal posterior}}$

Choice of conditional posterior $p(\mu|\sigma^2, y)$

Estimating parameters of a Gaussian (unknown μ, known σ^2)

- Given a training data $y = \{y_1, \ldots, y_n\}$ drawn *i.i.d* from a Gaussian $\mathcal{N}(y|\mu, \sigma^2)$ with unknown mean μ and a given variance σ^2

$$\mathcal{N}(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big(-\frac{1}{2\sigma^2}(y-\mu)^2\Big)$$

- Choosing a Gaussian prior over μ

$$p(\mu) = (2\pi\sigma_0^2)^{-n/2} \exp\Big[\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\Big]$$

- Our posterior over parameter μ

$$\mathcal{N}(\mu|\mu_p, \sigma_p^2) = \frac{1}{\sqrt{2\pi\sigma_p^2}} exp\Big(-\frac{1}{2\sigma_p^2}(\mu-\mu_p)^2\Big)$$

where

$$\mu_p = \sigma_p^2\Big(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_i y_i}{\sigma^2}\Big); \qquad \sigma_p^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

Marginal posterior distribution $p(\sigma^2|y)$

- We can derive this as $p(\sigma^2|y) = \int p(\mu, \sigma^2|y)d\mu$

$$= \int (\sigma^2)^{-(n+2)/2} \exp\Big(-\frac{1}{2\sigma^2}\Big[(n-1)s^2 + \underbrace{n(\bar{y}-\mu)^2}_{\text{simple normal integral}}\Big]\Big)d\mu$$

$$\begin{aligned} p(\sigma^2|y) &\propto (\sigma^2)^{-(n+2)/2} exp\Big(-\tfrac{1}{2\sigma^2}(n-1)s^2\Big)\sqrt{2\pi\sigma^2/n} \\ &\propto (\sigma^2)^{-(n+1)/2} exp\Big(-\tfrac{(n-1)s^2}{2\sigma^2}\Big) \end{aligned}$$

This can be mapped to an Inverse-Gamma distribution

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{-(\alpha+1)} e^{-1/x\beta}$$

$$\text{where } \alpha = -1 + (n+1)/2 = \frac{n-1}{2} \text{ and } \beta = \frac{2}{s^2(n-1)}$$

Marginal posterior distribution $p(\mu|y)$

To determine the marginal posterior for μ, we need to do marginalization

$$p(\mu|y) = \int p(\mu, \sigma^2|y)d\sigma^2 = \int p(\mu|\sigma^2, y)p(\sigma^2|y)d\sigma^2$$

$$p(\mu|\sigma^2, y) = \mathcal{N}(\mu|\mu_p, \sigma_p^2); \qquad \mu_p = \sigma_p^2\Big(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_i y_i}{\sigma^2}\Big); \qquad \sigma_p^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

$$p(\sigma^2|y) = InvGamma(\alpha, \beta); \qquad \alpha = \frac{n-1}{2} \qquad \beta = \frac{2}{s^2(n-1)}$$

- While it may not always be possible, in this case $p(\mu|y)$ can be derived analytically
- We will use a sampling approach that is more widely applicable

Sampling Algorithm

$$p(\mu|y) = \int p(\mu, \sigma^2|y)d\sigma^2 = \int p(\mu|\sigma^2, y)p(\sigma^2|y)d\sigma^2$$

$$p(\mu|\sigma^2, y) = \mathcal{N}(\mu|\mu_p, \sigma_p^2); \qquad \mu_p = \sigma_p^2\Big(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_i y_i}{\sigma^2}\Big); \qquad \sigma_p^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

$$p(\sigma^2|y) = InvGamma(\alpha, \beta); \qquad \alpha = \frac{n-1}{2} \qquad \beta = \frac{2}{s^2(n-1)}$$

Algorithm:

- *Step 1*: Sampling σ^2
  - Compute α and β
  - Sample 1000 values of $\sigma^2|y \sim InvGamma(\alpha, \beta)$
- *Step 2*: Sampling μ
  - Assume a prior μ_0, σ_0^2
  - repeat for sample of $\sigma^2|y$
    - Compute posterior parameters μ_p, σ_p^2
    - Sample a value of μ from $\mu|\sigma^2, y \sim \mathcal{N}(\mu|\mu_p, \sigma_p^2)$

Julia code: Function to generate μ samples

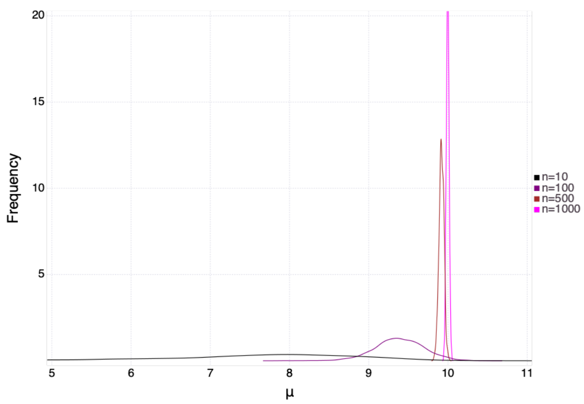
```julia
function generate_mu_samples(y,mu_0, sigma_0_sq)
    n = length(y);
    # Generating samples of Sigma^2
    s_sq = (sum((y.-mean(y)).^2))/(n-1);
    a = (n-1)/2;
    b = s_sq^2*(n-1)/2;
    d_sig_sq_given_y = InverseGamma(a,b);
    sample_sig_sq = rand(d_sig_sq_given_y,1000);
    ## using Sigma^2 to generate mu samples
    sigma_p_sq = 1./((1/sigma_0_sq).+(n./(sample_sig_sq)));
    mu_p = sigma_p_sq.*(mu_0/sigma_0_sq + sum(y)./(sample_sig_sq));
    sample_mu = zeros(1000);
    for i=1:1000
        d_mu_given_sig_sq_y = Normal(mu_p[i],sigma_p_sq[i]);
        sample = rand(d_mu_given_sig_sq_y,1);
        sample_mu[i] = sample[1];
    end
    return sample_mu;
end
```


Julia code: Generation and visualization of μ samples

```julia
sample_mu_n_10 = generate_mu_samples(rand(Normal(10,2),10),0,4);
sample_mu_n_100 = generate_mu_samples(rand(Normal(10,2),100),0,4);
sample_mu_n_500 = generate_mu_samples(rand(Normal(10,2),500),0,4);
sample_mu_n_1000 = generate_mu_samples(rand(Normal(10,2),1000),0,4);

myplot = Gadfly.plot(
layer(x=sample_mu_n_10, Geom.density,
      Theme(default_color=colorant"black")),
layer(x=sample_mu_n_100, Geom.density,
      Theme(default_color=colorant"purple")),
layer(x=sample_mu_n_500, Geom.density,
      Theme(default_color=colorant"brown")),
layer(x=sample_mu_n_1000, Geom.density,
      Theme(default_color=colorant"magenta")),
Coord.Cartesian(xmin=5, xmax=11,ymax=20),
Guide.manual_color_key("", ["n=10", "n=100", "n=500", "n=1000"],
["black","purple","brown","magenta"]), Guide.xlabel(" "),
Guide.ylabel("Frequency"), white_panel);
#draw(PNG("./figs/gaussian_mu_samples.png", 10inch, 7inch), myplot);
```

Visualizing density of μ samples



Multivariate Gaussian

Let $\boldsymbol{y} = (y_1, \ldots, y_k)$ have a multivariate Gaussian distribution, i.e. $\boldsymbol{y} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$.

The probability density function (pdf) for $\boldsymbol{y}$ is

$$p(\boldsymbol{y}) = (2\pi)^{-k/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)$$

Modelling height and weight of subjects in a survey.

Samples	Weight	Height
1	77.4	182.6
2	58.5	161.3
3	63.1	161.2
4	68.6	177.7
5	59.3	157.8
6	76.7	170.4

Bayesian Estimation for a Multivariate Gaussian

Assume we know $\boldsymbol{\Sigma}$, and we want to estimate $\boldsymbol{\mu} = (\mu_W, \mu_H)$,

- We first begin with a prior $p(\boldsymbol{\mu})$
  - preferably a natural conjugate prior $\mu \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$
- We write the likelihood

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(y_i - \boldsymbol{\mu})\right)$$

- We derive the posterior (similar to the univariate Gaussian case)

$$p(\boldsymbol{\mu}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k, , \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}_p)$$

$$\text{where } \boldsymbol{\Lambda}_p^{-1} = \boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}^{-1} \qquad \boldsymbol{\mu}_p = (\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_o + n\boldsymbol{\Sigma}^{-1}\bar{y})$$

Bayesian Estimation for a Multivariate Gaussian

Posterior over $\boldsymbol{\mu}$: $p(\boldsymbol{\mu}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k, , \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}_p)$

- We need to derive the marginal posteriors and the point estimates from them
- Recall from MvGaussian discussion

$$\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \boldsymbol{z} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{pmatrix}$$

- We derive the marginal posteriors (using this MvGaussian property)
  - $p(\mu_W|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) \sim \mathcal{N}(\mu_1, \Sigma_{11})$ and $p(\mu_H|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) \sim \mathcal{N}(\mu_2, \Sigma_{22})$
- We compute point estiamtes for μ_W and μ_H

Summary

- Bayesian approach generalizes to multiparameter models
- Even though we estimate joint posterior, we are often interested in estimates for some of the parameters only
- We need to derive the marginal posterior

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)d\theta_2 = \int p(\theta_1|\theta_2, y)p(\theta_2|y)d\theta_2$$

- Sampling approach
  - Sample $\theta_2 \sim p(\theta_2|y)$
  - Sample $\theta_1 \sim p(\theta_1|\theta_2, y)$
  - Plot the histogram of θ_1 samples
- Example
  - Univariate Gaussian when both μ and σ^2 are unknown
- Multivariate Gaussian when $\boldsymbol{\mu}$ is unknown and $\boldsymbol{\Sigma}$ is known