Lecture 1: Introduction

STATS305C: Applied Statistics III

Scott Linderman

March 26, 2022

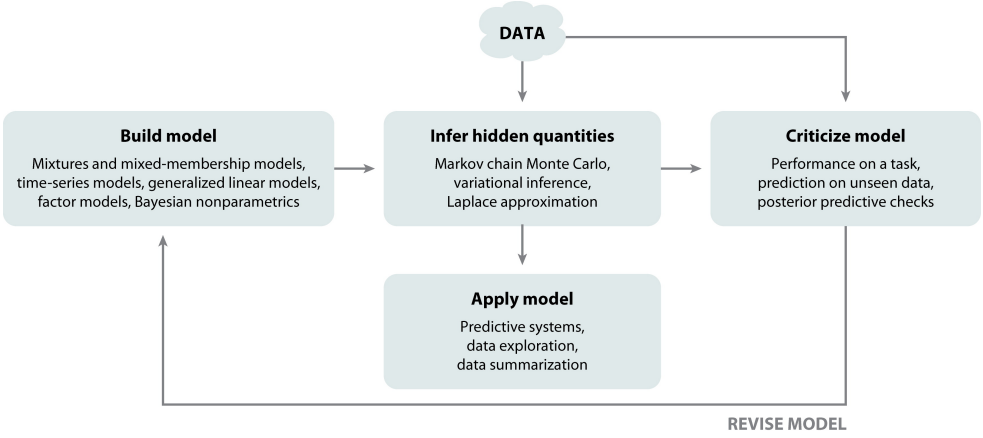
What Is This Course About?

This course is about modeling and inference with high dimensional data.

What is the end goal?

- **Predict:** given features, estimate labels or outputs
- **Simulate:** given partial observations, generate the rest
- **Summarize:** given high dimensional data, find low-dimensional factors of variation
- **Visualize:** given high dimensional data, find informative 2D/3D plots
- **Decide:** given past actions/outcomes, which choice is best?
- **Understand:** what generative mechanisms gave rise to this data?

Box's Loop



Blei DM. 2014.
Annu. Rev. Stat. Appl. 1:203–32

Blei [2014].

Bayesian Approach

1. A **model** is a **joint distribution** of parameters and data.
2. An **inference algorithm** computes the **posterior distribution** of parameters given data.
3. **Model criticism** and application are based on **posterior expectations**.

Notation

Let

- $\boldsymbol{\theta}$ denote parameters
- $\boldsymbol{\eta}$ denote hyperparameters
- $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ denote the set of data points
- $p(\boldsymbol{\theta}, \boldsymbol{X} \mid \boldsymbol{\eta})$ denote a model (i.e. joint distribution)
- $p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{\eta})$ denote the posterior distribution
- $p(\boldsymbol{X} \mid \boldsymbol{\eta})$ denote the marginal likelihood of the data

Generally, lowercase bold letters denote vectors, uppercase bold letters denote matrices, and regular characters denote scalars.

Bayes' Rule

$$\underbrace{p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{\eta})}_{\text{posterior distribution}} = \frac{\overbrace{p(\boldsymbol{\theta}, \boldsymbol{X} \mid \boldsymbol{\eta})}^{\text{joint distribution}}}{\underbrace{p(\boldsymbol{X} \mid \boldsymbol{\eta})}_{\text{marginal likelihood}}} = \frac{\overbrace{p(\boldsymbol{\theta} \mid \boldsymbol{\eta})}^{\text{prior}} \overbrace{p(\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{\eta})}^{\text{likelihood}}}{\int p(\boldsymbol{\theta}, \boldsymbol{X} \mid \boldsymbol{\eta}) \, \mathrm{d}\boldsymbol{\theta}} \tag{1}$$

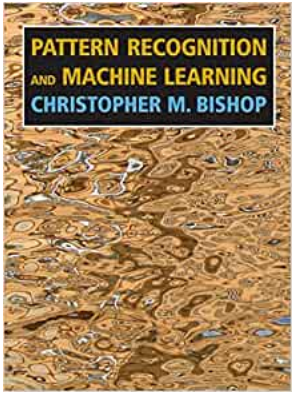
Tentative Course Outline

Model	Algorithm	Application
Multivariate Normal Models	Conjugate Inference	Bayesian Linear Regression
Hierarchical Models	MCMC (MH & Gibbs)	Modeling Polling Data
Probabilistic PCA & Factor Analysis	MCMC (HMC)	Images Reconstruction
Mixture Models	EM & Variational Inference	Image Segmentation
Mixed Membership Models	Coordinate Ascent VI	Topic Modeling
Variational Autoencoders	Black Box, Amortized VI	Image Generation
State Space Models	Message Passing	Segmenting Video Data
Bayesian Nonparametrics	Fancy MCMC	Modeling Neural Spike Trains

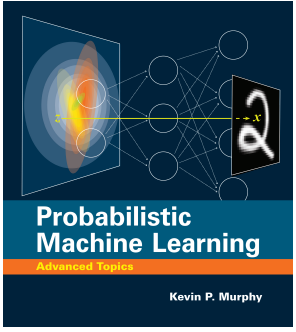
Logistics

- MW 11:30am-1pm Lecture
- Office hours Mon 1-2pm (Scott) and Tues 5:30-7pm (Matt)
- 8 weekly assignments, released Wednesday after class, due 11:59pm the following Wednesday.
- Assignments will be in Jupyter notebooks with both coding and math problems.
- You can run R or Julia in the notebooks if you want, but we will use Python for demos.
- Final project: details to come

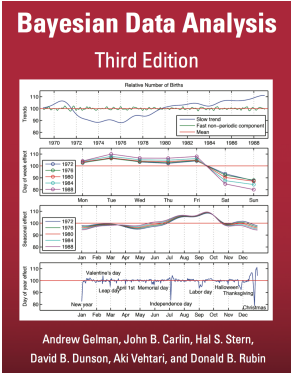
Books



https://tinyurl.com/yckjp266



https://probml.github.io/pml-book/book2.html



http://www.stat.columbia.edu/~gelman/book/

Grading

- 7 homeworks (we drop the lowest of the 8) × 10% each = 70%
- Final project = 25%
- Class participation = 5%
- You must do a project even if you take the course pass/fail.

Questions?

Warm-up: Normal Model with Unknown Mean

Example: Modeling SAT scores. Suppose we have scores of N students from one class. Assume the scores are well modeled as Gaussian random variables and that they are conditionally independent given the mean and variance. For now, assume we know the variance but not the mean.

Notation: Let,

- $x_n \in \mathbb{R}$ denote the score of the n-th student,
- $\mu \in \mathbb{R}$ denote the (unknown) mean of the distribution, and
- $\sigma^2 \in \mathbb{R}_+$ denote the (known) variance of the distribution.
- μ_0, σ_0^2 denote the mean and variance of the Gaussian prior on μ.

Model:

$$x_n \mid \mu, \sigma^2 \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2) \quad \text{for } n = 1, \ldots, N \tag{2}$$

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2) \tag{3}$$

Draw the Graphical Model

Warm-up: Normal Model with Unknown Mean II

Goal: Infer $p(\mu \mid \boldsymbol{X}, \boldsymbol{\eta})$, the posterior distribution over model parameters given data and hyperparameters $\boldsymbol{\eta} = (\sigma^2, \mu_0, \sigma_0^2)$.

$$p(\mu \mid \boldsymbol{X}, \boldsymbol{\eta}) \propto p(\mu \mid \boldsymbol{\eta}) \prod_{n=1}^{N} p(x_n \mid \mu, \boldsymbol{\eta}) \tag{4}$$

$$= \mathcal{N}(\mu \mid \mu_0, \sigma_0^2) \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \sigma^2) \tag{5}$$

$$\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\} \prod_{n=1}^{N} \exp\left\{-\frac{1}{2\sigma^2}(x_n - \mu)^2\right\} \tag{6}$$

$$\propto \exp\left\{-\frac{1}{2} J_N \mu^2 + h_N \mu\right\} \tag{7}$$

where

$$J_N = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \quad \text{and} \quad h_N = \frac{\mu_0}{\sigma_0^2} + \sum_{n=1}^{N} \frac{x_n}{\sigma^2} \tag{8}$$

Warm-up: Normal Model with Unknown Mean III

Completing the square: Show that

$$\exp\left\{-\frac{1}{2}J_N\mu^2 + h_N\mu\right\} \propto \mathcal{N}(\mu \mid \mu_N, \sigma_N^2) \tag{9}$$

where $\sigma_N^2 = J_N^{-1}$ and $\mu_N = J_N^{-1}h_N$.

Warm-up: Normal Model with Unknown Mean IV

Thus, $p(\mu \mid \boldsymbol{X}, \eta) = \mathcal{N}(\mu \mid \mu_N, \sigma_N^2)$ where

$$\mu_N = \frac{\sigma_0^2 \sigma^2}{\sigma^2 + N\sigma_0^2} \left(\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2} \sum_{n=1}^{N} x_n \right) \tag{10}$$

$$= \frac{\sigma^2}{\sigma^2 + N\sigma_0^2} \mu_0 + \frac{N\sigma_0^2}{\sigma^2 + N\sigma_0^2} \mu_{\mathrm{ML}}. \tag{11}$$

and $\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n$ is the maximum likelihood estimate.

Questions:

1. What is the posterior mean in the limit where $N \to \infty$?
2. What happens when $\mu_0 = 0$ and $\sigma_0^2 \to \infty$ (i.e. when the prior is uninformative)?

Warm-up: Normal Model with Unknown Mean V

What about the posterior variance?

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \tag{12}$$

Same questions: what happens when $N \to \infty$ or when $\sigma_0^2 \to \infty$?

Normal Model with Unknown *Precision*

Now suppose we know the mean μ but not the variance σ^2. Our calculations will be a little simpler if we work with the *precision* instead, $\lambda = 1/\sigma^2$. Then,

$$p(x \mid \mu, \lambda) = \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda}{2}(x-\mu)^2\right\} \tag{13}$$

What would be a nice prior distribution for the precision?

- Support for the non-negative reals
- Control of mean and variance
- Conjugate with the Gaussian likelihood

Chi-Squared (χ^2) Distribution

Let $z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for $i = 1, \ldots, \nu$ and $\lambda = \sum_{i=1}^{\nu} z_i^2$. (Assume ν is an integer.) Then,

$$\lambda \sim \chi^2(\nu_0) \tag{14}$$

where ν is called the *degrees of freedom*.

The χ^2 pdf is,

$$\chi^2(\lambda \mid \nu_0) = \frac{1}{2^{\frac{\nu_0}{2}}\Gamma(\frac{\nu_0}{2})} \lambda^{\frac{\nu_0}{2}-1} e^{-\frac{\lambda}{2}}. \tag{15}$$

The chi-squared distribution is a special case of the gamma distribution,

$$\chi^2(\nu_0) = \text{Ga}\left(\frac{\nu_0}{2}, \frac{1}{2}\right). \tag{16}$$

(Here, using the rate parameterization of the gamma distirbution.)

Adding a scale parameter

We can add a scale parameter by considering $z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \lambda_0)$ where λ_0 is the variance and defining $\lambda = \frac{1}{\nu_0} \sum_{i=1}^{\nu_0} z_i^2$. (Note that we have defined it to be the **average** sum of squares!)

We say λ is a *scaled* chi-squared random variable,

$$\lambda \sim \chi^2(\nu_0, \lambda_0) \tag{17}$$

where λ_0 is called the *scale* parameter.

Question: What is the mean of the scaled chi-squared distribution? What is its density?

Scaled χ^2 Distribution

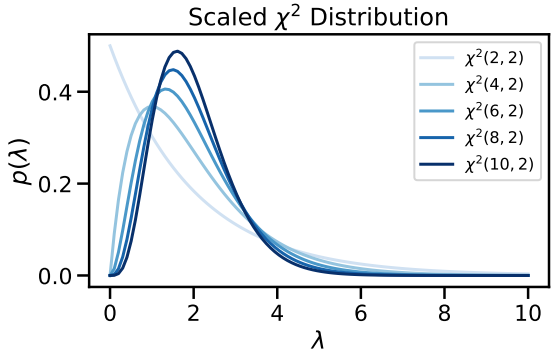


Figure: The $\chi^2(\nu_0, \lambda_0)$ pdf for $\lambda_0 = 2$ and varying degrees of freedom ν_0. In all cases, the mean is $\mathbb{E}[\lambda] = \lambda_0$, but the variance shrinks as ν_0 increases.

Normal Model with Unknown Precision II

Model:

$$\lambda \sim \chi^2(\nu_0, \lambda_0) \tag{18}$$

$$x_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1/\lambda) \tag{19}$$

Exercise: Draw the graphical model.

Normal Model with Unknown Precision III

The χ^2 distribution is conjugate with the Gaussian likelihood. Letting $\boldsymbol{\eta} = (\mu, \nu_0, \lambda_0)$, we have,

$$p(\lambda \mid \boldsymbol{X}, \boldsymbol{\eta}) \propto \chi^2(\lambda \mid \nu_0, \lambda_0) \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \tfrac{1}{\lambda}) \tag{20}$$

$$\propto \lambda^{\frac{\nu_0}{2}-1} e^{-\frac{\nu_0 \lambda}{2\lambda_0}} \prod_{n=1}^{N} \lambda^{\frac{1}{2}} e^{-\frac{\lambda}{2}(x_n-\mu)^2} \tag{21}$$

$$\propto \lambda^{\frac{\nu_N}{2}-1} e^{-\frac{\nu_N \lambda}{2\lambda_N}} \tag{22}$$

where

$$\nu_N = \nu_0 + N \quad \text{and} \quad \lambda_N = \nu_N \left(\frac{\nu_0}{\lambda_0} + \sum_{n=1}^{N} (x_n - \mu)^2 \right)^{-1}. \tag{23}$$

Question: What does the posterior mean converge to as $\nu_0 \to 0$?

Normal Model with Unknown *Variance*

If the precision $\lambda \sim \chi^2(\nu_0, \lambda_0)$, then the variance $\sigma^2 = 1/\lambda$ is a scaled inverse χ^2 random variable,

$$\sigma^2 \sim \chi^{-2}(\nu_0, \sigma_0^2) \tag{24}$$

where $\sigma_0^2 = 1/\lambda_0$.

Its pdf can be found with the change of measure formula. Let $f(\sigma^2) = 1/\sigma^2$.

$$p(\sigma^2 \mid \nu_0, \sigma_0^2) = \left| \frac{\mathrm{d}f(\sigma^2)}{\mathrm{d}\sigma^2} \right| \chi^2(f(\sigma^2) \mid \nu_0, 1/\sigma_0^2) \tag{25}$$

$$= \frac{\left(\frac{\nu_0 \sigma_0^2}{2}\right)^{\nu_0/2}}{\Gamma(\frac{\nu_0}{2})} (\sigma^2)^{-\frac{\nu_0}{2}-1} e^{-\frac{\nu_0 \sigma_0^2}{2\sigma^2}} \tag{26}$$

$$\triangleq \chi^{-2}(\sigma^2 \mid \nu_0, \sigma_0^2). \tag{27}$$

The scaled inverse chi-squared is a special case of the inverse gamma distribution, $\chi^{-2}(\nu_0, \sigma_0^2) \equiv \mathrm{IGa}(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$, again using the rate parameterization.

Scaled Inverse χ^2 Distribution

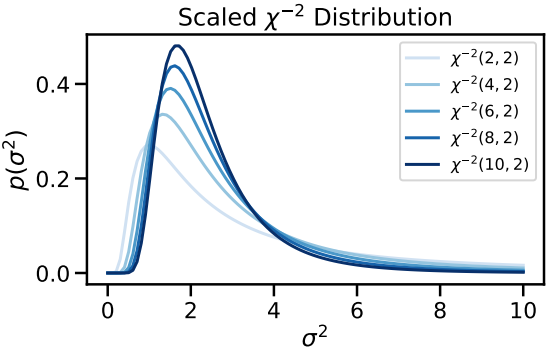


Figure: The $\chi^{-2}(\nu_0, \sigma_0^2)$ pdf for $\sigma_0^2 = 2$ and varying degrees of freedom ν_0. In all cases, the mean is $\mathbb{E}[\sigma^2] = \frac{\nu_0}{\nu_0 - 2}\sigma_0^2$ (for $\nu_0 > 2$), but the variance shrinks as ν_0 increases.

Normal Model with Unknown Variance II

Now let's parameterize the model in terms of the variance,

$$\sigma^2 \sim \chi^{-2}(\nu_0, \sigma_0^2) \tag{28}$$

$$x_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2) \tag{29}$$

Exercise: Draw the graphical model.

Normal Model with Unknown Variance III

The χ^{-2} distribution is conjugate with the Gaussian likelihood. Letting $\boldsymbol{\eta} = (\mu, \nu_0, \sigma_0^2)$,

$$p(\sigma^2 \mid \boldsymbol{X}, \boldsymbol{\eta}) \propto \chi^{-2}(\sigma^2 \mid \nu_0, \sigma_0^2) \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \sigma^2) \tag{30}$$

$$\propto (\sigma^2)^{-\frac{\nu_0}{2}-1} e^{-\frac{\nu_0 \sigma_0^2}{2\sigma^2}} \prod_{n=1}^{N} (\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2} \tag{31}$$

$$\propto \chi^{-2}(\sigma^2 \mid \nu_N, \sigma_N^2) \tag{32}$$

where

$$\nu_N = \nu_0 + N \quad \text{and} \quad \sigma_N^2 = \frac{1}{\nu_N}\left(\nu_0 \sigma_0^2 + \sum_{n=1}^{N} (x_n - \mu)^2\right) \tag{33}$$

Question: What does σ_N^2 converge to as $\nu_0 \to 0$?

Normal Model with Unknown Mean and Variance

Finally, let's assume both the mean and the variance are unknown. Then, the conjugate prior is a *normal inverse chi-squared* (NIX) distribution.

$$
\begin{aligned}
p(\mu, \sigma^2) &= p(\sigma^2)\, p(\mu \mid \sigma^2) && (34)\\
&= \chi^{-2}(\sigma^2 \mid \nu_0, \sigma_0^2)\, \mathcal{N}(\mu \mid \mu_0, \sigma^2/\kappa_0) && (35)\\
&\triangleq \mathrm{NIX}(\mu, \sigma^2 \mid \mu_0, \kappa_0, \nu_0, \sigma_0^2) && (36)
\end{aligned}
$$

Our first multivariate distribution!

Exercise: Draw the graphical model.

Normal Inverse χ^2 Distribution

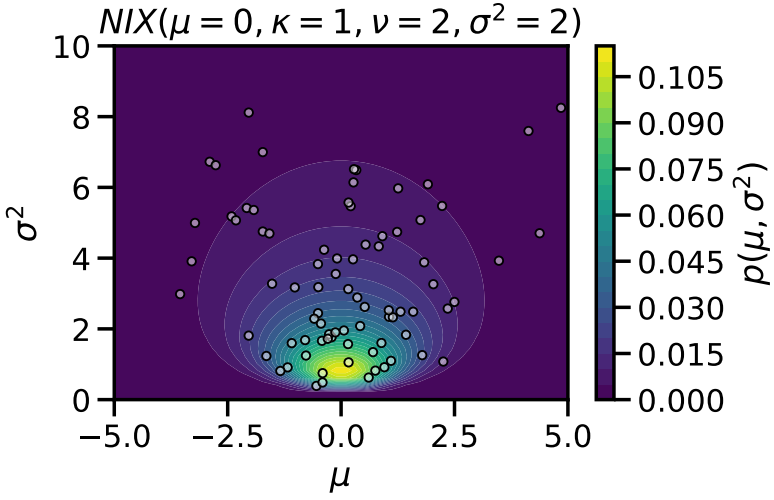


Figure: The $\text{NIX}(\mu_0, \kappa_0, \nu_0, \sigma_0^2)$ pdf. Note the dependence: large values of σ^2 imply larger variance in $p(\mu \mid \sigma^2)$.

Normal Model with Unknown Mean and Variance II

Exercise: Show that $p(\mu, \sigma^2 \mid \boldsymbol{X}, \boldsymbol{\eta}) = \mathrm{NIX}(\mu, \sigma^2 \mid \mu_N, \kappa_N, \nu_N, \sigma_N^2)$ where,

$$\nu_N = \nu_0 + N \tag{37}$$

$$\kappa_N = \kappa_0 + N \tag{38}$$

$$\mu_N = \frac{1}{\kappa_N}\left(\kappa_0 \mu_0 + \sum_{n=1}^{N} x_n\right) \tag{39}$$

$$\sigma_N^2 = \frac{1}{\nu_N}\left(\nu_0 \sigma_0^2 + \kappa_0 \mu_0^2 + \sum_{n=1}^{N} x_n^2 - \kappa_N \mu_N^2\right) \tag{40}$$

Question: Take the uninformative limit where $\nu_0 \to 0$ and $\kappa_0 \to 0$. What is the posterior mean?

The Posterior Marginals

Question: What is the posterior marginal distribution over the variance, $p(\sigma^2 \mid \boldsymbol{X}, \boldsymbol{\eta})$?

The posterior marginal distribution over the mean is,

$$p(\mu \mid \boldsymbol{X}, \boldsymbol{\eta}) = \int p(\mu, \sigma^2 \mid \boldsymbol{X}, \boldsymbol{\eta}) \, \mathrm{d}\sigma^2 \tag{41}$$

$$= \int \chi^{-2}(\sigma^2 \mid \nu_N, \sigma_N^2) \, \mathcal{N}(\mu \mid \mu_N, \sigma^2/\kappa_N) \tag{42}$$

$$= \mathrm{St}\big(\mu \mid \nu_N, \mu_N, \sigma_N^2/\kappa_N\big), \tag{43}$$

where St denotes a Student's t distribution with ν d.o.f., location μ, and scale σ. Its density is,

$$\mathrm{St}(x \mid \nu, \mu, \sigma^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\pi \nu \sigma^2}} \left[1 + \frac{\Delta^2}{\nu}\right]^{-\frac{\nu+1}{2}} \tag{44}$$

where $\Delta^2 = \left(\frac{x-\mu}{\sigma}\right)^2$ is the squared Mahalanobis distance. Its mean is μ and its variance is $\frac{\nu}{\nu+2}\sigma^2$.

Student's t Distribution

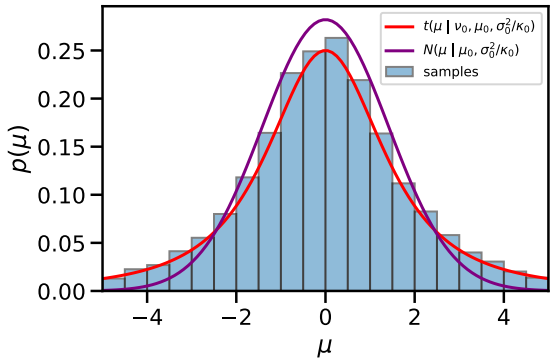


Figure: The $\mathrm{St}(\mu \mid \nu, \mu, \sigma^2/\kappa)$ is heavy tailed since it mixes over Gaussian distributions with varying scales.

Question: What are the mean and variance of the posterior marginal distribution over μ under an uninformative NIX prior?

Posterior Credible Intervals

Under an uninformative prior,

$$\nu_N = N \qquad \mu_N = \frac{1}{N}\sum_{n=1}^{N} x_n = \mu_{\mathrm{ML}} \tag{45}$$

$$\kappa_N = N \qquad \sigma_N^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_N)^2 = \sigma_{\mathrm{ML}}^2. \tag{46}$$

so $\mu \mid \boldsymbol{X}, \boldsymbol{\eta} \sim \mathrm{St}(N, \mu_{\mathrm{ML}}, \sigma_{\mathrm{ML}}^2/N)$.

The Bayesian way of testing whether $\mu \neq \mu^*$ is to see if μ^* is in the $1-\alpha$ central **posterior credible interval**,

$$\mathscr{I}_\alpha = \left[F_{\mathrm{St}}^{-1}(\tfrac{\alpha}{2} \mid N, \mu_{\mathrm{ML}}, \sigma_{\mathrm{ML}}^2/N),\ F_{\mathrm{St}}^{-1}(1-\tfrac{\alpha}{2} \mid N, \mu_{\mathrm{ML}}, \sigma_{\mathrm{ML}}^2/N)\right] \tag{47}$$

where F_{St}^{-1} is the quantile function of the Student's t distribution.

Posterior Credible Intervals II

Alternatively, note that

$$\frac{\mu - \mu_{\mathrm{ML}}}{\sigma_{\mathrm{ML}}/\sqrt{N}} \mid \boldsymbol{X}, \boldsymbol{\eta} \sim \mathrm{St}(N, 0, 1). \tag{48}$$

Testing if $\mu^* \in \mathscr{I}$ is equivalent to testing if $t \triangleq \frac{\mu^* - \mu_{\mathrm{ML}}}{\sigma_{\mathrm{ML}}/\sqrt{N}}$ is in,

$$\mathscr{I}_\alpha = \left[F_{\mathrm{St}}^{-1}\left(\tfrac{\alpha}{2} \mid N, 0, 1\right),\ F_{\mathrm{St}}^{-1}\left(1 - \tfrac{\alpha}{2} \mid N, 0, 1\right) \right] \tag{49}$$

Question: How does this compare to frequentist hypothesis testing?

References I

David M Blei. Build, compute, critique, repeat: Data analysis with latent variable models. *Annu. Rev. Stat. Appl.*, 1(1):203–232, January 2014.