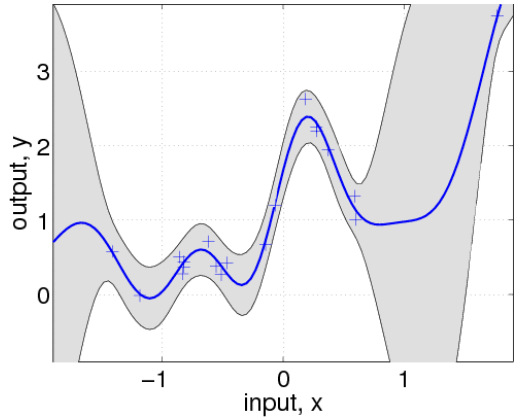
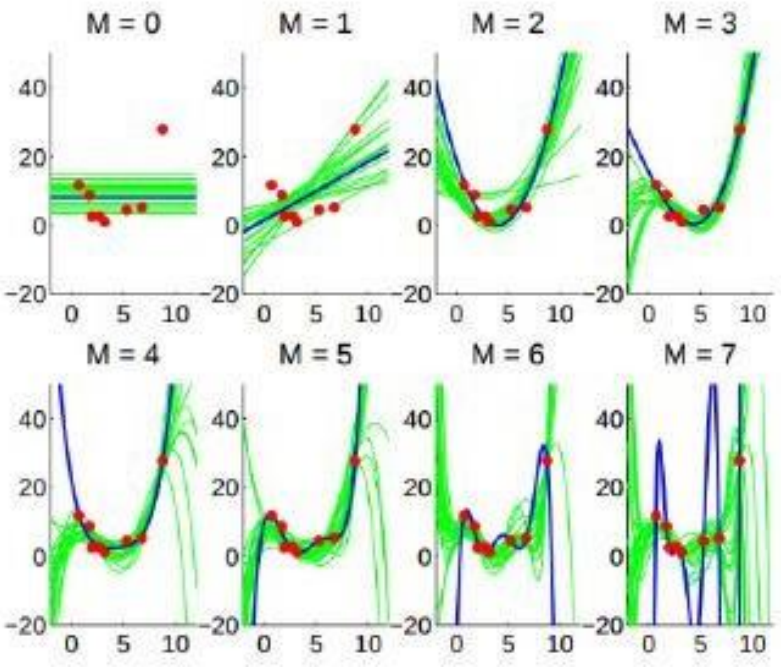
Probabilistic Bayesian Modelling

Debaditya Roy

Probabilistic Model

- x – an observation (random variable/vector)
- $X = \{x_1, x_2, \ldots, x_n\}$, set of observations, evidence, data
- Probabilistic model – a mathematical form which provides stochastic information about the random variable X
- θ - parameters of a model
- M – hyperparameters of a model





Modelling Goals

- Estimation (of the underlying model parameters) - $p(\theta, m|X)$
  - Understand
  - Generate new data

- Prediction $-\ p(x^*|\theta)$ or $p(x^*|X)$, x^* is a new observation

- Model comparison – $p(X|\ \theta_1)\ >\ p(X|\ \theta_2)$

- Solving the first goal helps solve the second and third goals

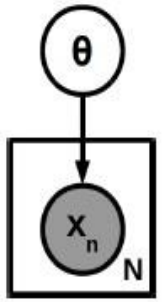
Some probabilities of interest

- Likelihood function $p(\boldsymbol{x}|\theta)$ or the "observation model" specifies how data is generated
  - Measures data fit (or "loss") w.r.t. the given parameter θ
- Prior distribution $p(\theta)$ specifies how likely different parameter values are *a priori*
  - Also corresponds to imposing a "regularizer" over θ
- Domain knowledge can help in the specification of the likelihood and the prior

Note: We are talking about probability distributions and not single (point) probabilities

Maximum Likelihood Estimation

- Perhaps the simplest way is to find θ that makes the observed data most likely or most probable



- Formally, find θ that maximizes the probability of the observed data

$$\hat{\theta} = \arg\max_{\theta} \log p(\mathbf{X}|\theta)$$

- However, this gives a single "point" estimate of θ. Doesn't tell us about the uncertainty in θ

Rules of Probability

- Keep in mind these two simple rules of probability: sum rule and product rule

$$P(a) = \sum_b P(a,b) \quad \text{(Sum Rule)}$$

$$P(a,b) = P(a)P(b|a) = P(b)P(a|b) \quad \text{(Product Rule)}$$

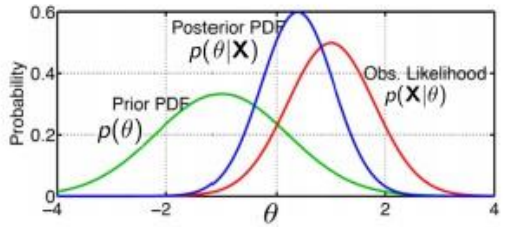
- Note: For continuous random variables, sum is replaced by integral: $P(a) = \int P(a,b)db$
- Another rule is the Bayes rule (can be easily obtained from the above two rules)

$$P(b|a) = \frac{P(b)P(a|b)}{P(a)} = \frac{P(b)P(a|b)}{\int P(a,b)db} = \frac{P(b)P(a|b)}{\int P(b)P(a|b)db}$$

Posterior Distribution

- Can infer the parameters by computing the posterior distribution (Bayesian inference)

$$p(\theta|\mathbf{X}, m) = \frac{p(\mathbf{X}, \theta|m)}{p(\mathbf{X}|m)} = \frac{p(\mathbf{X}|\theta, m)p(\theta|m)}{\int p(\mathbf{X}|\theta, m)p(\theta|m)d\theta} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$



- Cheaper alternative: Point Estimation of the parameters. E.g.,
  - **Maximum likelihood estimation (MLE):** Find θ that makes the observed data most probable

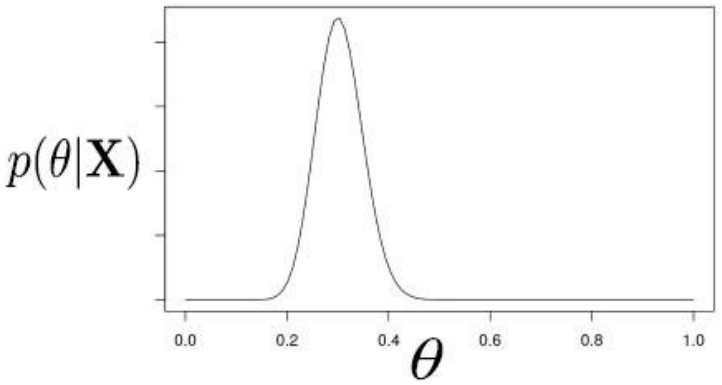
$$\hat{\theta}_{ML} = \arg\max_{\theta} \log p(\mathbf{X}|\theta)$$

  - **Maximum-a-Posteriori (MAP) estimation:** Find θ that has the largest posterior probability

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \log p(\theta|\mathbf{X}) = \arg\max_{\theta} [\log p(\mathbf{X}|\theta) + \log p(\theta)]$$

Posterior Distribution

- Posterior provides us a holistic view about θ given observed data
- A simple unimodal posterior distribution for a scalar parameter θ might look something like



- Various types of estimates regarding θ can be obtained from the posterior, e.g.,
  - Mode of the posterior (same as the MAP estimate)
  - Mean and median of the posterior
  - Variance/spread of the posterior (uncertainty in our estimate of the parameters)

Posterior Predictive Distribution

- Posterior can be used to compute the posterior predictive distribution (PPD) of new observation
- The PPD of a new observation $\boldsymbol{x}_*$ given previous observations

$$p(\boldsymbol{x}_*|\mathbf{X}, m) = \int p(\boldsymbol{x}_*, \theta|\mathbf{X}, m)d\theta = \int p(\boldsymbol{x}_*|\theta, \mathbf{X}, m)p(\theta|\mathbf{X}, m)d\theta$$
$$= \int p(\boldsymbol{x}_*|\theta, m)p(\theta|\mathbf{X}, m)d\theta$$

- Note: In the above, we assume that the observations are i.i.d. given θ
- Computing PPD requires doing a posterior-weighted averaging over all values of θ
- If the integral in PPD is intractable, we can approximate the PPD by plug-in predictive

$$p(\boldsymbol{x}_*|\mathbf{X}, m) \approx p(\boldsymbol{x}_*|\hat{\theta}, m)$$

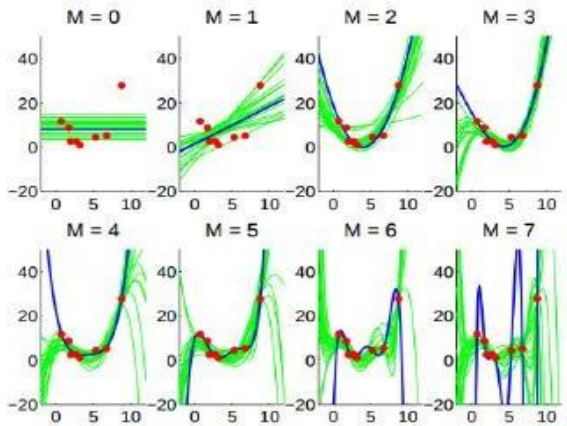
.. where $\hat{\theta}$ is a point estimate of θ (e.g., MLE/MAP)

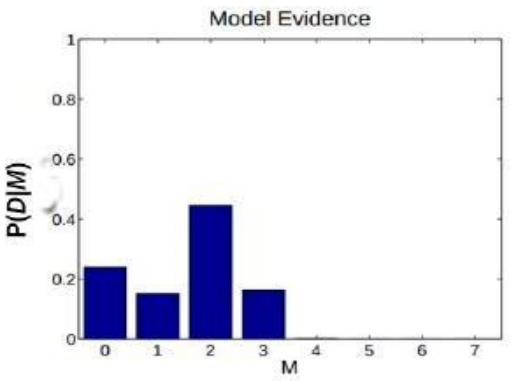
Marginal Likelihood

- Recall the Bayes rule for computing the posterior

$$p(\theta|\mathbf{X}, m) = \frac{p(\mathbf{X}, \theta|m)}{p(\mathbf{X}|m)} = \frac{p(\mathbf{X}|\theta, m)p(\theta|m)}{\int p(\mathbf{X}|\theta, m)p(\theta|m)d\theta} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$

- The denominator in the Bayes rule is the marginal likelihood (a.k.a. "model evidence")
- Note that $p(\mathbf{X}|m) = \mathbb{E}_{p(\theta|m)}[p(\mathbf{X}|\theta, m)]$ is the average/expected likelihood under model m
- For a good model, we would expect this "averaged" quantity to be large (most θ's will be good)





Model Comparison/Averaging

- Marginal likelihood is hard-to-compute (due to integral) but a very useful quantity
- It can be used for doing model selection
  - Choose model m that has largest posterior probability

$$\hat{m} = \arg\max_m p(m|\mathbf{X}) = \arg\max_m \frac{p(\mathbf{X}|m)p(m)}{p(\mathbf{X})} = \arg\max_m p(\mathbf{X}|m)p(m)$$

  - If all models are equally likely a priori then $\hat{m} = \arg\max_m p(\mathbf{X}|m)$
  - If m is a hyperparam, then $\arg\max_m p(\mathbf{X}|m)$ is MLE-II based hyperparameter estimation
- Marginal likelihood can be used to compute $p(m|\mathbf{X})$ and then perform Bayesian Model Averaging

$$p(\boldsymbol{x}_*|\mathbf{X}) = \sum_{m=1}^{M} p(\boldsymbol{x}_*|\mathbf{X}, m)p(m|\mathbf{X})$$

A Simple Parameter Estimation Problem

- for a single-parameter model
- hyperparameter if any will be assumed to be fixed/known

Simple Example (MLE)

- Consider a sequence of *N* coin tosses (call head = 1, tail = 0)
- The n^{th} outcome x_{n} is a binary random variable $\in \{0, 1\}$
- Assume θ to be probability of a head (parameter we wish to estimate)

- Each likelihood term $p(x_{\text{n}} \mid \theta)$ is Bernoulli: $p(x_n|\theta) = \theta^{x_n}(1-\theta)^{1-x_n}$

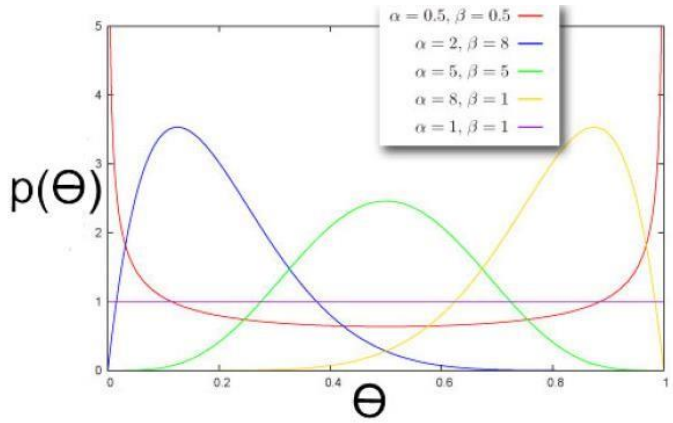
- Log-likelihood: $\sum_{n=1}^{N} \log p(x_n|\theta) = \sum_{n=1}^{N} x_n \log\theta + (1-x_n)\log(1-\theta)$

- Taking derivative of the log-likelihood w.r.t. θ, and setting it to zero gives:

$$\widehat{\boldsymbol{\theta}}_{\text{mle}} = \frac{\sum_{n=1}^{N} x_n}{N}$$

$\widehat{\boldsymbol{\theta}}_{\text{mle}}$ in this example is simply the fraction of heads!

MAP Estimate



- MAP estimation can incorporate a prior $p(\theta)$ on θ
- Since $\theta \in (0, 1)$, one possibility can be to assume a Beta prior

$$p(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- α, β are called hyperparameters of the prior (these can have intuitive meaning; we'll see shortly)
  - Note that each likelihood term is still a Bernoulli: $p(\boldsymbol{x}_n|\theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$
    - The log posterior probability $= \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta) + \log p(\theta)$
    - Ignoring the constants w.r.t. θ, the log posterior probability:

$$\sum_{n=1}^{N} \{\boldsymbol{x}_n \log \theta + (1-\boldsymbol{x}_n) \log(1-\theta)\} + (\alpha - 1) \log \theta + (\beta - 1) \log(1-\theta)$$

    - Taking derivative w.r.t. θ and setting to zero gives

$$\hat{\theta}_{MAP} = \frac{\sum_{n=1}^{N} \boldsymbol{x}_n + \alpha - 1}{N + \alpha + \beta - 2}$$

    - Note: For $\alpha = 1, \beta = 1$, i.e., $p(\theta) = \text{Beta}(1, 1)$ (equivalent to a uniform prior), $\hat{\theta}_{MAP} = \hat{\theta}_{MLE}$

Posterior Distribution

- Recall that each likelihood term was Bernoulli: $p(\boldsymbol{x}_n|\theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$
- Let's again choose the prior $p(\theta)$ as Beta: $p(\theta) = \text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- The posterior distribution will be proportional to the product of likelihood and prior

$$
\begin{aligned}
p(\theta|\mathbf{X}) &\propto \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)p(\theta) \\
&\propto \theta^{\alpha+\sum_{n=1}^{N}\boldsymbol{x}_n-1}(1-\theta)^{\beta+N-\sum_{n=1}^{N}\boldsymbol{x}_n-1}
\end{aligned}
$$

- From simple inspection, note that the posterior $p(\theta|\mathbf{X}) = \text{Beta}(\alpha + \sum_{n=1}^{N}\boldsymbol{x}_n, \beta + N - \sum_{n=1}^{N}\boldsymbol{x}_n)$

Posterior has the same form as prior – conjugate prior

Posterior Predictive Distribution

- Let's say we want to compute the probability that the next outcome $\boldsymbol{x}_{N+1} \in \{0, 1\}$ will be a head
- The plug-in predictive distribution using a point estimate $\hat{\theta}$ (e.g., using MLE/MAP)

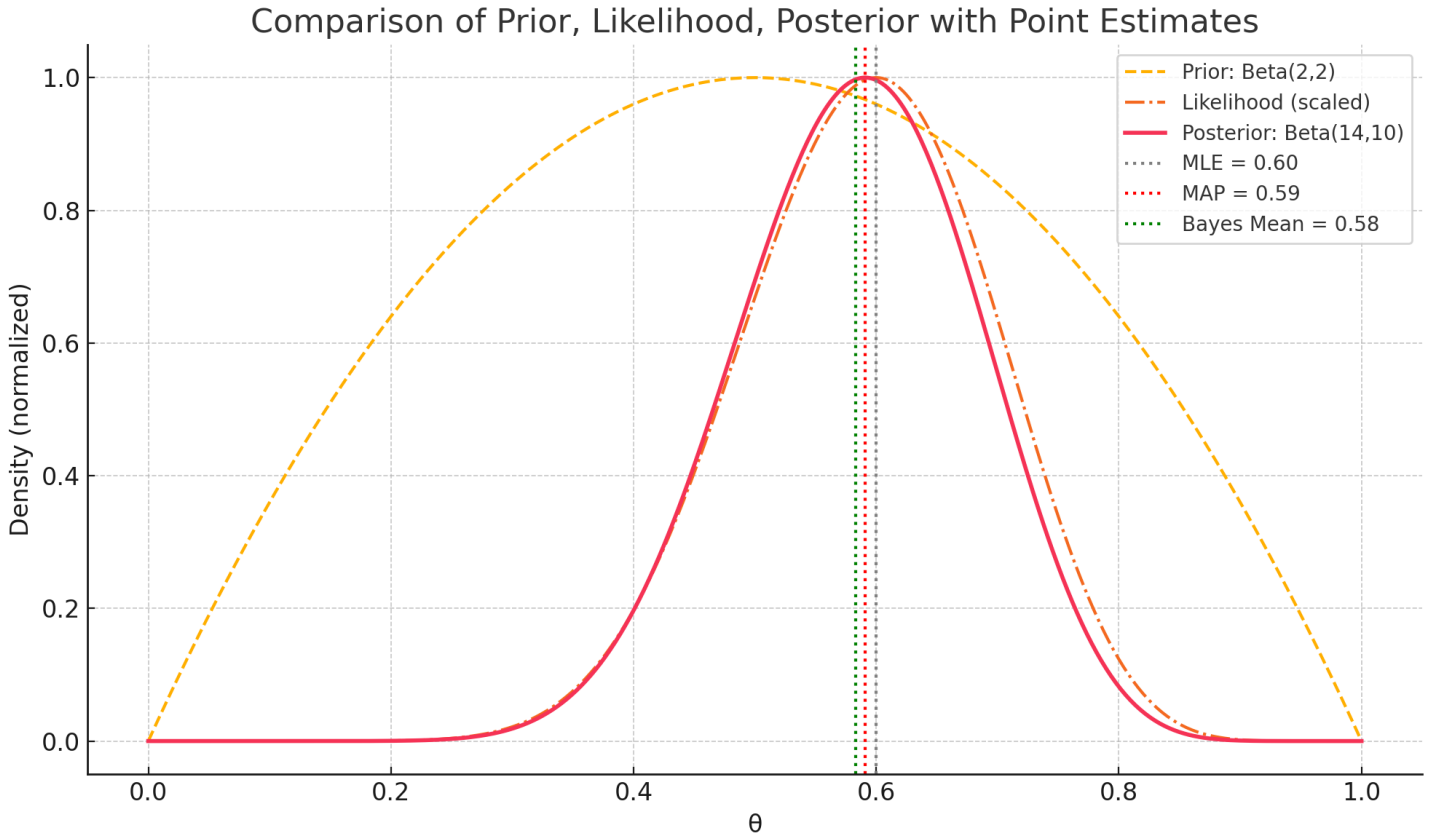
$$p(\boldsymbol{x}_{N+1} = 1|\mathbf{X}) \approx p(\boldsymbol{x}_{N+1} = 1|\hat{\theta}) = \hat{\theta} \quad \underline{\text{or equivalently}} \quad p(\boldsymbol{x}_{N+1}|\mathbf{X}) \approx \text{Bernoulli}(\boldsymbol{x}_{N+1} \mid \hat{\theta})$$

- The posterior predictive distribution (averaging over all θ weighted by their posterior probabilities):

$$\begin{aligned} p(\boldsymbol{x}_{N+1} = 1|\mathbf{X}) &= \int_0^1 P(\boldsymbol{x}_{N+1} = 1|\theta) p(\theta|\mathbf{X}) d\theta \\ &= \int_0^1 \theta \times \text{Beta}(\theta|\alpha + N_1, \beta + N_0) d\theta \\ &= \mathbb{E}[\theta|\mathbf{X}] \\ &= \frac{\alpha + N_1}{\alpha + \beta + N} \end{aligned}$$

- Therefore the posterior predictive distribution: $p(\boldsymbol{x}_{N+1}|\mathbf{X}) = \text{Bernoulli}(\boldsymbol{x}_{N+1} \mid \mathbb{E}[\theta|\mathbf{X}])$

Visualization



- **Prior** $Beta(2,2)$
- **Likelihood** (scaled)
- **Posterior** $Beta(14,10)$

Vertical lines for:

- **MLE**: $\theta = \frac{12}{20} = 0.60$
- **MAP**: $\theta = \frac{13}{22} \approx 0.59$
- **Bayesian Mean**: $\theta = \frac{14}{24} \approx 0.58$

Bayesian mean and **MAP** are pulled slightly toward the prior compared to the **MLE**.

Multinoulli Observation Model

Multinoulli Model

- Assume N discrete-valued observations $\{x_1, \ldots, x_N\}$ with each $x_n \in \{1, \ldots, K\}$, e.g.,
  - x_n represents the outcome of a dice roll with K faces
  - x_n represents the class label of the n-th example (total K classes)
  - x_n represents the identity of the n-th word in a sequence of words
- Assume likelihood to be multinoulli with unknown params $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$ s.t. $\sum_{k=1}^{K} \pi_k = 1$

$$p(x_n|\boldsymbol{\pi}) = \text{multinoulli}(x_n|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{\mathbb{I}[x_n=k]}$$

- $\boldsymbol{\pi}$ is a vector of probabilities ("probability vector"), e.g.,
  - Biases of the K sides of the dice
  - Prior class probabilities in multi-class classification
  - Probabilities of observing each words in the vocabulary
- Assume a conjugate Dirichlet prior on $\boldsymbol{\pi}$ with hyperparams $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$ (also, $\alpha_k \geq 0, \forall k$)

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

Detour: Dirichlet Distribution

A Bag of Proportions
Imagine you're trying to model the proportions of K **different categories** (say: red, green, blue marbles in a bag). But instead of knowing the exact proportions, you're uncertain — and you want a *probabilistic guess* of what those proportions might be.

The **Dirichlet distribution** gives you a way to describe that uncertainty:

- Each sample from a Dirichlet distribution gives you a possible set of proportions (like: 60% red, 30% green, 10% blue).
- Different parameters of the Dirichlet control what kinds of proportions you're more likely to see.

Detour: Dirichlet Distribution

Dirichlet distribution has a parameter **$\alpha = [\alpha_1, \alpha_2, ..., \alpha_k]$** — one for each category.
Here's what those parameters intuitively do:

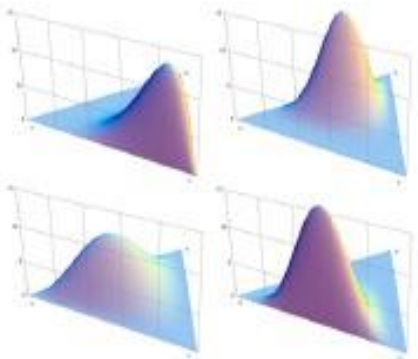
- **$\alpha_i > 1$** → "I believe the i^{th} category will have a **large** proportion."
- **$\alpha_i < 1$** → "I believe the i^{th} category will have a **small** proportion (or maybe even zero)."
- **$\alpha_i = 1$** → "I have **no strong preference** for the i^{th} category."

Sum of αs, often denoted $\alpha_0 = \sum \alpha_i$, controls concentration:

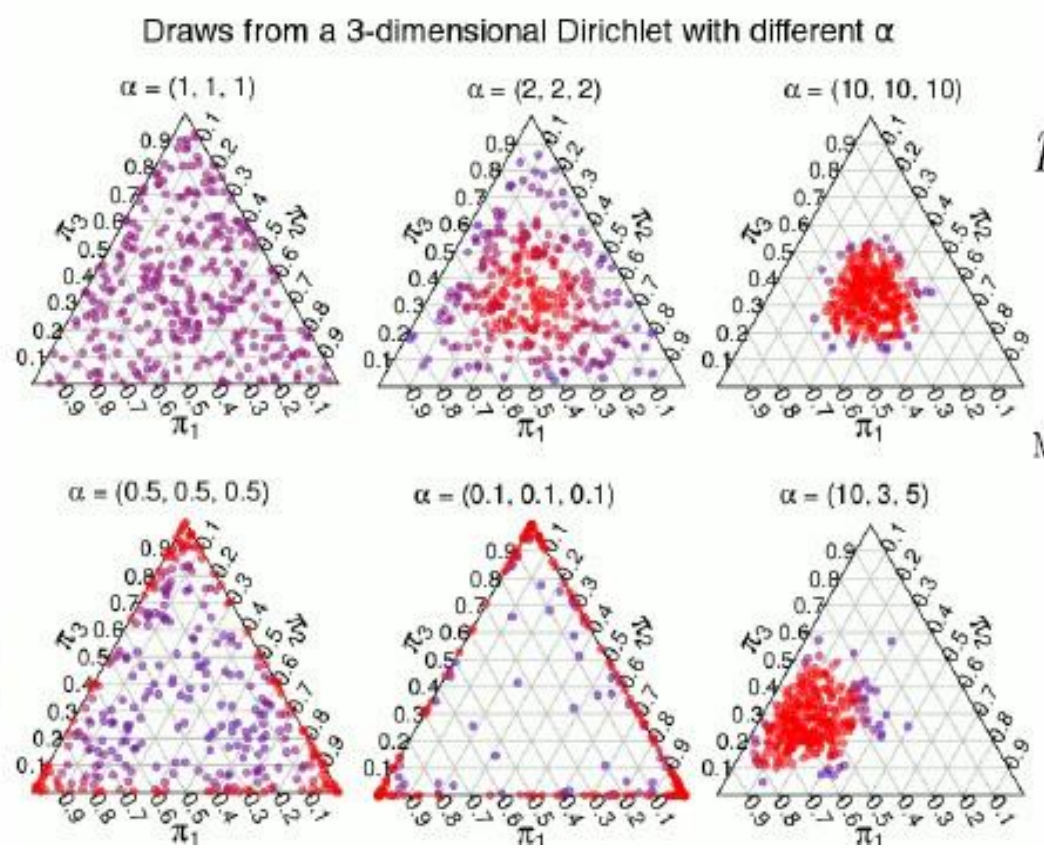
- **High α_0** (e.g. all $\alpha_i = 10$): samples are tightly clustered around the mean (less variability).
- **Low α_0** (e.g. all $\alpha_i = 0.2$): samples are sparse — most of the probability mass goes to just one or two categories in each sample.

Detour: Dirichlet Distribution

PDF for a 3-dim Dirichlet



Red dots denote regions of high probability denity



$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

$$\text{Mean} = \left[\frac{\alpha_1}{\sum_{k=1}^{K} \alpha_k}, \ldots, \frac{\alpha_K}{\sum_{k=1}^{K} \alpha_k} \right]$$

$$\text{Mode} = \left[\frac{\alpha_1 - 1}{\sum_{k=1}^{K} \alpha_k - K}, \ldots, \frac{\alpha_K - 1}{\sum_{k=1}^{K} \alpha_k - K} \right] (\alpha_k > 1)$$

$$\text{var}(\pi_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}$$

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k$$

Posterior Distribution

- The posterior over $\boldsymbol{\pi}$ is easy to compute in this case due to conjugacy b/w multinoulli and Dirichlet

$$p(\boldsymbol{\pi}|\mathbf{X}, \boldsymbol{\alpha}) = \frac{p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\alpha})p(\boldsymbol{\pi}|\boldsymbol{\alpha})}{p(\mathbf{X}|\boldsymbol{\alpha})} = \frac{p(\mathbf{X}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\boldsymbol{\alpha})}{p(\mathbf{X}|\boldsymbol{\alpha})}$$

- Assuming x_n's are i.i.d. given π, $p(\mathbf{X}|\boldsymbol{\pi}) = \prod_{n=1}^{N} p(x_n|\boldsymbol{\pi})$, therefore

$$p(\boldsymbol{\pi}|\mathbf{X}, \boldsymbol{\alpha}) \propto \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{\mathbb{I}[x_n=k]} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} = \prod_{k=1}^{K} \pi_k^{\alpha_k + \sum_{n=1}^{N} \mathbb{I}[x_n=k] - 1}$$

- Even without computing the normalization constant $p(\mathbf{X}|\boldsymbol{\alpha})$, we can see that it's a Dirichlet! :-)
- Denoting $N_k = \sum_{n=1}^{N} \mathbb{I}[x_n = k]$, i.e., number of observations with value k, the posterior will be

$$p(\boldsymbol{\pi}|\mathbf{X}, \boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1 + N_1, \ldots, \alpha_K + N_K)$$

Exercise

For Multinoulli Likelihood and Dirichlet Prior

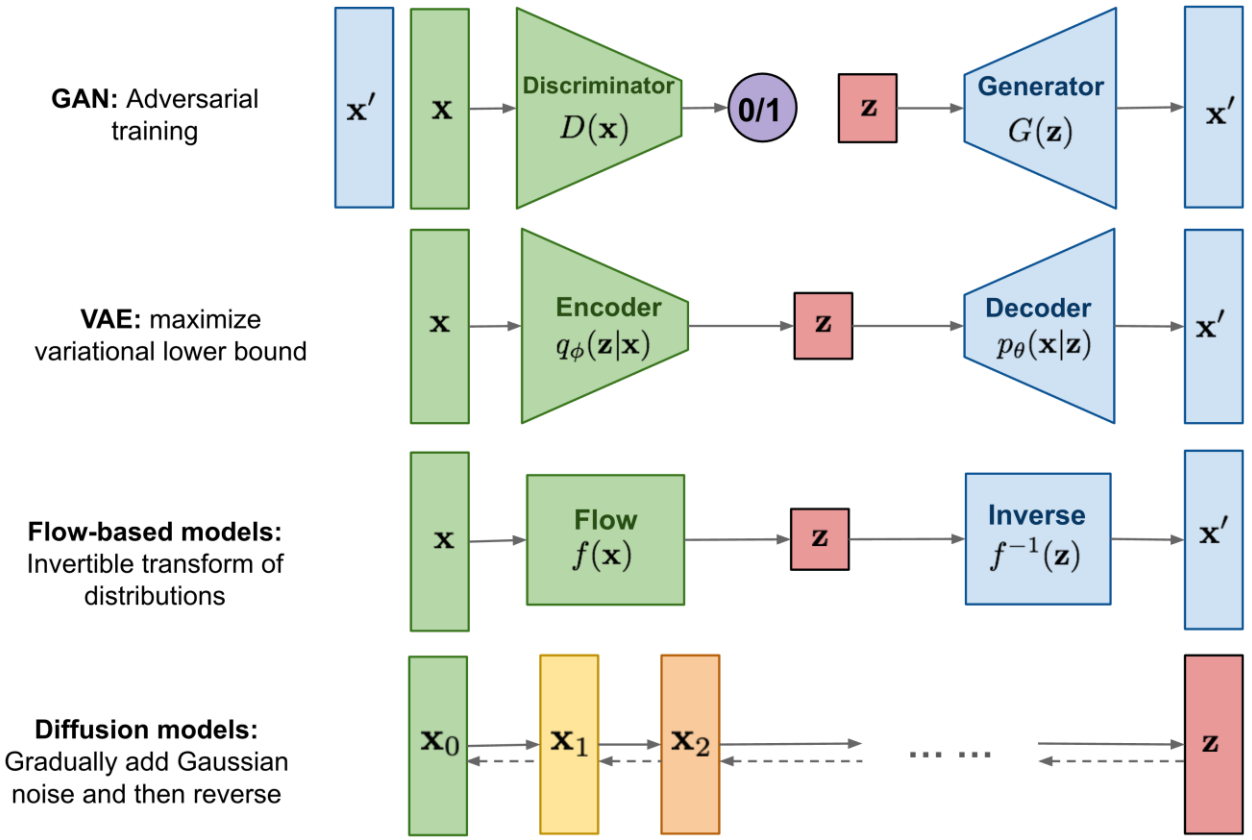
- What is the MLE/MAP?
- Posterior Predictive Distribution?

Gaussian Models

- Univariate with fixed variance
- Univariate with fixed mean
- Univariate with varying mean and variance
- Multivariate

Detour: Generative Models

Generative models invariably are also probabilistic models



- Image-to-image translation
- Deepfake generation

- Anomaly Detection in Medical Imaging
- Generating synthetic but interpretable data

- High-fidelity Audio Generation
  e.g., WaveGlow for speech synthesis

- Text-to-Image Generation

Figure credit: Lilian Weng

Fixed Variance Gaussian Model

- Consider N i.i.d. observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \propto \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]$$

$$p(\mathbf{X}|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

- Assume the mean $\mu \in \mathbb{R}$ of the Gaussian is unknown and assume variance σ^2 to be known/fixed
- We wish to estimate the unknown μ given the data $\mathbf{X}$
- Let's choose a Gaussian prior on μ, i.e., $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$ with μ_0, σ_0^2 as fixed

Bayesian Inference for Mean of a Gaussian

- The posterior distribution for the unknown mean parameter μ

$$p(\mu|\mathbf{X}) = \frac{p(\mathbf{X}|\mu)p(\mu)}{p(\mathbf{X})} \propto \prod_{n=1}^{N} \exp\left[-\frac{(x_n-\mu)^2}{2\sigma^2}\right] \times \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]$$

- Simplifying the above (using completing the squares trick) gives $p(\mu|\mathbf{X}) \propto \exp\left[-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right]$ with

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

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\bar{x} \qquad (\text{where } \bar{x} = \frac{\sum_{n=1}^{N} x_n}{N})$$

Notion of Sufficient Statistics

We only need sufficient statistics to estimate the parameters and values of individual observations aren't needed

Likelihood

$$\sum_{n=1}^{N} \frac{(x_n - \mu)^2}{2\sigma^2} = \sum_{n=1}^{N} \frac{x_n^2 - 2\mu x_n + \mu^2}{2\sigma^2} = \frac{1}{2\sigma^2}\left(\sum x_n^2 - 2\mu \sum x_n + N\mu^2\right)$$

Prior

$$\frac{(\mu - \mu_0)^2}{2\sigma_0^2} = \frac{1}{2\sigma_0^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2)$$

Group all terms with respect to μ:

$$\log p(\mu \mid \mathbf{X}) \propto -\frac{1}{2}\left[\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{\sum x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)\mu\right] + (\text{constants})$$

Let:

- $A = \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)$
- $B = \left(\frac{\sum x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)$

Completing the square

We rewrite:

$$A\mu^2 - 2B\mu = A\left(\mu^2 - \frac{2B}{A}\mu\right) = A\left(\mu - \frac{B}{A}\right)^2 - \frac{B^2}{A}$$

So:

$$\log p(\mu \mid \mathbf{X}) \propto -\frac{A}{2}\left(\mu - \frac{B}{A}\right)^2$$

Resulting Posterior

$$p(\mu \mid \mathbf{X}) = \mathcal{N}\left(\mu_N, \sigma_N^2\right)$$

Where:

- $\mu_N = \frac{B}{A} = \frac{\frac{\sum x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{\sigma^2\mu_0 + N\sigma_0^2\bar{x}}{N\sigma_0^2 + \sigma^2}$
- $\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}$

Posterior Predictive Distribution

- What is the posterior predictive distribution $p(x_*|\mathbf{X})$ of a new observation x_*?
- Using the inferred posterior $p(\mu|\mathbf{X})$, we can find the posterior predictive distribution

$$p(x_*|\mathbf{X}) = \int p(x_*|\mu, \sigma^2)p(\mu|\mathbf{X})d\mu = \int \mathcal{N}(x_*|\mu, \sigma^2)\mathcal{N}(\mu|\mu_N, \sigma_N^2)d\mu = \mathcal{N}(x_*|\mu_N, \sigma^2 + \sigma_N^2)$$

Convolution of Gaussians

If:

- $X \sim \mathcal{N}(\mu, \sigma^2)$
- $\mu \sim \mathcal{N}(\mu_N, \sigma_N^2)$

Then the **marginal** distribution of X (i.e., integrating out μ) is:

$$X \sim \mathcal{N}(\mu_N, \sigma^2 + \sigma_N^2)$$

That is:

$$\int \mathcal{N}(x|\mu, \sigma^2) \cdot \mathcal{N}(\mu|\mu_N, \sigma_N^2)\, d\mu = \mathcal{N}(x|\mu_N, \sigma^2 + \sigma_N^2)$$

Posterior Predictive Distribution

- What is the posterior predictive distribution $p(x_*|\mathbf{X})$ of a new observation x_*?
- Using the inferred posterior $p(\mu|\mathbf{X})$, we can find the posterior predictive distribution

$$p(x_*|\mathbf{X}) = \int p(x_*|\mu, \sigma^2)p(\mu|\mathbf{X})d\mu = \int \mathcal{N}(x_*|\mu, \sigma^2)\mathcal{N}(\mu|\mu_N, \sigma_N^2)d\mu = \mathcal{N}(x_*|\mu_N, \sigma^2 + \sigma_N^2)$$

Why? Because you are adding uncertainty

- μ is uncertain with variance σ_N^2
- Observations have noise σ^2

So the total variance for x_* is:

$$\mathrm{Var}(x_*) = \underbrace{\mathbb{E}_\mu[\mathrm{Var}(x_*|\mu)]}_{=\sigma^2} + \underbrace{\mathrm{Var}_\mu[\mathbb{E}(x_*|\mu)]}_{=\sigma_N^2} = \sigma^2 + \sigma_N^2$$

Fixed Mean Gaussian Model

- Again consider N i.i.d. observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \quad \text{and} \quad p(\mathbf{X}|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

- Assume the variance $\sigma^2 \in \mathbb{R}_+$ of the Gaussian is unknown and assume mean μ to be known/fixed
- Let's estimate σ^2 given the data $\mathbf{X}$ using fully Bayesian inference (not MLE/MAP)
- We first need a prior distribution for σ^2. What prior $p(\sigma^2)$ to choose in this case?
- If we want a conjugate prior, it should have the same form as the likelihood

$$p(x_n|\mu, \sigma^2) \propto (\sigma^2)^{-1/2} \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]$$

Choosing a Conjugate Prior for σ^2

Goal: Find a prior $p(\sigma^2)$ that makes posterior inference tractable (i.e., conjugate prior).

- Likelihood from Gaussian:

$$p(\mathbf{X} \mid \mu, \sigma^2) \propto (\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2\right)$$

- Choose prior to match this form:

$$\boxed{\sigma^2 \sim \text{Inverse-Gamma}(\alpha_0, \beta_0)}$$

$$p(\sigma^2) \propto (\sigma^2)^{-\alpha_0 - 1} \exp\left(-\frac{\beta_0}{\sigma^2}\right)$$

- This ensures:

$$p(\sigma^2 \mid \mathbf{X}) \propto (\text{likelihood}) \cdot (\text{prior}) \quad \rightarrow \text{same functional form}$$

Posterior Distribution over σ^2 or Precision λ

Likelihood:

$$p(\mathbf{x} \mid \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{S}{2\sigma^2}\right), \quad S = \sum_{i=1}^{N}(x_i-\mu)^2$$

sum of squared deviations

Prior (conjugate):

$$\sigma^2 \sim \text{Inverse-Gamma}(\alpha_0, \beta_0) \quad \Rightarrow \quad p(\sigma^2) \propto (\sigma^2)^{-\alpha_0-1} \exp\left(-\frac{\beta_0}{\sigma^2}\right)$$

Posterior (via Bayes' Rule):

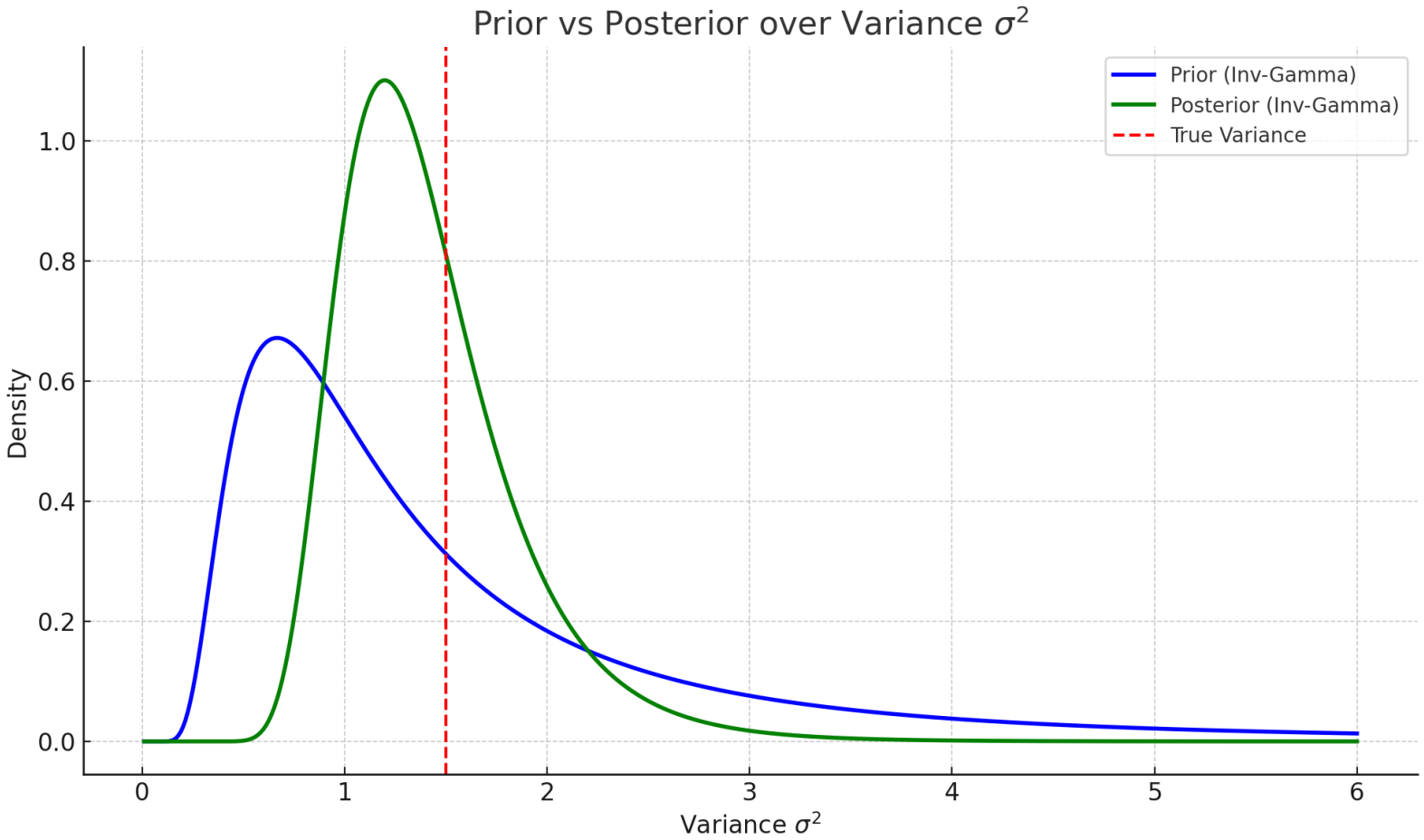
$$p(\sigma^2 \mid \mathbf{x}) \propto p(\mathbf{x} \mid \sigma^2) \cdot p(\sigma^2)$$

$$\propto (\sigma^2)^{-N/2} \exp\left(-\frac{S}{2\sigma^2}\right) \cdot (\sigma^2)^{-\alpha_0-1} \exp\left(-\frac{\beta_0}{\sigma^2}\right)$$

$$= (\sigma^2)^{-(\alpha_0+\frac{N}{2})-1} \exp\left(-\frac{\beta_0 + \frac{1}{2}S}{\sigma^2}\right)$$

$$\sigma^2 \mid \mathbf{x} \sim \text{Inverse-Gamma}\left(\alpha_0 + \frac{N}{2},\ \beta_0 + \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right)$$

Visualization



- Posterior sharpens around the true variance
- Bayesian inference updates our belief after observing data.

Univariate Gaussian — Unknown Mean & Variance

Model:

- $x_1, \ldots, x_N \sim \mathcal{N}(\mu, \sigma^2)$
- Both μ and σ^2 unknown

Likelihood:

$$p(\mathbf{x} \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2\right)$$

Conjugate Prior: Normal-Inverse-Gamma $p(\mu, \sigma^2) = p(\mu \mid \sigma^2) \cdot p(\sigma^2)$

$$\sigma^2 \sim \text{Inv-Gamma}(\alpha_0, \beta_0)$$

$$\mu \mid \sigma^2 \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$$

κ_0 is a **scaling parameter** that determines confidence in prior belief about μ

Posterior Derivation — Normal-Inverse-Gamma

Posterior Parameters:

Let $\bar{x} = \frac{1}{N}\sum x_i$, and $S = \sum (x_i - \bar{x})^2$

$$\kappa_N = \kappa_0 + N \qquad \mu_N = \frac{\kappa_0 \mu_0 + N\bar{x}}{\kappa_0 + N}$$

$$\alpha_N = \alpha_0 + \frac{N}{2} \qquad \beta_N = \beta_0 + \frac{1}{2}S + \frac{\kappa_0 N}{2(\kappa_0 + N)}(\bar{x} - \mu_0)^2$$

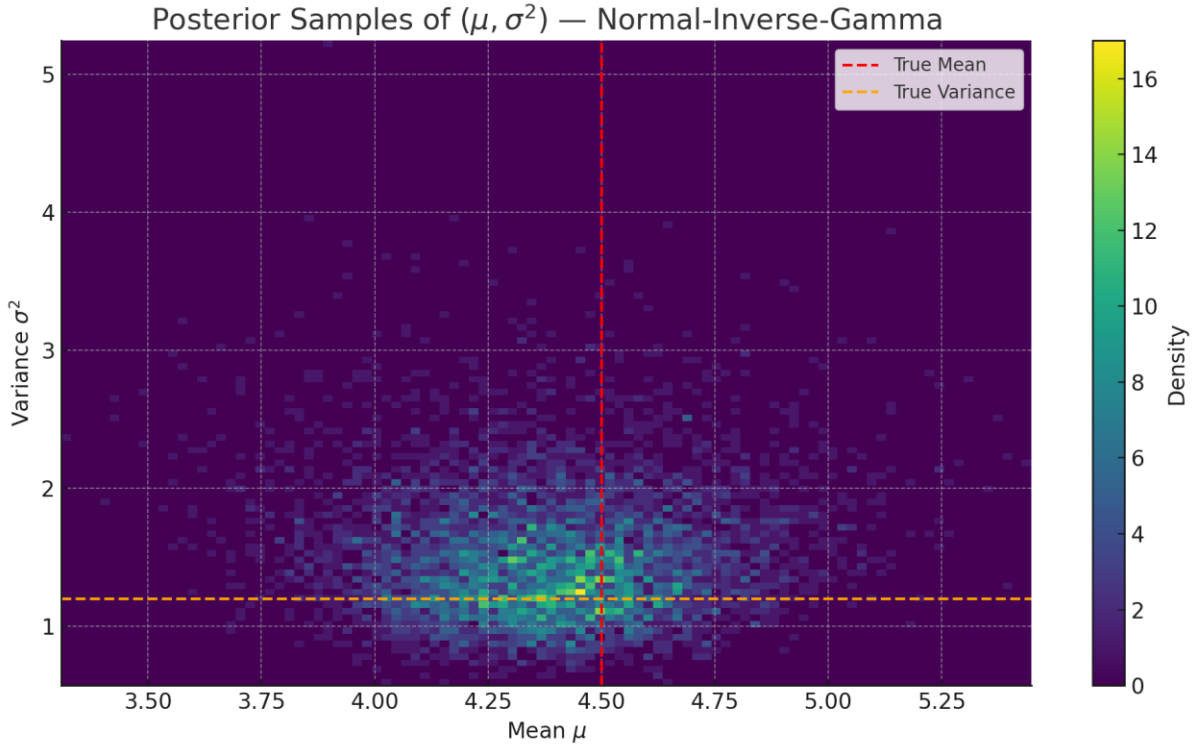
Posterior Distributions:

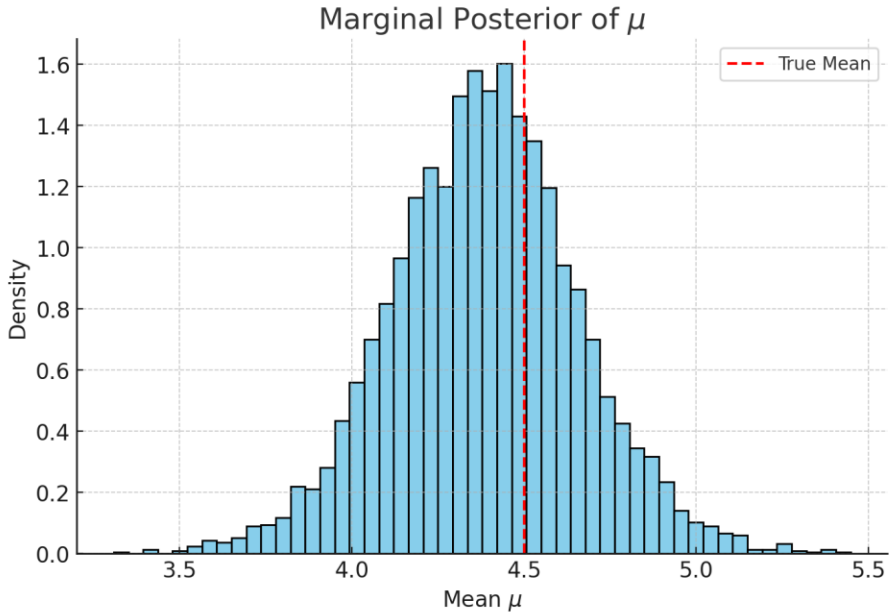
$$\sigma^2 \mid \mathbf{x} \sim \text{Inv-Gamma}(\alpha_N, \beta_N)$$

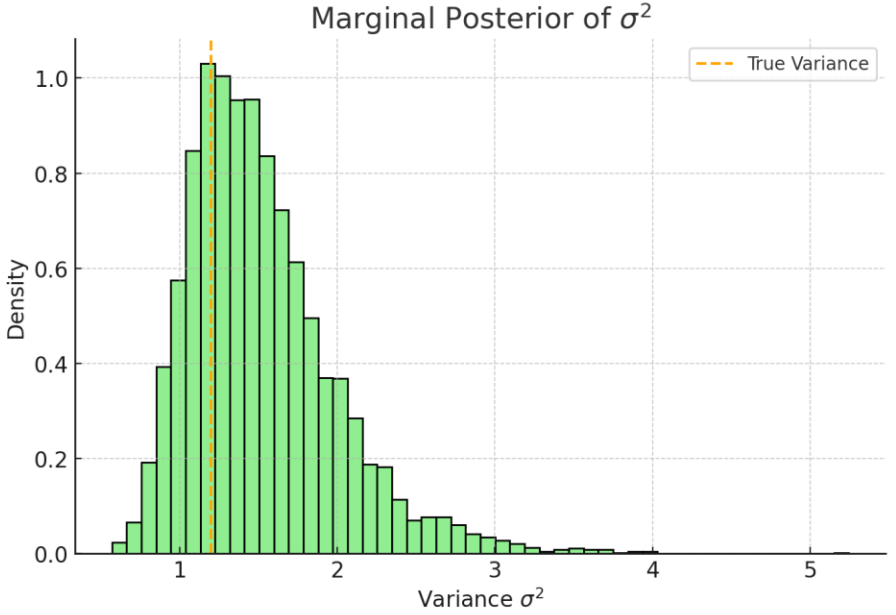
$$\mu \mid \sigma^2, \mathbf{x} \sim \mathcal{N}\left(\mu_N, \frac{\sigma^2}{\kappa_N}\right)$$

$$\boxed{p(\mu, \sigma^2 \mid \mathbf{x}) = \text{Normal-Inverse-Gamma}(\mu_N, \kappa_N, \alpha_N, \beta_N)}$$

Visualization







Multivariate Gaussian

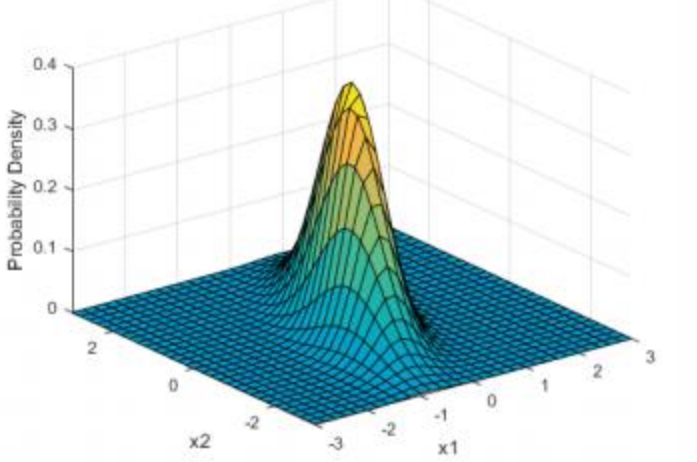
- A **Multivariate Gaussian** describes a vector of real-valued random variables:

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Parameters:
  - $\boldsymbol{\mu} \in \mathbb{R}^d$: Mean vector
  - $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$: Covariance matrix
- PDF:

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

A two-dimensional Gaussian

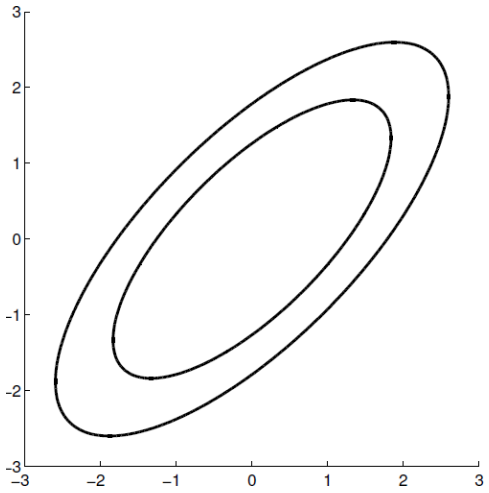


Multivariate Gaussian: Examples

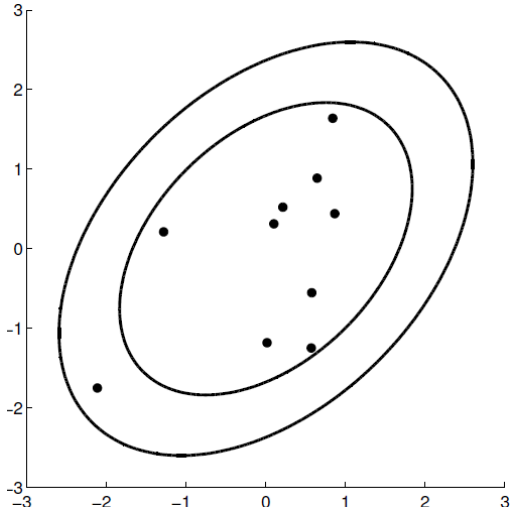
Covariance matrix Σ determines:

- **Shape** of the distribution
- **Orientation** and **spread**

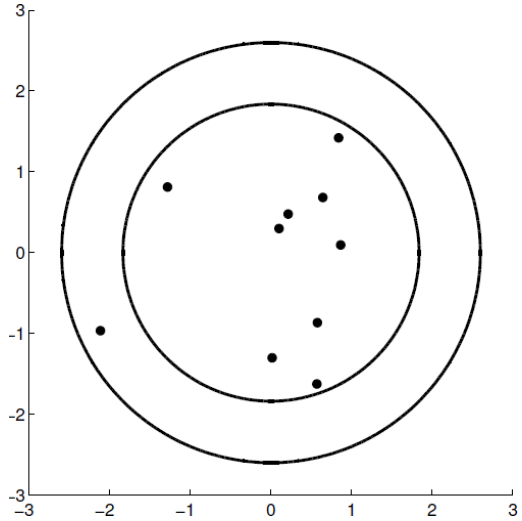
- **Identity matrix** → spherical shape
- **Large off-diagonal** σ_{12} → correlated variables (elliptical orientation)



$$\Sigma = \begin{bmatrix} 1 & .7 \\ .7 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & .4 \\ .4 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Multivariate Gaussian: Marginals and Conditionals

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ with $\Lambda = \Sigma^{-1}$. Suppose

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

- The marginal distribution is simply

$$p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \Sigma_{aa})$$

- The conditional distribution is given by

$$\begin{aligned} p(\mathbf{x}_a|\mathbf{x}_b) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \end{aligned}$$

Thus marginals and conditionals of Gaussians are Gaussians

Multivariate Gaussian : Full Bayesian Estimation

Case 1: Unknown Mean $\boldsymbol{\mu}$, Known Covariance $\boldsymbol{\Sigma}$

- **Prior:** The mean vector $\boldsymbol{\mu}$ follows a multivariate normal distribution:

$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa_0)$$

- **Posterior:** Given the data, the posterior for the mean is also multivariate normal:

$$\boldsymbol{\mu} \mid \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}/\kappa_n)$$

- **Posterior Parameters:**

$$\boldsymbol{\mu}_n = \frac{\kappa_0 \boldsymbol{\mu}_0 + n\bar{x}}{\kappa_0 + n}, \quad \kappa_n = \kappa_0 + n$$

Multivariate Gaussian : Full Bayesian Estimation

Case 2: Known Mean $\boldsymbol{\mu}$, Unknown Covariance $\boldsymbol{\Sigma}$

- **Prior:** The covariance matrix $\boldsymbol{\Sigma}$ follows an inverse-Wishart distribution:

$$\boldsymbol{\Sigma} \sim \mathcal{W}^{-1}(\Psi_0, \nu_0)$$

- **Posterior:** The posterior for the covariance is:

$$\boldsymbol{\Sigma} \mid \mathbf{X} \sim \mathcal{W}^{-1}(\Psi_n, \nu_n)$$

- **Posterior Parameters:**

$$\Psi_n = \Psi_0 + S, \quad \nu_n = \nu_0 + n$$

where S is the sample scatter matrix.

$$p(\boldsymbol{\Sigma}) = \frac{|\Psi|^{\nu/2}}{2^{\nu d/2}\Gamma_d(\nu/2)} |\boldsymbol{\Sigma}|^{-(\nu+d+1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\Psi\boldsymbol{\Sigma}^{-1})\right)$$

Multivariate Gaussian : Full Bayesian Estimation

Case 3: Unknown Mean $\boldsymbol{\mu}$ and Covariance $\boldsymbol{\Sigma}$

- **Prior:**

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa_0), \quad \boldsymbol{\Sigma} \sim \mathcal{W}^{-1}(\Psi_0, \nu_0)$$

- **Posterior:** The joint posterior is a **Normal-Inverse-Wishart distribution**:

$$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mid \mathbf{X} \sim \mathrm{NIW}(\boldsymbol{\mu}_n, \kappa_n, \Psi_n, \nu_n)$$

- **Posterior Parameters:**

$$\boldsymbol{\mu}_n = \frac{\kappa_0 \boldsymbol{\mu}_0 + n\bar{x}}{\kappa_0 + n}$$

$$\kappa_n = \kappa_0 + n$$

$$\nu_n = \nu_0 + n$$

$$\Psi_n = \Psi_0 + S + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{x} - \boldsymbol{\mu}_0)(\bar{x} - \boldsymbol{\mu}_0)^T$$

Linear Gaussian Model Formulation

Model the data as:

$$\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

This is a **Linear Gaussian Model**:

- Latent variable: $\boldsymbol{\mu}$
- Linear transformation: identity
- Noise: multivariate Gaussian

LGM $\leftrightarrow$ Bayesian Inference Mapping

Bayesian Inference	**Linear Gaussian Model**
Unknown $\boldsymbol{\mu}$	Latent variable
$\mathbf{x}_i = \mu + \epsilon$	LGM equation
Gaussian noise $\epsilon \sim \mathcal{N}(0, \boldsymbol{\Sigma})$	Measurement uncertainty
Gaussian prior on $\boldsymbol{\mu}$	Conjugate prior
Posterior of $\boldsymbol{\mu}$	LGM inference result

Bayesian inference in this setup is equivalent to inference in a Linear Gaussian Model where parameters (like the mean vector) are **latent variables** and observations are generated through a **linear-Gaussian transformation**.

Gaussian Observation Model

- MLE/MAP for μ, σ^2 (or both) is straightforward in Gaussian observation models.

- Posterior also straightforward in most situations for such models
  - (As we saw) computing posterior of μ is easy (using Gaussian prior) if variance σ^2 is known
  - Likewise, computing posterior of σ^2 is easy (using gamma prior on σ^2) if mean μ is known

- If μ, σ^2 both are unknown, posterior computation requires computing $p(\mu, \sigma^2|x)$
  - Computing joint posterior $p(\mu, \sigma^2|\boldsymbol{x})$ exactly requires a jointly conjugate prior $p(\mu, \sigma^2)$
  - "Gaussian-gamma" ("Normal-gamma") is such a conjugate prior – a product of normal and gamma
  - Note: Computing joint posteriors exactly is possible only in rare cases such this one

- If each observation $\boldsymbol{x}_n \in \mathbb{R}^D$, can assume a likelihood/observation model $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Need to estimate a vector-valued mean $\boldsymbol{\mu} \in \mathbb{R}^D$. Can use a multivariate Gaussian prior
  - Need to estimate a $D \times D$ positive definite covariance matrix $\boldsymbol{\Sigma}$. Can use a Wishart prior
  - If $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ both are unknown, can use Normal-Wishart as a conjugate prior

References

- Review article on Ghahramani, **Probabilistic machine learning and artificial intelligence** *Nature*, *521*(7553), 452-459. (freely available online)

- **Section 4.6 and Section 11.7** Kevin Murphy, Probabilistic Machine Learning: An Introduction (PML-1), MIT Press, 2022 (freely available online)

- **Chapter 2 and Appendix B** of Christopher Bishop, Pattern Recognition and Machine Learning (PRML), Springer, 2007 (freely available online)

- Kevin Murphy, Conjugate Bayesian analysis of the Gaussian distribution https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf

- Probabilistic Machine Learning (CS772A), Piyush Rai