Exponential Family and Generalized Linear Models
&
Bayesian Inference for the Multivariate Gaussian

Prof. Nicholas Zabaras
University of Notre Dame
Notre Dame, IN, USA

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

September 15, 2017

Contents

- Exponential Family, Computing the Moments, Moment Parametrization, Sufficiency and Neymann Factorization, Sufficient Statistics and MLE Estimates, MLE and Kullback-Leibler Distance, Conjugate Priors, Posterior Predictive, Maximum Entropy and the Exponential Family

- Generalized Linear Models, Canonical Response Function, Batch IRLS, Sequential Estimation – LMS

- Gaussian mixtures challenges, Posterior responsibilities, MLE of Mixture Model, Examples of Mixture of 2D Gaussians

- Chris Bishops' PRML book, Chapters 2 and 9
- 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

- Chris Bishop, Pattern Recognition and Machine Learning, Chapter 2
- Kevin Murphy, Machine Learning: A probabilistic Perspective, Chapter 4

Exponential Family of Distributions

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)
- We will briefly introduce *the conditional setting in which we have a directed model X→Y with both X & Y observed and with p(Y|X) being an exponential family distribution parametrized using* *Generalized Linear Models (GLIM's)* (more on the topic on a forthcoming lecture).

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. ***η are the natural parameters and u(x) is referred to as a sufficient statistic.***

- *g(η) 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:

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

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

- The space of η 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 θ enters the exponential family as $\eta(\theta)$, we write the probability density of the exponential family as follows:

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

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

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

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

- *θ is the parameter vector of some distribution that can be written in the exponential family format*

Joint Probability Distribution on Discrete RVs

- Any joint probability distribution on discrete random variables lies on the exponential family.* Indeed:

$$p(\boldsymbol{x}\mid\Psi)=\frac{1}{Z(\Psi)}\prod_{C\in\mathcal{C}}\Psi_C(\boldsymbol{x}_C)=\exp\left(\sum_{C\in\mathcal{C}}\log\Psi_C(\boldsymbol{x}_C)-\log Z(\Psi)\right)$$

$$\text{But for discrete rv's}:\Psi_C(\boldsymbol{x}_C)=\prod_{v_1^{i_1},v_2^{i_2},\ldots,v_k^{i_k}}\Psi_C\left(v_1^{i_1},v_2^{i_2},\ldots,v_k^{i_k}\right)^{\delta\left(x_1,v_1^{i_1}\right)\delta\left(x_2,v_2^{i_2}\right)\ldots\delta\left(x_k,v_k^{i_k}\right)}$$

- Substitution to the 1st Eq. gives:

$$p(\boldsymbol{x}\mid\Psi)=\frac{1}{Z(\Psi)}\prod_{C\in\mathcal{C}}\Psi_C(\boldsymbol{x}_C)=\exp\left(\sum_{C\in\mathcal{C}}\sum_{v_1^{i_1},v_2^{i_2},\ldots,v_k^{i_k}}\log\Psi_C\left(v_1^{i_1},v_2^{i_2},\ldots,v_k^{i_k}\right)\delta\left(x_1,v_1^{i_1}\right)\delta\left(x_2,v_2^{i_2}\right)\ldots\delta\left(x_k,v_k^{i_k}\right)-\log Z(\Psi)\right)\Rightarrow$$

$$p(\boldsymbol{x}\mid\Psi)=\frac{1}{Z(\Psi)}\prod_{C\in\mathcal{C}}\Psi_C(\boldsymbol{x}_C)=\exp\left(\sum_{v_1^{i_1},v_2^{i_2},\ldots,v_k^{i_k}}\sum_{C\in\mathcal{C}}\log\Psi_C\left(v_1^{i_1},v_2^{i_2},\ldots,v_k^{i_k}\right)\delta\left(x_1,v_1^{i_1}\right)\delta\left(x_2,v_2^{i_2}\right)\ldots\delta\left(x_k,v_k^{i_k}\right)-\log Z(\Psi)\right)$$

- This is in the exponential family with $\log\Psi_C\left(v_1^{i_1},v_2^{i_2},\ldots,v_k^{i_k}\right)$ corresponding to each component of $\boldsymbol{\eta}$, and $\delta\left(x_1,v_1^{i_1}\right)\delta\left(x_1,v_2^{i_2}\right)\ldots\delta\left(x_k,v_k^{i_k}\right)$ corresponding to components of the sufficient statistic $\mathbf{u}(x)$.

* We consider here the joint distribution of $\mathbf{x}$ written in terms of potentials Ψ's. We will see that this representation arises for rv's defined in undirected graphs where the potentials are defined over the random variables in each maximal clique c.

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*)

$$\eta=\ln\left(\frac{\mu}{1-\mu}\right)\Rightarrow$$

$$\mu=\sigma(\eta)=\frac{1}{1+e^{-\eta}}$$

Logistic sigmoid function

and

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

- *Finally:*

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

Exponential Family: The Poisson Distribution

- Consider the Poisson distribution with parameter λ:

$$p(x \mid \lambda) = \textit{Poisson}(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \underbrace{\frac{1}{x!}}_{h(x)} \exp\left\{ \underbrace{\ln \lambda}_{\eta} \underbrace{x}_{u(x)} - \underbrace{\lambda}_{A(\eta)} \right\}$$

- Recall that λ is the mean of the distribution and observe once more that *the relation $\lambda(\eta)$ is invertible*:

$$\eta = \ln(\lambda) \Rightarrow \lambda = e^{\eta}$$

Exponential Family: The Multinoulli Distribution

- Consider the multinomial distribution:

$$p(\boldsymbol{x}\,|\,\boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} = \exp\left(\boldsymbol{\eta}^T \boldsymbol{x}\right),$$

$$\boldsymbol{x} = \left\{x_1, \ldots, x_M\right\}^T, \boldsymbol{\eta} = \left\{\eta_1, \ldots, \eta_M\right\}^T, \eta_k = \ln \mu_k$$

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

- From this expression we see that $h(\boldsymbol{x})=1$, $u(\boldsymbol{x})=\boldsymbol{x}$, $g(\boldsymbol{\eta})=1$. It appears also that $A(\eta)=0$!

- We can resolve this problem by accounting for the dependence of μ_k, i.e. $\sum_{k=1}^{M} \mu_k = 1.$

Exponential Family: The Multinoulli Distribution

- We will express the distribution in terms of μ_k, k=1,…,M-1 subject to:

$$0 \le \mu_k \le 1, \quad \sum_{k=1}^{M-1} \mu_k \le 1$$

- The multinomial distribution becomes:

$$\exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} = \exp\left\{\sum_{k=1}^{M-1} x_k \ln \mu_k + \left(1 - \sum_{k=1}^{M-1} x_k\right) \ln\left(1 - \sum_{k=1}^{M-1} \mu_k\right)\right\} =$$

$$\exp\left\{\sum_{k=1}^{M-1} x_k \underbrace{\ln \frac{\mu_k}{1 - \sum_{k=1}^{M-1} \mu_k}}_{\eta_k} + \ln\left(1 - \sum_{k=1}^{M-1} \mu_k\right)\right\}$$

Exponential Family: The Multinoulli Distribution

- We identify

$$\eta_k = \ln \frac{\mu_k}{1 - \sum_{k=1}^{M-1} \mu_k} = \ln \frac{\mu_k}{\mu_M}, k = 1, \ldots, M-1$$

- Can also define:

$$\eta_M = \ln \frac{\mu_M}{\mu_M} = 0$$

- This equation can be inverted as:

$$\exp(\eta_k) = \frac{\mu_k}{1 - \sum_{k=1}^{M-1} \mu_k} \Rightarrow \sum_{k=1}^{M-1} \exp(\eta_k) = \frac{\sum_{k=1}^{M-1} \mu_k}{1 - \sum_{k=1}^{M-1} \mu_k} \Rightarrow$$

$$1 + \sum_{k=1}^{M-1} \exp(\eta_k) = \frac{1}{1 - \sum_{k=1}^{M-1} \mu_k} \Rightarrow \sum_{k=1}^{M-1} \mu_k = \frac{\sum_{k=1}^{M-1} \exp(\eta_k)}{1 + \sum_{k} \exp(\eta_k)} \Rightarrow 1 - \sum_{k=1}^{M-1} \mu_k = \left(1 + \sum_{k} \exp(\eta_k)\right)^{-1}$$

- Substitution intro the expression on the top of the slide:

$$\eta_k = \ln \frac{\mu_k}{1 - \sum_{k=1}^{M-1} \mu_k} = \ln \left[\mu_k \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k) \right) \right] \Rightarrow \mu_k = \frac{\exp(\eta_k)}{1 + \sum_{k=1}^{M-1} \exp(\eta_k)}$$

Exponential Family: The Multinoulli Distribution

- This is the so called the softmax function (note again *the relation $\mu(\eta)$ is invertible)*:

$$\mu_k = \frac{\exp(\eta_k)}{1+\sum_{k=1}^{M-1}\exp(\eta_k)}$$

Softmax function

- In this reduced representation, the distribution takes the form:

$$p(\boldsymbol{x}\,|\,\boldsymbol{\eta}) = \exp\left\{\sum_{k=1}^{M-1} x_k\eta_k + \ln\left(1-\sum_{k=1}^{M-1}\mu_k\right)\right\} = \left(1+\sum_{k=1}^{M-1}\exp(\eta_k)\right)^{-1}\exp(\boldsymbol{\eta}^T\boldsymbol{x})$$

- Comparing with the generic form of the exponential family:

$$\boldsymbol{\eta} = (\eta_1,\ldots,\eta_{M-1},0)^T,\ u(\boldsymbol{x}) = \boldsymbol{x},\ h(\boldsymbol{x}) = 1,\ g(\boldsymbol{\eta}) = \left(1+\sum_{k=1}^{M-1}\exp(\eta_k)\right)^{-1}$$

$$A = -\ln g(\boldsymbol{\eta}) = \ln\left(1+\sum_{k=1}^{M-1}\exp(\eta_k)\right) = \ln\left(\sum_{k=1}^{M}\exp(\eta_k)\right)$$

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:

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

$$A(a,b) = \ln\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Exponential Family: Gamma Distribution

- Consider the Gamma distribution

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

- Comparing this with our exponential family:

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

we can easily indentify:

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

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):

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

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

Exponential Family: von Mises Distribution

- Consider the von Mises distribution

$$p(\theta \mid \theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left(m\cos(\theta - \theta_0)\right) = \frac{1}{2\pi I_0(m)} \exp\left(m\cos\theta\cos\theta_0 + m\sin\theta\sin\theta_0\right)$$

- 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 identify that:

$$u(\theta) = (\cos\theta, \sin\theta)^T, \boldsymbol{\eta} = (m\cos\theta_0, m\sin\theta_0)^T, h(\theta) = 1, g(m, \theta_0) = \frac{1}{2\pi I_0(m)},$$

$$A(m, \theta_0) = \ln\left(2\pi I_0(m)\right)$$

The Multivariate Gaussian

- The exponent in the multivariate Gaussian is:

$$-\frac{1}{2}tr\left(\Lambda xx^T\right)+\mu^T\Lambda x-\frac{1}{2}\mu^T\Lambda\mu=-\frac{1}{2}tr\left(\Lambda xx^T\right)+\xi^T x-\frac{1}{2}\xi^T\Lambda^{-1}\xi,\ where:\ \Lambda=\Sigma^{-1},\ \xi=\Lambda\mu$$

- We need to put this in the form $p(\boldsymbol{x}\,|\,\boldsymbol{\eta})=h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}$
- The 3nd term contributes to g(η) whereas the 2nd term is directly an inner product between $\boldsymbol{x}$ and $\boldsymbol{\xi}=\boldsymbol{\Lambda\mu}$.
- For the 1st term, define two D^2-dimensional vector vec(Λ) and vec($\mathbf{xx}^T$) that consist of the columns of Λ and $\mathbf{xx}^T$, respectively. Then the 1st term has the form of an inner product between these two vectors. In summary:

$$\boldsymbol{\eta}=\begin{pmatrix}\boldsymbol{\xi}\\ vec(\Lambda)\end{pmatrix},\ u(\boldsymbol{x})=\begin{pmatrix}\boldsymbol{x}\\ -\frac{1}{2}vec(\boldsymbol{xx}^T)\end{pmatrix},\ g(\boldsymbol{\eta})=|\Lambda|^{1/2}\exp\left(-\frac{1}{2}\boldsymbol{\xi}^T\Lambda^{-1}\boldsymbol{\xi}\right),\ h(\boldsymbol{x})=(2\pi)^{-D/2},\ \boldsymbol{\xi}=\boldsymbol{\Lambda\mu}$$

$$A=-\ln g(\boldsymbol{\eta})=-\frac{1}{2}\ln|\Lambda|+\frac{1}{2}\boldsymbol{\xi}^T\Lambda^{-1}\boldsymbol{\xi}$$

Computing Moments of Sufficient Statistics u(x)

- Differentiate wrt η the $\int p(\boldsymbol{x}\,|\,\boldsymbol{\eta})d\boldsymbol{x}=1$ for the exponential family:

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

$$\nabla g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x})\right\}d\boldsymbol{x}+g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}\boldsymbol{u}(\boldsymbol{x})d\boldsymbol{x}=0\Rightarrow$$

$$-\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}=g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\}\boldsymbol{u}(\boldsymbol{x})d\boldsymbol{x}=\mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right]$$

- The above equation can be further simplified if written in terms of the partition function $Z=1/g(\eta)$ or $A=\log Z=-\log g(\eta)$:

$$\nabla A(\boldsymbol{\eta})=\mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right]$$

- Let us re-write explicitly the above equation as:

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

- We can compute the variance of $\mathbf{u}(\mathbf{x})$ by differentiating the Eq. above with respect to $\boldsymbol{\eta}$.

Computing Moments of Sufficient Statistics u(x)

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

$$\nabla^2 A(\boldsymbol{\eta}) = \underbrace{\nabla g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\} \boldsymbol{u}(\boldsymbol{x}) d\boldsymbol{x}}_{-\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})]\mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})^T\right]} + \underbrace{g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T u(\boldsymbol{x})\right\} \boldsymbol{u}(\boldsymbol{x}) \boldsymbol{u}(\boldsymbol{x})^T d\boldsymbol{x}}_{\mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x})^T\right]}$$

- Thus the covariance of u(**x**) can be expressed in terms of the 2nd derivatives of $A(\boldsymbol{\eta})$ and similarly for higher order moments.

$$\nabla^2 A(\boldsymbol{\eta}) = \mathrm{cov}\left[\boldsymbol{u}(\boldsymbol{x})\right] \; so \; A(\boldsymbol{\eta}) \; is \; convex$$

- Provided we can normalize a distribution from the exponential family, we can always find its moments by simple differentiation.

Computing Moments of Sufficient Statistics u(x)

$$\nabla A(\boldsymbol{\eta}) = \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right]$$

$$\nabla^2 A(\boldsymbol{\eta}) = \mathrm{cov}\left[\boldsymbol{u}(\boldsymbol{x})\right] \; so \; A(\boldsymbol{\eta}) \; is \; convex$$

- Let us check these relations for the Univariate Gaussian:

$$A(\boldsymbol{\eta}) = -\frac{1}{2}\ln\left(-2\eta_2\right) - \frac{\eta_1^2}{4\eta_2}, \boldsymbol{\eta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^T, u(x) = (x, x^2)^T$$

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta_1} = -\frac{\eta_1}{2\eta_2} = \mu = \mathbb{E}\left[X\right], \frac{\partial A(\boldsymbol{\eta})}{\partial \eta_2} = -\frac{1}{2\eta_2} + \frac{\eta_1^2}{4\eta_2^2} = \mu^2 + \sigma^2 = \mathbb{E}\left[X^2\right]$$

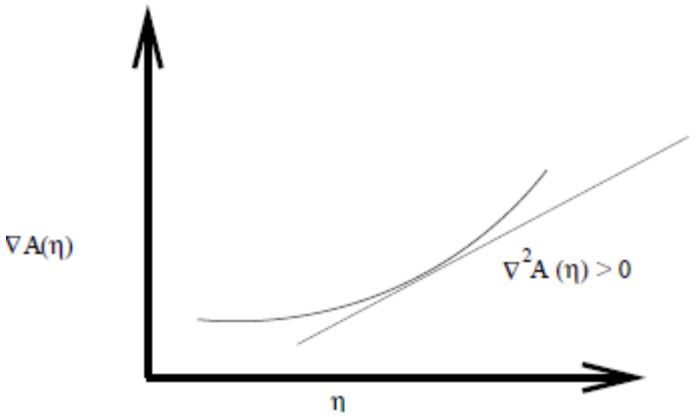
$$\frac{\partial^2 A(\boldsymbol{\eta})}{\partial \eta_1^2} = -\frac{1}{2\eta_2} = \sigma^2 = \mathrm{var}\left[X\right], etc.$$

Moment Parametrization

- We have shown that we can compute the mean of the distribution $\mu=\mathbb{E}[u(\mathbf{x})]$ in terms of the canonical parameter η:

$$\mu = \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right] = \nabla A(\boldsymbol{\eta})$$

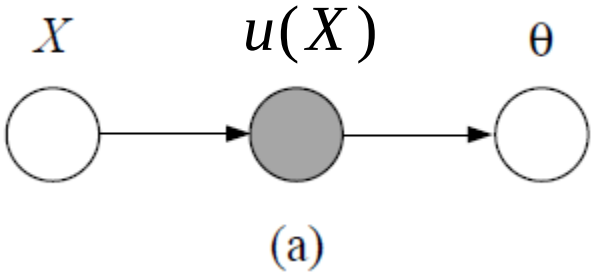
- We have also shown that $A(\eta)$ is a convex function. Since for a convex function there is one-to-one relation between the argument of the function and its derivative, the mapping $\mu(\eta)$ is invertable.



- Thus the exponential family of distributions can also be parameterized in terms of μ (*moment parametrization*) exactly as we started this course.

Sufficiency

- *u(X) is sufficient for θ if there is no information in X regarding θ beyond that in u(X).* Having observed u(X), we can throw away X for the purposes of inference with respect to θ.
- In the Bayesian approach in the Fig shown, we treat θ as an rv and say that u(X) is sufficient for θ if the following CI statement holds:



(a)

$$\theta \perp X \mid u(X)$$

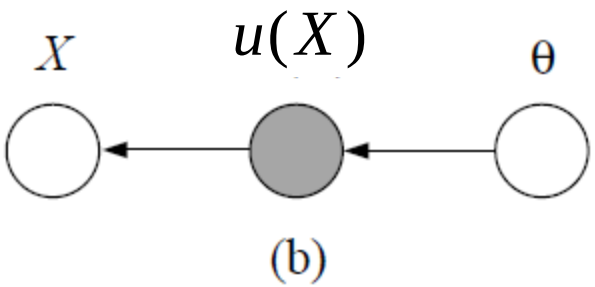
$$p\left(\theta \mid u(x), x\right) = p\left(\theta \mid u(x)\right)$$

- Thus, *u(X) contains all the needed information in X about θ.*

Frequentist Definition: Sufficiency

- The model in Fig b shown asserts the same CI relations as shown in Fig a earlier but has different parametrization.

$$p\left(x \mid u(x), \theta\right) = p\left(x \mid u(x)\right)$$



(b)

- Treating θ as a label, we can see the above CI statement as a frequentist definition of sufficiency.
- *u(X) is sufficient for θ if the p(x|u(x)) is not a function of θ.*
- The two approaches discussed imply a particular factorization of p(x|θ).

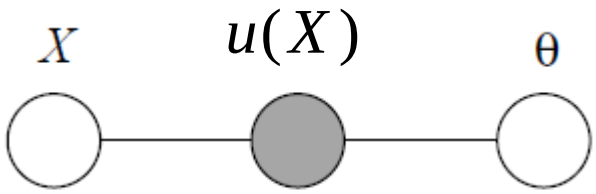
Neymann Factorization Theorem

- From the undirected graph (that expresses the same CI relations as the two earlier graphs), we can factorize as:

$$p\left(x,u(x),\theta\right)=g_1\left(u(x),\theta\right)g_2\left(x,u(x)\right)$$

- On the left, u(x) is a deterministic function of x and can be dropped as an argument:

$$p\left(x,\theta\right)=g_1\left(u(x),\theta\right)g_2\left(x,u(x)\right)$$



(c)

- One can derive for given ψ_1, ψ_2:

$$p\left(x\,|\,\theta\right)=p\left(x,\theta\right)/\,p\left(\theta\right)=\psi_1\left(u(x),\theta\right)\psi_2\left(x,u(x)\right)$$

- We can now see why u(x) was sufficient statistic for $\boldsymbol{\eta}$ in the exponential family:

$$p(\boldsymbol{x}\,|\,\theta)=\underbrace{h(\boldsymbol{x})}_{\psi_2(u(x),x)}\underbrace{\exp\left\{\boldsymbol{\eta}(\theta)^T u(\boldsymbol{x})-A\left(\boldsymbol{\eta}(\theta)\right)\right\}}_{\psi_1(u(x),\theta)}$$

MLE for the Exponential Family

- The joint density for a data $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ *is itself an exp. distribution with sufficient statistics* $\sum_{n=1}^{N} u(\boldsymbol{x}_n)$

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

$$\ln p(\boldsymbol{X}\,|\,\boldsymbol{\eta}) = \sum_{n=1}^{N} h(\boldsymbol{x}_n) + N\ln g(\boldsymbol{\eta}) + \boldsymbol{\eta}^T \sum_{n=1}^{N} u(\boldsymbol{x}_n) = \sum_{n=1}^{N} h(\boldsymbol{x}_n) - NA(\boldsymbol{\eta}) + \boldsymbol{\eta}^T \sum_{n=1}^{N} u(\boldsymbol{x}_n)$$

- *The exponential family is the only family of distributions* ***with finite sufficient statistics*** *(size independent of the data set size).*
- *The log likelihood is concave (A convex) and has a unique maximum.*
- Maximizing wrt $\boldsymbol{\eta}$ gives: $\nabla A(\boldsymbol{\eta}_{ML}) = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{u}(\boldsymbol{x}_n) \Rightarrow \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right] = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{u}(\boldsymbol{x}_n)$
- *At the MLE, the empirical average of the sufficient statistic is equal the model's theoretical expected sufficient statistics (moment matching).*
- *Thus to find the expected value of the sufficient statistics, one can use directly the data without having to estimate* $\boldsymbol{\eta}$. When $\mathbf{u}(\mathbf{x})=\mathbf{x}$, the above allows us to compute the expectation of x directly from the data.

MLE for the Exponential Family

$$\nabla A(\boldsymbol{\eta}_{ML}) = \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right] = \frac{1}{N}\sum_{n=1}^{N} u(\boldsymbol{x}_n)$$

- Using the sufficient statistic, one can in principle invert the above equ. to compute $\boldsymbol{\eta}_{MLE}$. For example, for the Bernoulli distribution,

$$p(x\,|\,\eta) = g(\eta)\exp\{\eta x\},\ u(x) = x,\ h(x) = 1,$$

$$\mu = \frac{1}{1+e^{-\eta}},\ g(\eta) = \frac{1}{1+e^{\eta}},\ \eta = \ln\left(\frac{\mu}{1-\mu}\right)$$

and thus:

$$\mathbb{E}[X] = p(X=1) = \bar{\mu} \equiv \mu_{MLE} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}(x_n = 1)$$

and

$$\eta_{MLE} = \ln\left(\frac{\bar{\mu}}{1-\bar{\mu}}\right)$$

MLE and Kullback-Leibler Distance

- A useful property for the MLE (and not just a property for the exponential family of distributions) is the following:
- Minimizing the KL distance to the empirical distribution is equivalent to maximizing the likelihood.
- Indeed, let us consider the model logp(x|θ) and the empirical distribution:

$$p_{emp}(x) = \frac{1}{N}\sum_{n=1}^{N}\delta\left(x, x_n\right)$$

- We can then derive the following:

$$\sum_x p_{emp}(x)\log p(x\,|\,\theta) = \frac{1}{N}\sum_{n=1}^{N}\sum_x \delta\left(x, x_n\right)\log p(x\,|\,\theta) = \frac{1}{N}\sum_{n=1}^{N}\log p(x_n\,|\,\theta) = \frac{1}{N}\ell(\theta\,|\,\mathcal{D})$$

and from this:

$$KL\left(p_{emp}(x), p(x\,|\,\theta)\right) = \sum_x p_{emp}(x)\log\frac{p_{emp}(x)}{p(x\,|\,\theta)} = \sum_x p_{emp}(x)\log p_{emp}(x) - \sum_x p_{emp}(x)\log p(x\,|\,\theta)$$

$$= \sum_x p_{emp}(x)\log p_{emp}(x) - \frac{1}{N}\ell(\theta\,|\,\mathcal{D})$$

- Since the 1st term is independent of θ, the assertion is proved.

Conjugate Priors

- In general, for a given probability distribution $p(\mathbf{x}|\boldsymbol{\eta})$, 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,

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

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 $\boldsymbol{X}'$, 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:

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

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

Maximum Entropy and Exponential Family

- If nothing is known about a distribution except that it belongs to a certain class, then the distribution with the largest entropy should be chosen as the default.[a]
- The entropy is defined as
  - discrete case $\mathbb{H}(\pi) = -\sum_k \pi(\theta_k) \log\left(\pi(\theta_k)\right)$
- When some statistics (moments) of the distribution are known,

$$\mathbb{E}_\pi\left[g_k(\theta)\right] = w_k, k = 1,..,K$$

the maximum entropy distribution is of the form (λ's are the Lagrange multipliers enforcing the constraints):

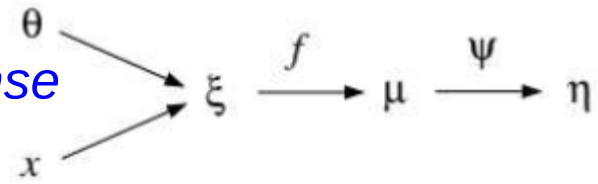
$$\pi(\theta_i) = \frac{\exp\left(-\sum_{k=1}^{K} \lambda_k g_k(\theta_i)\right)}{\sum_j \exp\left(-\sum_{k=1}^{K} \lambda_k g_k(\theta_j)\right)}, \lambda_k = Lagrange\ multipliers$$

- *Thus the MaxEnt distribution has the form of the exponential family.*

[a] C. P. Robert, *The Bayesian Choice*, *Springer, 2nd edition,* chapter 3 (full text available)

Generalized Linear Models

- We now study regression given data X and Y and a GLIM.
- We choose a particular conditional expectation of Y. We denote the modeled value of conditional expectation as $\mu = f(\theta^T x)$, $\xi = \boldsymbol{\theta}^T \boldsymbol{x}$.
- For linear regression, *GLIM extends these ideas beyond the Gaussian, Bernoulli and multinomial setting to the more general exponential family.*
- X enters linearly as $\theta^T x$ and *f is called a response function.* Ψ is a one-to-one map of μ to η.

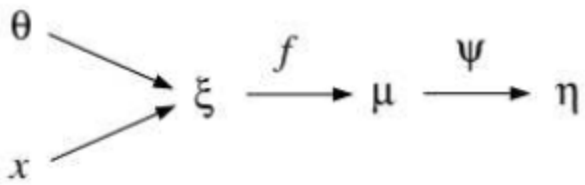


- To specify a GLIM we need (a) a choice of exponential family distribution, and (b) a choice of the response function $f(\cdot)$.
- Choosing the exponential family distribution is strongly constrained by the nature of the data.
- Note that $f(\cdot)$ needs to be both monotonic and differentiable. However*, there is a particular response function (canonical response function) that is uniquely associated with a given exponential family distribution.*

Canonical Response Function

- Canonical response function:

$$f(\cdot) = \Psi^{-1}(\cdot)$$
$$\xi = \eta$$



- *If we decide to use the canonical response function, the choice of the exponential family density completely determines the GLIM.*

$$\xi = f^{-1}(\mu) = \Psi(\mu) = \eta$$

MLE & Canonical Response Function

- Consider a regression problem with data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i), i=1,\ldots N\}$. *The log likelihood for a GLIM is as:*

$$\ell(\theta, \mathcal{D}) = \sum_{n=1}^{N} \log h(y_n) + \sum_{n=1}^{N} \underbrace{\eta_n \; y_n - A(\eta_n)}_{\psi(\mu_n)}, \; where: \; \mu_n = f(\xi_n) \; with \; \xi_n = \boldsymbol{\theta}^T \boldsymbol{x}_n$$

- *For a canonical response,* $\eta = \xi = \boldsymbol{\theta}^T \boldsymbol{x}$ *, and this is simplified as:*

$$\ell(\theta, \mathcal{D}) = \sum_{n=1}^{N} \log h(y_n) + \boldsymbol{\theta}^T \underbrace{\sum_{n=1}^{N} \boldsymbol{x}_n y_n}_{Sufficient\ statistic\ for\ \boldsymbol{\theta}} - \sum_{n=1}^{N} A(\eta_n)$$

- Regardless of *N*, the size of the sufficient statistic is fixed: the dimension of x_n - important reason for using canonical response.

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \mathcal{D}) = \sum_{n=1}^{N} (y_n - A'(\eta_n)) \nabla_{\boldsymbol{\theta}} \eta_n = \sum_{n=1}^{N} (y_n - \mu_n) \nabla_{\boldsymbol{\theta}} \eta_n = \sum_{n=1}^{N} (y_n - \mu_n) \underbrace{\boldsymbol{x}_n}_{\boldsymbol{x}_n} \; or \; \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \mathcal{D}) = \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{\mu})$$

- *This is a general expression for GLM with exponential family distributions and the canonical response function*.

Iterative Reweighted Least Squares (IRLS)

- The Hessian can now be computed from

$$\nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta},\mathcal{D}\right)=\sum_{n=1}^{N}\left(y_n-\mu_n\right)\boldsymbol{x}_n \text{ or } \nabla_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta},\mathcal{D}\right)=\boldsymbol{X}^T\left(\boldsymbol{y}-\boldsymbol{\mu}\right)$$

as:

$$\boldsymbol{H}=\nabla^2_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta},\mathcal{D}\right)=-\sum_{n=1}^{N}\frac{d\mu_n}{d\eta_n}\boldsymbol{x}_n\boldsymbol{x}_n^T \text{ or } \nabla^2_{\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta},\mathcal{D}\right)=-\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}, \text{where}: \boldsymbol{W}=\left\{\frac{d\mu_1}{d\eta_1},\ldots,\frac{d\mu_n}{d\eta_n}\right\}$$

- To estimate parameters in the canonical response function choice, one can use the iteratively reweighted least squares (IRLS) algorithm
- *The batch Newton algorithm now takes the familiar IRLS form:*

$$\boldsymbol{\theta}^{t+1}=\boldsymbol{\theta}^t+\left(\boldsymbol{X}^T\boldsymbol{W}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\left(\boldsymbol{y}-\boldsymbol{\mu}^t\right)=\left(\boldsymbol{X}^T\boldsymbol{W}^t\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{W}^t\boldsymbol{X}\boldsymbol{\theta}^t+\boldsymbol{X}^T\left(\boldsymbol{y}-\boldsymbol{\mu}^t\right)\right)$$

$$=\left(\boldsymbol{X}^T\boldsymbol{W}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^t\left(\underbrace{\boldsymbol{X}\boldsymbol{\theta}^t}_{\eta}+\boldsymbol{W}^{t^{-1}}\left(\boldsymbol{y}-\boldsymbol{\mu}^t\right)\right)=\left(\boldsymbol{X}^T\boldsymbol{W}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^t\left(\eta+\boldsymbol{W}^{t^{-1}}\left(\boldsymbol{y}-\boldsymbol{\mu}^t\right)\right)$$

- For non-canonical response functions, the Hessian has an extra term that contains the factor (y-μ). When we take expectations this term vanishes! So using the expected Hessian in the Newton method the algorithm looks essentially the same (Fisher Scoring algorithm).

Sequential Estimation - LMS

- An on-line estimation algorithm can be obtained by following the stochastic gradient of the log likelihood function.

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t} + \rho\left(y_n - \mu_n^t\right)x_n, \mu_n^t = f\left(\boldsymbol{\theta}^{t^T} x_n\right)$$

- If we do not use the canonical response function, then the gradient also includes the derivatives of $f(\cdot)$ and $\Psi(\cdot)$. These can be viewed as scaling coefficients that alter the step size, but otherwise leave the general LMS form intact.
- *The LMS algorithm is the generic stochastic gradient algorithm for models throughout the GLIM family.*

Bayesian Inference for the Multivariate Gaussian

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.

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

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

- The gamma distribution has a finite integral if a > 0, and the distribution itself is finite if a ≥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 dummy 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 $\beta > 0$

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

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

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

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

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:

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

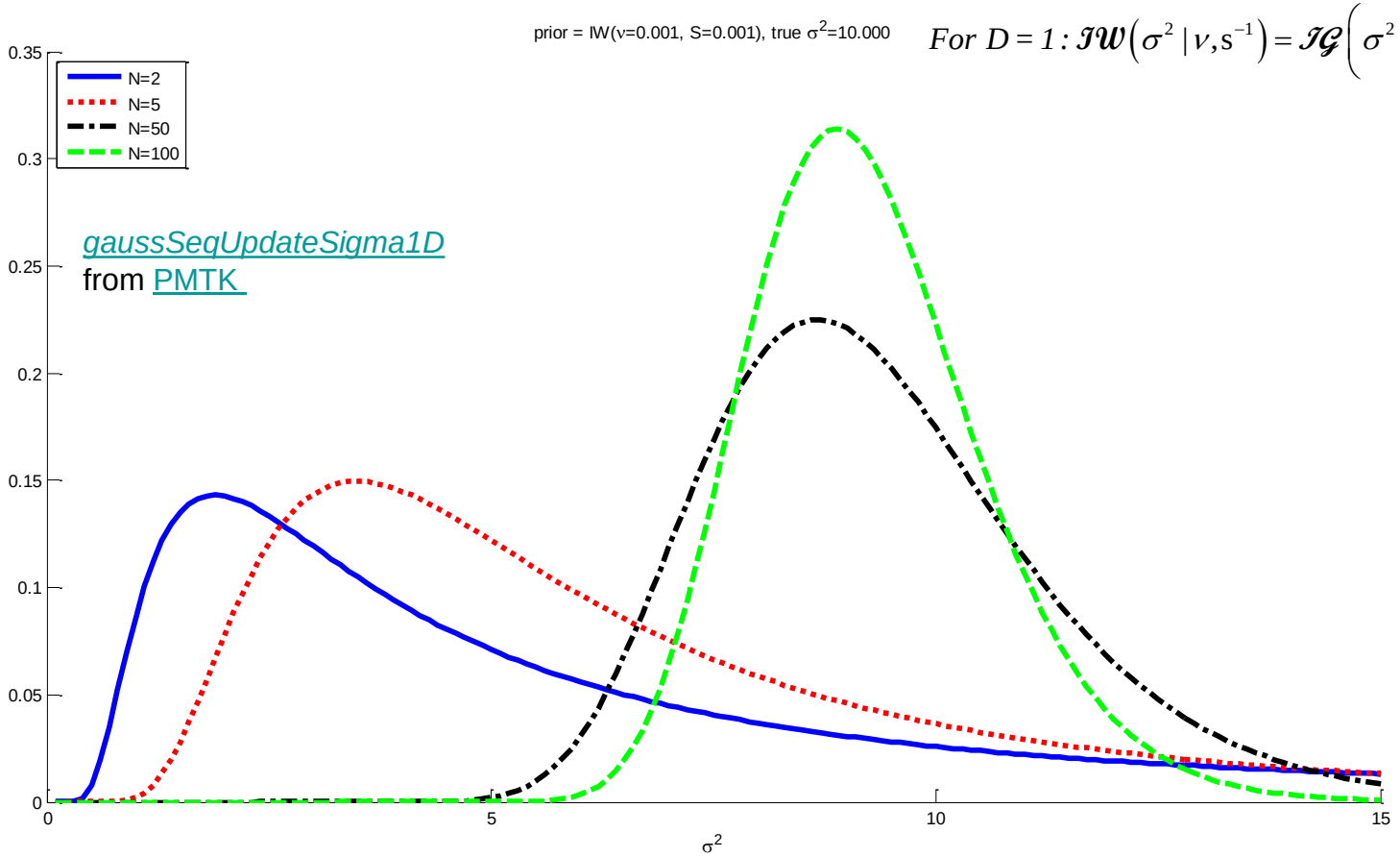
- The posterior dof ν_N is the prior dof plus N. The posterior sum of squares $\sigma_N^2\nu_N = \nu_0\sigma_0^2 + \sum_{i=1}^{N}\left(x_i - \mu\right)^2$ is the sum of the prior sum of squares plus the data sum of squares. An ***uninformative prior*** corresponds ***to zero virtual sample size, ν_0=0***. 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 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:

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

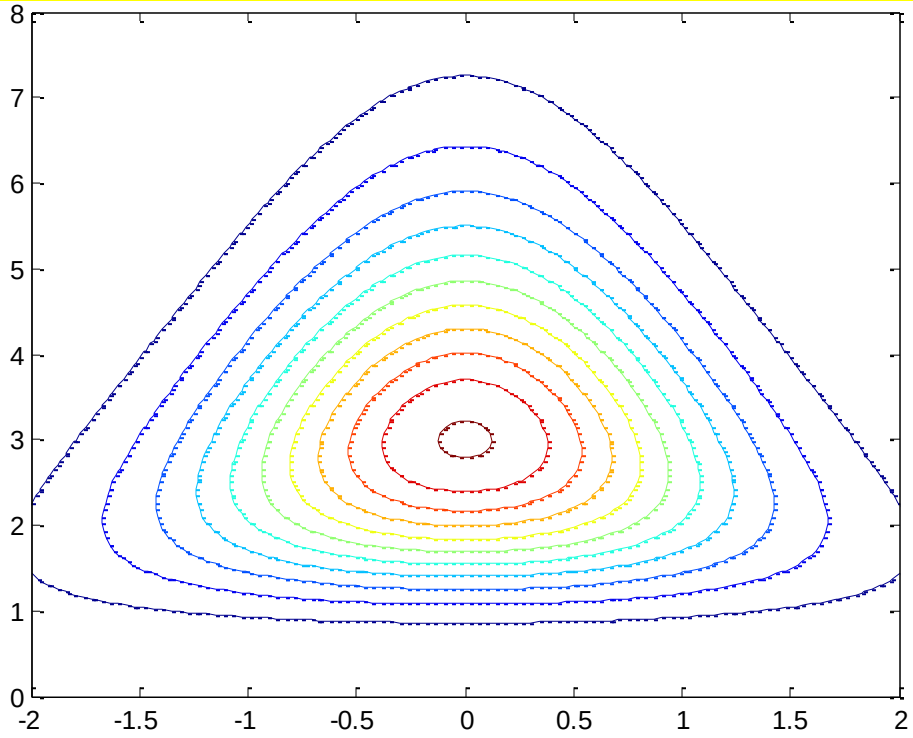
$$\left(\lambda(N+\beta)\right)^{1/2}\exp\left(-\frac{\lambda\left(N+\beta\right)}{2}\left(\mu-\frac{\beta\mu_0+\sum_{n=1}^{N}x_n}{N+\beta}\right)^2\right)$$

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

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

The Normal-Gamma Distribution

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



$$a = 5, b = 6,$$
$$\mu_0 = 0, \beta = 2$$

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:

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

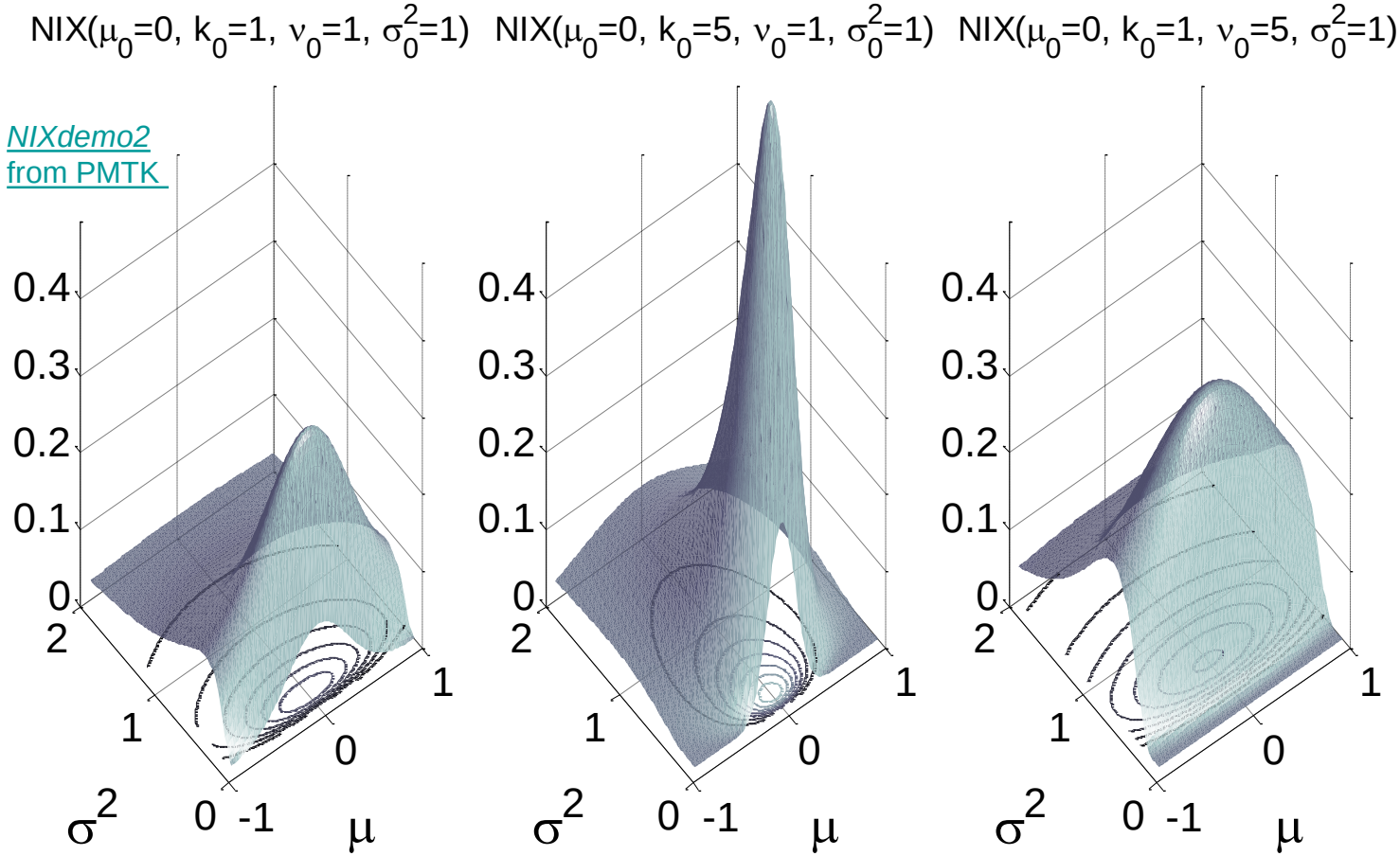
$$\kappa_N=\kappa_0+N,\nu_N=\nu_0+N,\nu_N\sigma_N^2=\nu_0\sigma_0^2+\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2+\frac{N\kappa_0}{\kappa_0+N}\left(m_0-\bar{x}\right)^2$$

- 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



- The $NI\chi^2(m_0, \kappa_0, v_0, \sigma^2_0)$ distribution. m_0 is the prior mean and κ_0 is how strongly we believe this; σ^2_0 is the prior variance and v_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*:

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

- 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: $I_{0.95}[\mu \mid \mathcal{D}] \approx \bar{x} \pm 2\frac{s}{\sqrt{N}}$

Bayesian T-Test

- We want to test the hypothesis $\mu \neq \mu_0$ for some known value μ_0 (often 0), given $x_i \sim \mathcal{N}(\mu, \sigma^2)$. *This is called a two-sided, one-sample* ***t-test****.*

- Check if $\mu_0 \in I_{0.95}(\mu|\mathcal{D})$. If it is not, then we can be 95% sure that $\mu \neq \mu_0$.

- A more common scenario is when we want to test if two paired samples have the same mean. Suppose $y_i \sim \mathcal{N}(\mu_1, \sigma^2)$ and $z_i \sim \mathcal{N}(\mu_2, \sigma^2)$. We want to determine if $\mu = \mu_1 - \mu_2 > 0$, using $x_i = y_i - z_i$ as our data. We can evaluate this as follows:

$$p(\mu > \mu_0 \mid \mathcal{D}) = \int_{\mu_0}^{\infty} p(\mu \mid \mathcal{D}) d\mu$$

- This is called *a one-sided,* ***paired t-test***. To calculate the posterior, we must specify a prior. Suppose we use an uninformative prior. As shown earlier, the posterior marginal is:

$$p(\mu \mid \mathcal{D}) = \mathcal{T}\left(\mu \mid \bar{x}, \frac{s^2}{N}, N-1\right)$$

Bayesian T-Test

- We *define the t statistic* as: $t = \frac{\mu_0 - \bar{x}}{s/\sqrt{N}}$
- The denominator is the standard error of the mean. With this definition note that

$$p(\mu > \mu_0 \mid \mathcal{D}) = \int_{\mu_0}^{\infty} p(\mu \mid \mathcal{D})\, d\mu = P\left(\frac{\mu - \bar{x}}{s/\sqrt{N}} > \frac{\mu_0 - \bar{x}}{s/\sqrt{N}} \right) = 1 - F_{N-1}(t)$$

  where *$F_v(t)$ is the CDF of the standard Student's $\mathcal{T}$ distribution* $\mathcal{T}(0,1,v)$

- *Note:* The posterior of μ has a form $\frac{\mu - \bar{x}}{s/\sqrt{N}} \mid \mathcal{D} \sim \mathcal{T}_{N-1}$. From a frequentist point of view*, this is identical to the sampling distribution of the MLE:* $\frac{\mu - \bar{x}}{s/\sqrt{N}} \mid \mu \sim \mathcal{T}_{N-1}$. Thus *the one sided p value in a frequentist test is the same as the Bayesian estimate* $p(\mu > \mu_0 \mid \mathcal{D})$*. The interpretation of the results in the two approaches is however very different.*

bayesTtestDemo from PMTK

- Box, G. and G. Tiao (1973). *Bayesian inference in statistical analysis*. Addison-Wesley.
- Gonen, M., W. Johnson, Y. Lu, and P. Westfall (2005, August). The Bayesian Two-Sample t Test. *The American Statistician 59*(3), 252–257.
- Rouder, J., P. Speckman, D. Sun, and R. Morey (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Pyschonomic Bulletin & Review 16*(2), 225–237.

Sensor Fusion with Unknown Parameters

- Consider a sensor fusion problem where *the precision of each measurement device is unknown.*
- *The unknown precision case turns* out to give qualitatively different results from the case of known precision, *yielding a potentially multi-modal posterior.*
- Suppose we want to ***pool data from two sources*** *x and y to estimate some quantity* $\mu \in R$, ***but the reliability of the sources is unknown***. Specifically, suppose we have two different measurement

$$x_i \mid \mu \sim \mathcal{N}\left(\mu, \lambda_x^{-1}\right),\ y_i \mid \mu \sim \mathcal{N}\left(\mu, \lambda_y^{-1}\right)$$

- We make two independent measurements with each device, which turn out to be $x_1 = 1.1$, $x_2 = 1.9$, $y_1 = 2.9$, $y_2 = 4.1$
- We use a non-informative prior for μ, $p(\mu) \propto 1$, which we can emulate using a Gaussian,

$$p(\mu) = \mathcal{N}\left(\mu \mid m_0 = 0, \lambda_0^{-1} = \infty\right)$$

- Minka, T. (2000). Estimating a Dirichlet distribution. Technical report, MIT.

Sensor Fusion with Unknown Parameters

- *For known precisions*, the posterior is Gaussian:

$$p(\mu \mid \mathcal{D}) = \mathcal{N}\left(\mu \mid m_N, \lambda_N^{-1}\right),$$

$$m_N = \frac{\lambda_x N_x \bar{x} + \lambda_y N_y \bar{y}}{\lambda_x N_x + \lambda_y N_y}, \; N_x = N_y = 2, \bar{x} = \frac{\sum_{i=1}^{N_x} x_i}{N_x} = 1.5, \bar{y} = \frac{\sum_{i=1}^{N_y} y_i}{N_y} = 3.5$$

$$\lambda_N = \lambda_0 + \lambda_x N_x + \lambda_y N_y$$

- However, the measurement precisions are not known. Initially we will estimate them by MLE. The log-likelihood is given by

$$\ell\left(\mu, \lambda_x, \lambda_y\right) = \frac{N_x}{2}\log\lambda_x - \frac{\lambda_x}{2}\sum_{i=1}^{N}\left(x_i - \mu\right)^2 + \frac{N_y}{2}\log\lambda_y - \frac{\lambda_y}{2}\sum_{i=1}^{N}\left(y_i - \mu\right)^2 + const.$$

- The MLE is obtained by solving the following coupled equations:

$$\frac{1}{\bar{\lambda}_x} = \frac{1}{N_x}\sum_{i=1}^{N_x}\left(x_i - \bar{\mu}\right)^2, \frac{1}{\bar{\lambda}_y} = \frac{1}{N_y}\sum_{i=1}^{N_y}\left(y_i - \bar{\mu}\right)^2, \bar{\mu} = \frac{\bar{\lambda}_x N_x \bar{x} + \bar{\lambda}_y N_y \bar{y}}{\bar{\lambda}_x N_x + \bar{\lambda}_y N_y}$$

Sensor Fusion with Unknown Parameters

$$\frac{1}{\bar{\lambda}_x} = \frac{1}{N_x}\sum_{i=1}^{N_x}\left(x_i - \bar{\mu}\right)^2, \frac{1}{\bar{\lambda}_y} = \frac{1}{N_y}\sum_{i=1}^{N_y}\left(y_i - \bar{\mu}\right)^2, \bar{\mu} = \frac{\bar{\lambda}_x N_x \bar{x} + \bar{\lambda}_y N_y \bar{y}}{\bar{\lambda}_x N_x + \bar{\lambda}_y N_y}$$

- We *solve these equs by iteration* starting with

$$\bar{\lambda}_x = \frac{1}{s_x^2} = \frac{N_x}{\sum_{i=1}^{N_x}\left(x_i - \bar{x}\right)^2} = \frac{1}{0.16}, \bar{\lambda}_y = \frac{1}{s_y^2} = \frac{N_y}{\sum_{i=1}^{N_y}\left(y_i - \bar{y}\right)^2} = \frac{1}{0.36}$$

- Upon convergence,

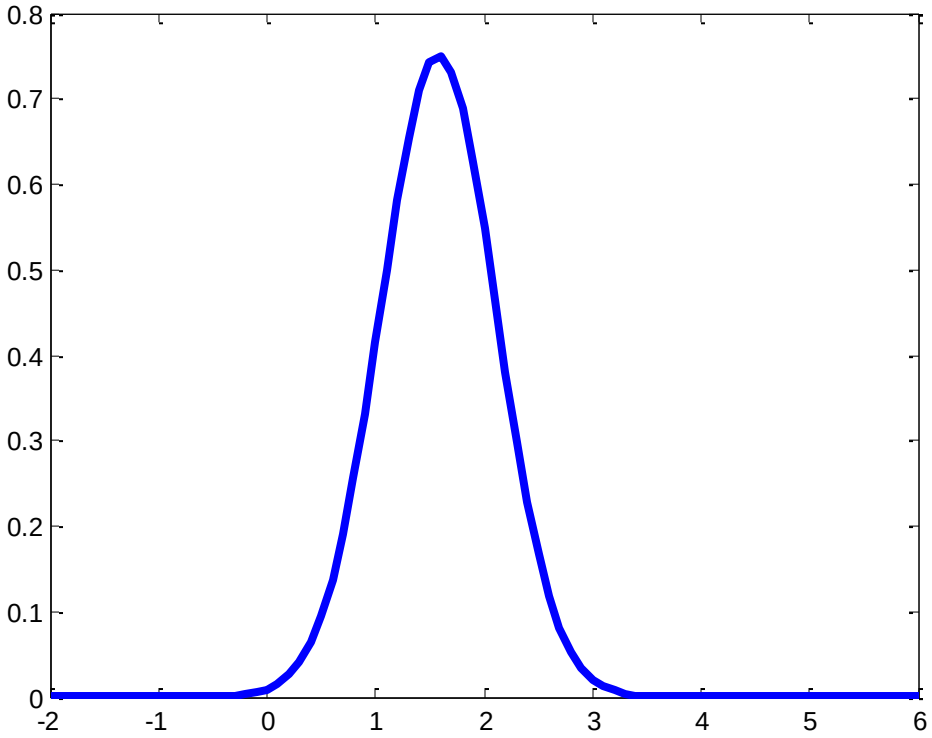
$$\bar{\lambda}_x = \frac{1}{0.1662}, \bar{\lambda}_y = \frac{1}{4.0509}, \; p\left(\mu \mid \mathcal{D}, \bar{\lambda}_x, \bar{\lambda}_y\right) = \mathcal{N}\left(\mu \mid 1.5788, 0.0798\right)$$

- The plug-in approximation to the posterior is plotted next.

Plug-in Approximation to the Posterior

- Posterior for μ. Plug-in approximation.

sensorFusionUnknownPrec
from Kevin Murphys' PMTK



- This weights each sensor according to its estimated precision.
- Since sensor *y* was estimated to be much less reliable than sensor *x*, we have

$$\mathbb{E}\left(\mu \mid \mathcal{D}, \bar{\lambda}_x, \bar{\lambda}_y\right) \approx \bar{x}$$

- Effectively with this approximation we ignore the *y* sensor.

Sensor Fusion with Unknown Parameters

- We will adopt *a Bayesian approach and integrate out the unknown precisions, rather than trying to estimate them*. That is, we compute

$$p(\mu \mid \mathcal{D}) \propto p(\mu) \int p(\mathcal{D}_x \mid \mu, \lambda_x)\, p(\lambda_x \mid \mu)\, d\lambda_x \int p(\mathcal{D}_y \mid \mu, \lambda_y)\, p(\lambda_y \mid \mu)\, d\lambda_y$$

- We will use *uninformative Jeffrey's priors*, $p(\mu) \propto 1$, $p(\lambda_x|\mu) \propto 1/\lambda_x$ and $p(\lambda_y|\mu) \propto 1/\lambda_y$.

- The first integral becomes using the likelihood derived earlier:

$$p(\boldsymbol{X} \mid \mu, \lambda) = \left[\lambda^{1/2} \exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left(\lambda\mu \sum_{n=1}^{N} x_n - \frac{1}{2}\lambda \sum_{n=1}^{N} x_n^2\right), \text{ where } \sum_{n=1}^{N} x_n^2 = N\left(s^2 + \bar{x}^2\right)$$

$$I = \int p(\mathcal{D}_x \mid \mu, \lambda_x)\, p(\lambda_x \mid \mu)\, d\lambda_x \propto \int \frac{1}{\lambda_x} \left(N_x \lambda_x\right)^{N_x/2} \exp\left(-\frac{N_x}{2}\lambda_x \left(\bar{x} - \mu\right)^2 - \frac{N_x}{2} s_x^2 \lambda_x\right) d\lambda_x$$

- For N_x=2, it simplifies to the normalizing factor of a Gamma distribution

$$I \propto \int \lambda_x^{1-1} \exp\left(-\lambda_x \underbrace{\left(\left(\bar{x} - \mu\right)^2 + s_x^2\right)}_{b}\right) d\lambda_x = \int \lambda_x^{a-1} \exp\left(-\lambda_x b\right) d\lambda_x = \Gamma(a) b^{-a}$$

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

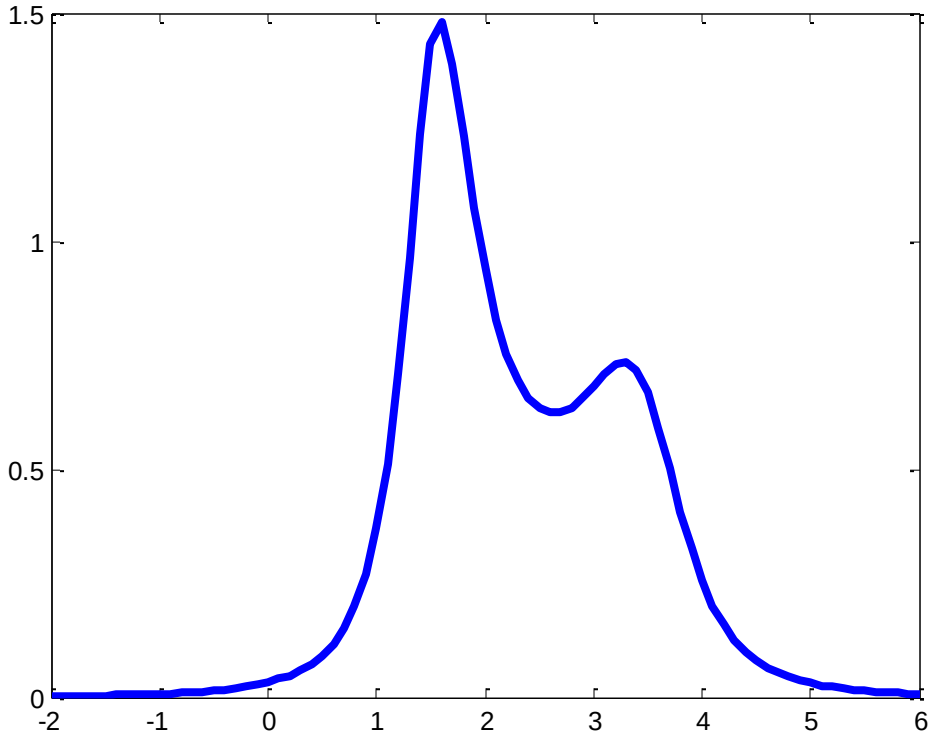
$$I \propto \left(\left(\bar{x} - \mu\right)^2 + s_x^2\right)^{-1}$$

Sensor Fusion with Unknown Parameters

- Finally:

$$p(\mu \mid \mathcal{D}) \propto \frac{1}{(\bar{x}-\mu)^2 + s_x^2} \frac{1}{(\bar{y}-\mu)^2 + s_y^2}$$

- The exact posterior is plotted below.



sensorFusionUnknownPrec
from Kevin Murphys' PMTK

- The posterior has two modes at

$$\bar{x} = 1.5, \bar{y} = 3.5$$

- *The weight of the 1st mode is larger, since the data from the x sensor agree more with each other. It seems likely that the x sensor is the reliable one.*
- The Bayesian solution makes it possible that the *y* sensor is the more reliable one; from two measurements, we cannot tell, and choosing just the *x* sensor, as the plug-in approximation does, results in over confidence (a narrow posterior)

Multivariate Gaussian: Posterior of μ

- Consider a known variance Σ and a Gaussian prior $\mathcal{N}(\mu,\Sigma_0)$ with the posterior for the unknown mean μ taking the form:

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

- This posterior is the exponential of a quadratic in μ:

$$-\frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)-\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\mu)^T\Sigma^{-1}(\boldsymbol{x}_n-\mu)=$$

$$-\frac{1}{2}\mu^T\underbrace{\left(\Sigma_0^{-1}+N\Sigma^{-1}\right)}_{\Sigma_N^{-1}}\mu+\mu^T\underbrace{\left(\Sigma_0^{-1}\mu_0+\Sigma^{-1}\sum_{n=1}^{N}\boldsymbol{x}_n\right)}_{\Sigma_N^{-1}\mu_N}+const$$

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

$$\Sigma_N^{-1}=\Sigma_0^{-1}+N\Sigma^{-1},$$

$$\mu_N=\left(\Sigma_0^{-1}+N\Sigma^{-1}\right)^{-1}\left(\Sigma_0^{-1}\mu_0+\Sigma^{-1}\sum_{n=1}^{N}\boldsymbol{x}_n\right)=\left(\Sigma_0^{-1}+N\Sigma^{-1}\right)^{-1}\left(\Sigma_0^{-1}\mu_0+N\Sigma^{-1}\mu_{ML}\right)$$

- For uninformative prior, $\Sigma_0=\infty\boldsymbol{I}\Rightarrow p(\mu \mid \boldsymbol{X},\Sigma)\to\mathcal{N}\left(\mu \mid \mu_{ML},\frac{1}{N}\Sigma\right)$

Posterior Distribution of Precision Λ

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

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

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

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

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

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

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

$$For\ D=1, \mathcal{Wi}\left(\lambda\,|\,v,\mathrm{s}^{-1}\right) = \mathcal{Gamma}\left(\lambda\,|\,\frac{v}{2},\frac{\mathrm{s}}{2}\right)$$

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 $\mathbf{S}_0$.

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

- 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 $v_N = v_0 + N$, is the prior strength v_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

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

we conclude that the MAP estimate is:

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

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

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

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

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, $\mathbf{S}_0$, 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:

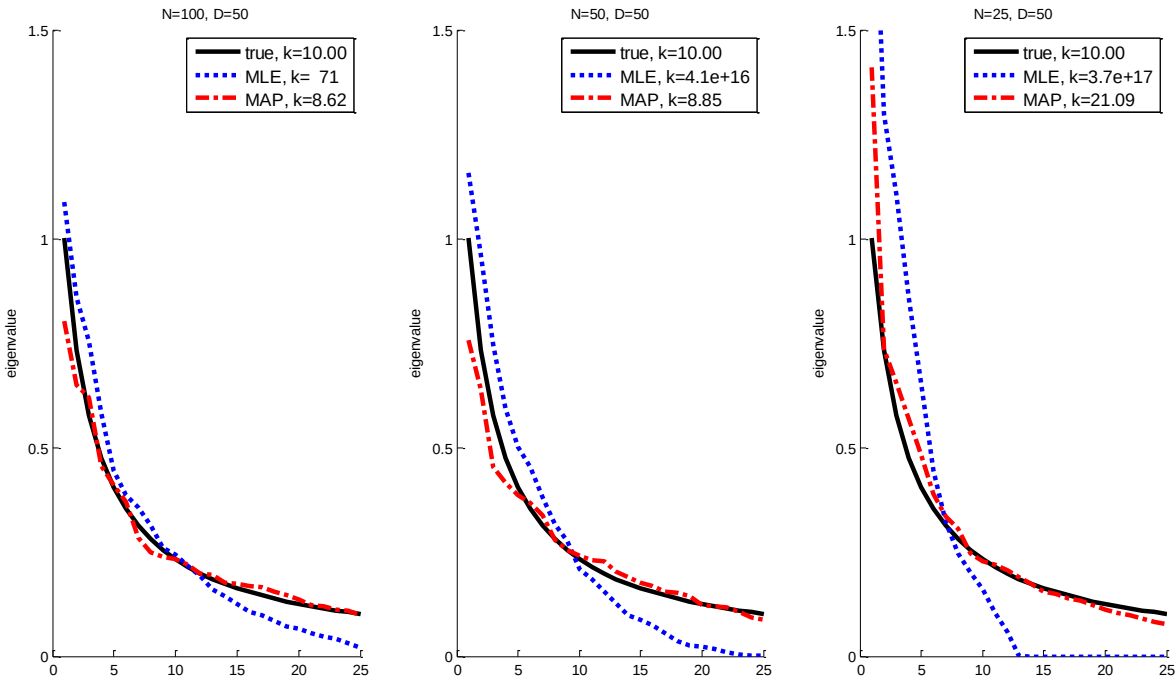
$$\bar{\Sigma}_{MAP} = \begin{cases} \bar{\Sigma}_{MLE}(i,j) \; if \; i = j \\ (1-\lambda)\bar{\Sigma}_{MLE}(i,j) \; otherwise \end{cases}$$

- 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 $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N \sim$(i.i.d) $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$. 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

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

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 *U* has the Wishart distribution, then U^{-1} has the inverse-Wishart distribution. *The resulting* $p(\mu, \Lambda | \boldsymbol{\mu}_0, \kappa, \mathbf{T}, \nu)$ is the ***Gaussian-Wishart distribution***.
- The quantity ν *is a positive scalar, while* $\mathbf{T}$ *is a* positive definite matrix. They play roles analogous to α and β, respectively, in the Gamma *distribution*.
- Other parameters of the prior are the mean vector $\boldsymbol{\mu}_0$ and $\boldsymbol{\kappa}$ which represents the `a priori number of observations'.

Inference for Both μ and Λ

- The likelihood and prior distributions are given explicitly as:

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

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

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

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

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

- Combining gives:

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

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

- 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 μ 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:

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

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

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

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

Inference for Both μ and Λ

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

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

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

- The posterior marginals can be derived as:

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

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

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

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

Inference for Both μ and Λ

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

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

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

- 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: $p(\boldsymbol{x} \mid \mathcal{D}) = \mathcal{T}_{\nu_N - D + 1}\left(\boldsymbol{\mu}_N, \underbrace{\frac{\kappa_N (\nu_N - D + 1)\boldsymbol{T}_N}{(\kappa_N + 1)}}_{\text{precision}}\right)$
- The marginal likelihood can be computed as a ratio of normalization constants:

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

- 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:

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

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

Inference for μ and Σ

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

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

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

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

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

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

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

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

Inference for μ and Σ

$$p(\mu,\Sigma\mid\mathcal{D})\propto|\Sigma|^{-\frac{N}{2}-\frac{1}{2}-\frac{\nu_0+D+1}{2}}\exp\left(-\frac{1}{2}\Sigma^{-1}\left(S_{\bar{x}}+S_0+N(\mu-\bar{x})(\mu-\bar{x})^T+\kappa_0(\mu-\mu_0)(\mu-\mu_0)^T\right)\right)$$

$$\propto|\Sigma|^{-\frac{\nu_0+D+2+N}{2}}\exp\left(-\frac{1}{2}\Sigma^{-1}\left(S_{\bar{x}}+S_0+N(\mu-\bar{x})(\mu-\bar{x})^T+\kappa_0(\mu-\mu_0)(\mu-\mu_0)^T\right)\right)$$

- 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:

$$p(\mu,\Sigma\mid\mathcal{D})\propto|\Sigma|^{-\frac{\nu_N+D+2}{2}}\exp\left(-\frac{1}{2}\Sigma^{-1}\left(S_{\bar{x}}+S_0+\kappa_N(\mu-\mu_N)(\mu-\mu_N)^T+\frac{\kappa_0 N}{\kappa_N}(\bar{x}-\mu_0)(\bar{x}-\mu_0)^T\right)\right),$$

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

The Posterior of μ and Σ

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

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

$$\boldsymbol{\mu}_N = \frac{\kappa_0\boldsymbol{\mu}_0 + N\bar{\boldsymbol{x}}}{\kappa_N} = \frac{\kappa_0}{\kappa_0+N}\boldsymbol{\mu}_0 + \frac{N}{\kappa_0+N}\bar{\boldsymbol{x}}$$

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

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

- *The posterior mean is a convex combination of the prior mean and the MLE with strength κ_0+N.*

- *The posterior scatter matrix $\boldsymbol{S}_N$ is the prior scatter matrix $\boldsymbol{S}_0$ plus the empirical scatter matrix $\boldsymbol{S}_{\bar{x}}$ 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 κ_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 computed earlier – 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 $\mathbf{S}_N$ 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 $p(\mathbf{x}|\mathcal{D})=p(\mathbf{x},\mathcal{D})/p(\mathcal{D})$ can be evaluated as:

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

- 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:

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

- 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

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

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

- 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:

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

- Also the posterior predictive is:

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

- 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

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

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

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

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

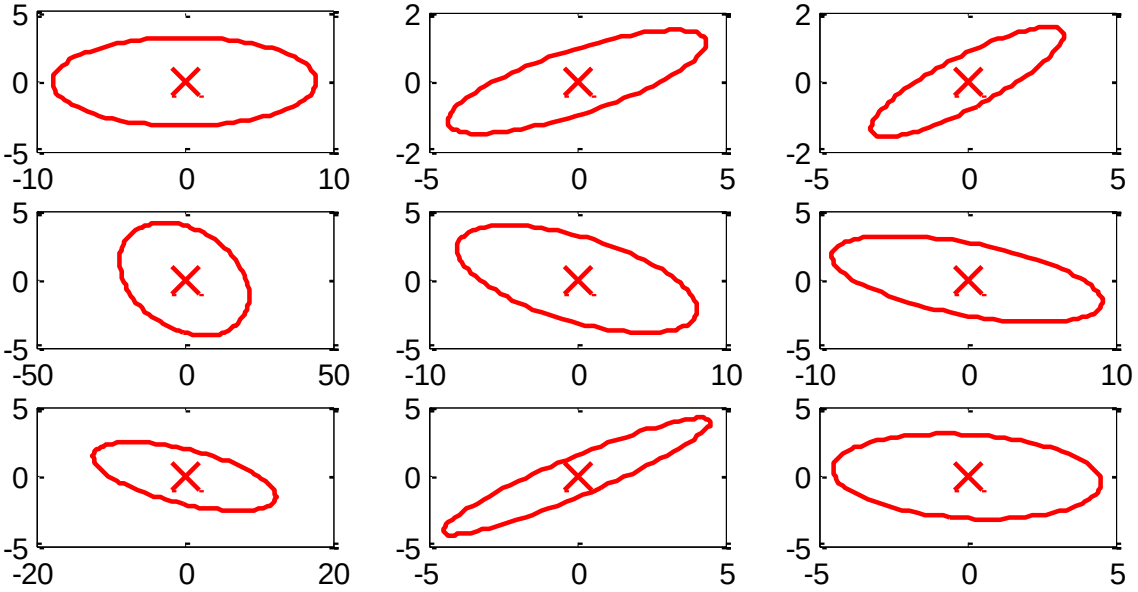
$$\mu_0 = \bar{x}, \kappa_0 = 0.01$$

- 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



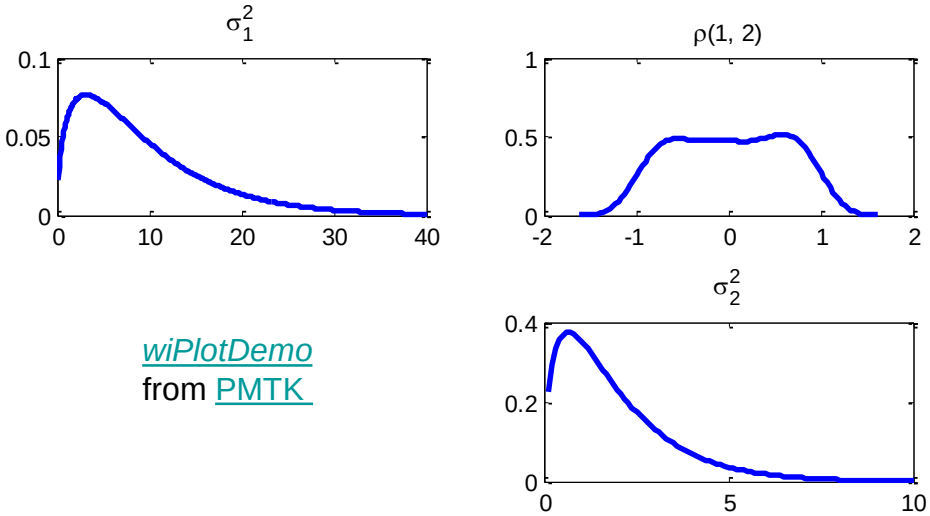
- Above: Samples from $\mathbf{\Sigma} \sim \text{Wi}(\mathbf{S}, \nu)$, where $\mathbf{S}=3.1653, -0.0262; -0.0262, 0.6477]$ and $\nu = 3$.
- The sampled matrices are highly variable, and some are nearly singular. As ν increases, the sampled matrices are more concentrated on the prior $\mathbf{S}$.

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

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

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



wiPlotDemo from PMTK

Gaussian Mixtures

Mixture of Gaussians: Challenges

- Consider a superposition of *K* Gaussian densities as follows:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x} \middle| \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \quad 0 \le \pi_k \le 1, \quad \sum_{k=1}^{K} \pi_k = 1$$

- The last condition comes from the normalization of p(**x**).

- The Log likelihood function takes the form

$$\ln p\left(\mathcal{D} \middle| \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \left\{ \pi_k \mathcal{N}\left(x_n \middle| \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right\}$$

- Note **that the sum over components appears *inside* the log**

- The unknown parameters are π_k, k=1,..K, and $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$, k=1,.. K. There is no closed form MLE solution. The EM (Expectation-Maximization) algorithm provides a solution (to be discussed in detail in a forthcoming lecture).

Posterior Responsibilities

- Using the sum and product rules we can also write:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} p(k)\, \underbrace{p(\boldsymbol{x} \mid k)}_{\mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)} = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right),$$

$$p(k) = \textit{prior probability for model } k$$

- We will see in follow up lectures that of significant interest are the posterior probabilities (responsibilities)

$$\gamma_k(\boldsymbol{x}) = p(k \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid k)\, p(k)}{\sum_{l=1}^{K} p(\boldsymbol{x} \mid l)\, p(l)} = \frac{\mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \pi_k}{\sum_{l=1}^{K} \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right) \pi_l}$$

MLE of Mixture Model: EM Algorithm

- Let us return to the log-likelihood: $\ln L = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}_{nk} \right\}$

$$where: \mathcal{N}_{nk} \equiv \mathcal{N}\left(x_n \middle| \mu_k, \Sigma_k\right)$$

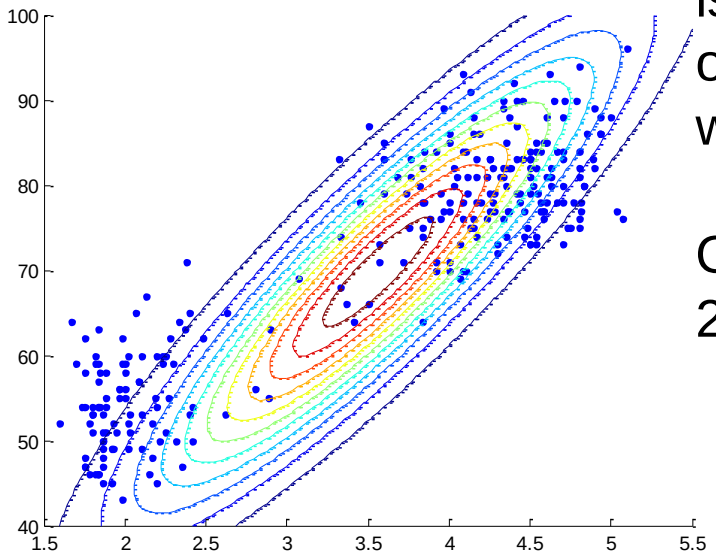
$$0 = \frac{\partial \ln L}{\partial \mu_j} = \sum_{n=1}^{N} \underbrace{\frac{\pi_j \mathcal{N}_{nj}}{\sum_k \pi_k \mathcal{N}_{nk}}}_{\text{Responsibilities: } \gamma_j(x_n) \equiv \gamma_{jn}} \frac{1}{\mathcal{N}_{nj}} \frac{\partial \mathcal{N}_{nj}}{\partial \mu_j} = \sum_{n=1}^{N} \gamma_j(x_n) \frac{\partial \ln \mathcal{N}_{nj}}{\partial \mu_j}$$

$$= -\Sigma_j^{-1} \sum_{n=1}^{N} \gamma_j(x_n)\left(x_n - \mu_j\right) \Rightarrow \mu_j = \frac{\sum_{n=1}^{N} \gamma_j(x_n) x_n}{\sum_{n=1}^{N} \gamma_j(x_n)} \quad (\textit{mean of the } j \textit{ component})$$

- Note that we don't have a closed form solution since the γ_j's contain the μ_j's in a complex way.

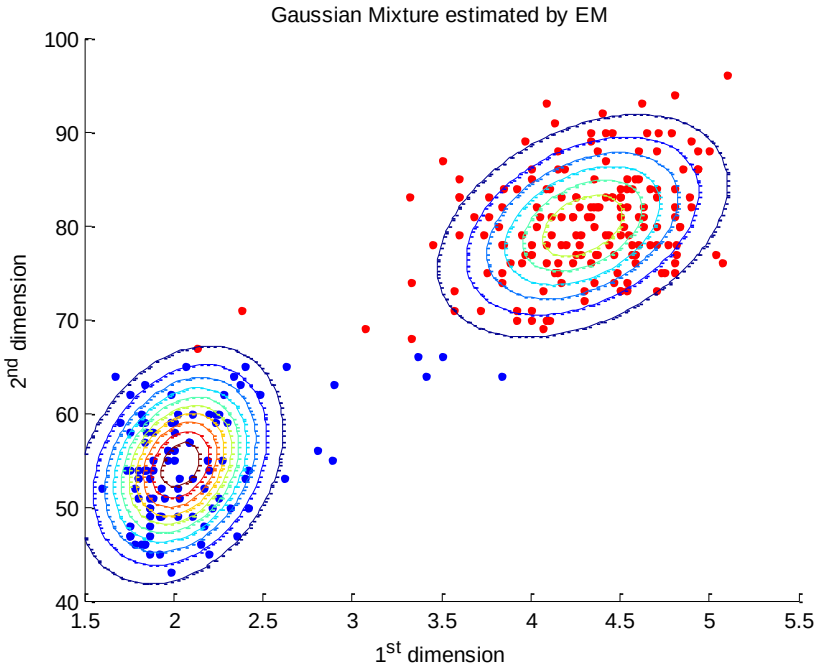
- This is the 1st step in the EM algorithm to be examined later on. We will see that the responsibilities γ_j represent posterior probabilities.

Gaussian Mixture Example



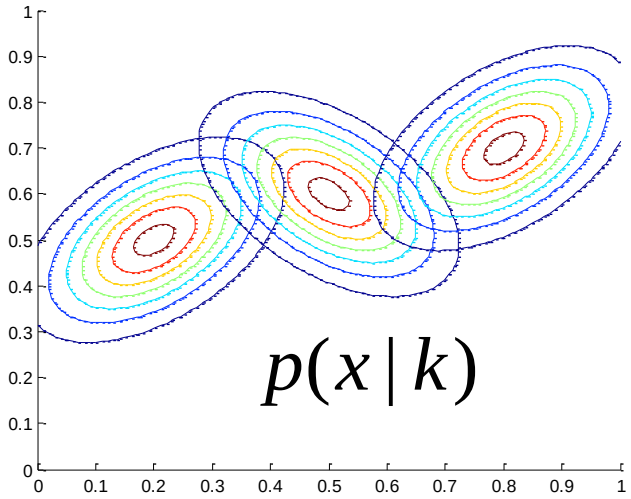
Old Faithful data. On the left a single Gaussian is fit using MLE. The approximation places much of the probability mass on the central region where the data are sparse.

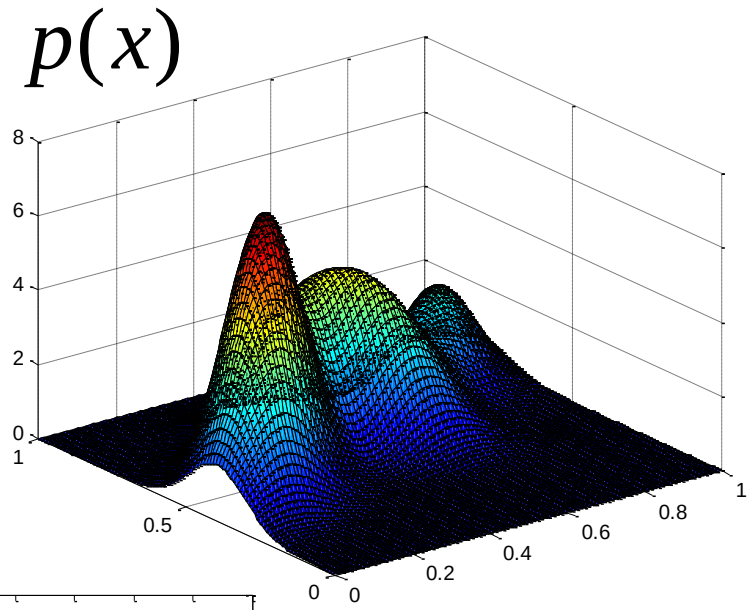
On the plot below the EM algorithm is used to fit 2 Gaussians to the same data



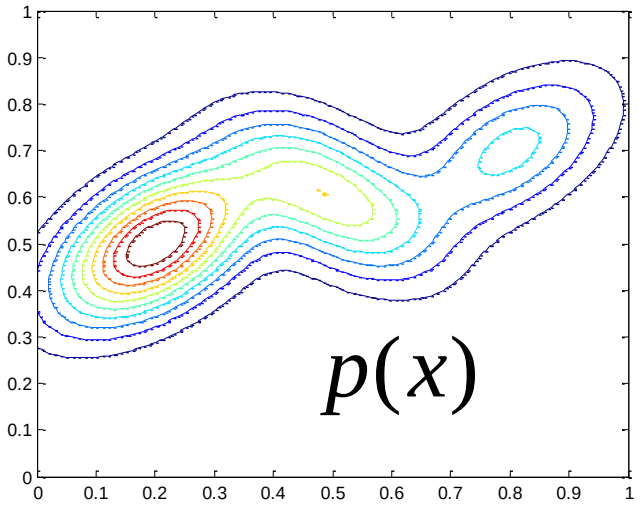
Run main file in this MatLab directory

Gaussian Mixture Example





Mixture of 3 Gaussians in a 2D space. (a) Contours of constant density for each of the mixture components. (b) Contours of the marginal probability density $p(\boldsymbol{x})$ of the mixture distribution. (c) A surface plot of the distribution $p(\boldsymbol{x})$.



MatLab Code