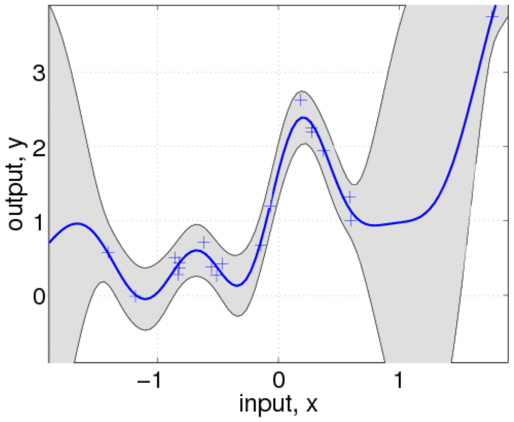
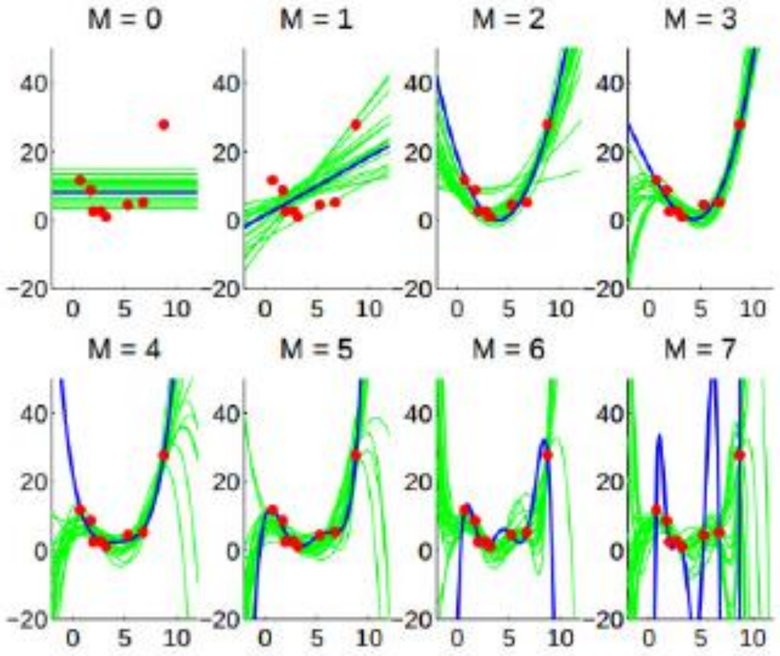
Probabilistic Bayesian Modelling

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

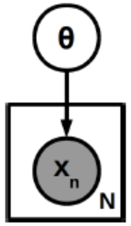
Some probabilities of interest

- 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

NB: 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

$$
\begin{aligned}
P(a) &= \sum_b P(a, b) \quad \text{(Sum Rule)} \\
P(a, b) &= P(a)P(b|a) = P(b)P(a|b) \quad \text{(Product Rule)}
\end{aligned}
$$

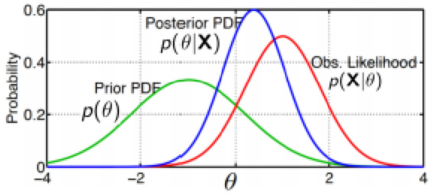
- Note: For continuous random variables, sum is replaced by integral: $P(a) = \int P(a, b)db$
- Another rule is the Bayes rule (can be easily obtained from the above two rules)

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

Bayesian Estimation

- 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

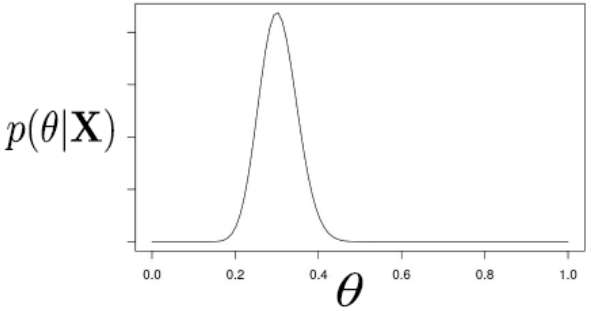
$$\hat{\theta}_{ML} = \arg\max_{\theta} \log p(\mathbf{X}|\theta)$$

  - **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)

Predictions

- 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

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

- 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

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

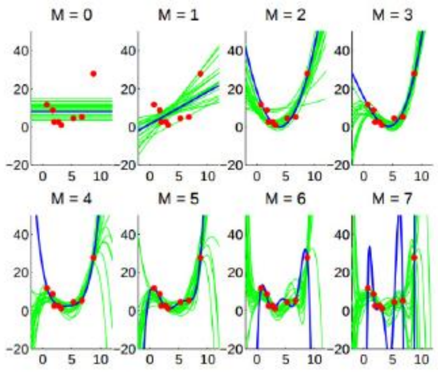
.. where $\hat{\theta}$ is a point estimate of θ (e.g., MLE/MAP)

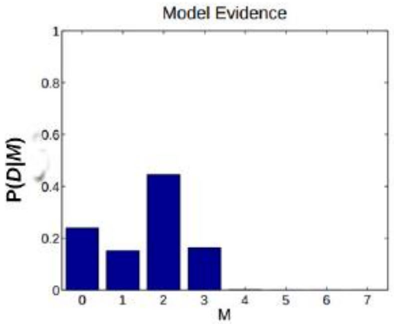
Marginal Likelihood

- 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

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

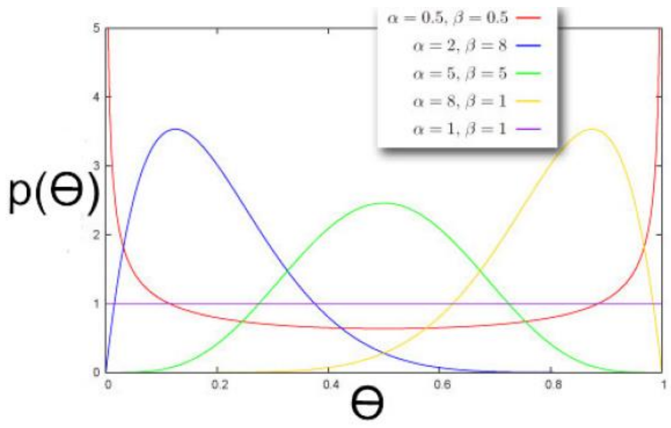
Simple Example (MLE)

- Consider a sequence of N coin tosses (call head $= 0$, tail $= 1$)
- The n^{th} outcome $\boldsymbol{x}_n$ is a binary random variable $\in \{0, 1\}$
- Assume θ to be probability of a head (parameter we wish to estimate)
- Each likelihood term $p(\boldsymbol{x}_n \mid \theta)$ is Bernoulli: $p(\boldsymbol{x}_n \mid \theta) = \theta^{\boldsymbol{x}_n}(1 - \theta)^{1-\boldsymbol{x}_n}$
- Log-likelihood: $\sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \theta) = \sum_{n=1}^{N} \boldsymbol{x}_n \log \theta + (1 - \boldsymbol{x}_n) \log(1 - \theta)$
- Taking derivative of the log-likelihood w.r.t. θ, and setting it to zero gives

$$\hat{\theta}_{MLE} = \frac{\sum_{n=1}^{N} \boldsymbol{x}_n}{N}$$

- $\hat{\theta}_{MLE}$ 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), $\hat{\theta}_{MAP} = \hat{\theta}_{MLE}$

Bayesian Estimate

- 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 <u>proportional</u> 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

Predictions

- 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)

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

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

Multinomial 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. $\sum_{k=1}^{K} \pi_k = 1$

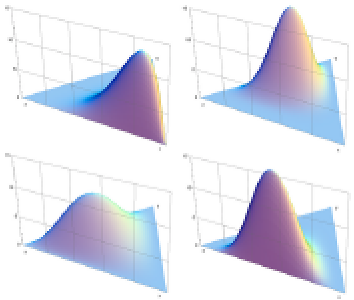
$$p(x_n|\boldsymbol{\pi}) = \text{multinoulli}(x_n|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{\mathbb{I}[x_n=k]}$$

- $\boldsymbol{\pi}$ 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, $\alpha_k \geq 0, \forall k$)

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

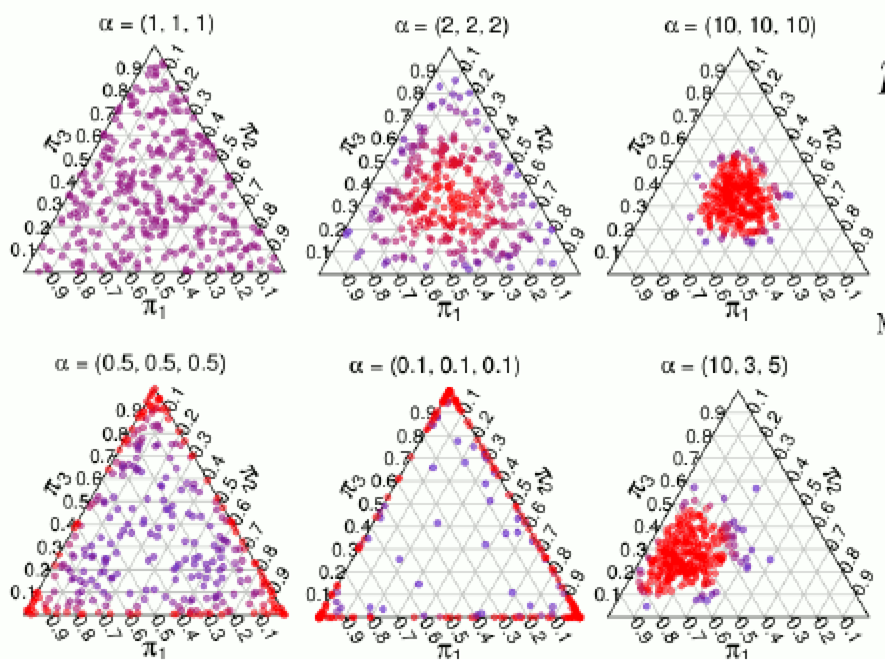
Dirichlet Distribution

PDF for a 3-dim Dirichlet



Red dots denote regions of high probability denity

Draws from a 3-dimensional Dirichlet with different α



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

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

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

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

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

Estimation

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

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

- 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

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

Gaussian Models

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

Fixed Variance Gaussian Model

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

$$
\begin{aligned}
p(x_n|\mu, \sigma^2) &= \mathcal{N}(x|\mu, \sigma^2) \propto \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right] \\
p(\mathbf{X}|\mu, \sigma^2) &= \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)
\end{aligned}
$$

- 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 Estimate of Mean

- The posterior distribution for the unknown mean parameter μ

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

- 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

$$\begin{aligned}\frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \\ \mu_N &= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\bar{x}\end{aligned}$$

(where $\bar{x} = \frac{\sum_{n=1}^{N} x_n}{N}$)

Notion of Sufficient Statistics

Prediction

- 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

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

- Note; Can also get the above result by thinking of x_* as $x_* = \mu + \epsilon$ where $\mu \sim \mathcal{N}(\mu_N, \sigma_N^2)$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independently added observation noise
- Note that, as per the above, the uncertainty in distribution of x_* now has two components
  - σ^2: Due to the noisy observation model, σ_N^2: Due to the uncertainty in μ
- In contrast, the plug-in predictive posterior, given a point estimate $\hat{\mu}$ (e.g., MLE/MAP) would be

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

  - Note that as $N \to \infty$, both approaches would give the same $p(x_*|\mathbf{X})$ since $\sigma_N^2 \to 0$

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

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

  - An inverse-gamma prior $IG(\alpha, \beta)$ has this form (α, β are shape and scale hyperparams, resp)

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

The posterior $p(\sigma^2|\mathbf{X}) = IG(\alpha + \frac{N}{2}, \beta + \frac{\sum_{n=1}^{N}(x_n-\mu)^2}{2})$.

The posterior $p(\sigma^2|\mathbf{X}) = IG(\alpha + \frac{N}{2}, \beta + \frac{\sum_{n=1}^{N}(x_n-\mu)^2}{2})$. Again IG due to conjugacy.

Gaussian Model: Mean and Variance

- Goal: Infer the mean and precision of a univariate Gaussian $\mathcal{N}(x|\mu, \lambda^{-1})$
- Given N i.i.d. observations $\mathbf{X} = \{x_1, \ldots, x_N\}$, the likelihood will be

$$p(\mathbf{X}|\mu, \lambda) = \prod_{n=1}^{N} \sqrt{\frac{\lambda}{2\pi}} \exp\left[-\frac{\lambda}{2}(x_n - \mu)^2\right] \propto \left[\lambda^{1/2} \exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left[\lambda\mu \sum_{n=1}^{N} x_n - \frac{\lambda}{2}\sum_{n=1}^{N} x_n^2\right]$$

- Let's choose the following joint distribution as the prior (compare its form with $p(\mathbf{X}|\mu, \lambda)$)

$$p(\mu, \lambda) \propto \left[\lambda^{1/2} \exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\kappa_0} \exp\left[\lambda\mu c - \lambda d\right] = \underbrace{\exp\left[-\frac{\kappa_0\lambda}{2}(\mu - c/\kappa_0)^2\right]}_{\text{prop. to a Gaussian}} \underbrace{\lambda^{\kappa_0/2} \exp\left[-\left(d - \frac{c^2}{2\kappa_0}\right)\lambda\right]}_{\text{prop. to a gamma}}$$

- The above is known as the Normal-gamma (NG) distribution (product of a Normal and a gamma)

$$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\kappa_0\lambda)^{-1})\text{Gamma}(\lambda|\alpha_0, \beta_0) = \text{NG}(\mu, \lambda|\mu_0, \kappa_0, \alpha_0, \beta_0) \qquad \text{(note: } \mu \text{ and } \lambda \text{ are coupled in the Gaussian part)}$$

where $\mu_0 = c/\kappa_0$, $\alpha_0 = 1 + \kappa_0/2$, $\beta_0 = d - c^2/2\kappa_0$ are prior's hyperparameters

  - NG is conjugate to Gaussian when both mean & precision are unknown

Gaussian Model: Mean and Variance

- Due to conjugacy, $p(\mu, \lambda|\mathbf{X})$ will also be NG: $p(\mu, \lambda|\mathbf{X}) \propto p(\mathbf{X}|\mu, \lambda)p(\mu, \lambda)$

$$p(\mu, \lambda|\mathbf{X}) = \mathrm{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N) = \mathcal{N}(\mu|\mu_N, (\kappa_N\lambda)^{-1})\mathrm{Gamma}(\lambda|\alpha_N, \beta_N)$$

where the updated posterior hyperparameters are given by[1]

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

$$\alpha_N = \alpha_0 + N/2, \quad \beta_N = \beta_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n - \bar{x})^2 + \frac{\kappa_0 N(\bar{x} - \mu_0)^2}{2(\kappa_0 + N)}$$

Posterior Predictive Distribution:

$$p(x_*|\mathbf{X}) = \int \underbrace{p(x_*|\mu, \lambda)}_{\text{Gaussian}} \underbrace{p(\mu, \lambda|\mathbf{X})}_{\text{Normal-Gamma}} d\mu d\lambda = t_{2\alpha_N}\left(x_*|\mu_N, \frac{\beta_N(\kappa_N + 1)}{\alpha_N\kappa_N}\right)$$

Multivariate Gaussian

- The (multivariate) Gaussian with mean $\boldsymbol{\mu}$ and cov. matrix $\boldsymbol{\Sigma}$

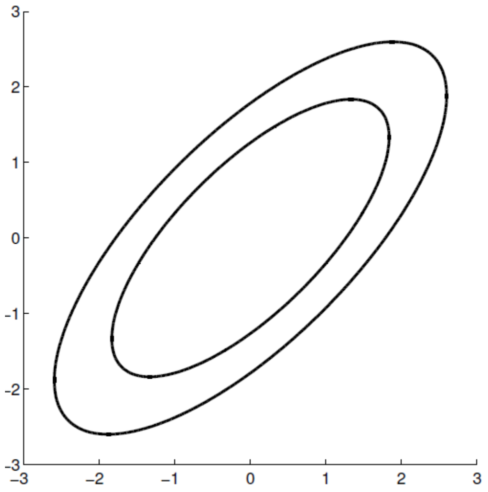
$$\begin{aligned}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\\ &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}\text{trace}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]\right\} \qquad \text{where } \mathbf{S}=(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\end{aligned}$$

- An alternate representation: The "information form"

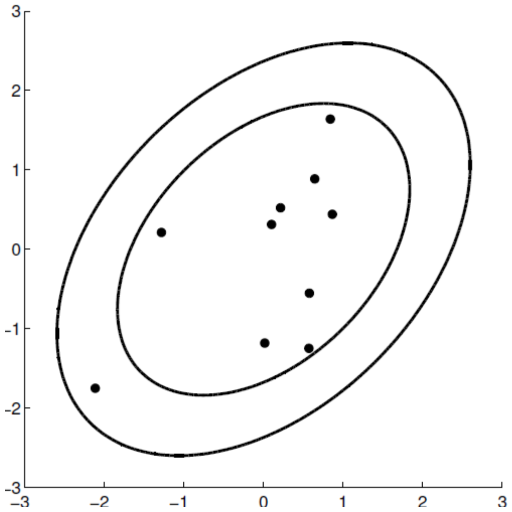
$$\mathcal{N}_c(\boldsymbol{x}|\xi,\boldsymbol{\Lambda}) = (2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}^\top\boldsymbol{\Lambda}\boldsymbol{x}+\xi^\top\boldsymbol{\Lambda}^{-1}\xi-2\boldsymbol{x}^\top\xi\right)\right\}$$

where $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{-1}$ and $\xi=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ are the "natural parameters" (more when we discuss exp. family).

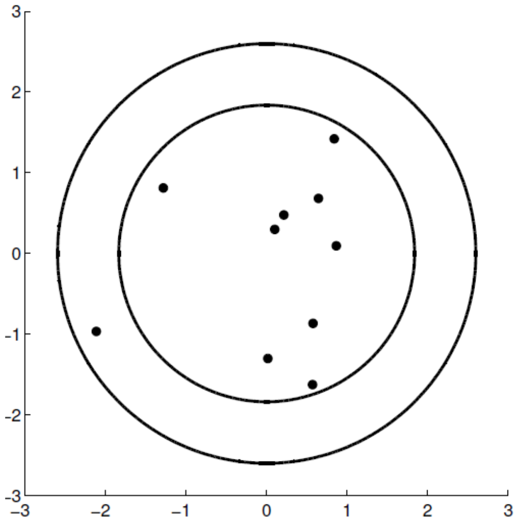
Multivariate Gaussians



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

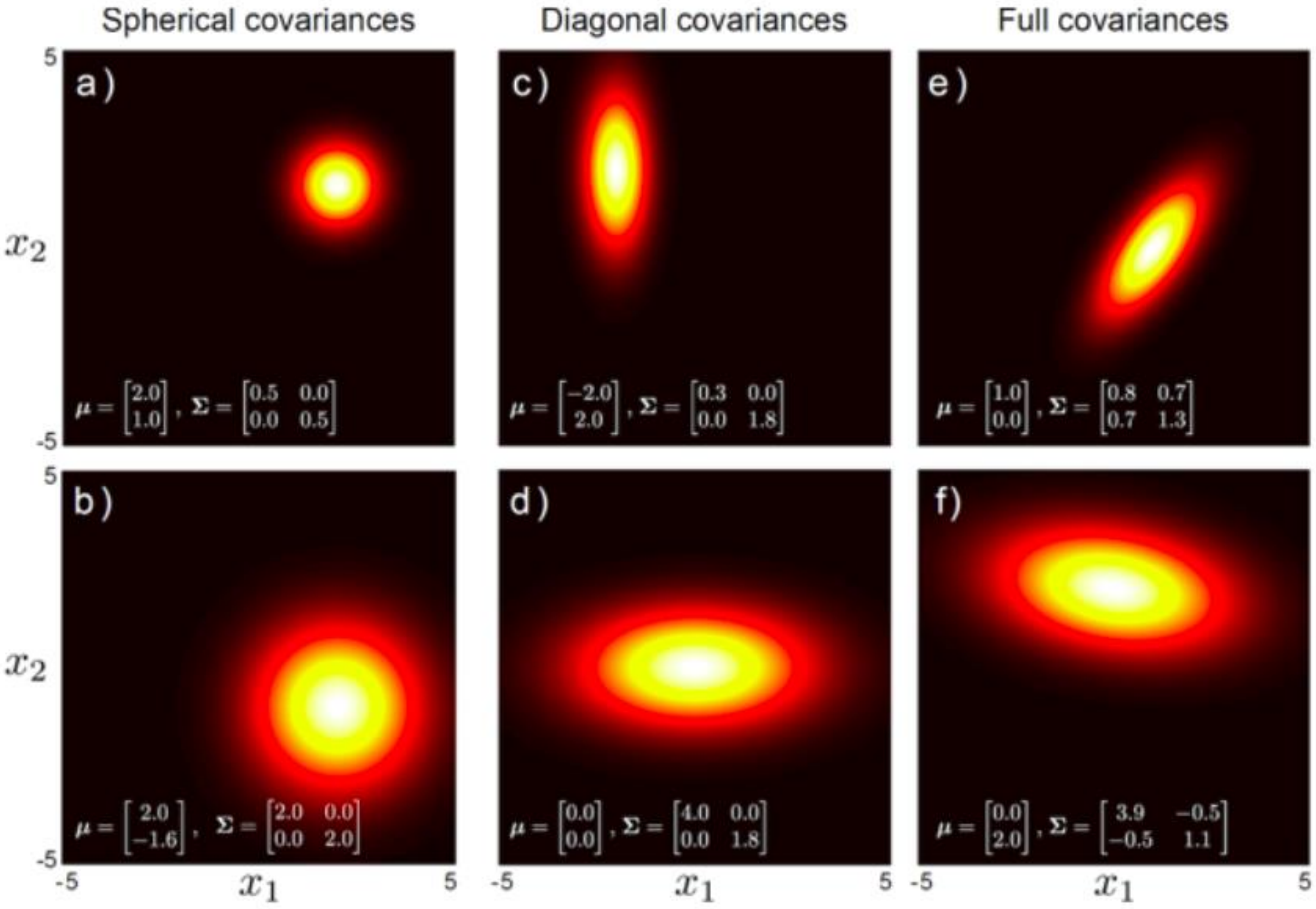


$$\Sigma = \begin{bmatrix} 1 & .4 \\ .4 & 1 \end{bmatrix}$$



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

Covariance Matrix



Multivariate Gaussians: Grouped Variables

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

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}$$

- The marginal distribution of one block, say $\boldsymbol{x}_a$, is a Gaussian

$$p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_b = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

- The conditional distribution of $\boldsymbol{x}_a$ given $\boldsymbol{x}_b$, is Gaussian, i.e., $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$ where

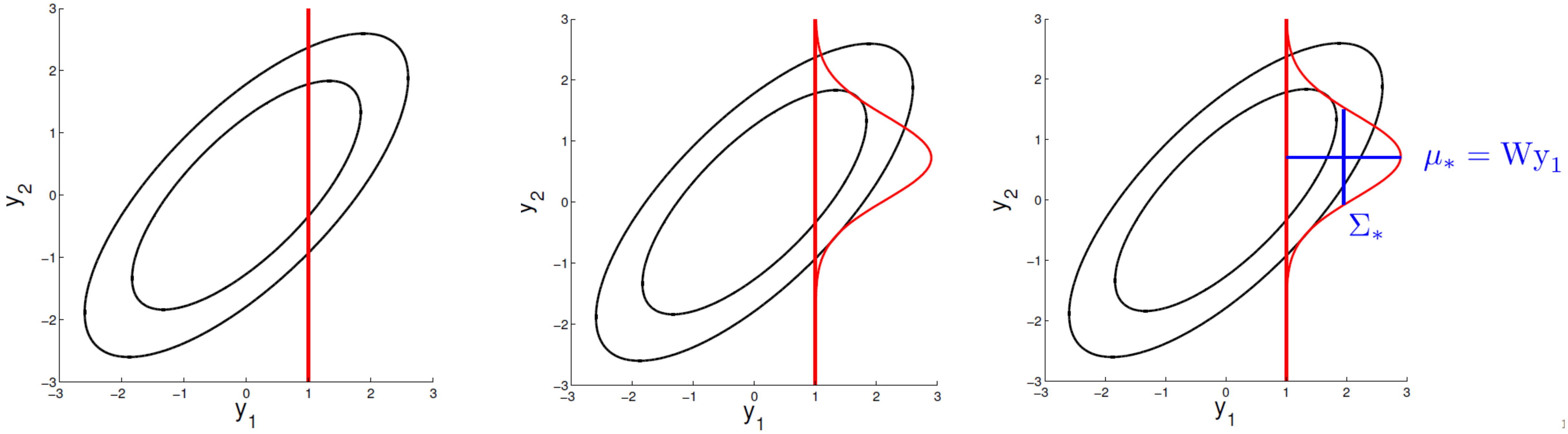
$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$$

("smaller" than Σ_{aa}; makes sense intuitively)

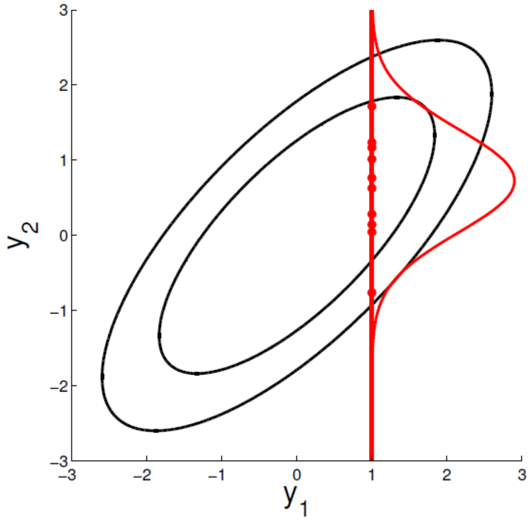
$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)\} \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \\ &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \end{aligned}$$

Conditional Distributions

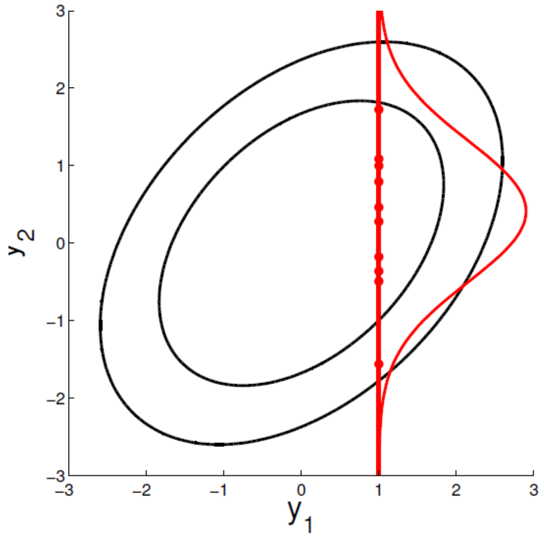
$$p(\mathrm{y}_2|\mathrm{y}_1, \Sigma) \propto \exp\left(-\tfrac{1}{2}(\mathrm{y}_2 - \mu_*)\Sigma_*{}^{-1}(\mathrm{y}_2 - \mu_*)\right)$$



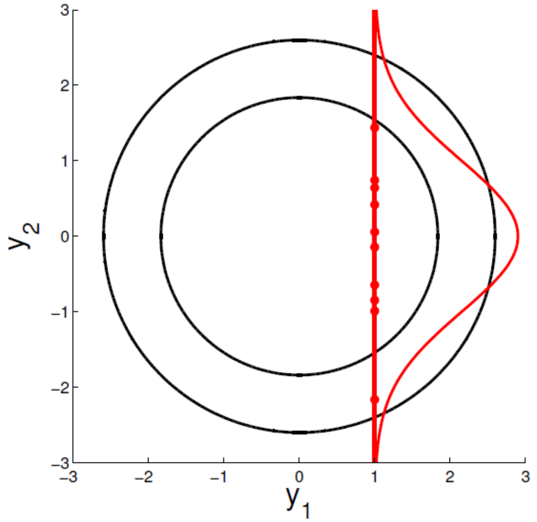
Conditional Distributions



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



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



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

Multivariate Gaussian

- The parameters are now the mean vector and the covariance/precision matrix
- Posterior updates for these have forms similar to that in the univariate case
- For the mean, commonly a multivariate Gaussian prior is used
  - Posterior is also Gaussian due to conjugacy
- For the covariance matrix (with mean fixed), commonly an inverse-Wishart prior is used
  - Posterior is also inverse-Wishart due to conjugacy
- For the precision matrix (with mean fixed), commonly a Wishart prior is used
  - Posterior is also Wishart due to conjugacy
- When both parameters are unknown, there still exist conjugate joint priors
  - Normal-Inverse Wishart for mean + cov matrix, Normal-Wishart for mean + precision matrix

Wishart Distribution: Multidimensional extension of Gamma distribution

Linear Transformation of Random Variables

- Suppose $\boldsymbol{x} = f(\boldsymbol{z}) = \mathbf{A}\boldsymbol{z} + \mathbf{b}$ be a linear function of an r.v. $\boldsymbol{z}$ (not necessarily Gaussian)
- Suppose $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{\mu}$ and $\text{cov}[\boldsymbol{z}] = \boldsymbol{\Sigma}$
  - Expectation of $\boldsymbol{x}$

$$\mathbb{E}[\boldsymbol{x}] = \mathbb{E}[\mathbf{A}\boldsymbol{z} + \mathbf{b}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

  - Covariance of $\boldsymbol{x}$

$$\text{cov}[\boldsymbol{x}] = \text{cov}[\mathbf{A}\boldsymbol{z} + \mathbf{b}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$$

- Likewise if $x = f(\boldsymbol{z}) = \boldsymbol{a}^T\boldsymbol{z} + b$ is a scalar-valued linear function of an r.v. $\boldsymbol{z}$:
  - $\mathbb{E}[x] = \mathbb{E}[\boldsymbol{a}^T\boldsymbol{z} + b] = \boldsymbol{a}^T\boldsymbol{\mu} + b$
  - $\text{var}[x] = \text{var}[\boldsymbol{a}^T\boldsymbol{z} + b] = \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}$

- These properties are often helpful in obtaining the marginal distribution $p(\boldsymbol{x})$ from $p(\boldsymbol{z})$

Linear Gaussian Model

- Consider linear transformation of a Gaussian r.v. $\boldsymbol{z}$ with $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\mu, \boldsymbol{\Lambda}^{-1})$, plus Gaussian noise

$$\boxed{\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \epsilon} \qquad \text{where} \quad p(\epsilon) = \mathcal{N}(\epsilon|\mathbf{0}, \mathbf{L}^{-1})$$

- Easy to see that, conditioned on $\boldsymbol{z}$, $\boldsymbol{x}$ too has a Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$$

- This is called a Linear Gaussian Model. Very commonly encountered in probabilistic modeling
- The following two distributions are of particular interest. Defining $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}$, we have

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{z})} = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top \mathbf{L}(\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{\Lambda}\mu\right\}, \boldsymbol{\Sigma})$$

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\mu + \boldsymbol{b}, \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top + \mathbf{L}^{-1})$$

Exponential Family Distributions

- Defines a class of distributions. An Exponential Family distribution is of the form

$$\boxed{p(\boldsymbol{x}|\theta) = \frac{1}{Z(\theta)} h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] \quad = \quad h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]}$$

- $\boldsymbol{x} \in \mathcal{X}^m$ is the random variable being modeled (where $\mathcal{X}$ denotes some space, e.g., $\mathbb{R}$ or $\{0, 1\}$)
- $\theta \in \mathbb{R}^d$: Natural parameters or canonical parameters defining the distribution
- $\phi(\boldsymbol{x}) \in \mathbb{R}^d$: Sufficient statistics (another random variable)
  - Why "sufficient": $p(\boldsymbol{x}|\theta)$ <u>as a function of θ</u> depends on $\boldsymbol{x}$ only via $\phi(\boldsymbol{x})$
- $Z(\theta) = \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$: Partition function
- $A(\theta) = \log Z(\theta)$: Log-partition function (also called the <u>cumulant function</u>)
- $h(\boldsymbol{x})$: A constant (doesn't depend on θ)

Expressing a Distribution in Exp-family form

- Recall the form of exp-fam distribution: $h(\boldsymbol{x})\exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$
- To write any exp-fam dist $p()$ in the above form, write it as $\exp(\log p())$, e.g., for Binomial

$$
\begin{aligned}
\exp\left(\log \text{Binomial}(x|N,\mu)\right) &= \exp\left(\log \binom{N}{x} \mu^x (1-\mu)^{N-x}\right) \\
&= \exp\left(\log \binom{N}{x} + x\log\mu + (N-x)\log(1-\mu)\right) \\
&= \binom{N}{x} \exp\left(x \log \frac{\mu}{1-\mu} - N\log(1-\mu)\right)
\end{aligned}
$$

- Now compare the resulting expression with the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\exp(\theta^\top \phi(\boldsymbol{x}) - A(\theta))$$

Gaussian as Exponential Form

- Let's try to write a univariate Gaussian in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma\right]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left[\begin{bmatrix}\frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{bmatrix}^\top \begin{bmatrix} x \\ x^2 \end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)\right]$$

- $h(x) = \frac{1}{\sqrt{2\pi}}$
- $\theta = \begin{bmatrix}\frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{bmatrix} = \begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix}$, and $\begin{bmatrix}\mu \\ \sigma^2\end{bmatrix} = \begin{bmatrix}-\frac{\theta_1}{2\theta_2} \\ -\frac{1}{2\theta_2}\end{bmatrix}$
- $\phi(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$
- $A(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma = \frac{-\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2) - \frac{1}{2}\log(2\pi)$

- Many other distribution belong to the exponential family
  - Bernoulli
  - Beta
  - Gamma
  - Multinoulli/Multinomial
  - Dirichlet
  - Multivariate Gaussian
  - .. and many more (`https://en.wikipedia.org/wiki/Exponential_family`)
- Note: Not all distributions belong to the exponential family, e.g.,
  - Uniform distribution ($x \sim \text{Unif}(a, b)$)

MLE on Exponential Families

- Suppose we have data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ drawn i.i.d. from an exponential family distribution

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp\left[\theta^\top \phi(\boldsymbol{x}) - A(\theta)\right]$$

- To do MLE, we need the overall likelihood. This is simply a product of the individual likelihoods

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \sum_{i=1}^{N} \phi(\boldsymbol{x}_i) - NA(\theta)\right] = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$$

- To estimate θ (as we'll see shortly), we only need $\phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$ and N
- Size of $\phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$ does not grow with N (same as the size of each $\phi(\boldsymbol{x}_i)$)
- Only exponential family distributions have finite-sized sufficient statistics
  - No need to store all the data; can simply store and recursively update the sufficient statistics

- The likelihood is of the form $p(\mathcal{D}|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$
- The log-likelihood is (ignoring constant w.r.t. θ): $\log p(\mathcal{D}|\theta) = \theta^\top \phi(\mathcal{D}) - NA(\theta)$
- Note: This is concave in θ (since $-A(\theta)$ is concave). Maximization will yield a global maxima of θ
- MLE for exp-fam distributions can also be seen as doing moment-matching. To see this, note that

$$\nabla_\theta \left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] = \phi(\mathcal{D}) - N\nabla_\theta[A(\theta)] = \phi(\mathcal{D}) - N\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i) - N\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]$$

- Therefore, at the "optimal" (i.e., MLE) $\hat{\theta}$, where the derivative is 0, the following must hold

$$\boxed{\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \frac{1}{N}\sum_{i=1}^{N} \phi(\boldsymbol{x}_i)}$$

matching the expected moments of the distribution with empirical moments

Bayesian Estimate in Exponential Families

- We saw that the total likelihood given N i.i.d. observations $\mathcal{D}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- Let's choose the following prior (note: it looks similar in terms of θ within the exponent)

$$\boxed{p(\theta|\nu_0, \boldsymbol{\tau}_0) = h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta) - A_c(\nu_0, \boldsymbol{\tau}_0)\right]}$$

- Ignoring the prior's log-partition function $A_c(\nu_0, \boldsymbol{\tau}_0) = \log \int_\theta h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right] d\theta$

$$\boxed{p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]}$$

- Comparing the prior's form with the likelihood, we notice that
  - ν_0 is like the number of "pseudo-observations" coming from the prior
  - τ_0 is the total sufficient statistics of these ν_0 pseudo-observations

Posterior Distribution

- As we saw, the likelihood is

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\mathbf{x}_i)$$

- And the prior we chose is

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$$

- For this form of the prior, the posterior $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$ will be

$$\boxed{p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]}$$

- Note that the posterior has the same form as the prior; such a prior is called a **conjugate prior** (note: all exponential family distributions have a conjugate prior having a form shown as above)

- Thus posterior hyperparams $\nu_0{}', \tau_0{}'$ are obtained

$$\begin{aligned} \nu_0{}' &\leftarrow \nu_0 + N \\ \tau_0{}' &\leftarrow \tau_0 + \phi(\mathcal{D}) \end{aligned}$$

Contd..

- Assuming the prior $p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$, the posterior was
$$p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]$$
- Assuming $\tau_0 = \nu_0 \bar{\tau}_0$, we can also write the prior as $p(\theta|\nu_0, \bar{\tau}_0) \propto \exp\left[\theta^\top \nu_0 \bar{\tau}_0 - \nu_0 A(\theta)\right]$
- Can think of $\bar{\tau}_0 = \tau_0/\nu_0$ as the average sufficient statistics per pseudo-observation
- The posterior can be written as
$$p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\nu_0 + N) \frac{\nu_0 \bar{\tau}_0 + \phi(\mathcal{D})}{\nu_0 + N} - (\nu_0 + N)A(\theta)\right]$$
- Denoting $\bar{\phi} = \frac{\phi(D)}{N}$ as the average suff-stats per real observation, the posterior updates are
$$\begin{aligned} {\nu_0}' &\leftarrow \nu_0 + N \\ \bar{\tau}_0' &\leftarrow \frac{\nu_0 \bar{\tau}_0 + N\bar{\phi}}{\nu_0 + N} \end{aligned}$$

Posterior Predictive Distribution

- Assume some past (training) data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ generated from an exp. family distribution
- Assme some test data $\mathcal{D}' = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{N'}\}$ from the same distribution ($N' \geq 1$)
- The posterior predictive distribution of $\mathcal{D}'$ (probability distribution of new data given old data)

$$p(\mathcal{D}'|\mathcal{D}) = \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta$$

$$\begin{aligned} p(\mathcal{D}'|\mathcal{D}) &= \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta \\ &= \int \underbrace{\left[\prod_{i=1}^{N'} h(\tilde{\boldsymbol{x}}_i)\right]}_{\text{constant w.r.t. } \theta} \exp\left[\theta^\top \phi(\mathcal{D}') - N'A(\theta)\right] h(\theta) \exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta) - \underbrace{A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))}_{\text{constant w.r.t. } \theta}\right] d\theta \end{aligned}$$

Summary of Single Node Models

- Likelihood, Prior, Posterior, Predictive, Model averaging
- Hyperparameters (Parametric/Non-parametric models)
- Conjugate priors and closed form expression
- Point estimates (MLE, MAP), Distribution Estimates (Bayesian)
- Generative models

- Bernoulli (coin)
- Multinomial (dice)
- Gaussians (continuous variables)
- Exponential families

Questions