Workshop: Intro to Bayesian Learning
Lecture 2 - Multi-parameter models, Marginalization, Priors and Prediction

Mattias Villani

Department of Statistics
Stockholm University

mattiasvillani.com   @matvil   @matvil   mattiasvillani

Overview

- Marginalization
- Normal model with both parameters unknown
- Monte Carlo simulation - dipping a toe
- Prior elicitation
- Prediction

Marginalization

- Models with **multiple parameters** $\theta_1, \theta_2, ...$
- Examples: $y_i \overset{iid}{\sim} N(\theta, \sigma^2)$; multiple regression ...
- **Joint posterior distribution**

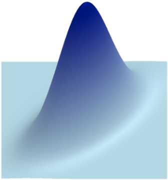
$$p(\theta_1, \theta_2, ..., \theta_p|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta_1, \theta_2, ..., \theta_p)p(\theta_1, \theta_2, ..., \theta_p).$$

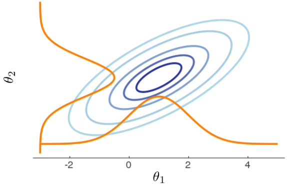
- In vector form

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

- **Marginalize** out parameters. **Marginal posterior** of θ_1:

$$p(\theta_1|\boldsymbol{y}) \quad = \quad \int p(\theta_1, \theta_2|\boldsymbol{y})d\theta_2$$





Normal model - normal prior

- **Model**

$$y_1, ..., y_n|\theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$$

- **Conjugate prior**

$$\theta|\sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$$
$$\sigma^2 \sim Inv\text{-}\chi^2(\nu_0, \sigma_0^2)$$

Scaled inverse chi-squared distribution $\mathrm{Inv}-\chi^2(\nu,\tau^2)$

- Variant of **inverse Gamma**.

$$\mathrm{Inv}\text{−}\chi^2(\nu,\tau^2) \quad \Longleftrightarrow \quad \nu\tau^2\frac{1}{X} \text{ where } X \sim \chi^2_\nu$$

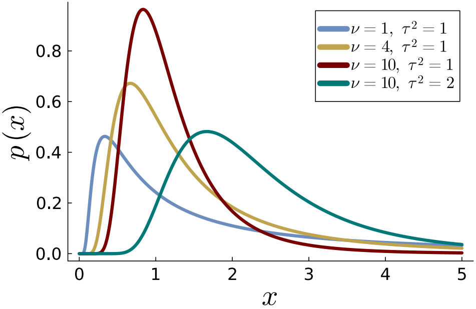
Inv-χ^2 distribution

$X \sim \mathrm{Inv}\text{−}\chi^2(\nu,\tau^2)$, $X \in (0,\infty)$

$$p(x) = \frac{(\tau^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \; \frac{\exp\left(\frac{-\nu\tau^2}{2x}\right)}{x^{1+\nu/2}}$$

$$\mathbb{E}(X) = \frac{\nu}{\nu-2}\tau^2$$

$$\mathbb{V}(X) = \frac{2\nu^2\tau^4}{(\nu-2)^2(\nu-4)}$$



Normal model with normal prior

Posterior

$$\theta|\mathbf{y}, \sigma^2 \sim N\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right)$$
$$\sigma^2|\mathbf{y} \sim \mathit{Inv}\text{-}\chi^2(\nu_n, \sigma_n^2).$$

where

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\sigma_n^2 &= \text{see book}
\end{aligned}$$

Normal model with normal prior

Posterior

$$\theta|\mathbf{y}, \sigma^2 \sim N\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right)$$
$$\sigma^2|\mathbf{y} \sim \mathit{Inv}\text{-}\chi^2(\nu_n, \sigma_n^2).$$

where

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\sigma_n^2 &= \text{see book}
\end{aligned}$$

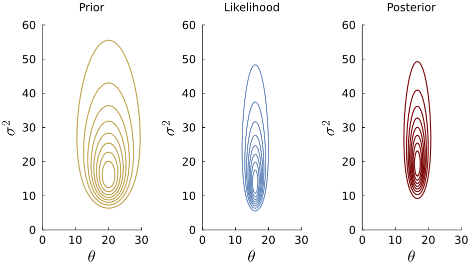
Marginal posterior

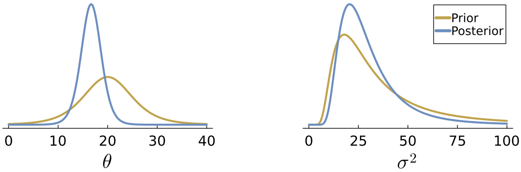
$$\theta|\mathbf{y} \sim t_{\nu_n}\left(\mu_n, \sigma_n^2/\kappa_n\right)$$

Internet speed data - joint and marginal posteriors

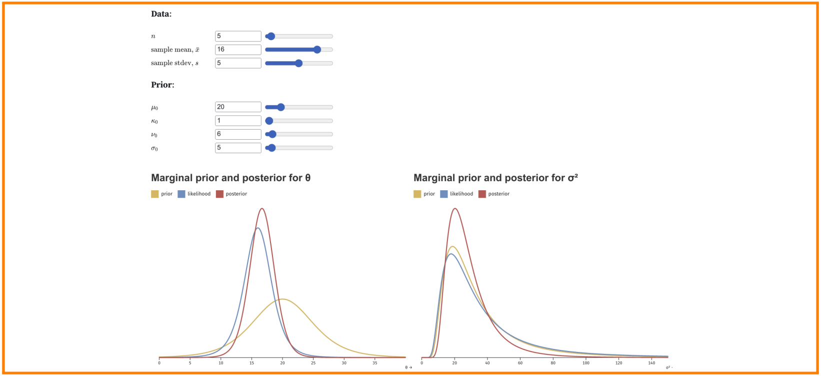
- Prior:

$$\theta|\sigma^2 \sim N\left(20, \frac{\sigma^2}{1}\right) \text{ and } \sigma^2 \sim \text{Inv-}\chi^2\left(\nu_0 = 5, \sigma_0^2 = 5^2\right)$$

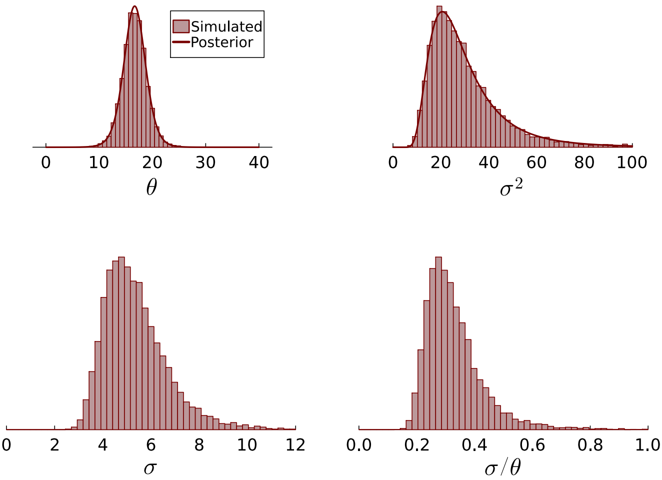




Gaussian model - conjugate prior



Monte Carlo simulation



Simulating from posterior - pseudo code

Posterior simulation - iid Gaussian with conjugate prior.

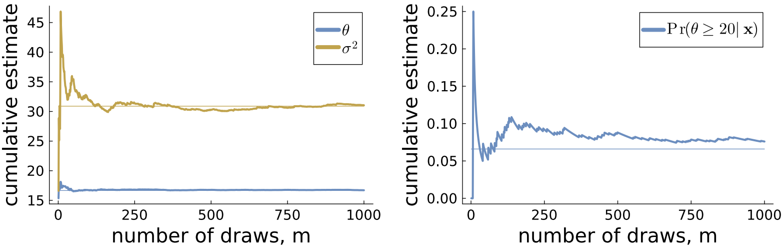
Input: data $\mathbf{x} = (x_1, \ldots, x_n)$
number of posterior draws m.
compute μ_n, σ_n^2, κ_n and ν_n using Figure 50.
for i in *1:m* **do**
  $\sigma^2 \leftarrow$ rInvChi2(ν_n,σ_n^2)
  $\theta \leftarrow$ rNormal(μ_n,σ^2/κ_n)
end
Output: m draws for θ and σ^2 from joint posterior.

Function rInvChi2(ν,τ^2)
  x = rChi2(ν)
  y = $\nu\tau^2/x$
  return y

Simulating from posterior - output

draw	θ	σ^2	σ/θ	$\theta \geq 20$
1	18.165	18.451	0.236	0
2	20.431	29.943	0.267	1
3	15.565	29.094	0.346	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
10,000	16.400	21.668	0.283	0
Mean	16.645	30.813	0.330	0.066

Monte Carlo simulation



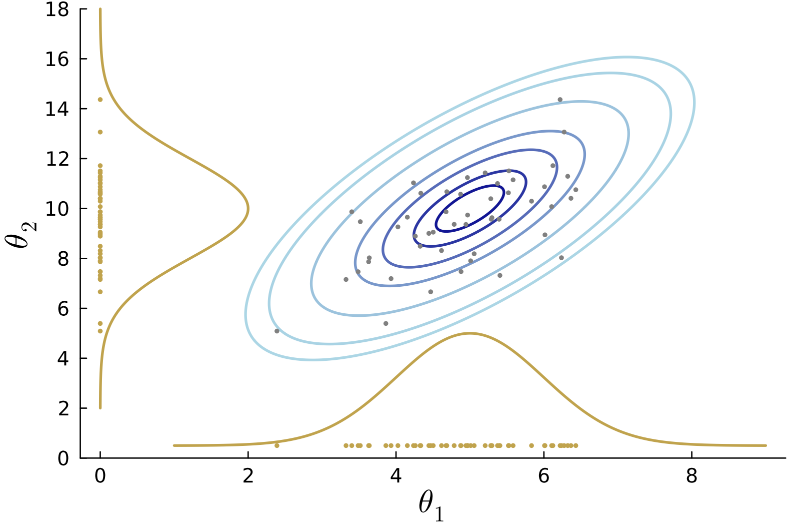
- Law of large numbers for **consistency**:

$$\bar{\theta}_{1:m} \equiv \frac{1}{m}\sum_{i=1}^{m}\theta^{(i)} \overset{\text{a.s.}}{\to} \mathbb{E}(\theta|\mathbf{x}) \text{ as } m \to \infty$$

- Central limit theorem for the **accuracy**:

$$\bar{\theta}_{1:m} \sim N\left(\mathbb{E}(\theta|\mathbf{x}), \frac{\mathbb{V}(\theta|\mathbf{x})}{m}\right)$$

Simulation from marginals by selection



Multinomial model - Dirichlet prior

Prior

α_1	3
α_2	3
α_3	5

Data

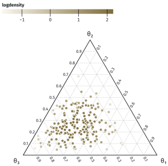
n_1	5
n_2	10
n_3	3
number of draws in plot	200

Simulate new draws!

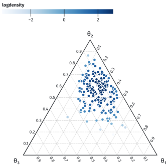
Prior distribution

$\mathbb{E}(\boldsymbol{\theta}) = [0.2727, 0.2727, 0.4545]$

$\mathbb{S}(\boldsymbol{\theta}) = [0.1286, 0.1286, 0.1437]$



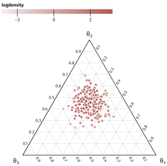
Likelihood (normalized)



Posterior distribution

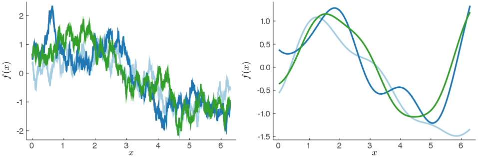
$\mathbb{E}(\boldsymbol{\theta}|\boldsymbol{n}) = [0.2759, 0.4483, 0.2759]$

$\mathbb{S}(\boldsymbol{\theta}|\boldsymbol{n}) = [0.0816, 0.0908, 0.0816]$



Prior - where to get them?

- **Expert knowledge**
- **Past data**, **other data**.
- **Smoothness priors**
- **Regularization priors** (Ridge and Lasso are priors)
- **Non-informative priors**
- **Invariant priors**

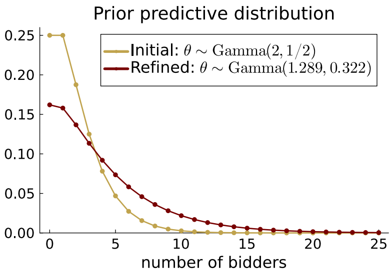


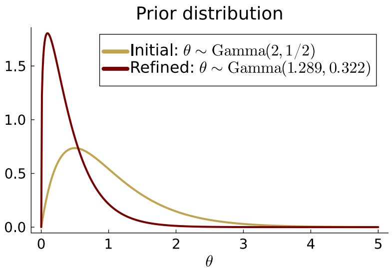
Prior elication from prior predictive distribution

- Easier to **reason about data** than model parameters.
- **Prior predictive distribution**:
  - generate data from the model $p(y|\theta)$
  - with parameters generated from the prior $p(\theta)$

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

- Example: Poisson model for eBay auctions. Experts says:
  - Average number of bidders in an auction is $\mathbb{E}(y) = 4$
  - Only 2 of auctions have more than 15 bidders.





Prior predictive distribution - Poisson model

α : 2

β : 0.5

Quantile: 15.01

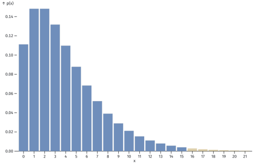
Prior predictive properties:

$$\mathrm{E}(X) = \frac{\alpha}{\beta} = 4.00$$

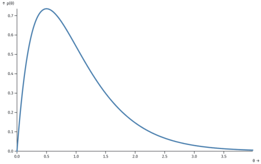
$$\mathrm{StDev}(X) = \frac{\alpha(1+\beta)}{\beta^2} = 12.0$$

$$P(X \leq 15.01) = 0.9904$$

Prior predictive distribution



Prior distribution



Prediction/Forecasting

- **Posterior predictive density** for new $\tilde{y}$ given observed iid data $\boldsymbol{y} = (y_1, \ldots, y_n)$

$$p(\tilde{y}|\boldsymbol{y}) = \int_\theta p(\tilde{y}|\theta)p(\theta|\boldsymbol{y})d\theta$$

- **Parameter uncertainty** in $p(\tilde{y}|\boldsymbol{y})$ by **averaging over** $p(\theta|\boldsymbol{y})$.

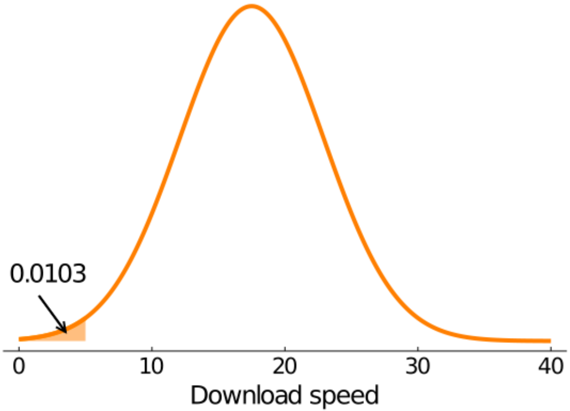
- Predictive distribution in model $y_1, \ldots, y_n|\theta \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$

$$\tilde{y}|\boldsymbol{y} \sim N(\mu_n, \sigma^2 + \tau_n^2)$$

where μ_n and τ_n^2 are the posterior mean and variance of θ.

Predictive distribution - Internet speed data

- My Netflix starts to buffer at speeds < 5Mbit. 🤬



Prediction by simulation

- The integral in the predictive distribution is often intractable.

Simulation algorithm:

1. Generate a **posterior draw** $\theta^{(1)} \sim N(\mu_n, \tau_n^2)$
2. Generate a **predictive draw** $\tilde{y}^{(1)} \sim N(\theta^{(1)}, \sigma^2)$
3. Repeat Steps 1 and 2 N times to output:

- Sequence of posterior draws: $\theta^{(1)}, ..., \theta^{(N)}$
- Sequence of predictive draws: $\tilde{y}^{(1)}, ..., \tilde{y}^{(N)}$.

Bayesian decision making

- Let θ be an **unknown quantity**. **State of nature**.
  - Future inflation
  - Disease.

- Let $a \in \mathcal{A}$ be an **action**.
  - Interest rate
  - Treatment.

- Choosing action a when state of nature is θ gives **utility**

$$U(a, \theta)$$

- Choose action that **maximizes posterior expected utility**

$$a_{\mathrm{opt}} = \mathrm{argmax}_{a \in \mathcal{A}} \, \mathbb{E}_{p(\theta|y)} \left(U(a, \theta) \right),$$