Introduction to Bayesian Statistics: Sufficiency and Likelihood Principles, Prior,Posterior and Posterior Predictive Distributions, Gaussian Examples, Exponential Family of Distributions

Prof. Nicholas Zabaras
Center for Informatics and Computational Science
https://cics.nd.edu/
University of Notre Dame
Notre Dame, Indiana, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

September 3, 2018

References

- C P Robert, The Bayesian Choice: From Decision-Theoretic Motivations to Compulational Implementation, Springer-Verlag, NY, 2001 (online resource)
- A Gelman, JB Carlin, HS Stern and DB Rubin, Bayesian Data Analysis, Chapman and Hall CRC Press, 2nd Edition, 2003.
- J M Marin and C P Robert, The Bayesian Core, Spring Verlag, 2007 (online resource)
- D. Sivia and J Skilling, Data Analysis: A Bayesian Tutorial, Oxford University Press, 2006.
- Bayesian Statistics for Engineering, Online Course at Georgia Tech, B. Vidakovic.
- Chris Bishops' PRML book, Chapter 2
- M. Jordan, An introduction to Probabilistic Graphical Models, Chapter 8 (pre-print)
- Kevin Murphy's, Machine Learning: A probablistic perspective, Chapters 2 and 4

Contents

Parametric Modeling

- **Statistical theory** derives from observations of a random phenomenon an inference about the probability distribution underlying this phenomenon.[a]

- We consider **parametric modeling**: The observations x are the realizations of a random variable X of known probability density function $f(x|\theta)$ where

  - θ is unknown and belongs to a space Θ of finite dimension.

  - The function $f(x|\theta)$ considered as a function of θ for a fixed realization of the observation $X = x$ is called **the likelihood function**.

$$\ell(\theta \mid x) = f(x \mid \theta)$$

[a] Here we follow closely:
- C. P. Robert, *The Bayesian Choice, Springer, 2nd edition,* chapter 1 (full text available)
- Brani Vidakovic, Bayesian Statistics for Engineers, online course.

Example of Parametric Modeling

- Consider the problem of forest fires. Ecological and meteorological factors influence their eruption. Determining the probability p of fire as a function of these factors can be useful in the prevention of forest fires.

- We assume a parametrized shape for the function p.

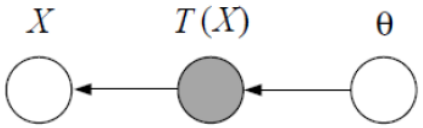
- Denoting by h the humidity rate, t the average temperature, x the degree of management of the forest, a logistic model (Bernoulli random variable of parameter p) could be proposed as:

$$p = \exp(\alpha_1 h + \alpha_2 t + \alpha_3 x) / [1 + \exp(\alpha_1 h + \alpha_2 t + \alpha_3 x)]$$

- The statistical step is now dealing with the evaluation of $\alpha_1, \alpha_2, \alpha_3$.

Sufficiency Principle

- Consider $X \sim f(x \mid \theta)$. **A function T of X (called a statistic of X) is said to be sufficient if the distribution of X conditional upon $T(X)$ is independent of θ.**



$$f(x \mid \theta) = h(T(x) \mid \theta)\, g(x \mid T(x))^*$$

- **A sufficient statistic $T(x)$ contains the whole information brought by x about θ.**

- Let us consider a simple example. Let $X = (X_1, X_2, \ldots, X_n)$ be i.i.d. from $\mathcal{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$. Then we can write:

$$f(x \mid \theta) = \prod_{j=1}^{N} \mathcal{N}(x_j \mid \theta) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_j - \mu)^2\right) = \frac{1}{(2\pi\theta_2)^{N/2}} \exp\left(-\frac{1}{2\theta_2}\sum_{j=1}^{N}(x_j - \theta_1)^2\right) =$$

$$= \frac{1}{(2\pi\theta_2)^{N/2}} \exp\left(-\frac{1}{2\theta_2}\sum_{j=1}^{N} x_j^2 + \frac{\theta_1}{\theta_2}\sum_{j=1}^{N} x_j - \frac{N\theta_1^2}{2\theta_2}\right)$$

* $f(x|\theta) = f(x, T(x)|\theta) = h(T(x)|\theta)\, g(x \mid T(x), \theta) = h(T(x)|\theta)\, g(x \mid T(x))$

Sufficiency Principle

$$f(x\,|\,\theta)=\frac{1}{\left(2\pi\theta_2\right)^{N/2}}\exp\left(-\frac{1}{2\theta_2}\sum_{j=1}^{N}x_j^2+\frac{\theta_1}{\theta_2}\sum_{j=1}^{N}x_j-\frac{N\theta_1^2}{2\theta_2}\right)$$

- We can see that $f(x\,|\,\theta)$ depends only on $T(x)=\left(\sum_{j=1}^{N}x_j,\sum_{j=1}^{N}x_j^2\right)$ which is our set of sufficient statistics.
- Introducing $\bar{x}=\frac{\sum_{j=1}^{N}x_j}{N}$, $s^2=\sum_{j=1}^{N}(x_j-\bar{x})^2$, we can also re-write the above equation:

$$f(x|\theta)=\frac{1}{(2\pi\theta_2)^{N/2}}\exp\left(-\frac{s^2+N\bar{x}^2}{2\theta_2}+\frac{\theta_1 N\bar{x}}{\theta_2}-\frac{N\theta_1^2}{2\theta_2}\right)$$

So $\bar{x}, s^2$ is an alternative set of sufficient statistics.

Sufficiency Principle

- **Sufficiency principle:** Two observations x and y have the same values of $T(x) = T(y)$ *of statistics sufficient for* $f(.|\theta)$. Then the inferences about θ based on x and y should be the same.

- Consider the model $X_i \sim \mathcal{N}(\mu, 1)$ and we want to estimate μ (our θ) based on N data. In this case, the sufficient statistic is

$$T(x_{1:N}) = \sum_{j=1}^{N} x_j$$

- Consider the (MLE) estimate of θ: $\hat{\mu}_1 = \sum_{j=1}^{N} x_j \Big/ N$. It satisfies the sufficiency principle because if we have another dataset $x'_{1:N}$ such that

$$T(x_{1:N}) = T(x'_{1:N}) \text{ then } \hat{\mu}_2 = \frac{1}{N}\sum_{j=1}^{N} x'_j = \frac{1}{N}\sum_{j=1}^{N} x_j = \hat{\mu}_1$$

- On the other hand, the estimate $\hat{\mu}_1 = x_1$ does not satisfy the sufficiency principle for $n > 1$ because if we have another dataset $x'_{1:N}$ such that $T(x_{1:N}) = T(x'_{1:N})$, then $\hat{\mu}_2 = x'_1 \neq \hat{\mu}_1$, *if* $x'_1 \neq x_1$.

The Likelihood Principle

- **Likelihood Principle.** In the inference about θ, the information brought by an observation is entirely contained in the likelihood function $\ell(\theta \mid x) = f(x \mid \theta)$.

- Also, two likelihood functions contain the same information about θ if they are proportional to each other; i.e.

$$\ell_1(\theta \mid x) = c(x)\,\ell 2(\theta \mid x)$$

- It is straight forward to show that the MLE (maximum likelihood procedure) satisfies the likelihood principle

$$arg\ \max_{\theta} \ell_1(\theta \mid x) = arg\ \max_{\theta} \ell_2(\theta \mid x)$$

- Classical approaches do not necessarily satisfy the likelihood principle.

MLE: Summary

- The likelihood principle is fairly vague since it does not lead to the selection of a particular procedure.

- Maximum likelihood estimation is one way to implement the sufficiency and likelihood principles

$$\hat{\theta} = \arg\sup_{\theta} \ell(\theta|x)$$

- Indeed:

$$\arg\sup_{\theta} \ell(\theta \mid x) = \arg\sup_{\theta} h(x) g(T(x) \mid \theta) = \arg\sup_{\theta} g(T(x) \mid \theta)$$

$$\ell_1(\theta \mid x) = c(x)\ell_2(\theta \mid x) \Rightarrow \arg\sup_{\theta} \ell_1(\theta \mid x) = \arg\sup_{\theta} \ell_2(\theta \mid x)$$

Maximum Likelihood for a Gaussian

- Suppose that we have a data set of observations $\boldsymbol{\mathcal{D}} = (x_1, \ldots, x_N)^{\boldsymbol{T}}$, representing N observations of the scalar random variable X. The observations are drawn independently from a Gaussian distribution whose mean μ and variance σ^2 are unknown.

- We would like to determine these parameters from the data set.

- *i.i.d. data: Data points that are drawn independently from the same distribution are said to be independent and identically distributed, which is often abbreviated to i.i.d.*

Maximum Likelihood for a Gaussian

- Because our data set $\mathcal{D}$ is i.i.d., we can write the probability of the data set, given μ *and* σ^2*, in the form*

Likelihood function: $p(\mathbf{x} \mid \mu, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(x_i \mid \mu, \sigma^2)$

This is seen as a function of μ, σ^2

Max Likelihood for a Gaussian Distribution

$$Likelihood\ \ function: p(\boldsymbol{x} \mid \mu, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(x_i \mid \mu, \sigma^2)$$

- One common criterion for determining the parameters in a probability distribution using an observed data set is to find the parameter values that *maximize the likelihood function, i.e. maximizing the probability of the data given the parameters* (contrast this with maximizing the probability of the parameters given the data).

- We can equivalently maximize the log-likelihood:

$$\max_{\mu,\sigma^2} \ln p(\boldsymbol{x} \mid \mu, \sigma^2) = \max_{\mu,\sigma^2} \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) \right) \Rightarrow$$

$$\mu_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i, \ \sigma^2_{ML} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_{ML})^2$$

Maximum Likelihood for a Gaussian Distribution

$$\mu_{ML} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} x_i}_{\text{Sample mean}}, \quad \sigma^2_{ML} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ML})^2}_{\text{Sample variance wrt MLE mean (not the exact mean)}}$$

- The MLE underestimates the variance (*bias* due to overfitting) because μ_{ML} fitted some of the noise in the data.
- The maximum likelihood solutions μ_{ML}, σ^2_{ML} are functions of the data set values $x_1, \ldots, x_N$. Consider the expectations of these quantities with respect to the data set values, which come from a Gaussian.
- Using the equations above you can show that :

$$\mathbb{E}\left[\mu_{ML}\right] = \mu, \quad \mathbb{E}\left[\sigma^2_{ML}\right] = \frac{N-1}{N}\sigma^2$$

In this derivation you need to use :

$$\mathbb{E}\left[x_i x_j\right] = \mu^2 \quad \text{for } i \neq j$$

$$\mathbb{E}\left[x_i^2\right] = \sigma^2 + \mu^2$$

Maximum Likelihood for a Gaussian Distribution

$$\mathbb{E}\left[\sigma^2_{ML}\right] = \frac{N-1}{N}\sigma^2$$

We use :

$$\mathbb{E}\left[x_i x_j\right] = \mu^2 \quad for\ i \neq j$$

$$\mathbb{E}\left[x_i^2\right] = \sigma^2 + \mu^2$$

$$\mathbb{E}\left[\sigma^2_{ML}\right] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2\right] = \frac{1}{N}\mathbb{E}\left[\sum_{n=1}^{N}(x_n - \frac{1}{N}\sum_{m=1}^{N}x_m)^2\right]$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - \frac{2}{N}x_n\sum_{m=1}^{N}x_m + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}x_m x_l\right]$$

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

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

$$= \frac{(N-1)}{N}\sigma^2$$

Maximum Likelihood for a Gaussian Distribution

$$\mathbb{E}[\mu_{ML}] = \mu, \sigma^2_{ML} = \frac{N-1}{N}\sigma^2, \mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i, \sigma^2_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ML})^2$$

- On average the MLE estimate obtains the correct mean but will *underestimate the true variance by a factor* $(N-1)/N$.
- An unbiased estimate of the variance is given as:

$$\bar{\sigma}^2 = \frac{N}{N-1}\sigma^2_{ML} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_{ML})^2$$

For large N, the bias is not a problem

- This result can be obtained from "a Bayesian treatment" in which we *marginalize over the unknown mean*.
- The $N-1$ factor takes account the fact that "1 *degree of freedom has been used in fitting the mean" and removes the bias of the MLE*.

MLE for the Multivariate Gaussian

- We can easily generalize the earlier MLE results for a multivariate Gaussian. The log-likelihood takes the form:

$$\ln p(\boldsymbol{X} \mid \mathcal{D}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})$$

- Setting the derivatives wrt $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ equal to zero gives the following:

$$\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n, \boldsymbol{\Sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^T$$

- We provide a proof of the calculation of $\boldsymbol{\Sigma}_{ML}$ next.

MLE for the Multivariate Guassian

$$\ln p(\boldsymbol{X} \mid \mathcal{D}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})$$

- We differentiate the log likelihood wrt $\boldsymbol{\Sigma}^{-1}$. Each contributing term is:

$$-\frac{N}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}}\ln|\boldsymbol{\Sigma}| = \frac{N}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}}\ln|\boldsymbol{\Sigma}^{-1}| = \frac{N}{2}\boldsymbol{\Sigma}^T = \boxed{\frac{N}{2}\boldsymbol{\Sigma}}$$

A useful trick!

$$-\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) = -\frac{1}{2}N\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}}Tr\left(\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}\frac{1}{N}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^T\right)$$

$$= -\frac{1}{2}N\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}}Tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right)$$

S symmetric

$$= \boxed{-\frac{1}{2}N\boldsymbol{S}}, \quad where\ \boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^T$$

- Setting the derivative equal to zero leads to: $\boldsymbol{\Sigma}_{ML} = \boldsymbol{S}$

- Here we used:

$$\frac{\partial}{\partial \mathbf{A}}Tr(\mathbf{AB}) = \mathbf{B}^T, \frac{\partial}{\partial \boldsymbol{A}}\ln|\boldsymbol{A}| = \left(\boldsymbol{A}^{-1}\right)^T,$$

$$|\boldsymbol{A}^{-1}| = |\boldsymbol{A}|^{-1}, tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$$

Appendix: Some Useful Matrix Operations

- Show that

$$\frac{\partial}{\partial \mathbf{A}} Tr(\mathbf{AB}) = \mathbf{B}^T \text{ and } \frac{\partial}{\partial \mathbf{B}} Tr(\mathbf{AB}) = \mathbf{A}^T$$

Indeed

$$\frac{\partial}{\partial A_{mn}} Tr(\mathbf{AB}) = \frac{\partial}{\partial A_{mn}} (A_{ik} B_{ki}) = B_{nm} \Rightarrow \frac{\partial}{\partial \mathbf{A}} Tr(\mathbf{AB}) = \mathbf{B}^T$$

- Show that

$$\frac{\partial}{\partial \boldsymbol{A}} \ln |\boldsymbol{A}| = \left(\boldsymbol{A}^{-1}\right)^T$$

Using the cofactor expansion of the det:

$$\frac{\partial}{\partial A_{mn}} \ln |\boldsymbol{A}| = \frac{1}{|\boldsymbol{A}|} \frac{\partial}{\partial A_{mn}} |\boldsymbol{A}| = \frac{1}{|\boldsymbol{A}|} \frac{\partial}{\partial A_{mn}} \sum_j (-1)^{i+j} A_{ij} M_{ij} = \frac{1}{|\boldsymbol{A}|} (-1)^{m+n} M_{mn} = \left(\boldsymbol{A}^{-1}\right)_{nm}$$

where in the last step we used Cramer's rule.

MLE for a Multivariate Gaussian

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \equiv \bar{x},\ \Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})(x_n - \mu_{ML})^T = \frac{1}{N}\sum_{n=1}^{N} x_n x_n^T - \bar{x}\bar{x}^T$$

- Note that *the unconstrained maximization of the log-likelihood gives a symmetric Σ.*

- As for the univariate case, we can define an unbiased covariance as:

$$\bar{\Sigma}_{ML} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{ML})(x_n - \mu_{ML})^T,\ \mathbb{E}[\bar{\Sigma}_{ML}] = \Sigma$$

- To prove this, you will need to use that:

$$\mathbb{E}\left[\boldsymbol{x}_n \boldsymbol{x}_m^T\right] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \delta_{mn}\boldsymbol{\Sigma}$$

Bayesian Statistics

- A Bayesian model is made of a parametric statistical model $(\mathcal{X}, f\ (x|\ \theta))$ and a prior distribution on the parameters $(\Theta, \pi\ (\theta))$.

- The unknown parameters are now considered as random.

  - Some statisticians question this approach but most accept the probabilistic modeling on the observations.

- Example: Assume you want to measure the speed of light given some observations. Why should you put a prior on a physical constant?

  - Due to the limited accuracy of the measurement, this constant will never be known exactly.

  - It is thus justified to put a (e.g. uniform) prior on this parameter reflecting this uncertainty.

Recall Bayes' rule

- Coming back from a trip, you feel sick and your doctor thinks you might have contracted a rare disease (0.01% of the population has the disease).

- A test is available but not perfect.

  - If a tested patient has the disease, 100% of the time the test will be positive.

  - If a tested patient does not have the disease, 95% of the the time the test will be negative (5% false positive).

- Your test is positive, should you really care?

Medical Diagnosis Example

- Let A=`the patient has the disease'
- Let B=`the test returns a positive result'

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} = \frac{1 \times 0.0001}{1 \times 0.0001 + 0.05 \times 0.9999} \approx 0.002$$

- Such a test would be a complete waste of money

Prior, Likelihood and Posterior

In the previous example, we can identify the following:

- Data x *(e.g.` the test is positive')*
- Hypothesis h (e.g. `do you have the disease?'). We want to make inferences about h.

In Bayesian settings all variables are random and all inferences are probabilistic.

We identify three key ingredients of a Bayesian inference approach:

- Prior $\pi(h)$*:* How likely is hypothesis h before looking at the data
- Likelihood $f(x \mid h)$*:* How likely is to observe x assuming h is true.
- Posterior $\pi(h \mid x)$*:* How likely is h after data x have been observed.

$$\pi(h \mid x) = \frac{f(x \mid h)\pi(h)}{m(x)}$$

Prior $\pi(\theta)$

- We use the prior to introduce quantitatively some insights on the parameters of interest.
- This can be as subjective or as objective as you want it to be – and that's why frequentists do not like Bayesian approaches!
- There is no such a thing as a true prior!
- Even when prior information is heavily subjective, the Bayesian inference model is honest.

Likelihood $f(x|\theta)$

- The likelihood encapsulates the mathematical model of the physical phenomena you are investigating.

- If you know the input $X = x$ to your problem, the likelihood can represent the computed output $y = f(x)$.

- It is the most computational expensive part of Bayesian approaches to inference problems (inverse problems).

Posterior $\pi(\theta|x)$: Inference and Prediction

- It combines the prior and likelihood.
- It weights the data and the prior information in making probabilistic inferences
- The posterior distribution is also useful in estimating the probability of observing a future outcome (prediction)

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{m(x)}$$

- The normalizing factor is given as:

$$m(x) = \int f(x|\theta)\pi(\theta)d\theta$$

Posterior Inference: Point Estimates

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{m(x)}$$

Maximum A Posteriori estimate (MAP)

$$\theta^* = \arg\max_{\theta} \log\left(\pi(\theta \mid x)\right) = \arg\max_{\theta} \left(\log \pi(x \mid \theta) + \log \pi(\theta)\right)$$

Posterior Mean

$$\hat{\theta} = \mathbb{E}_{p(\theta|x)}[\theta] = \int \theta \pi(\theta|x) d\theta$$

Posterior Quantiles

$$\Pr\left[\theta > a\right] = \int_a^{\infty} \pi(\theta \mid x) d\theta$$

Prediction

Suppose we have observed $\boldsymbol{x}$ and we want to make a prediction about (future) unknown observables: What is the probability of observing data $\hat{\boldsymbol{x}}$ If we already have observed data $\boldsymbol{x}$?

This means finding $g(\hat{\boldsymbol{x}}|\boldsymbol{x})$

We have:

$$g(\hat{\boldsymbol{x}}|\boldsymbol{x}) = \int g(\hat{\boldsymbol{x}}, \theta|\boldsymbol{x})d\theta = \int \frac{\pi(\hat{\boldsymbol{x}}, \theta, \boldsymbol{x})}{m(\boldsymbol{x})} d\theta = \int \frac{\pi(\hat{\boldsymbol{x}}, \theta, \boldsymbol{x})}{\phi(\theta, \boldsymbol{x})} \frac{\phi(\theta, \boldsymbol{x})}{m(\boldsymbol{x})} d\theta =$$

$$= \int f(\hat{\boldsymbol{x}}|\theta, \boldsymbol{x})\pi(\theta|\boldsymbol{x})d\theta = \int f(\hat{\boldsymbol{x}}|\theta)\pi(\theta|\boldsymbol{x})d\theta$$

Compare this with the normalizing factor in Bayes' rule:

$$m(\hat{\boldsymbol{x}}) = \int f(\hat{\boldsymbol{x}}|\theta)\pi(\theta)d\theta$$

A Gaussian Example

- Consider $X_1 \mid \theta \sim \mathcal{N}(\theta, \sigma^2)$, *with prior* $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$.
- Then we can derive the following:

$$\pi(\theta \mid x_1) \propto f(x_1 \mid \theta)\pi(\theta) \propto \exp\left(-\frac{(x_1-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right) \Rightarrow$$

$$\pi(\theta \mid x_1) \propto \exp\left(-\frac{\theta^2}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right)+\theta\left(\frac{x_1}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\right) \propto \exp\left(-\frac{1}{2\sigma_1^2}(\theta-\mu_1)^2\right) \Rightarrow$$

With prior $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$ *and observation* x_1

$\theta \mid x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ *with*

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2} \Rightarrow \sigma_1^2 = \frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2}, \text{ and}$$

$$\mu_1 = \sigma_1^2\left(\frac{x_1}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)$$

A Gaussian Example: Continued

- To predict the distribution of a new observation $X \mid \theta \sim \mathcal{N}(\theta, \sigma^2)$ in light of x_1, we use the predictive distribution as follows:

$$f(x \mid x_1) = \int \underbrace{f(x \mid \theta)}_{Likelihood} \underbrace{\pi(\theta \mid x_1)}_{Posterior} d\theta \propto \int e^{-\frac{(x-\theta)^2}{2\sigma^2}} e^{-\frac{(\theta-\mu_1)^2}{2\sigma_1^2}} d\theta = \int e^{-\frac{1}{2}\left(\frac{(x-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_1)^2}{\sigma_1^2}\right)} d\theta$$

- You can verify with direct substitution the following:

$$-\frac{1}{2}\left(\frac{(x-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_1)^2}{\sigma_1^2}\right) = -\frac{1}{2}(x-\mu_1 \quad \theta-m_1)\frac{1}{\sigma_1^2\sigma^2}\begin{bmatrix}\sigma_1^2 & -\sigma_1^2 \\ -\sigma_1^2 & \sigma_1^2+\sigma^2\end{bmatrix}\begin{pmatrix}x-\mu_1 \\ \theta-\mu_1\end{pmatrix}+\cdots$$

$$= -\frac{1}{2}(x-\mu_1 \quad \theta-\mu_1)\begin{bmatrix}\sigma_1^2+\sigma^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2\end{bmatrix}^{-1}\begin{pmatrix}x-\mu_1 \\ \theta-\mu_1\end{pmatrix}+..$$

- This is a bivariate Gaussian and thus $f(x|x_1)$ is the marginal in x, i.e.

$$X \mid x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2+\sigma^2)$$

Bayesian Inference for the Gaussian

- Consider $\boldsymbol{X} = \{x_1, x_2, ..., x_N\} \sim \mathcal{N}(\mu, \sigma^2)$, *with prior* $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$.

- The likelihood takes the form:

$$p(\boldsymbol{X} \mid \mu) = \prod_{n=1}^{N} f(x_n \mid \mu) = \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(-\frac{\sum_{n=1}^{N} (x_n - \mu)^2}{2\sigma^2} \right)$$

- Note that in terms of μ this is not a probability density and is not normalized. Introducing the conjugate (Gaussian) prior on μ leads to:

$$\pi(\mu \mid \boldsymbol{X}) = \prod_{n=1}^{N} f(x_n \mid \mu)\pi(\mu) \propto \exp\left(-\frac{\sum_{n=1}^{N} (x_n - \mu)^2}{2\sigma^2} - \frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right) \Rightarrow$$

$$\pi(\mu \mid \boldsymbol{X}) \propto \exp\left(-\frac{\mu^2}{2}\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \right) + \mu\left(\frac{\sum_{n=1}^{N} x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) \right) \propto \exp\left(-\frac{1}{2\sigma_N^2}(\mu - \mu_N)^2 \right)$$

Bayesian Inference for the Gaussian

$$\pi(\mu \mid \boldsymbol{X}) \propto \exp\left(-\frac{\mu^2}{2}\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \right) + \mu \left(\frac{\sum_{n=1}^{N} x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) \right) \propto \exp\left(-\frac{1}{2\sigma_N^2}(\mu - \mu_N)^2 \right)$$

- So the posterior is a Gaussian as before with

$$\mu \mid \boldsymbol{X} \sim \mathcal{N}(\mu_N, \sigma_N^2) \text{ with}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \Rightarrow \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}, \text{ and}$$

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

Bayesian Inference for the Gaussian

$$\mu \mid \boldsymbol{X} \sim \mathcal{N}(\mu_N, \sigma_N^2) \text{ with}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \Rightarrow \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}, \text{ and } \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0$$

- *The posterior precision is the sum of the precision of the prior plus one contribution of the data precision for each observed data point*.
- Observe the posterior mean for $N \to \infty$ and $N \to 0$.
- For $N \to \infty$ the posterior peaks around the μ_{ML} and the posterior variance goes to zero, i.e. the point MLE estimate is recovered within the Bayesian paradigm for infinite data.
- How about when $\sigma_0^2 \to \infty$? In this case note that $\sigma_N^2 \to \frac{\sigma^2}{N}$ *and* $\mu_N \to \mu_{ML}$

Gaussian Example: Bayes' versus MLE

- We have seen that the ML estimate of θ at data point N is simply:

$$\theta_{ML} = \arg\sup_{\theta} \prod_{i=1}^{N} f(x_i \mid \theta) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- The posterior of θ at time N is (simply generalizing the earlier result):

$$\theta \mid x_1, \ldots, x_N \sim \mathcal{N}(\mu_N, \sigma_N^2)$$

where

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \Rightarrow \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2} \underset{N\to\infty}{\sim} \frac{\sigma^2}{N}$$

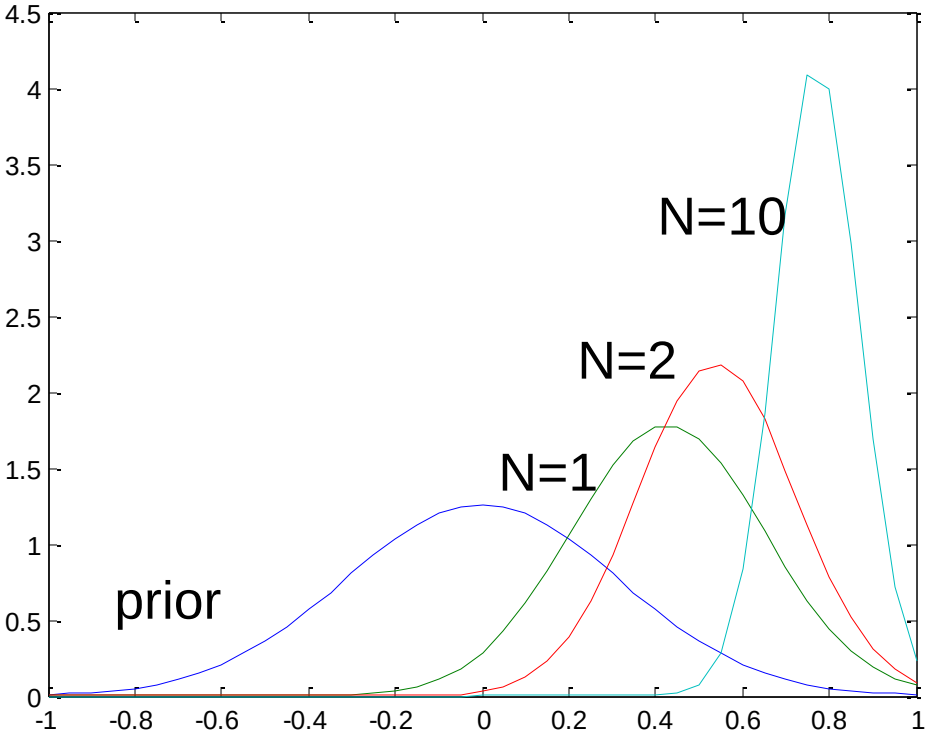
$$\mu_N = \sigma_N^2 \left(\frac{\sum_{i=1}^{N} x_i}{\sigma^2} + \frac{m_0}{\sigma_0^2} \right) \underset{N\to\infty}{\sim} \frac{\sum_{i=1}^{N} x_i}{N}$$

- As N→∞, the prior is washed out by the data and the posterior mean is the MLE estimate: $\mathbb{E}[\theta | x_1, \ldots, x_N] = \mu_N \simeq \theta_{ML}$

Bayesian Inference for the Gaussian

$$\mu \mid \boldsymbol{X} \sim \mathcal{N}(\mu_N, \sigma_N^2) \; with \; \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}, \; and \; \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0$$

MatLab implementation



$$\boldsymbol{X} = \{x_1, x_2, ..., x_N\} \sim \mathcal{N}(0.8, 0.1), \; with \; prior \; \mu \sim \mathcal{N}(0, 0.1).$$

Sequential Bayesian Inference

$$\mu \mid \boldsymbol{X} \sim \mathcal{N}(\mu_N, \sigma_N^2) \text{ with}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \Rightarrow \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}, \text{ and } \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0$$

- We can derive sequential estimates of the posterior variance and mean. They are as follows:

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}, \text{ and } \mu_N = \frac{\sigma_N^2}{\sigma_{N-1}^2} \mu_{N-1} + \frac{\sigma_N^2}{\sigma^2} x_N$$

- Show this is by recognizing the sequential nature of Bayesian inference (as we collect one data at a time, the posterior at the previous step becomes the new prior) and recalling:

$$\text{With prior } \theta \sim \mathcal{N}(\mu_0, \sigma_0^2) \text{ and observation } x_1 \Rightarrow \theta \mid x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2) \text{ with}$$

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2} \Rightarrow \sigma_1^2 = \frac{\sigma_0^2 \sigma^2}{\sigma_0^2 + \sigma^2}, \text{ and } \mu_1 = \sigma_1^2 \left(\frac{x_1}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right)$$

Example of Linear Gaussian Systems: Inferring the Mean

- Revisit the Bayesian inference problem for the Gaussian. Consider

$$\boldsymbol{y} = \{y_1, y_2, ..., y_N\} \sim \mathcal{N}(y \mid x, \sigma^2 = \lambda_y^{-1}), \text{ with prior } x \sim \mathcal{N}(x \mid \mu_0, \sigma_0^2 = \lambda_0^{-1}).$$

- Likelihood for the N –dataset in the form of a linear Gaussian system:

$$p(\boldsymbol{y} \mid x) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}x + \boldsymbol{b}, \boldsymbol{\Sigma}_y), \ \boldsymbol{A} = \mathbf{1}_N \ (\text{column vector of } 1's), \ \boldsymbol{b} = \mathbf{0}, \ \boldsymbol{\Sigma}_y^{-1} = diag(\lambda_y \boldsymbol{I})$$

- Applying conditional Gaussian results (to be reviewed later on):

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1})$$

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{x} \mid \left(\boldsymbol{\Lambda} + \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{A}\right)^{-1} \left(\boldsymbol{\Lambda}\boldsymbol{\mu} + \boldsymbol{A}^T \boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b})\right), \left(\boldsymbol{\Lambda} + \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{A}\right)^{-1}\right)$$

$$p(x \mid \boldsymbol{y}) = \mathcal{N}\left(x \mid \left(\lambda_0 + \mathbf{1}_N^T \lambda_y \boldsymbol{I} \mathbf{1}_N\right)^{-1} \left(\lambda_0 \mu_0 + \mathbf{1}_N^T \lambda_y \boldsymbol{I}(\boldsymbol{y} - \boldsymbol{0})\right), \left(\lambda_0 + \mathbf{1}_N^T \lambda_y \boldsymbol{I} \mathbf{1}_N\right)^{-1}\right)$$

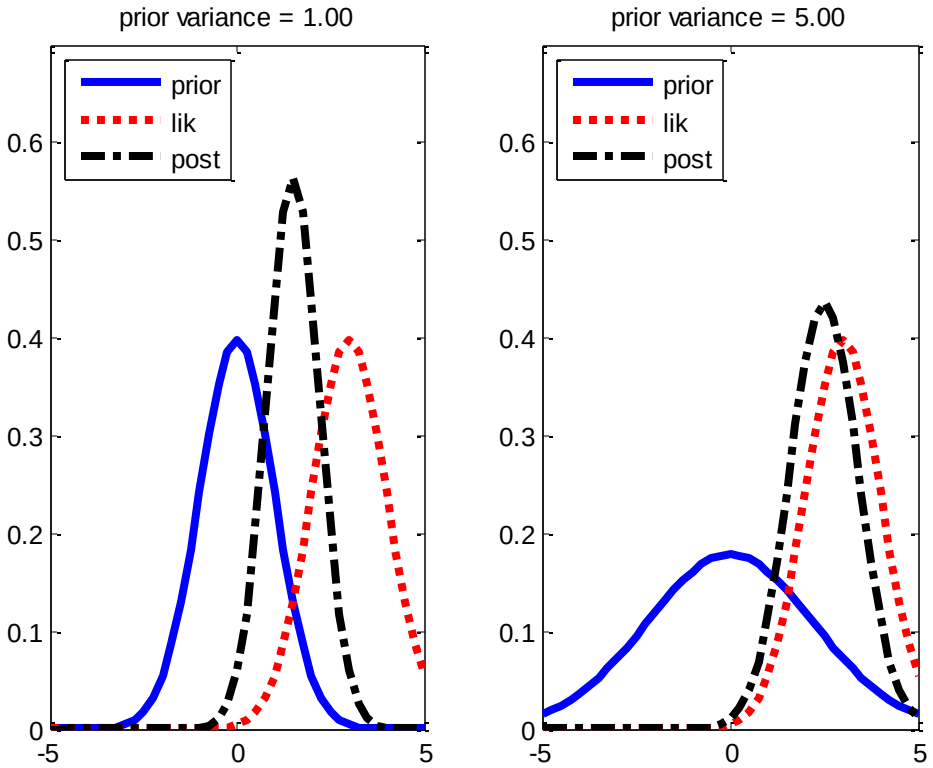
- This can be simplified as $p(x|y) = \mathcal{N}(x|\mu_N, \lambda_N^{-1})$ where:

$$p(x|\boldsymbol{y}) = \mathcal{N}\left(x \mid \frac{N\lambda_y}{\lambda_0 + N\lambda_y} \bar{y} + \frac{\lambda_0}{\lambda_0 + N\lambda_y} \mu_0, \left(\lambda_0 + N\lambda_y\right)^{-1}\right)$$

- The precision is the prior precision + N measurement precisions. The mean is the weighted average of the MLE and prior mean. These are identical results to those obtained earlier.

Inferring the Mean of a Gayssian

gaussInferParamsMean1d
from PMTK



- Inference about x given a single noisy observation $y = 3$.
  - (a) Strong prior $\mathcal{N}(0, 1)$. The posterior mean is "shrunk" towards the prior mean, which is 0.
  - (b) Weak prior $\mathcal{N}(0, 5)$. The posterior mean is similar to the MLE

Shrinkage and Signal-To-Noise Ratio

$$\sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}, \text{ and } \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0$$

- The posterior precision is the sum of the precision of the prior plus one contribution of the data precision for each observed data point. For $N \to \infty$ the posterior peaks around the μ_{ML} and the posterior variance goes to zero, i.e. MLE estimate is recovered within the Bayesian paradigm.

- If we apply the data sequentially, we can write for the posterior mean *after the collection of one data point* $(N = 1)$, *i.e.* $\mu_{ML} = y$ the following:

$$\mu_1 = y - (y - \mu_0)\frac{\sigma^2}{\sigma_0^2 + \sigma^2} \; (\textit{shrinkage of the data } y \textit{ towards the prior mean } \mu_0)$$

- Shrinkage is often measured also with the *signal-to-noise ratio*:

$$SNR = \frac{\mathbb{E}\left[X^2\right]}{\mathbb{E}\left[\varepsilon^2\right]} = \frac{\sigma_0^2 + \mu_0^2}{\sigma^2}, \; \textit{for } y = x + \varepsilon \; (\textit{observed signal}), \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

- How about when $\sigma_0^2 \to \infty$? In this case note that $\sigma_N^2 \to \frac{\sigma^2}{N}$ *and* $\mu_N \to \mu_{ML}$

Inference of Precision with Known Mean

- Consider $x_n \sim \mathcal{N}(x_n \mid \mu, \lambda^{-1}), n = 1, ..., N$. We want to infer the precision $\lambda = 1/\sigma^2$ with the mean μ taken as known.

- The likelihood takes the form:

$$p(\boldsymbol{X} \mid \lambda) = \prod_{n=1}^{N} f(x_n \mid \mu) \propto \lambda^{N/2} \exp\left(-\frac{1}{2} \lambda \sum_{n=1}^{N} (x_n - \mu)^2 \right)$$

- The corresponding "conjugate prior" (a prior that results in a posterior of the same family as the prior) should be proportional to the product of a power of λ and the exponential of a linear function of λ. This corresponds to the gamma distribution.

$$\mathcal{Gamma}(\lambda \mid a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}, \; x \in [0, \infty]$$

$$\mathbb{E}[\lambda] = \frac{a}{b}, \; var[\lambda] = \frac{a}{b^2}$$

- The gamma distribution has a finite integral if $a > 0$, and the distribution itself is finite if $a \geq 1$.

Inference of Precision with Known Mean

- The posterior takes the form:

$$p(\lambda \mid \boldsymbol{X}, \mu) = \prod_{n=1}^{N} f(x_n \mid \mu)\mathcal{Gamma}(\lambda \mid a_0, b_0) \propto \lambda^{N/2+a_0-1} \exp\left(-b_0\lambda - \frac{1}{2}\lambda\sum_{n=1}^{N}(x_n-\mu)^2\right)$$

- We can immediately see that the posterior is also a Gamma distribution:

$$p(\lambda \mid \boldsymbol{X}, \mu) = \mathcal{Gamma}(\lambda \mid a_N, b_N),\ a_N = N/2 + a_0,\ b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma_{ML}^2$$

- Here σ_{ML}^2 is the MLE of the variance.

Inference of Precision with Known Mean

- The effect of observing N data points is to increase the value of a by $N/2$ (i.e. ½ for each data point). Thus we interpret the parameter a_0 as $2a_0$ 'effective' prior observations.

$$p(\lambda \mid \boldsymbol{X}, \mu) = \mathcal{Gamma}(\lambda \mid a_N, b_N),\ a_N = N/2 + a_0,\ b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^2 = b_0 + \frac{N}{2}\sigma_{ML}^2$$

- Each measurement contributes to the parameter b a variance $\sigma_{ML}^2/2$. Since we have $2a_0$ effective prior measurements, each of them contributes to b an effective prior variance

$$b_0 = \frac{2a_0}{2}\sigma^2 \Rightarrow \sigma^2 = \frac{b_0}{a_0}$$

- The interpretation of a conjugate prior in terms of effective data points is typical for the exponential family of distributions.

- The results above are identical with inference directly of the variance σ^2 using as prior $\mathcal{InvGamma}(\sigma^2 \mid a_0, b_0)$ resulting in a posterior

$$\mathcal{InvGamma}(\sigma^2 \mid a_N, b_N)$$

Gamma and Inverse Gamma

Gamma	$\theta \sim \text{Gamma}(\alpha, \beta)$ $p(\theta) = \text{Gamma}(\theta \mid \alpha, \beta)$	shape $\alpha > 0$ inverse scale $\beta > 0$

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}, \quad \theta > 0$$

$$\text{E}(\theta) = \frac{\alpha}{\beta}$$
$$\text{var}(\theta) = \frac{\alpha}{\beta^2}$$
$$\text{mode}(\theta) = \frac{\alpha-1}{\beta}, \text{ for } \alpha \geq 1$$

Inverse-gamma	$\theta \sim \text{Inv-gamma}(\alpha, \beta)$ $p(\theta) = \text{Inv-gamma}(\theta \mid \alpha, \beta)$	shape $\alpha > 0$ scale $\beta > 0$

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\beta/\theta}, \quad \theta > 0$$

$$\text{E}(\theta) = \frac{\beta}{\alpha-1}, \text{ for } \alpha > 1$$
$$\text{var}(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \alpha > 2$$
$$\text{mode}(\theta) = \frac{\beta}{\alpha+1}$$

$$If \ \theta \sim \mathcal{Gamma}(\theta \mid a,b) \ then \ \theta^{-1} \sim \mathcal{InvGamma}(\theta^{-1} \mid a,b)$$

$$Here: \lambda \sim \mathcal{Gamma}(\lambda \mid a,b) \, then \, \sigma^2 \sim \mathcal{InvGamma}(\sigma^2 \mid a,b)$$

Bayesian Data Analysis, A. Gelman, J. Carlin, H. Stern and D. Rubin, 2004

Univariate Posterior – Inverse Chi Squared Prior

- Alternative prior for σ^2 is the Scaled *Inverse Chi-Squared Distribution**

$$\chi^{-2}\left(\sigma^2 \mid \nu_0, \sigma_0^2\right) \equiv \mathit{InvGamma}\left(\sigma^2 \mid \frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right) \propto \left(\sigma^2\right)^{-\nu_0/2-1} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right)$$

- Here ν_0 represents the strength of the prior and σ_0^2 encodes the value of the prior. With this, the posterior takes the form:

$$\chi^{-2}\left(\sigma^2 \mid \mathcal{D}, \mu\right) = \chi^{-2}\left(\sigma^2 \mid \nu_N, \sigma_N^2\right), \nu_N = \nu_0 + N, \sigma_N^2 = \frac{\nu_0\sigma_0^2 + \sum_{i=1}^{N}\left(x_i - \mu\right)^2}{\nu_N}$$

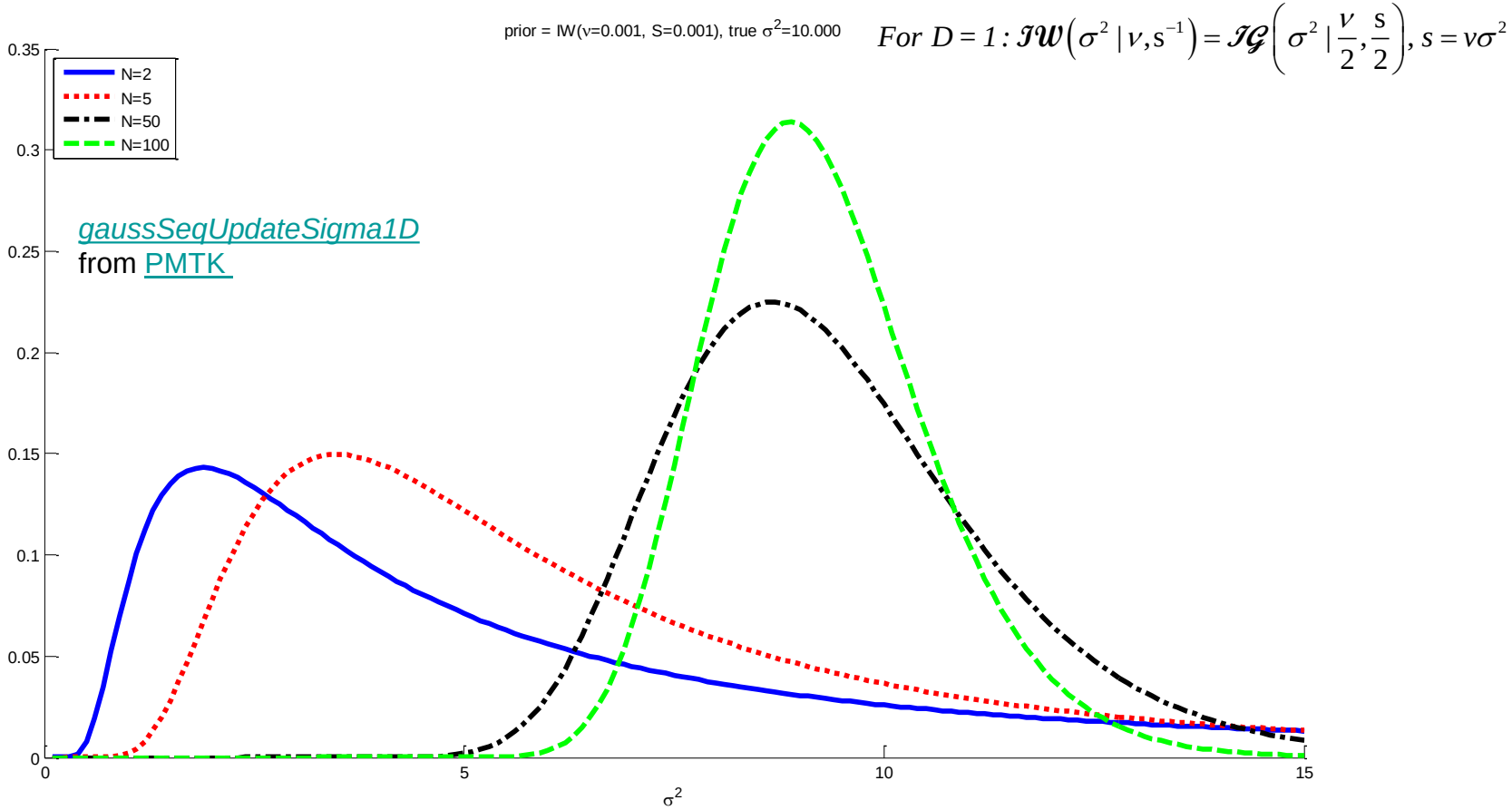
- The posterior dof ν_N is the prior dof plus N. The posterior sum of squares $\sigma_N^2\nu_N = \nu_0\sigma_0^2 + \sum_{i=1}^{N}\left(x_i - \mu\right)^2$ is the sum of the prior sum of squares plus the data sum of squares. An ***uninformative prior*** corresponds ***to zero virtual sample size, $\nu_0 = 0$***. This is the prior $p\left(\sigma^2\right) \propto \sigma^{-2}$

- This approach is certainly more appealing.

* Often denoted as $\mathit{Scale\text{-}Inv}\text{-}\chi^2\left(\sigma^2 \mid \nu_0, \sigma_0^2\right), mean = \frac{\nu_0\sigma_0^2}{\nu_0 - 2}$ *for* $\nu_0 > 2$

$$\text{var} = \frac{2\nu_0^2\sigma_0^4}{\left(\nu_0 - 2\right)^2\left(\nu_0 - 4\right)} \text{ for } \nu_0 > 4, mode = \frac{\nu_0\sigma_0^2}{\nu_0 + 2}$$

Sequential Update of the Posterior for σ^2



- Sequential update of the posterior for σ^2 starting from an uninformative prior $\mathcal{IW}\left(\sigma^2 \mid \nu_0 = 0.001, s_0 = \nu_0\sigma_0^2 = 0.001\right)$. The scalar inverse-Wishart here is the same as the inverse Gamma. The data were generated from $\mathcal{N}(5,10)$.
  - Gelman 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1(3):515–533

Bayesian Inference: Unknown Mean and Precision

- Consider $x_n \sim \mathcal{N}(x_n \mid \mu, \lambda^{-1}), n = 1, ..., N$. We want to infer both the precision $\lambda = 1/\sigma^2$ and the mean μ.

- The likelihood takes the form:

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

$$= \left[\lambda^{1/2} \exp\left(-\frac{\lambda \mu^2}{2} \right) \right]^N \exp\left(\lambda\mu \sum_{n=1}^{N} x_n - \frac{1}{2}\lambda \sum_{n=1}^{N} x_n^2 \right)$$

- We need a prior that has a similar functional form in terms of λ and μ.

$$p(\mu, \lambda) \propto \left[\lambda^{1/2} \exp\left(-\frac{\lambda \mu^2}{2} \right) \right]^{\beta} \exp\left(\lambda\mu c - \lambda d \right)$$

$$= \underbrace{\exp\left(-\frac{\beta\lambda}{2} \left(\mu - \frac{c}{\beta} \right)^2 \right)}_{p(\mu \mid \lambda)} \underbrace{\lambda^{\beta/2} \exp\left(-\left(d - \frac{c^2}{2\beta} \right) \lambda \right)}_{p(\lambda)}$$

Bayesian Inference: Unknown Mean and Precision

$$p(\mu,\lambda) \propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\beta}\exp\left(\lambda\mu c-\lambda d\right)$$

$$=\underbrace{(\beta\lambda)^{1/2}\exp\left(-\frac{\beta\lambda}{2}\left(\mu-\frac{c}{\beta}\right)^2\right)}_{p(\mu|\lambda)}\underbrace{\lambda^{(\beta-1)/2}\exp\left(-\left(d-\frac{c^2}{2\beta}\right)\lambda\right)}_{p(\lambda)}$$

- We can easily identify that the prior is of the form (Normal-Gamma):

$$p(\mu,\lambda)=\mathcal{N}\left(\mu\,|\,\mu_0=\frac{c}{b},(\beta\lambda)^{-1}\right)\mathcal{Gamma}\left(\lambda\,|\,a=\frac{1+\beta}{2},b=d-\frac{c^2}{2\beta}\right)$$

- Recall the form of the Gamma distribution:

$$\mathcal{Gamma}(\lambda\,|\,a,b)=\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}$$

Bayesian Inference: Unknown Mean and Precision

- Combining the likelihood and prior, we can re-arrange and write:

$$p(\mu,\lambda \mid \boldsymbol{X}) \propto \lambda^{N/2}\lambda^{a-1}\exp\left(-\left(b+\frac{1}{2}\sum_{n=1}^{N}x_n^2+\frac{\beta}{2}\mu_0^2\right)\lambda\right)\times$$

$$\left(\lambda(N+\beta)\right)^{1/2}\exp\left(-\frac{\lambda\left(N+\beta\right)}{2}\left(\mu^2-\frac{2}{N+\beta}\left(\beta\mu_0+\sum_{n=1}^{N}x_n\right)\mu\right)\right)$$

- Completing the square on the 2nd argument gives:

$$p(\mu,\lambda \mid \boldsymbol{X}) \propto \lambda^{N/2}\lambda^{a-1}\exp\left(-\underbrace{\left(b+\frac{1}{2}\sum_{n=1}^{N}x_n^2+\frac{\beta}{2}\mu_0^2-\underbrace{\frac{\left(\beta\mu_0+\sum_{n=1}^{N}x_n\right)^2}{2\left(N+\beta\right)}}_{\frac{(N+\beta)}{2}\mu_N^2}\right)}_{b_N}\lambda\right)$$

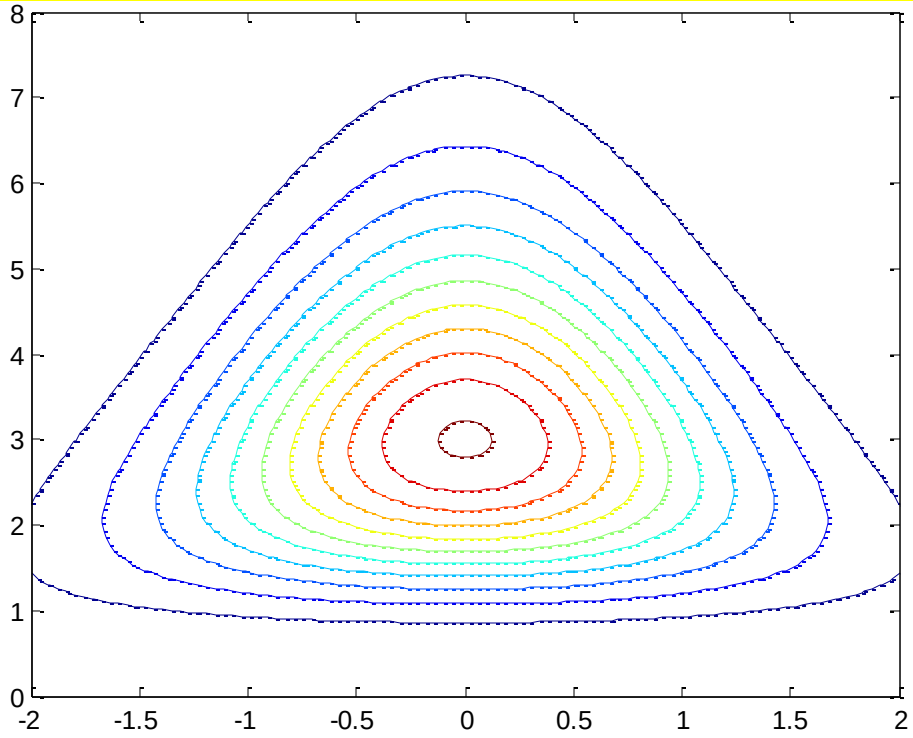
$$\left(\lambda(N+\beta)\right)^{1/2}\exp\left(-\frac{\lambda\left(N+\beta\right)}{2}\left(\mu-\frac{\beta\mu_0+\sum_{n=1}^{N}x_n}{N+\beta}\right)^2\right)$$

$$\Leftarrow \mathcal{G}amma\left(\lambda \mid a_N=\frac{N}{2}+a, b_N\right)$$

$$\Leftarrow \mathcal{N}\left(\mu \mid \mu_N=\frac{\beta\mu_0+\sum_{n=1}^{N}x_n}{N+\beta}, \left(\underbrace{\left(N+\beta\right)\lambda}_{\beta_N}\right)^{-1}\right)$$

The Normal-Gamma Distribution

$$p(\mu,\lambda)=\mathcal{N}\left(\mu\,|\,\mu_0=\frac{c}{b},(\beta\lambda)^{-1}\right)\mathcal{Gamma}\left(\lambda\,|\,a=1+\frac{\beta}{2},b=d-\frac{c^2}{2\beta}\right)$$



$a=5, b=6,$

$\mu_0=0, \beta=2$

MatLab Code

- K. Murphy, Conjugate Baeysian Analysis of the Gaussian Distribution, 2007 (provides additional results for posterior marginals, posterior predictive, and reference results for an uninformative prior)

Posterior for μ and σ for Scalar Data

- We can also work directly with σ^2. We use the normal inverse chi-squared distribution (NIX)

$$\mathcal{NI}\chi^2\left(\mu,\sigma^2 \mid m_0,\kappa_0,\nu_0,\sigma_0^2\right)=\mathcal{N}\left(\mu \mid m_0,\sigma^2/\kappa_0\right)\chi^{-2}\left(\sigma^2 \mid \nu_0,\sigma_0^2\right)$$

$$\propto\left(\frac{1}{\sigma^2}\right)^{(\nu_0+3)/2}\exp\left(-\frac{\nu_0\sigma_0^2+\kappa_0\left(\mu-m_0\right)^2}{2\sigma^2}\right)$$

$$\chi^{-2}\left(\sigma^2 \mid \nu_0,\sigma_0^2\right)=\mathcal{IG}\left(\sigma^2 \mid \frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2}\right)$$

$$\propto\left(\sigma^2\right)^{-\nu_0/2-1}\exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right)$$

- Similarly to our earlier calculations, the posterior is given as:

$$p(\mu,\sigma^2 \mid \mathcal{D})=\mathcal{NI}\chi^2\left(\mu,\sigma^2 \mid m_N,\kappa_N,\nu_N,\sigma_N^2\right), m_N=\frac{\kappa_0 m_0+N\bar{x}}{\kappa_N}=\frac{\kappa_0}{\kappa_0+N}m_0+\frac{N}{\kappa_0+N}\bar{x}$$

$$\kappa_N=\kappa_0+N,\nu_N=\nu_0+N,\nu_N\sigma_N^2=\nu_0\sigma_0^2+\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2+\frac{N\kappa_0}{\kappa_0+N}\left(m_0-\bar{x}\right)^2$$

- The posterior marginal for σ^2 and posterior expectation are:

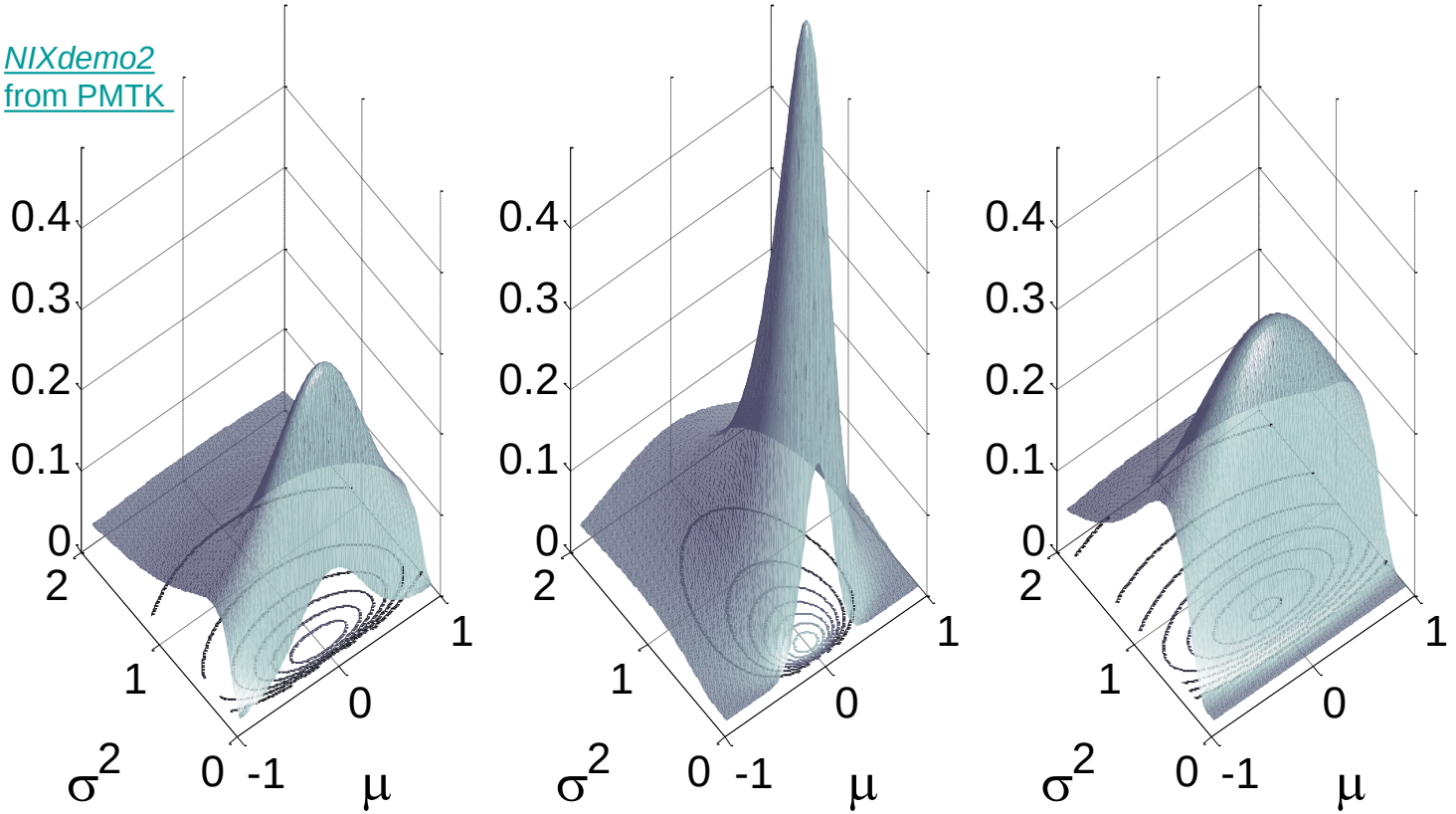
$$p(\sigma^2 \mid \mathcal{D})=\int p(\mu,\sigma^2 \mid \mathcal{D})d\mu=\chi^{-2}\left(\sigma^2 \mid \nu_N,\sigma_N^2\right),\mathbb{E}\left[\sigma^2 \mid \mathcal{D}\right]=\frac{\nu_N}{\nu_N-2}\sigma_N^2$$

  - K. Murphy, Conjugate Baeysian Analysis of the Gaussian Distribution, 2007 (provides additional results for posterior marginals, posterior predictive, and reference results for an uninformative prior, also Section 6 provides the analysis for a normal-inverse-Gamma prior)

Normal Inverse x^2 Distribution

NIX(μ_0=0, k_0=1, ν_0=1, σ_0^2=1) NIX(μ_0=0, k_0=5, ν_0=1, σ_0^2=1) NIX(μ_0=0, k_0=1, ν_0=5, σ_0^2=1)

NIXdemo2 from PMTK



- The $NI\chi^2(\mu_0, \kappa_0, \nu_0, \sigma_0^2)$ distribution. μ_0 is the prior mean and κ_0 is how strongly we believe this;σ_0^2 is the prior variance and ν_0 is how strongly we believe this. (a) The contour plot (underneath the surface) is shaped like a "squashed egg". (b) We increase the strength of our belief in the mean, so it gets narrower (c) We increase the strength of our belief in the variance, so it gets narrower.

Posterior for μ and σ for Scalar Data

- The posterior for μ is Students' $\mathcal{T}$:

$$p(x \mid \mu, \lambda, \upsilon) = \frac{\Gamma(\frac{\upsilon}{2}+\frac{1}{2})}{\Gamma(\frac{\upsilon}{2})}\left(\frac{\lambda}{\pi\upsilon}\right)^{1/2}\left[1+\frac{\lambda(x-\mu)^2}{\upsilon}\right]^{-\upsilon/2-1/2}$$

$$p(\mu \mid \mathcal{D}) = \int p(\mu, \sigma^2 \mid \mathcal{D}) d\sigma^2 = \mathcal{T}\left(\mu \mid m_N, \sigma_N^2/\kappa_N, \nu_N\right), \mathbb{E}[\mu \mid \mathcal{D}] = m_N$$

$$Mean: \mu, \upsilon > 1$$

$$Mode: \mu$$

$$Var: \frac{\upsilon\sigma^2}{\upsilon-2} = \frac{\upsilon}{\lambda(\upsilon-2)}, \upsilon > 2$$

$$\lambda = \sigma^{-2}$$

- Let us revisit these results with *an uninformative prior*:

$$p(\mu, \sigma^2) \propto p(\mu)p(\sigma^2) \propto \sigma^{-2} \propto \mathcal{NI}\chi^2\left(\mu, \sigma^2 \mid \mu_0 = 0, \kappa_0 = 0, \nu_0 = -1, \sigma_0^2 = 0\right)$$

- With this prior, the posterior becomes:

$$p(\mu, \sigma^2 \mid \mathcal{D}) = \mathcal{NI}\chi^2\left(\mu, \sigma^2 \mid m_N = \bar{x}, \kappa_N = N, \nu_N = N-1, \sigma_N^2 = s^2\right), s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 = \frac{N}{N-1}\bar{\sigma}_{MLE}^2$$

- s is the sample std. Thus the marginal posterior for μ becomes:

$$p(\mu \mid \mathcal{D}) = \mathcal{T}\left(\mu \mid \bar{x}, s^2/N, N-1\right), \text{var}[\mu \mid \mathcal{D}] = \frac{\nu_N}{\nu_N - 2}\sigma_N^2 = \frac{N-1}{N-3}\frac{s^2}{N} \to \frac{s^2}{N}$$

- The standard error of the mean is defined as $\sqrt{\text{var}[\mu \mid \mathcal{D}]} \approx \frac{s}{\sqrt{N}}$

- An approximate 95% *posterior credible interval* is thus: $I_{0.95}[\mu \mid \mathcal{D}] \approx \bar{x} \pm 2\frac{s}{\sqrt{N}}$

Multivariate Gaussian: Posterior of μ

- Consider a known variance $\boldsymbol{\Sigma}$ and a Gaussian prior $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$ with the posterior for the unknown mean $\boldsymbol{\mu}$ taking the form:

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

- This posterior is the exponential of a quadratic in $\boldsymbol{\mu}$:

$$-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}) =$$

$$-\frac{1}{2}\boldsymbol{\mu}^T \underbrace{\left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)}_{\boldsymbol{\Sigma}_N^{-1}} \boldsymbol{\mu} + \boldsymbol{\mu}^T \underbrace{\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}\boldsymbol{x}_n\right)}_{\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu}_N} + const$$

- So the variance and mean of the posterior $p(\boldsymbol{\mu} \mid \boldsymbol{X}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$ are:

$$\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1},$$

$$\boldsymbol{\mu}_N = \left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}\boldsymbol{x}_n\right) = \left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + N\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{ML}\right)$$

- For uninformative prior, $\boldsymbol{\Sigma}_0 = \infty \boldsymbol{I} \Rightarrow p(\boldsymbol{\mu} \mid \boldsymbol{X}, \boldsymbol{\Sigma}) \to \mathcal{N}\left(\boldsymbol{\mu} \mid \boldsymbol{\mu}_{ML}, \frac{1}{N}\boldsymbol{\Sigma}\right)$

Posterior Distribution of Precision Λ

- We now discuss how to compute $p(\boldsymbol{\Lambda}|\mathcal{D},\boldsymbol{\mu})$ for a D-dimensional Gaussian. The likelihood has the form

$$p(\mathcal{D}\,|\,\boldsymbol{\mu},\boldsymbol{\Lambda}) \propto |\,\boldsymbol{\Lambda}\,|^{N/2}\exp\left(-\frac{1}{2}Tr\left(\mathbf{S}_{\mu}\boldsymbol{\Lambda}\right)\right),\ \mathbf{S}_{\mu}=\sum_{n}^{N}\left(\boldsymbol{x}_n-\boldsymbol{\mu}\right)\left(\boldsymbol{x}_n-\boldsymbol{\mu}\right)^{T}$$

- The corresponding conjugate prior is known as *the Wishart distribution*

$$\mathcal{W}i\left(\boldsymbol{\Lambda}\,|\,\boldsymbol{W},v\right)=B\,|\,\boldsymbol{\Lambda}\,|^{(v-D-1)/2}\exp\left(-\frac{1}{2}Tr\left(\boldsymbol{W}^{-1}\boldsymbol{\Lambda}\right)\right),v>D-1\,(dof),\boldsymbol{W}\ sym\ pos.def.\,(D\times D)$$

$$B\left(\boldsymbol{W},v\right)=|\,\boldsymbol{W}\,|^{-v/2}\left(2^{vD/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{v+1-i}{2}\right)\right)^{-1}$$

$$\Gamma_D\left(\frac{v}{2}\right)=\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{v+1-i}{2}\right)(multivariate\,Gamma\ Function)$$

- The following slide shows the similarities of this distribution with the Gamma prior for λ used earlier for univariate Gaussian distributions.

$$For\ D=1,\mathcal{W}i\left(\lambda\,|\,v,\mathrm{s}^{-1}\right)=\mathcal{G}amma\left(\lambda\,|\,\frac{v}{2},\frac{\mathrm{s}}{2}\right)$$

Wishart Distribution

Wishart	$W \sim \text{Wishart}_\nu(S)$ $p(W) = \text{Wishart}_\nu(W\|S)$ (implicit dimension $k \times k$)	degrees of freedom ν symmetric, pos. definite $k \times k$ scale matrix S

$$p(W) = \left(2^{\nu k/2}\pi^{k(k-1)/4}\prod_{i=1}^{k}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1} \times |S|^{-\nu/2}|W|^{(\nu-k-1)/2} \times \exp\left(-\frac{1}{2}\text{tr}(S^{-1}W)\right), \; W \text{ pos. definite}$$

$$\text{E}(W) = \nu S$$

$$\text{mode} = (\nu - k - 1)S \quad \text{for } \nu \geq k+1$$

Gamma	$\theta \sim \text{Gamma}(\alpha, \beta)$ $p(\theta) = \text{Gamma}(\theta\|\alpha, \beta)$	shape $\alpha > 0$ inverse scale $\beta > 0$

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}, \quad \theta > 0$$

$$\text{E}(\theta) = \frac{\alpha}{\beta}$$

$$\text{var}(\theta) = \frac{\alpha}{\beta^2}$$

$$\text{mode}(\theta) = \frac{\alpha-1}{\beta}, \text{ for } \alpha \geq 1$$

$$\text{For } \mathbf{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \; \mathbf{S} = \sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T \text{ (scatter matrix)} \sim \mathcal{W}i(\mathbf{S}\,|\,\boldsymbol{\Sigma}, N) \Rightarrow \mathbb{E}[\mathbf{S}] = N\boldsymbol{\Sigma}$$

Bayesian Data Analysis, A. Gelman, J. Carlin, H. Stern and D. Rubin, 2004

Posterior Distribution of Σ

- We similarly discuss computing $p(\boldsymbol{\Sigma}|\mathcal{D},\boldsymbol{\mu})$. The likelihood has the form

$$p(\mathcal{D}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) \propto |\,\boldsymbol{\Sigma}\,|^{-N/2} \exp\left(-\frac{1}{2}Tr\left(\mathbf{S}_{\mu}\boldsymbol{\Sigma}^{-1}\right)\right),\ \mathbf{S}_{\mu} = \sum_{n}^{N}\left(\boldsymbol{x}_n-\boldsymbol{\mu}\right)\left(\boldsymbol{x}_n-\boldsymbol{\mu}\right)^{T}$$

- The corresponding conjugate prior is known as the *inverse Wishart*

$$\mathcal{InvWi}\left(\boldsymbol{\Sigma}\,|\,\boldsymbol{S}_0,\nu_0\right) \propto |\,\boldsymbol{\Sigma}\,|^{-(\nu_0+D+1)/2} \exp\left(-\frac{1}{2}Tr\left(\mathbf{S}_0\boldsymbol{\Sigma}^{-1}\right)\right), \mathbf{S}_0\ sym\ pos.\ def.$$

- $\nu_0 + D + 1$ controls the strength of the prior, and hence plays a role analogous to the sample size N.

- The prior scatter matrix is here $\boldsymbol{S}_0$.

$$If\ \boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{-1} \sim \mathcal{Wi}\left(\boldsymbol{\Lambda}\,|\,\boldsymbol{S},\nu\right) then\ \boldsymbol{\Sigma} \sim \mathcal{InvWi}\left(\boldsymbol{S}^{-1},\nu\right)$$

- Note: There are many parametrizations of the $\mathcal{InvWi}$. We here follow the notation from Gelman et al. with the same ν for both $\mathcal{Wi}$ and $\mathcal{InvW}$i in the Eq. above. In some literature (e.g. K. Murphy's book), the distribution is denoted as $\mathcal{InvWi}\left(\boldsymbol{\Sigma}\,|\,\boldsymbol{S}_0^{-1},\nu_0\right)$

  - Steven W. Nydick, The Wishart and Inverse Wishart Distributions, Report, 2012.
  - A. Gelman, J. Carlin, H. Stern and D. Rubin, Bayesian Data Analysis, 2004

Inverse Wishart Distribution

Inverse-Wishart

$$W \sim Inv-Wishart_{\nu}(S)$$

$$p(W) = Inv-Wishart_{\nu}(W \mid S)$$

degrees of freedom ν
symmetric, pos. definite $k \times k$ scale matrix S

$$p(W) = \left(2^{\nu k/2}\pi^{k(k-1)/4}\prod_{i=1}^{k}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1} \times |S|^{\nu/2}|W|^{-(\nu+k+1)/2} \times \exp\left(-\tfrac{1}{2}\mathrm{tr}(SW^{-1})\right),\ W \text{ pos. definite}$$

$$\mathrm{E}(W) = (\nu - k - 1)^{-1}S$$

$$\text{mode} = (\nu + k + 1)^{-1}S$$

Inverse-gamma

$$\theta \sim \text{Inv-gamma}(\alpha, \beta)$$

$$p(\theta) = \text{Inv-gamma}(\theta|\alpha, \beta)$$

shape $\alpha > 0$
scale $\beta > 0$

$$p(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{-(\alpha+1)}e^{-\beta/\theta},\quad \theta > 0$$

$$\mathrm{E}(\theta) = \frac{\beta}{\alpha-1},\ \text{for } \alpha > 1$$

$$\mathrm{var}(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)},\ \alpha > 2$$

$$\mathrm{mode}(\theta) = \frac{\beta}{\alpha+1}$$

$$For\ k=1, \mathcal{InvWi}\left(\sigma^2 \mid \nu, \mathrm{S}\right) = \mathcal{InvGamma}\left(\sigma^2 \mid \frac{\nu}{2}, \frac{\mathrm{S}}{2}\right)$$

$$If\ \lambda \sim \mathcal{Gamma}(a,b) \Rightarrow \frac{1}{\lambda} \sim \mathcal{InvGamma}(a,b)$$

$$If\ \boldsymbol{\Sigma}^{-1} \sim \mathcal{Wi}(\nu, \mathbf{S}) \Rightarrow \boldsymbol{\Sigma} \sim \mathcal{InvWi}\left(\nu, \mathbf{S}^{-1}\right)$$

Bayesian Data Analysis, A. Gelman, J. Carlin, H. Stern and D. Rubin, 2004

Posterior Distribution of Σ

- Multiplying the likelihood and prior, we find that the posterior is also inverse Wishart:

$$p(\Sigma \mid \mathcal{D}, \boldsymbol{\mu}) \propto |\Sigma|^{-N/2} \exp\left(-\frac{1}{2} Tr\left(\mathbf{S}_{\mu}\Sigma^{-1}\right)\right) |\Sigma|^{-(\nu_0+D+1)/2} \exp\left(-\frac{1}{2} Tr\left(\mathbf{S}_0\Sigma^{-1}\right)\right),$$

$$\propto |\Sigma|^{-[N+(\nu_0+D+1)]/2} \exp\left(-\frac{1}{2} Tr\left(\left(\mathbf{S}_{\mu}+\mathbf{S}_0\right)\Sigma^{-1}\right)\right)$$

$$p(\Sigma \mid \mathcal{D}, \boldsymbol{\mu}) = \mathcal{InvWi}\left(\Sigma \mid N+\nu_0, \boldsymbol{S}_N\right), \boldsymbol{S}_N = \mathbf{S}_{\mu}+\mathbf{S}_0$$

- *The posterior strength $\nu_N = \nu_0 + N$, is the prior strength ν_0 plus the number of observations N.*

- *The posterior scatter matrix $\boldsymbol{S}_N$ is the prior scatter matrix $\boldsymbol{S}_0$ plus the data scatter matrix $\boldsymbol{S}_{\mu}$.*

MAP Estimation

- From the mode of the inverse Wishart and

$$p\left(\boldsymbol{\Sigma} \mid \mathcal{D}, \boldsymbol{\mu}\right) = \mathcal{I}\mathcal{n}\mathcal{W}\mathcal{i}\left(\boldsymbol{\Sigma} \mid \nu_N, \boldsymbol{S}_N\right), \boldsymbol{S}_N = \mathbf{S}_{\boldsymbol{\mu}} + \mathbf{S}_0, \nu_N = N + \nu_0$$

we conclude that the MAP estimate is:

$$\bar{\boldsymbol{\Sigma}}_{MAP} = \frac{\boldsymbol{S}_N}{\nu_N + D + 1} = \frac{\mathbf{S}_{\boldsymbol{\mu}} + \mathbf{S}_0}{N + \underbrace{\nu_0 + D + 1}_{N_0}} = \frac{\mathbf{S}_{\boldsymbol{\mu}} + \mathbf{S}_0}{N + N_0}$$

- For an improper prior, $\boldsymbol{S}_0 = \mathbf{0}$ and $N_0 = 0$, $\bar{\boldsymbol{\Sigma}}_{MAP} \rightarrow \frac{\mathbf{S}_{\boldsymbol{\mu}}}{N} = \bar{\boldsymbol{\Sigma}}_{MLE}$

- Consider now the use of a proper informative prior, which is necessary whenever D/N is large. Let $\boldsymbol{\mu} = \bar{\boldsymbol{x}} \Rightarrow \mathbf{S}_{\boldsymbol{\mu}} = \mathbf{S}_{\bar{\boldsymbol{x}}}$ .Rewrite the *MAP estimate as a convex combination of the prior mode and MLE*

$$\bar{\boldsymbol{\Sigma}}_{MAP} = \frac{\mathbf{S}_{\bar{\boldsymbol{x}}} + \mathbf{S}_0}{N + N_0} = \underbrace{\frac{N_0}{N + N_0}}_{\lambda} \frac{\mathbf{S}_0}{N_0} + \underbrace{\frac{N}{N + N_0}}_{1-\lambda} \frac{\mathbf{S}_{\bar{\boldsymbol{x}}}}{N} = \lambda \boldsymbol{\Sigma}_0 + (1-\lambda)\bar{\boldsymbol{\Sigma}}_{MLE}, \; \boldsymbol{\Sigma}_0 \equiv \frac{\mathbf{S}_0}{N_0} \; (prior\ mode)$$

where λ controls the amount of shrinkage towards the prior.

MAP Estimation

$$\bar{\Sigma}_{MAP} = \lambda \Sigma_0 + (1-\lambda)\bar{\Sigma}_{MLE}, \; \Sigma_0 \equiv \frac{\mathbf{S}_0}{N_0} \; (prior \; mode)$$

- *Can set λ by cross validation.*

Alternatively, we can use the formula provided in Ledoit & Wolf and Schaefer & Strimmer which is the optimal frequentist estimate (for squared loss).

This loss function for covariance matrices ignores the positive definite constraint but results in a simple estimator (see PMTK function shrinkcov).

- For the prior covariance matrix, $\boldsymbol{S}_0$, it is common to use the following *(data dependent) prior:* $\Sigma_0 = diag\left(\bar{\Sigma}_{MLE}\right)$

- Ledoit, O. and M. Wolf (2004b). A well conditioned estimator for large dimensional covariance matrices. *J. of Multivariate Analysis 88*(2), 365– 411.
- Ledoit, O. and M. Wolf (2004a). Honey, I Shrunk the Sample Covariance Matrix. *J. of Portfolio Management 31*(1).
- Schaefer, J. and K. Strimmer (2005). A shrinkage approach to largescale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol 4*(32).

MAP Shrinkage Estimation

$$\Sigma_0 = diag\left(\bar{\Sigma}_{MLE}\right)$$

- In this case, the MAP estimate is:

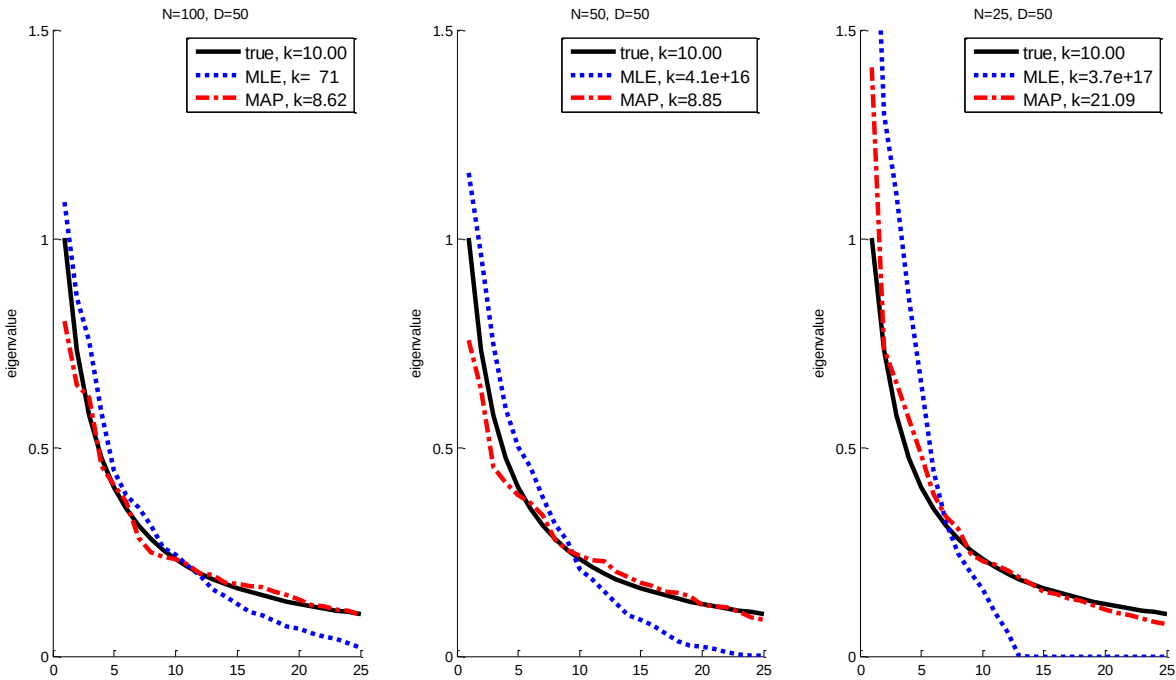
$$\bar{\Sigma}_{MAP} = \begin{cases} \bar{\Sigma}_{MLE}(i,j)\ if\ \ i=j \\ (1-\lambda)\bar{\Sigma}_{MLE}(i,j)\ otherwise \end{cases}$$

- Thus we see that the diagonal entries are equal to their MLE estimates, and *the off diagonal elements are "shrunk" somewhat towards 0* (*shrinkage estimation, or regularized estimation).*

- The benefits of MAP estimation are illustrated next. We consider fitting a $50-\dim$ Gaussian to $N = 100, N = 50$ and $N = 25$ data points.
  - The MAP estimate is always well-conditioned, unlike the MLE.
  - The *eigenvalue spectrum* of the MAP estimate is much closer to that of the true matrix than the MLE's.
  - The eigenvectors, however, are unaffected.

Posterior Distribution of Σ

- Estimating a covariance matrix in $D = 50$ dimensions using $N \in \{100, 50, 25\}$ samples.
- Eigenvalues in descending order for the true covariance matrix (solid black), MLE (dotted blue) and MAP estimates (dashed red) with $\lambda = 0.9$.
- The condition number k of each matrix is also given in the legend.



shrinkcovDemo from PMTK

Inference for Both μ and Λ

- Suppose $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_N \sim$(i.i.d) $\mathcal{N}(\boldsymbol{\mu}, \Lambda^{-1})$. We do not know $\boldsymbol{\mu}$ or $\boldsymbol{\Lambda}$
- When both sets of parameters are unknown, a conjugate family of priors is one in which

$$\boldsymbol{\Lambda} \sim \mathcal{W}(\boldsymbol{\Lambda} \,|\nu, \boldsymbol{T})$$

and

$$\boldsymbol{\mu} \,|\, \boldsymbol{\Lambda} \sim \mathcal{N}(\boldsymbol{\mu_0}, (\kappa \boldsymbol{\Lambda})^{-1})$$

- The Wishart distribution is the multivariate analog of the Gamma *distribution* (*extension to positive definite matrices*). If matrix $\boldsymbol{U}$ has the Wishart distribution, then $\boldsymbol{U}^{-1}$ has the inverse-Wishart distribution. *The resulting* $p(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{\mu_0}, \kappa, \boldsymbol{T}, \nu)$ is the ***Gaussian-Wishart distribution***.
- The quantity ν *is a positive scalar, while* $\boldsymbol{T}$ *is a* positive definite matrix. They play roles analogous to a and β, respectively, in the Gamma *distribution*.
- Other parameters of the prior are the mean vector $\boldsymbol{\mu_0}$ and κ which represents the `a priori number of observations'.

Inference for Both μ and Λ

- The likelihood and prior distributions are given explicitly as:

$$p\left(\mathcal{D}\mid\boldsymbol{\mu},\boldsymbol{\Sigma}\right)=\left(2\pi\right)^{-ND/2}\mid\boldsymbol{\Lambda}\mid^{N/2}\exp\left(-\frac{1}{2}\sum_{n=1}^{N}\left(\boldsymbol{x}_n-\boldsymbol{\mu}\right)^T\boldsymbol{\Lambda}\left(\boldsymbol{x}_n-\boldsymbol{\mu}\right)\right)$$

$$=\left(2\pi\right)^{-ND/2}\mid\boldsymbol{\Lambda}\mid^{N/2}\exp\left(-\frac{N}{2}\left(\boldsymbol{\mu}-\bar{\boldsymbol{x}}\right)^T\boldsymbol{\Lambda}\left(\boldsymbol{\mu}-\bar{\boldsymbol{x}}\right)\right)\exp\left(-\frac{1}{2}Tr\left(\boldsymbol{\Lambda}\boldsymbol{S}_{\bar{x}}\right)\right)$$

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda})=\mathcal{NWi}\left(\boldsymbol{\mu},\boldsymbol{\Lambda}\mid\boldsymbol{\mu}_0,\kappa_0,\nu_0,\boldsymbol{T}_0\right)=\mathcal{N}\left(\boldsymbol{\mu}\mid\boldsymbol{\mu}_0,\left(\kappa_0\boldsymbol{\Lambda}\right)^{-1}\right)\mathcal{Wi}\left(\boldsymbol{\Lambda}\mid\boldsymbol{T}_0,\nu_0\right)=$$

$$=\frac{1}{Z}\mid\boldsymbol{\Lambda}\mid^{1/2}exp\left(-\frac{\kappa_0}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)^T\boldsymbol{\Lambda}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)\right)\mid\boldsymbol{\Lambda}\mid^{(\nu_0-D-1)/2}\exp\left(-\frac{1}{2}Tr\left(\boldsymbol{T}_0^{-1}\boldsymbol{\Lambda}\right)\right)$$

$$Z_{NIW}=\left(\frac{\kappa_0}{2\pi}\right)^{D/2}\left|\boldsymbol{T}_0\right|^{\nu_0/2}2^{D\nu_0/2}\Gamma_D\left(\frac{\nu_0}{2}\right),\ \Gamma_D\ \text{multivariate Gamma function}$$

- Combining gives:

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}\mid\mathcal{D})\propto\exp\left(-\frac{N}{2}\left(\boldsymbol{\mu}-\bar{\boldsymbol{x}}\right)^T\boldsymbol{\Lambda}\left(\boldsymbol{\mu}-\bar{\boldsymbol{x}}\right)-\frac{\kappa_0}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)^T\boldsymbol{\Lambda}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)\right)\times$$

$$\mid\boldsymbol{\Lambda}\mid^{(N+\nu_0-D)/2}\exp\left(-\frac{1}{2}Tr\left(\left(\boldsymbol{T}_0^{-1}+\boldsymbol{S}_{\bar{x}}\right)\boldsymbol{\Lambda}\right)\right)$$

- M. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, 1970.
- K. Murphy, Conjugate Baeysian Analysis of the Gaussian Distribution, 2007 (Section 8)

Inference for Both μ and Λ

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}\,|\,\mathcal{D}) \propto \exp\left(-\frac{N}{2}\left(\boldsymbol{\mu}-\bar{\boldsymbol{x}}\right)^T\boldsymbol{\Lambda}\left(\boldsymbol{\mu}-\bar{\boldsymbol{x}}\right)-\frac{\kappa_0}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)^T\boldsymbol{\Lambda}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)\right)\times$$

$$|\boldsymbol{\Lambda}|^{(N+\nu_0-D)/2}\exp\left(-\frac{1}{2}Tr\left(\left(\boldsymbol{T}_0^{-1}+\boldsymbol{S}_{\bar{x}}\right)\boldsymbol{\Lambda}\right)\right)$$

- Can close the square in $\boldsymbol{\mu}$ as follows:

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}\,|\,\mathcal{D}) \propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\mu}-\frac{\kappa_0\boldsymbol{\mu}_0+N\bar{\boldsymbol{x}}}{\kappa_0+N}\right)^T\left(\kappa_0+N\right)\boldsymbol{\Lambda}\left(\boldsymbol{\mu}-\frac{\kappa_0\boldsymbol{\mu}_0+N\bar{\boldsymbol{x}}}{\kappa_0+N}\right)\right)\times$$

$$|\boldsymbol{\Lambda}|^{(\nu_0+N-D-1)/2}\exp\left(-\frac{1}{2}Tr\left(\left(\boldsymbol{T}_0^{-1}+\boldsymbol{S}_{\bar{x}}+N\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T+\kappa_0\boldsymbol{\mu}_0\boldsymbol{\mu}_0^T-\frac{1}{\kappa_0+N}\left(\kappa_0\boldsymbol{\mu}_0+N\bar{\boldsymbol{x}}\right)\left(\kappa_0\boldsymbol{\mu}_0+N\bar{\boldsymbol{x}}\right)^T\right)\boldsymbol{\Lambda}\right)\right)$$

- We can simplify as:

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}\,|\,\mathcal{D}) \propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\mu}-\frac{\kappa_0\boldsymbol{\mu}_0+N\bar{\boldsymbol{x}}}{\kappa_0+N}\right)^T\left(\kappa_0+N\right)\boldsymbol{\Lambda}\left(\boldsymbol{\mu}-\frac{\kappa_0\boldsymbol{\mu}_0+N\bar{\boldsymbol{x}}}{\kappa_0+N}\right)\right)\times$$

$$|\boldsymbol{\Lambda}|^{(\nu_0+N-D-1)/2}\exp\left(-\frac{1}{2}Tr\left(\left(\boldsymbol{T}_0^{-1}+\boldsymbol{S}_{\bar{x}}+\frac{\kappa_0 N}{\kappa_0+N}\left(\boldsymbol{\mu}_0-\bar{\boldsymbol{x}}\right)\left(\boldsymbol{\mu}_0-\bar{\boldsymbol{x}}\right)^T\right)\boldsymbol{\Lambda}\right)\right)$$

Inference for Both μ and Λ

$$p(\mu,\Lambda\,|\,\mathcal{D}) \propto \exp\left(-\frac{1}{2}\left(\mu-\frac{\kappa_0\mu_0+N\bar{\mathbf{x}}}{\kappa_0+N}\right)^T(\kappa_0+N)\Lambda\left(\mu-\frac{\kappa_0\mu_0+N\bar{\mathbf{x}}}{\kappa_0+N}\right)\right)\times$$

$$|\Lambda|^{(\nu_0+N-D-1)/2}\exp\left(-\frac{1}{2}Tr\left(\left(\mathbf{T}_0^{-1}+\mathbf{S}_{\bar{x}}+\frac{\kappa_0 N}{\kappa_0+N}\left(\mu_0-\bar{\mathbf{x}}\right)\left(\mu_0-\bar{\mathbf{x}}\right)^T\right)\Lambda\right)\right)$$

- This can be written as:

$$p(\mu,\Lambda\,|\,\mathcal{D}) = \mathcal{NWi}\left(\mu,\Lambda\,|\,\mu_N,\kappa_N,\nu_N,\mathbf{T}_N\right) = \mathcal{N}\left(\mu\,|\,\mu_N,\left(\kappa_N\Lambda\right)^{-1}\right)\mathcal{Wi}\left(\Lambda\,|\,\mathbf{T}_N,\nu_N\right)$$

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

$$\mathbf{T}_N^{-1} = \mathbf{T}_0^{-1}+\mathbf{S}_{\bar{x}}+\frac{\kappa_0 N}{\kappa_0+N}\left(\mu_0-\bar{\mathbf{x}}\right)\left(\mu_0-\bar{\mathbf{x}}\right)^T,\ \text{where}\ \mathbf{S}_{\bar{x}} = \sum_{i=1}^{N}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)^T$$

$$\nu_N = \nu_0+N,\ \kappa_N = \kappa_0+N$$

Normalization Factors

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}\,|\,\mathcal{D}) \propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\mu}-\frac{\kappa_0\boldsymbol{\mu}_0+N\bar{\mathbf{x}}}{\kappa_0+N}\right)^T(\kappa_0+N)\boldsymbol{\Lambda}\left(\boldsymbol{\mu}-\frac{\kappa_0\boldsymbol{\mu}_0+N\bar{\mathbf{x}}}{\kappa_0+N}\right)\right)\times$$

$$|\boldsymbol{\Lambda}|^{(\nu_0+N-D-1)/2}\exp\left(-\frac{1}{2}Tr\left(\left(\mathbf{T}_0^{-1}+\mathbf{S}_{\bar{\mathbf{x}}}+\frac{\kappa_0 N}{\kappa_0+N}\left(\boldsymbol{\mu}_0-\bar{\mathbf{x}}\right)\left(\boldsymbol{\mu}_0-\bar{\mathbf{x}}\right)^T\right)\boldsymbol{\Lambda}\right)\right)$$

- Using the normalization constants of the multivariate Gaussian and the Wishart:

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}\,|\,\mathcal{D}) = \frac{1}{Z_N}\exp\left(-\frac{1}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right)^T\kappa_N\boldsymbol{\Lambda}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right)\right)|\boldsymbol{\Lambda}|^{(\nu_N-D-1)/2}\exp\left(-\frac{1}{2}Tr\left(\mathbf{T}_N^{-1}\boldsymbol{\Lambda}\right)\right)$$

$$Z_N = (2\pi)^{D/2}(\kappa_N)^{-D/2}\,2^{\nu_N D/2}\Gamma_D\left(\frac{\nu_N}{2}\right)|\mathbf{T}_N|^{\nu_N/2}$$

- Similarly for the prior the normalization factor is:

$$Z_0 = (2\pi)^{D/2}(\kappa_0)^{-D/2}\,2^{\nu_0 D/2}\Gamma_D\left(\frac{\nu_0}{2}\right)|\mathbf{T}_0|^{\nu_0/2}$$

Inference for Both μ and Λ

$$p(\mu,\Lambda\mid\mathcal{D})=\mathcal{NWi}\left(\mu,\Lambda\mid\mu_N,\kappa_N,\nu_N,\boldsymbol{T}_N\right)=\mathcal{N}\left(\mu\mid\mu_N,\left(\kappa_N\Lambda\right)^{-1}\right)\mathcal{Wi}\left(\Lambda\mid\boldsymbol{T}_N,\nu_N\right)$$

$$\mu_N=\frac{\kappa_0\mu_0+N\bar{\boldsymbol{x}}}{\kappa_0+N},\ \boldsymbol{T}_N^{-1}=\boldsymbol{T}_0^{-1}+\boldsymbol{S}_{\bar{x}}+\frac{\kappa_0 N}{\kappa_0+N}\left(\mu_0-\bar{\boldsymbol{x}}\right)\left(\mu_0-\bar{\boldsymbol{x}}\right)^T,\ \text{where } \boldsymbol{S}_{\bar{x}}=\sum_{i=1}^{N}\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}\right)^T$$

$$\nu_N=\nu_0+N,\ \kappa_N=\kappa_0+N$$

- The posterior marginals can be derived as:

$$p(\Lambda\mid\mathcal{D})=\mathcal{Wi}\left(\Lambda\mid\boldsymbol{T}_N,\nu_N\right)$$

$$p(\mu\mid\mathcal{D})=\mathcal{T}_{\nu_N-D+1}\left(\mu\mid\mu_N,\underbrace{\kappa_N\nu_N\boldsymbol{T}_N}_{\text{precision}}\right)$$

$$\mathcal{T}(\boldsymbol{x}\mid\mu,\Lambda,\upsilon)=\frac{\Gamma(\frac{D}{2}+\frac{(\nu_N-D+1)}{2})}{\Gamma(\frac{(\nu_N-D+1)}{2})}\frac{|\kappa_N\nu_N\boldsymbol{T}_N|^{1/2}}{\left(\pi(\nu_N-D+1)\right)^{D/2}}\left[1+\frac{\left(\boldsymbol{x}-\mu_N\right)^T\kappa_N\nu_N\boldsymbol{T}_N\left(\boldsymbol{x}-\mu_N\right)}{(\nu_N-D+1)}\right]^{-(\nu_N-D+1)/2-D/2}$$

* Refer to this report for these results based on Bernardo and Smith.

Inference for Both μ and Λ

$$p(\mu,\Lambda\mid\mathcal{D})=\mathcal{NWi}\left(\mu,\Lambda\mid\mu_N,\kappa_N,\nu_N,\boldsymbol{T}_N\right)=\mathcal{N}\left(\mu\mid\mu_N,\left(\kappa_N\Lambda\right)^{-1}\right)\mathcal{Wi}\left(\Lambda\mid\boldsymbol{T}_N,\nu_N\right)$$

$$\mu_N=\frac{\kappa_0\mu_0+N\bar{\boldsymbol{x}}}{\kappa_0+N},\ \boldsymbol{T}_N^{-1}=\boldsymbol{T}_0^{-1}+\boldsymbol{S}_{\bar{x}}+\frac{\kappa_0 N}{\kappa_0+N}\left(\mu_0-\bar{\boldsymbol{x}}\right)\left(\mu_0-\bar{\boldsymbol{x}}\right)^T,\ \text{where } \boldsymbol{S}_{\bar{x}}=\sum_{i=1}^{N}\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}\right)^T$$

$$\nu_N=\nu_0+N,\ \kappa_N=\kappa_0+N$$

- Differentiating the Eq. on the top of this slide, we can also derive the MAP estimates as:

$$\left(\bar{\mu},\bar{\Lambda}\right)=\arg\max_{\mu,\Lambda}\ p(\mu,\Lambda\mid\mathcal{D})$$

$$\bar{\mu}=\frac{\sum_{i=1}^{N}\boldsymbol{x}_i+\kappa_0\mu_0}{N+\kappa_0}$$

$$\bar{\Lambda}=\frac{\sum_{i=1}^{N}\left(\boldsymbol{x}_i-\bar{\mu}\right)\left(\boldsymbol{x}_i-\bar{\mu}\right)^T+\kappa_0\left(\bar{\mu}-\mu_0\right)\left(\bar{\mu}-\mu_0\right)^T+\boldsymbol{T}_0^{-1}}{N+\nu_0-D}$$

- These are reduced to the MLE by setting

$$\kappa_0=0,\ \nu_0=D,\ \left|\boldsymbol{T}_0^{-1}\right|=0$$

Inference for both μ and Λ

- The posterior predictive is: $p(\boldsymbol{x}\,|\,\mathcal{D}) = \mathcal{T}_{\nu_N - D+1}\left(\boldsymbol{\mu}_N, \underbrace{\frac{\kappa_N(\nu_N - D+1)\boldsymbol{T}_N}{(\kappa_N+1)}}_{precision}\right)$
- The marginal likelihood is computed as a ratio of normalization constants using $p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathcal{D}) = p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{\mu}, \boldsymbol{\Lambda})/p(\mathcal{D})$:

$$p(\mathcal{D}) = \frac{Z_N}{Z_0}\frac{1}{(2\pi)^{ND/2}} = \frac{1}{\pi^{ND/2}}\frac{\Gamma_D(\nu_N/2)}{\Gamma_D(\nu_0/2)}\frac{|\boldsymbol{T}_N|^{\nu_N/2}}{|\boldsymbol{T}_0|^{\nu/2}}\left(\frac{\kappa_0}{\kappa_N}\right)^{D/2}$$

- A useful reference analysis considers

$$\boldsymbol{\mu}_0 = 0, \kappa_0 = 0, \nu_0 = -1, \left|\boldsymbol{T}_0^{-1}\right| = 0$$

- This results in the following for the prior:

$$p(\boldsymbol{\mu}, \Lambda) \propto |\Lambda|^{-(D+1)/2}$$

- The posterior parameters are also simplified as:

$$\boldsymbol{\mu}_N = \bar{\boldsymbol{x}}, \boldsymbol{T}_N^{-1} = \boldsymbol{S}_{\bar{\boldsymbol{x}}}, \kappa_N = N, \nu_N = N-1$$

- The posterior marginals and posterior predictive are given as:

$$p(\Lambda\,|\,\mathcal{D}) = \mathcal{W}i_{N-D}\left(\Lambda\,|\,\boldsymbol{S}_{\bar{\boldsymbol{x}}}^{-1}\right),\ p(\boldsymbol{\mu}\,|\,\mathcal{D}) = \mathcal{T}_{N-D}\left(\boldsymbol{\mu}\,|\,\bar{\boldsymbol{x}}, \frac{\boldsymbol{S}_{\bar{\boldsymbol{x}}}^{-1}}{N(N-D)}\right)$$

$$p(\boldsymbol{x}\,|\,\mathcal{D}) = \mathcal{T}_{N-D}\left(\bar{\boldsymbol{x}}, \frac{\boldsymbol{S}_{\bar{\boldsymbol{x}}}^{-1}(N+1)}{N(N-D)}\right)$$

Inference for μ and Σ

- For the case of the multivariate Gaussian of a D-dim variable $\boldsymbol{x}, \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with both the mean and variance unknowns, the likelihood is

$$\prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \left(2\pi\right)^{-ND/2} |\boldsymbol{\Sigma}|^{-N/2} \exp\left(-\frac{1}{2}\sum_{n=1}^{N}\left(\boldsymbol{x}_n - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}_n - \boldsymbol{\mu}\right)\right)$$

$$= \left(2\pi\right)^{-ND/2} |\boldsymbol{\Sigma}|^{-N/2} \exp\left(-\frac{N}{2}\left(\boldsymbol{\mu} - \bar{\boldsymbol{x}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu} - \bar{\boldsymbol{x}}\right)\right) \exp\left(-\frac{1}{2} Tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_{\bar{x}}\right)\right)$$

- The conjugate prior is given as *the product of a Gaussian and the Inverse Wishart distribution*:

$$\mathcal{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}_0, \kappa_0, \boldsymbol{S}_0, \nu_0) = \mathcal{N}\left(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \frac{1}{\kappa_0}\boldsymbol{\Sigma}\right) \mathcal{IW}is\left(\boldsymbol{\Sigma} \mid \boldsymbol{S}_0, \nu_0\right) =$$

$$= \frac{1}{Z_{NIW}} |\boldsymbol{\Sigma}|^{-1/2} exp\left(-\frac{\kappa_0}{2}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)\right) |\boldsymbol{\Sigma}|^{-(\nu_0 + D + 1)/2} \exp\left(-\frac{1}{2} Tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_0\right)\right)$$

$$= \frac{1}{Z_{NIW}} exp\left(-\frac{\kappa_0}{2}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right) - \frac{1}{2} Tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_0\right)\right) |\boldsymbol{\Sigma}|^{-(\nu_0 + D + 2)/2}$$

$$Z_{NIW} = 2^{\nu_0 D/2} \Gamma_D\left(\frac{\nu_0}{2}\right)\left(\frac{2\pi}{\kappa_0}\right)^{D/2} |\boldsymbol{S}_0|^{-\nu_0/2}, \Gamma_D \text{ multivariate Gamma function}$$

Inference for μ and Σ

$$p(\mu,\Sigma\mid\mathcal{D})\propto|\Sigma|^{-\frac{N}{2}-\frac{1}{2}-\frac{\nu_0+D+1}{2}}\exp\left(-\frac{1}{2}\Sigma^{-1}\left(\boldsymbol{S}_{\bar{x}}+\boldsymbol{S}_0+N(\mu-\bar{x})(\mu-\bar{x})^T+\kappa_0(\mu-\mu_0)(\mu-\mu_0)^T\right)\right)$$

$$\propto|\Sigma|^{-\frac{\nu_0+D+2+N}{2}}\exp\left(-\frac{1}{2}\Sigma^{-1}\left(\boldsymbol{S}_{\bar{x}}+\boldsymbol{S}_0+N(\mu-\bar{x})(\mu-\bar{x})^T+\kappa_0(\mu-\mu_0)(\mu-\mu_0)^T\right)\right)$$

- One can show by completing the square in μ that:

$$N(\mu-\bar{x})(\mu-\bar{x})^T+\kappa_0(\mu-\mu_0)(\mu-\mu_0)^T=$$

$$\underbrace{(\kappa_0+N)}_{\kappa_N}\left(\mu-\underbrace{\frac{\kappa_0\mu_0+N\bar{x}}{\kappa_0+N}}_{\mu_N}\right)\left(\mu-\frac{\kappa_0\mu_0+N\bar{x}}{\kappa_0+N}\right)^T+\frac{\kappa_0 N}{\kappa_0+N}(\bar{x}-\mu_0)(\bar{x}-\mu_0)^T$$

- Thus:

$$p(\mu,\Sigma\mid\mathcal{D})\propto|\Sigma|^{-\frac{\nu_N+D+2}{2}}\exp\left(-\frac{1}{2}\Sigma^{-1}\left(\boldsymbol{S}_{\bar{x}}+\boldsymbol{S}_0+\kappa_N(\mu-\mu_N)(\mu-\mu_N)^T+\frac{\kappa_0 N}{\kappa_N}(\bar{x}-\mu_0)(\bar{x}-\mu_0)^T\right)\right),$$

$$where:\nu_N=\nu_0+N$$

The Posterior of μ and Σ

- The posterior is $\mathcal{NIW}$ given as:

$$p(\mu, \Sigma \mid \mathcal{D}, \mu_0, \kappa_0, \boldsymbol{S}_0, \nu_0) = \mathcal{NIW}(\mu, \Sigma \mid \mu_N, \kappa_N, \boldsymbol{S}_N, \nu_N)$$

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

$$\kappa_N = \kappa_0 + N, \nu_N = \nu_0 + N$$

$$\boldsymbol{S}_N = \boldsymbol{S}_0 + \boldsymbol{S}_{\bar{\boldsymbol{x}}} + \frac{\kappa_0 N}{\kappa_0 + N}(\bar{\boldsymbol{x}} - \mu_0)(\bar{\boldsymbol{x}} - \mu_0)^T = \boldsymbol{S}_0 + \boldsymbol{S} + \kappa_0 \mu_0 \mu_0^T - \kappa_N \mu_N \mu_N^T, \boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T$$

- *The posterior mean is a convex combination of the prior mean and the MLE with strength $\kappa_0 + N$.*

- *The posterior scatter matrix $\boldsymbol{S}_N$ is the prior scatter matrix $\boldsymbol{S}_0$ plus the empirical scatter matrix $\boldsymbol{S}_{\bar{\boldsymbol{x}}}$ plus an extra term due to the uncertainty in the mean which creates its own scatter matrix.*

- Minka, T. (2000). Inferring a Gaussian distribution. Technical report, MIT.
- Chipman, H., E. George, and R. Mc-Culloch (2001). The practical implementation of Bayesian Model Selection. *Model Selection*. IMS Lecture Notes.
- Fraley, C. and A. Raftery (2007). Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. *J. of Classification 24*, 155–181
- K. Murphy, Conjugate Baeysian Analysis of the Gaussian Distribution, 2007

MAP Estimate of μ and Σ

$$p(\mu,\Sigma \mid \mathcal{D},\mu_0,\kappa_0,\mathbf{S}_0,\nu_0) = \mathcal{NIW}(\mu,\Sigma \mid \mu_N,\kappa_N,\nu_N,\mathbf{S}_N)$$

$$\mu_N = \frac{\kappa_0\mu_0 + N\bar{\mathbf{x}}}{\kappa_N} = \frac{\kappa_0}{\kappa_0+N}\mu_0 + \frac{N}{\kappa_0+N}\bar{\mathbf{x}}, \kappa_N = \kappa_0 + N, \nu_N = \nu_0 + N$$

$$\mathbf{S}_N = \mathbf{S}_0 + \mathbf{S}_{\bar{x}} + \frac{\kappa_0 N}{\kappa_0+N}(\bar{\mathbf{x}}-\mu_0)(\bar{\mathbf{x}}-\mu_0)^T = \mathbf{S}_0 + \mathbf{S} + \kappa_0\mu_0\mu_0^T - \kappa_N\mu_N\mu_N^T, \mathbf{S} = \sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T$$

- The mode of the joint posterior is:

$$(\bar{\mu},\bar{\Sigma}) = \arg\max_{\mu,\Sigma} p(\mu,\Sigma \mid \mathcal{D},\mu_0,\kappa_0,\mathbf{S}_0,\nu_0) = \left(\mu_N, \frac{\mathbf{S}_N}{\nu_N + D + 2}\right)$$

- For $k_0 = 0$, this becomes:

$$(\bar{\mu},\bar{\Sigma}) = \arg\max_{\mu,\Sigma} p(\mu,\Sigma \mid \mathcal{D},\mu_0,\kappa_0 = 0,\mathbf{S}_0,\nu_0) = \left(\bar{\mathbf{x}}, \frac{\mathbf{S}_0 + \mathbf{S}_{\bar{x}}}{\nu_0 + N + D + 2}\right)$$

- It is interesting to note that this mode is almost the same as the MAP estimate shown next – it differs by 1 in the denominator as the mode above is the mode of the joint posterior and not of the marginal.

The Posterior Marginals of μ and Σ

- The posterior marginal for Σ and μ are:

$$p(\Sigma \mid \mathcal{D}) = \int p(\boldsymbol{\mu}, \Sigma \mid \mathcal{D}) d\boldsymbol{\mu} = \mathcal{IW}\left(\Sigma \mid \boldsymbol{S}_N, \nu_N\right)$$

$$\bar{\Sigma}_{MAP} = \frac{\boldsymbol{S}_N}{\nu_N + D + 1}, \mathbb{E}[\Sigma] = \frac{\boldsymbol{S}_N}{\nu_N - D - 1}$$

$$p(\boldsymbol{\mu} \mid \mathcal{D}) = \int p(\boldsymbol{\mu}, \Sigma \mid \mathcal{D}) d\Sigma = \mathcal{T}_{\nu_N - D + 1}\left(\boldsymbol{\mu} \mid \boldsymbol{\mu}_N, \frac{1}{\kappa_N(\nu_N - D + 1)} \boldsymbol{S}_N\right)$$

- It is not surprising that the last marginal is Student's $\mathcal{T}$ that we know can be represented as a mixture of Gaussians.

- To see the connection for the scalar case ($D = 1$), note that $\boldsymbol{S}_N$ plays the role of the posterior sum of squares $\nu_N \sigma_N^2$ (you may want to revisit the the results for the scalar case of simultaneously estimating μ and σ^2):

$$\frac{1}{\kappa_N(\nu_N - D + 1)} \boldsymbol{S}_N = \frac{\boldsymbol{S}_N}{\kappa_N \nu_N} = \frac{\sigma_N^2}{\kappa_N}$$

The Posterior Predictive of μ and Σ

- The posterior predictive $p(\boldsymbol{x}|\mathcal{D}) = p(\boldsymbol{x}, \mathcal{D})/p(\mathcal{D})$ can be evaluated as:

$$p(\mathbf{x}\,|\,\mathcal{D}) = \int\int \mathcal{N}(\mathbf{x}\,|\,\mu,\Sigma)\mathcal{NIW}\left(\mu,\Sigma\,|\,\mu_N,\kappa_N,\nu_N,\boldsymbol{S}_N\right)d\mu d\Sigma$$
$$= \mathcal{T}_{\nu_N-D+1}\left(\mathbf{x}\,|\,\mu_N, \frac{\kappa_N+1}{\kappa_N(\nu_N-D+1)}\boldsymbol{S}_N\right)$$

- Recall that the Student's $\mathcal{T}$ distribution has heavier tails than the Gaussian but rapidly becomes Gaussian like.

- To see the connection of the above expression with the scalar case, note:

$$\frac{\kappa_N+1}{\kappa_N(\nu_N-D+1)}\boldsymbol{S}_N = \frac{(\kappa_N+1)\nu_N\sigma_N^2}{\kappa_N\nu_N} = \frac{(\kappa_N+1)\sigma_N^2}{\kappa_N}$$

- K. Murphy, Conjugate Baeysian Analysis of the Gaussian Distribution, 2007 (Section 9)

Marginal Likelihood

- The posterior is given as:

$$p\left(\mu,\Sigma\,|\,\mathcal{D},\mu_0,\kappa_0,\boldsymbol{S}_0,\nu_0\right)=\frac{1}{p\left(\mathcal{D}\right)}\frac{1}{Z_0}\mathcal{NIW}'\left(\mu,\Sigma\,|\,a_0\right)\frac{1}{\left(2\pi\right)^{ND/2}}\mathcal{N}'\left(\mathcal{D}\,|\,\mu,\Sigma\right)=\frac{1}{Z_N}\mathcal{NIW}'\left(\mu,\Sigma\,|\,a_N\right)$$

$$\mathcal{NIW}'\left(\mu,\Sigma\,|\,a_0\right)=\left|\Sigma\right|^{-\left(\left(\nu_0+D\right)/2+1\right)}\exp\left(-\frac{1}{2}tr\left(\boldsymbol{S}_0\Sigma^{-1}\right)-\frac{\kappa_0}{2}\left(\mu-\mu_0\right)^T\Sigma^{-1}\left(\mu-\mu_0\right)\right)$$

$$\mathcal{N}'\left(\mathcal{D}\,|\,\mu,\Sigma\right)=\left|\Sigma\right|^{-N/2}\exp\left(-\frac{1}{2}tr\left(\Sigma^{-1}\boldsymbol{S}\right)\right)$$

- In the last two expressions ()' give the *unnormalized likelihood and prior*.

- The marginal likelihood $p(\mathcal{D})=\frac{Z_N}{Z_0}\frac{1}{(2\pi)^{ND/2}}$ is then:

$$p\left(\mathcal{D}\right)=\frac{2^{\nu_N D/2}\Gamma_D\left(\frac{\nu_N}{2}\right)\left(\frac{2\pi}{\kappa_N}\right)^{D/2}}{\left|\boldsymbol{S}_N\right|^{\nu_N/2}}\frac{\left|\boldsymbol{S}_0\right|^{\nu_0/2}}{2^{\nu_0 D/2}\Gamma_D\left(\frac{\nu_0}{2}\right)\left(\frac{2\pi}{\kappa_0}\right)^{D/2}}\frac{1}{\left(2\pi\right)^{ND/2}}=\frac{1}{\left(2\pi\right)^{ND/2}}\frac{2^{\nu_N D/2}}{2^{\nu_0 D/2}}\frac{\left(\frac{2\pi}{\kappa_N}\right)^{D/2}}{\left(\frac{2\pi}{\kappa_0}\right)^{D/2}}\frac{\left|\boldsymbol{S}_0\right|^{\nu_0/2}}{\left|\boldsymbol{S}_N\right|^{\nu_N/2}}\frac{\Gamma_D\left(\frac{\nu_N}{2}\right)}{\Gamma_D\left(\frac{\nu_0}{2}\right)}=\frac{1}{\pi^{ND/2}}\frac{\Gamma_D\left(\frac{\nu_N}{2}\right)}{\Gamma_D\left(\frac{\nu_0}{2}\right)}\frac{\left|\boldsymbol{S}_0\right|^{\nu_0/2}}{\left|\boldsymbol{S}_N\right|^{\nu_N/2}}\left(\frac{\kappa_0}{\kappa_N}\right)^{D/2}$$

- Note that for $D=1$, this reduces to the familiar Equ: $p(\mathcal{D})=\frac{1}{\pi^{N/2}}\frac{\Gamma_D\left(\frac{\nu_N}{2}\right)}{\Gamma_D\left(\frac{\nu_0}{2}\right)}\frac{\left(\nu_0\sigma_0^2\right)^{\nu_0/2}}{\left(\nu_N\sigma_N^2\right)^{\nu_N/2}}\left(\frac{\kappa_0}{\kappa_N}\right)^{1/2}$

  - K. Murphy, Conjugate Baeysian Analysis of the Gaussian Distribution, 2007 (See calculation of the marginal likelihood for 1D analysis of the Normal-Inverse-Chi-Squared prior on Section 5)

Non Informative Prior

- The uninformative Jeffrey's prior is $p(\mu,\Sigma)\propto|\Sigma|^{-(D+1)/2}$. This is obtained in the limit

$$\kappa_0\to 0,\ \nu_0\to -1,\ |\boldsymbol{S}_0|\to 0$$

$$p(\mu,\Sigma\,|\,\mathcal{D})\propto|\Sigma|^{-\frac{\nu_0+D+2}{2}}\exp\left(-\frac{1}{2}\Sigma^{-1}\left(\boldsymbol{S}_0+\kappa_0(\mu-\mu_0)(\mu-\mu_0)^T\right)\right)=|\Sigma|^{-\frac{D+1}{2}}$$

- In this case, we have:

$$\boldsymbol{\mu}_N=\bar{\boldsymbol{x}},\ \kappa_N=N,\ \nu_N=N-1,\ \boldsymbol{S}_N=\boldsymbol{S}_{\bar{x}}=\sum_{i=1}^{N}\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}\right)^T$$

- The posterior marginals are then given as:

$$p(\Sigma\,|\,\mathcal{D})=\mathcal{IW}\left(\Sigma\,|\,\boldsymbol{S},N-1\right),\ p(\mu\,|\,\mathcal{D})=\mathcal{T}_{N-D}\left(\mu\,|\,\bar{\boldsymbol{x}},\frac{1}{N(N-D)}\boldsymbol{S}\right)$$

- Also the posterior predictive is:

$$p(\boldsymbol{x}\,|\,\mathcal{D})=\mathcal{T}_{N-D}\left(\boldsymbol{x}\,|\,\bar{\boldsymbol{x}},\frac{N+1}{N(N-D)}\boldsymbol{S}\right)$$

- Bayesian Data Analysis, A. Gelman, J. Carlin, H. Stern and D. Rubin, 2004 (pp. 88)
- K. Murphy, Conjugate Baeysian Analysis of the Gaussian Distribution, 2007 (See Section 9)

Non Informative Prior

- Based on the report of Minka below, the uninformative prior should be instead

$$\lim_{\kappa \to 0} \mathcal{N}\left(\mu \mid \mu_0, \frac{1}{\kappa}\Sigma\right) \mathcal{InvWis}\left(\Sigma \mid \boldsymbol{S}_0, \kappa\right) \qquad v_0 = 0 \ \textit{instead of} \ v_0 \to -1$$

$$\propto |2\pi\Sigma|^{-1/2} |\Sigma|^{-(D+1)/2} \propto |\Sigma|^{-(D/2+1)} \propto \mathcal{NIW}\left(\mu, \Sigma \mid \mathbf{0}, 0, 0\boldsymbol{I}, 0\right)$$

- Often, a *data-dependent weakly informative prior is recommended* (see Chipman et al. and Fraley and Raftery):

$$\textit{Set}: \boldsymbol{S}_0 = \frac{diag\boldsymbol{S}_{\bar{x}}}{N}, v_0 = D + 2 \ \textit{to ensure} \ \mathbb{E}[\Sigma] = \boldsymbol{S}_0, \textit{and}$$

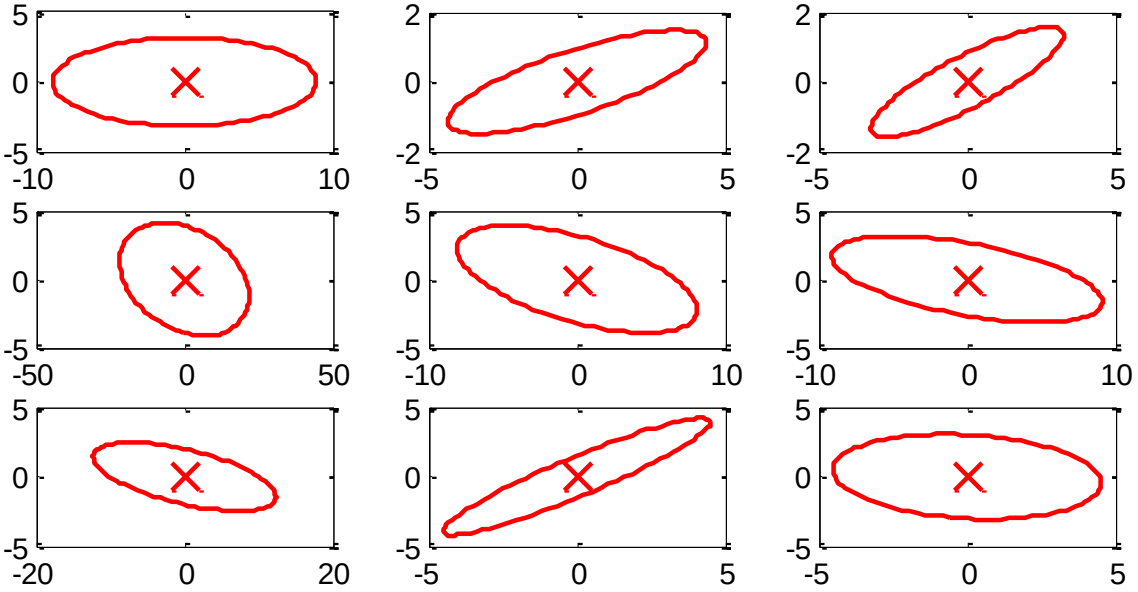
$$\mu_0 = \bar{x}, \kappa_0 = 0.01$$

- Minka, T. (2000). Inferring a Gaussian distribution. Technical report, MIT.
- Chipman, H., E. George, and R. Mc-Culloch (2001). The practical implementation of Bayesian Model Selection. *Model Selection*. IMS Lecture Notes.
- Fraley, C. and A. Raftery (2007). Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. *J. of Classification 24*, 155–181

Visualization of the Wishart Distribution

- $\mathcal{Wis}$ is a distribution over matrices thus difficult to plot the PDF. However, one can sample from it and in 2d use the eigen-vectors of the resulting sample to define an ellipse as we have done for the 2D Gaussian.

wiPlotDemo
from PMTK



- Above: Samples from $\boldsymbol{\Sigma} \sim Wi(\boldsymbol{S}, \nu)$, where $\boldsymbol{S} = 3.1653, -0.0262;\ -0.0262, 0.6477]$ and $\nu = 3$.

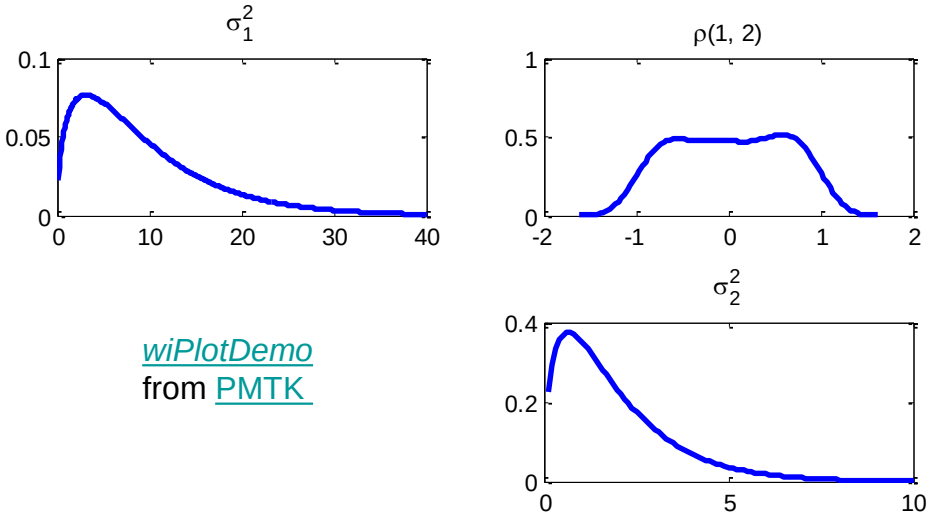
- The sampled matrices are highly variable, and some are nearly singular. As ν increases, the sampled matrices are more concentrated on the prior $\boldsymbol{S}$.

Visualization of the Wishart Distribution

- For off-diagonal elements, one can sample matrices from the distribution, and then compute their distribution empirically.
- We can *convert each sampled matrix to a correlation matrix*, and thus compute a Monte Carlo approximation

$$\mathbb{E}[R_{ij}] \approx \frac{1}{S}\sum_{s=1}^{S} R\left(\Sigma^{(s)}\right)_{ij}, \Sigma^{(s)} \sim \mathcal{Wi}(\Sigma, v), R\left(\Sigma\right)_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$$

- We can then use kernel density estimation to produce for plotting purposes a smooth approximation to the univariate density $\mathbb{E}[R_{ij}]$.
- On the right: Plots of the marginals (which are $\mathcal{Gamma}$), and the sample-based marginal on the correlation coefficient
- If $\nu = 3$ there is a lot of uncertainty about the the correlation coefficient ρ (nearly uniform on $[-1, 1]$).



wiPlotDemo from PMTK

Exponential Family

- Large family of useful distributions with common properties
  - Bernoulli, beta, binomial, chi-square, Dirichlet, gamma, Gaussian, geometric, multinomial, Poisson, Weibull, . .
- Not in the family:
  - ✓ Uniform,
  - ✓ Student's T,
  - ✓ Cauchy,
  - ✓ Laplace,
  - ✓ mixture of Gaussians,
  - ✓ . . .
- Variable can be discrete/continuous (or vectors thereof)

Exponential Family

- The exponential family of distributions over $\boldsymbol{x}$, given parameters $\boldsymbol{\eta}$, is defined to be the set of distributions of the form

$$p(\boldsymbol{x}\,|\,\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\} \; or$$

$$p(\boldsymbol{x}\,|\,\boldsymbol{\eta}) = h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x}) - A(\boldsymbol{\eta})\right\}, \; where: A(\boldsymbol{\eta}) = -\log g(\boldsymbol{\eta})$$

$\boldsymbol{x}$ is scalar/vector, discrete/continuous. ***$\boldsymbol{\eta}$ are the natural parameters and $u(\boldsymbol{x})$ is referred to as a sufficient statistic.***

- *$g(\boldsymbol{\eta})$ ensures that the distribution is normalized* and satisfies

$$g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}d\boldsymbol{x} = 1$$

- The normalization factor Z and the log of it A are defined as:

$$Z(\boldsymbol{\eta}) = \frac{1}{g(\boldsymbol{\eta})}, \; A(\boldsymbol{\eta}) = \ln Z(\boldsymbol{\eta}) = -\ln g(\boldsymbol{\eta}) = \ln \int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}d\boldsymbol{x}$$

$$p(\boldsymbol{x}\,|\,\boldsymbol{\eta}) = h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}\Big/ Z(\boldsymbol{\eta})$$

- The space of $\boldsymbol{\eta}$ for which $\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}d\boldsymbol{x} < \infty$ is the ***natural parameter space***.

Canonical or Natural Parameters

- When the parameter $\boldsymbol{\theta}$ enters the exponential family as $\boldsymbol{\eta}(\boldsymbol{\theta})$, we write the probability density of the exponential family as follows:

$$p(\mathbf{x}\,|\,\boldsymbol{\theta}) = h(\mathbf{x})\, g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right) \exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\, u(\mathbf{x})\right\} \; or$$

$$p(\mathbf{x}\,|\,\boldsymbol{\theta}) = h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\, u(\mathbf{x}) - A\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)\right\},$$

$$where: A\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right) = -\log g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)$$

- $\boldsymbol{\eta}(\boldsymbol{\theta})$ *are the canonical or natural parameters,*

- $\boldsymbol{\theta}$ *is the parameter vector of some distribution that can be written in the exponential family format*

Exponential Family: The Bernoulli Distribution

- Consider the Bernoulli distribution:

$$p(x\,|\,\mu) = \mathcal{Bern}(x\,|\,\mu) = \mu^x(1-\mu)^{1-x} = \exp\left\{x\ln\mu + (1-x)\ln(1-\mu)\right\} =$$

$$= \underbrace{(1-\mu)}_{g(\eta)} \exp\left\{\underbrace{\ln\left(\frac{\mu}{1-\mu}\right)}_{\eta} x\right\}$$

$$p(\boldsymbol{x}\,|\,\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}$$
$$= h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x}) - A(\boldsymbol{\eta})\right\}$$

- From this we see that (note that *the relation $\mu(\eta)$ is invertible*)

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right) \Rightarrow \qquad \mu = \sigma(\eta) = \frac{1}{1+e^{-\eta}}$$

Logistic sigmoid function

and

$$g(\eta) = 1-\mu = 1-\sigma(\eta) = \sigma(-\eta)$$

- *Finally:*

$$p(x\,|\,\eta) = g(\eta)\exp\{\eta x\},\ u(x) = x,\ h(x) = 1,\ g(\eta) = \sigma(-\eta),$$
$$A(\eta) = -\ln g(\boldsymbol{\eta}) = -\log(1-\mu) = \log(1+e^{\eta})$$

Exponential Family: The Beta Distribution

- Consider the Beta distribution

$$\mathcal{Beta}(\mu \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\exp\left[(a-1)\ln\mu + (b-1)\ln(1-\mu)\right]$$

- Comparing this with our exponential family:

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}$$
$$= h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x}) - A(\boldsymbol{\eta})\right\}$$

we can easily indentify:

$$u(\mu) = (\ln\mu, \ln(1-\mu))^T,\ \boldsymbol{\eta} = (a-1, b-1)^T,\ h(\mu) = 1,\ g(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)},$$

$$A(a,b) = -\ln g(\boldsymbol{\eta}) = \ln\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Exponential Family: The Gaussian

- Consider the univariate Gaussian

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}\mu^2 + \frac{\mu}{\sigma^2}x\right\}$$

- Comparing this with our exponential family:

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = h(\boldsymbol{x}) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\} = h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x}) - A(\boldsymbol{\eta})\right\}$$

we can indentify (this is a two parameter distribution):

$$u(x) = (x, x^2)^T, \boldsymbol{\eta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^T, h(x) = \frac{1}{\sqrt{2\pi}}, g(\boldsymbol{\eta}) = \left(-2\eta_2\right)^{1/2} \exp\frac{\eta_1^2}{4\eta_2}$$

$$A(\eta) = -\ln g(\boldsymbol{\eta}) = -\frac{1}{2}\ln\left(-2\eta_2\right) - \frac{\eta_1^2}{4\eta_2}$$

Conjugate Priors

- In general, for a given probability distribution $p(\boldsymbol{x}|\boldsymbol{\eta})$, we can seek a prior $p(\boldsymbol{\eta})$ that is conjugate to the likelihood function, so that the posterior distribution has the same functional form as the prior.
  - For the Bernoulli, the conjugate prior is the Beta distribution
  - For the Gaussian, the conjugate prior for the mean is a Gaussian, and the conjugate prior for the precision is the Wishart distribution

Conjugate Priors

- For any member of the exponential family with likelihood

$$p(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = h(\boldsymbol{x})\,g(\boldsymbol{\eta}(\boldsymbol{\theta}))\exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\,u(\boldsymbol{x})\right\}$$

there exists a conjugate prior that can be written in the form

$$p(\boldsymbol{\theta}|\nu_0,\boldsymbol{\tau}_0) \propto g(\boldsymbol{\eta}(\boldsymbol{\theta}))^{\nu_0}\exp\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\boldsymbol{\tau}_0\} = \exp\{\nu_0\boldsymbol{\eta}^T(\boldsymbol{\theta})\bar{\tau}_0 - A(\boldsymbol{\eta}(\boldsymbol{\theta}))\nu_0\},\ where:\ \boldsymbol{\tau}_0 \equiv \nu_0\bar{\boldsymbol{\tau}}_0$$

- In normalized form, we write:

$$p(\boldsymbol{\theta}|\nu_0,\boldsymbol{\tau}_0) = \frac{1}{Z(\nu_0,\boldsymbol{\tau}_0)}g(\boldsymbol{\eta}(\boldsymbol{\theta}))^{\nu_0}\exp\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\boldsymbol{\tau}_0\} = \frac{1}{Z(\nu_0,\boldsymbol{\tau}_0)}\exp\{\nu_0\boldsymbol{\eta}^T(\boldsymbol{\theta})\bar{\boldsymbol{\tau}}_0 - A(\boldsymbol{\eta}(\boldsymbol{\theta}))\nu_0\}$$

$$where: Z(\nu_0,\boldsymbol{\tau}_0) = \int \exp\{\nu_0\boldsymbol{\eta}^T(\boldsymbol{\theta})\bar{\boldsymbol{\tau}}_0 - A(\boldsymbol{\eta}(\boldsymbol{\theta}))\nu_0\}d\boldsymbol{\theta}$$

Conjugate Priors

$$p(\boldsymbol{X}\mid\boldsymbol{\theta})=\prod_{n=1}^{N}\left(h(\boldsymbol{x}_n)g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)\exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})u(\boldsymbol{x}_n)\right\}\right)=\prod_{n=1}^{N}\left(h(\boldsymbol{x}_n)\right)g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^N\exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\sum_{n=1}^{N}u(\boldsymbol{x}_n)\right\}$$

$$p(\boldsymbol{\theta}|\nu_0,\boldsymbol{\tau_0})=\frac{1}{Z(\nu_0,\boldsymbol{\tau}_0)}g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_0}\exp\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\boldsymbol{\tau}_0\}=\frac{1}{Z(\nu_0,\boldsymbol{\tau}_0)}\exp\{\nu_0\boldsymbol{\eta}^T(\boldsymbol{\theta})\bar{\boldsymbol{\tau}}_0-A\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)\nu_0\}$$

- Using $\bar{\boldsymbol{u}}=\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{u}(\boldsymbol{x}_n)$, the posterior becomes (this form justifies $\bar{\boldsymbol{\tau}}_0$):

$$p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{\chi},\nu)\propto g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_0+N}\exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\left(\sum_{n=1}^{N}\boldsymbol{u}(\boldsymbol{x}_n)+\nu_0\bar{\boldsymbol{\tau}}_0\right)\right\}=g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_0+N}\exp\{\boldsymbol{\eta}^{\boldsymbol{T}}(\boldsymbol{\theta})(N\bar{\boldsymbol{u}}+\nu_0\bar{\boldsymbol{\tau}}_0)\}$$

- The parameter ν_0 can be interpreted as *effective number of fictitious observations* in the prior each of which has a value for the sufficient statistic equal to $\bar{\boldsymbol{\tau}}_0$.

$$p(\boldsymbol{\theta}|\boldsymbol{X},\nu_N,\boldsymbol{\tau_N})=\frac{1}{Z(\nu_N,\boldsymbol{\tau}_N)}g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_N}\exp\left\{(N+\nu_0)\boldsymbol{\eta}^T(\boldsymbol{\theta})\frac{N\bar{\boldsymbol{u}}+\nu_0\bar{\boldsymbol{\tau}}_0}{N+\nu_0}\right\}=\frac{1}{Z(\nu_N,\tau_N)}g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_N}\exp\{\nu_N\boldsymbol{\eta}^T(\theta)\bar{\boldsymbol{\tau}}_N\},$$

$$where\ \nu_N=\nu_0+N,\ \bar{\boldsymbol{\tau}}_N=\frac{N\bar{\boldsymbol{u}}+\nu_0\bar{\boldsymbol{\tau}}_0}{N+\nu_0},\boldsymbol{\tau}_N=\nu_N\bar{\boldsymbol{\tau}}_N=N\bar{\boldsymbol{u}}+\nu_0\bar{\boldsymbol{\tau}}_0=\sum_{i=1}^{N}\boldsymbol{u}(x_i)+\boldsymbol{\tau}_0$$

Posterior Predictive

- Let $\boldsymbol{u}(\boldsymbol{X}) = \sum_{i=1}^{N} \boldsymbol{u}(\boldsymbol{x}_i), \boldsymbol{u}(\boldsymbol{X}') = \sum_{i=1}^{N'} \boldsymbol{u}(\boldsymbol{x}_i'),$ the posterior predictive is then:

$$p(\boldsymbol{X}' \mid \boldsymbol{X}) = \int p(\boldsymbol{X}' \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{X}) d\boldsymbol{\theta}$$
$$= \prod_{i=1}^{N'} h(\boldsymbol{x}_i') \int g(\boldsymbol{\eta})^{N'} \exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta}) \boldsymbol{u}(\boldsymbol{X}')\right\} \frac{1}{Z(\nu_0 + N, \boldsymbol{u}(\boldsymbol{X}) + \boldsymbol{\tau}_0)} g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{\nu_N} \exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\left(\boldsymbol{u}(\boldsymbol{X}) + \boldsymbol{\tau}_0\right)\right\} d\boldsymbol{\theta}$$

- This is simplified as follows:

$$p(\boldsymbol{X}' \mid \boldsymbol{X}) = \prod_{i=1}^{N'} h(\boldsymbol{x}_i') \frac{1}{Z\left(\nu_0 + N, \boldsymbol{u}(\boldsymbol{X}) + \boldsymbol{\tau}_0\right)} \int g\left(\boldsymbol{\eta}(\boldsymbol{\theta})\right)^{N' + \nu_N} \exp\left\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\left(\boldsymbol{u}(\boldsymbol{X}') + \boldsymbol{u}(\boldsymbol{X}) + \boldsymbol{\tau}_0\right)\right\} d\boldsymbol{\theta}$$
$$= \prod_{i=1}^{N'} h(\boldsymbol{x}_i') \frac{Z\left(\nu_0 + N + N', \boldsymbol{u}(\boldsymbol{X}') + \boldsymbol{u}(\boldsymbol{X}) + \boldsymbol{\tau}_0\right)}{Z\left(\nu_0 + N, \boldsymbol{u}(\boldsymbol{X}) + \boldsymbol{\tau}_0\right)}$$

- If $N = 0$, this becomes the marginal likelihood of X', which reduces to the normalizer of the posterior divided by the normalizer of the prior multiplied by a constant.

Beta/Bernoulli: Posterior Predictive

- Consider a Bernoulli likelihood with a Beta prior. The likelihood takes the familiar exponential distribution form:

$$p(\mathcal{D}\mid\theta)=\theta^{\sum_i x_i}(1-\theta)^{N-\sum_i x_i}=(1-\theta)^N\exp\left(\log\frac{\theta}{1-\theta}\sum_i x_i\right)$$

- The conjugate prior is a Beta: $p(\theta\mid\nu_0,\tau_0)\propto(1-\theta)^{\nu_0}\exp\left(\log\left(\frac{\theta}{1-\theta}\right)\tau_0\right)=\theta^{\tau_0}(1-\theta)^{\nu_0-\tau_0}$

$$p(\theta\mid\nu_0,\tau_0)=\mathcal{Beta}(\alpha,\beta),\alpha=\tau_0+1,\beta=\nu_0-\tau_0+1,$$

- Thus the posterior becomes: $p(\theta\mid\mathcal{D})\propto\theta^{\tau_0+s}(1-\theta)^{\nu_0-\tau_0+N-s}\Rightarrow$

$$p(\theta\mid\mathcal{D})=\mathcal{Beta}(\alpha_N,\beta_N),\alpha_N=\alpha+s,\beta_N=\beta+(N-s),s=\sum_i\mathbb{I}(x_i=1)$$

- Let s the number of heads in the past data. The probability of $s'=\sum_{i=1}^{m}\mathbb{I}(x_i'=1)$ future heads in m trials is then:

$$p(s'\mid\mathcal{D},m)=\int\theta^{s'}(1-\theta)^{m-s'}\frac{\Gamma(\alpha_N+\beta_N)}{\Gamma(\alpha_N)\Gamma(\beta_N)}\theta^{\alpha_N-1}(1-\theta)^{\beta_N-1}d\theta=\frac{\Gamma(\alpha_N+\beta_N)}{\Gamma(\alpha_N)\Gamma(\beta_N)}\frac{\Gamma(\alpha_{N+m})\Gamma(\beta_{N+m})}{\Gamma(\alpha_{N+m}+\beta_{N+m})}$$

$$\alpha_{N+m}=\alpha_N+s',\beta_{N+m}=\beta_N+(m-s')$$