Exponential Family

Contents

1 The Exponential Family

We are interested in the exponential family primarily because it makes inference easier. When a problem can be cast within the exponential family framework, inference can be tied to general principles, and parameter updates often have nice interpretations. This is true regardless of whether we're doing frequentist inference (such as maximum likelihood) or Bayesian inference. Bayesian inference with exponential family likelihoods tends to be especially nice, as all exponential family likelihoods have conjugate priors, and these are often also in the exponential family [**?**]. [1] More complicated models may not be in the exponential family, but may have exponential family complete conditional distributions; in such situation, we can appeal to exponential family formalisms to more easily work out inference schemes for expectation maximization, variational inference, or Gibbs sampling.

1.1 Definition

We define an *exponential family* of probability distributions as those distributions whose density has the following form

$$p(x \mid \theta) = h(x) \exp\{\eta(\theta)^T t(x) - a(\theta)\} \tag{1.1.1}$$

where we refer to h as the base measure, η as the natural parameter, t as the sufficient statistics, and a as the log normalizer. [2]

Remark 1.1.1. (*Non-uniqueness of natural parameter*) Note from (1.1.1) that natural parameters are not unique since, for example, η could be multiplied by a non-zero constant c if $t(x)$ is divided by c. [3] Thus, we should speak of *a* natural parameter, rather than *the* natural parameter.

1.2 Examples

1.2.1 Dirichlet distribution

Example 1.2.1. (Dirichlet Distribution) We can write the density of the Dirichlet distribution in exponential family form:

$$\begin{aligned} p(\pi \mid \alpha) &= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \pi_1^{\alpha_1 - 1} \cdots \pi_K^{\alpha_K - 1} \\ &= \exp\left\{ \sum_{k=1}^{K} (\alpha_k - 1) \log \pi_k - \left[\sum_k \log \Gamma(\alpha_k) - \log \Gamma(\sum \alpha_k) \right] \right\} \end{aligned}$$

with natural parameter $\eta(\alpha) = [\alpha_1 - 1, ..., \alpha_K - 1]^T$, sufficient statistics $t(\pi) = \log \pi = [\log \pi_1, ..., \log \pi_K]^T$, base measure $h(\pi) = 1$, and log normalizer $a(\alpha) = \sum_k \log \Gamma(\alpha_k) - \log \Gamma(\sum_k \alpha_k)$. □

For an example of how the natural parametrization can help provide insight into message passing, see Section A.

Remark 1.2.1. The exponential family representation of the Dirichlet, as given in Example 1.2.1, is useful when we want to compute the expectation of a log probability from a Dirichlet distributed probability vector (as happens in the derivation of LDA with variational inference; see my notes on variational inference).

[1] TODO: Get clearer on the relationship. There is a brief discussion on this in [**?**]. I am not clear on whether all likelihoods with conjugate priors need to be in the exponential family.

[2] TODO: It would be helpful to get more solid on integrating against probability measure here, so that I can set this up in a more precise way, as Jordan does. He remarks on this somewhere in his exponential family lecture notes. What also may be helpful is this beautiful excerpt from pp.38 of [**?**]: " [...] we represent the probability distribution as a density p absolutely continuous with respect to some measure η. This base measure η might be the counting measure on $\{0, 1, ..., r-1\}$, in which case p is a probability mass function; alternatively, for a continuous random vector, the base measure η could be the ordinary Lebesgue measure on $\mathbb{R}$."

[3] Are they unique up to scalar multiplication?

In those notes, we see

$$\begin{aligned}\mathbb{E}[\log \pi_k] = \mathbb{E}[t_k(p)] &\overset{1}{=} \frac{\partial}{\partial \eta_k} a(\eta) \\ &= \Psi(\alpha_k) - \Psi(\sum_k \alpha_k) \end{aligned} \tag{1.2.1}$$

where (1) uses a well-known exponential family property and where $\Psi(\cdot)$ is the first derivative of the $\log \Gamma$ function. It is known as the *digamma function*. □

TODO: Add multivariate Gaussian example, showing that the natural parameters are the precision Σ^{-1} and precision-weighted mean $\Sigma^{-1}\mu$, as we use this in Section 5.1 combined with the exponential family formalism to derive the updates to the mean for a Bayesian normal model with conditionally conjugate prior.

1.2.2 Truncated normal distribution

Example 1.2.2. (Truncated normal distribution) The univariate truncated normal distribution $\mathcal{TN}(\mu, \sigma^2, \Omega)$ results when a normal distribution $\mathcal{N}(\mu, \sigma^2)$ is truncated to some set $\Omega \in \mathbb{R}$. [4] Note that the parameters μ, σ^2 denote the mean and variance of the *parent* normal distribution; i.e. if $X \sim \mathcal{TN}(\mu, \sigma^2, \Omega)$ then $\mathbb{E}[X] \neq \mu$ (unless $\Omega = \mathbb{R}$).

If we assume that the truncation set is an interval $\Omega = (a, b)$ for $a, b \in \mathbb{R}$, then the distribution $\mathcal{TN}(\mu, \sigma^2, (a, b))$ has p.d.f.

$$f(x; \mu, \sigma^2, a, b) = \frac{\phi_{\mu,\sigma^2}(x)}{\Phi_{\mu,\sigma^2}(b) - \Phi_{\mu,\sigma^2}(a)} 1_{a \leq x \leq b} \tag{1.2.2}$$

where ϕ_{μ,σ^2} and Φ_{μ,σ^2} denote the pdf and cdf, respectively, of a univariate normal distribution with mean μ and variance σ^2.

If we write

$$\begin{aligned} f(x; \mu, \sigma^2, a, b) &= K \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) 1_{a \leq x \leq b} \\ &= K \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x + \frac{\mu^2}{\sigma^2} - \log\sigma\right) 1_{a \leq x \leq b} \end{aligned}$$

where $K := \left(\Phi_{\mu,\sigma^2}(b) - \Phi_{\mu,\sigma^2}(a)\right)^{-1}$, then we see that see that $\mathcal{TN}(\mu, \sigma^2, (a, b))$ belongs to the exponential family (1.1.1) where, in this case, we have natural parameter $\eta = (-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2})^T$, sufficient statistics function $t(x) = (x^2, x)^T$, and base measure $h(x) = K\frac{1}{\sqrt{2\pi}} 1_{a \leq x \leq b}$. □

1.2.3 Inverse Gamma distribution

Example 1.2.3. (Inverse Gamma Distribution) The Inverse Gamma distribution is the distribution of the reciprocal of a Gamma random variable.[5] We can write the density of the Inverse Gamma $\mathcal{IG}(\alpha, \beta)$ distribution in exponential family form:

$$\begin{aligned} p(x \mid \alpha, \beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right) \\ &= \exp\left\{(-\alpha - 1)\log x + (-\beta)\frac{1}{x} + \log\frac{\beta^\alpha}{\Gamma(\alpha)}\right\} \end{aligned}$$

[4] For more information on the truncated normal, see e.g. [**?**] or `http://parker.ad.siu.edu/Olive/ch4.pdf`.

[5] The density of the inverse gamma can easily be obtained from the gamma density by defining the transformation $Y = \frac{1}{X} := g(X)$ and then applying the change of variables formula, $f_Y(y) = f_X(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)|$.

with natural parameter $\eta(\alpha) = [-\alpha - 1, -\beta]^T$, sufficient statistics $t(x) = [\log x, \frac{1}{x}]^T$, base measure $h(x) = 1$, and log normalizer $a(\alpha, \beta) = \log \frac{\beta^\alpha}{\Gamma(\alpha)}$. □

1.2.4 Inverse Wishart distribution

Example 1.2.4. (Inverse Wishart distribution) The Inverse Wishart distribution is the distribution of the inverse of a Wishart random variable. We can write the density of the Inverse Wishart $\mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$ distribution in exponential family form:

$$\begin{aligned} p(\boldsymbol{X} \mid \boldsymbol{\Psi}, \nu) &\overset{1}{=} C(\boldsymbol{\Psi}, \nu)\, |\boldsymbol{X}|^{-(\nu+p+1)/2} \exp\left\{ -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{X}^{-1}) \right\} \\ &= \exp\left\{ \frac{-(\nu+p+1)}{2} \log|\boldsymbol{X}| \; -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{X}^{-1}) \; + \log C(\boldsymbol{\Psi}, \nu) \right\} \\ &\overset{2}{=} \exp\left\{ \frac{-(\nu+p+1)}{2} \log|\boldsymbol{X}| \; -\frac{1}{2}\sum_{i,j=1}^{p} \boldsymbol{\Psi}_{ij}\boldsymbol{X}^{-1}_{ij} \; + \log C(\boldsymbol{\Psi}, \nu) \right\} \end{aligned}$$

Equation (1) gives the standard representation of the $\mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$ density, where $C(\boldsymbol{\Psi}, \nu)$ is the normalizing constant, $|\cdot|$ refers to the determinant, $\boldsymbol{X}, \boldsymbol{\Psi} \in \mathbb{R}^{p\times p}$ are positive definite matrices, and $\nu > p - 1$. Equation (2) uses the fact that the trace of a matrix product behaves like a dot product (C.2.1).

As we see from the last line, in the exponential family representation, we have natural parameter $\eta = [\frac{-(\nu+p+1)}{2}, \; -\frac{1}{2}\mathrm{vec}(\boldsymbol{\Psi})]^T$, sufficient statistics $t(\boldsymbol{X}) = [\log|\boldsymbol{X}|, \mathrm{vec}(\boldsymbol{X}^{-1})]^T$, base measure $h(\boldsymbol{X}) = 1$, and log normalizer $\log C(\boldsymbol{\Psi}, \nu)$. □

1.3 i.i.d samples from an exponential family

If $\boldsymbol{x} = (x_1, ..., x_n)$ are n independent samples from the same exponential family, then

$$p(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} h(x_i) \exp\left\{ \eta(\theta)^T \sum_{i=1}^{n} t(x_i) - n\, a(\eta(\theta)) \right\} \tag{1.3.1}$$

2 Exponential Family: Maximum Likelihood Estimation

The goal for maximum likelihood is to determine the parameter

$$\theta_{ML} = \underset{\theta}{\mathrm{argmax}} \; \log p(\boldsymbol{x} \mid \theta) \tag{2.0.1}$$

Let us assume that $\boldsymbol{x} = (x_1, ..., x_n)$ are i.i.d observations from a fixed exponential family, so that the likelihood has form (1.3.1). Let us compute the gradient with respect to the natural parameter η of $\ell(\eta) := \log p(\boldsymbol{x} \mid \eta)$

$$\nabla_\eta \ell(\eta) = \sum_{i=1}^{n} t(x_i) - n\, \nabla_\eta a(\eta)$$

Setting the gradient to zero, we obtain

$$\nabla_\eta a(\eta) = \frac{1}{n}\sum_{i=1}^{n} t(x_i)$$

But $\nabla_\eta a(\eta) = \mathbb{E}[t(X)]$ **[?]**. Thus, we should set θ_{ML} such that

$$\mu(\theta_{ML}) = \frac{1}{n}\sum_{i=1}^{n} t(x_i)$$

where $\mu := \mathbb{E}[t(x)]$ refers to the mean parametrization of the likelihood. [6]

3 Exponential Family: Expectation Maximization

Some models have latent variables associated with each observation, and so maximum likelihood is not possible. Let us see how expectation maximization looks when the complete data likelihood is in the exponential family.

The expectation maximization algorithm is

$$\boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}} \, \mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})} \Big[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) \Big] \tag{3.0.1}$$

We see how this plays out in the exponential family by following the logic of Section 2. Let us assume that $(\boldsymbol{x}, \boldsymbol{z}) = ((x_1, z_1), ..., (x_n, z_n))$ are n independent samples from the same exponential family, where $\boldsymbol{x}$ is observed data and $\boldsymbol{z}$ is unobserved data. Moreover, let us assume that the complete data likelihood is in the exponential family

$$p(\boldsymbol{x}, \boldsymbol{z} \mid \theta) = \prod_{i=1}^{n} h(x_i, z_i) \exp \left\{ \eta(\theta)^T \sum_{i=1}^{n} t(x_i, z_i) - n \, a(\eta(\theta)) \right\} \tag{3.0.2}$$

Here we want to find $\boldsymbol{\theta}$ to optimize

$$f(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})} \Big[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) \Big]$$

Following the logic of Section 2, we determine that we should select $\theta^{(t+1)}$ such that

$$\mu(\theta^{(t+1)}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})} t(x_i, z_i)$$

where $\mu := \mathbb{E}[t(x_1, z_1)]$ refers to the mean parametrization of the likelihood.

This is why an EM iteration is often described and/or implemented as performing maximum likelihood with the expected sufficient statistics.

TODO: But is EM *always* equivalent to performing ML with ESS's? Or is this ONLY true if I'm working within the exponential family? I need to read up some more on EM theory.

TODO: Check this section, especially with respect to the fact that I am dealing with three parametrizations here - μ, θ, ν; that is, mean, arbitrary, and natural, respectively. Really the core problem is that it's not sufficiently clear in how head how and when reparametrizations affect things.

4 Conjugate Priors

TODO: State what a conjugate prior is without using the formalism of Section 4.4

4.1 Univariate normal model

4.1.1 Example: Normal prior on mean of univariate Gaussian with known covariance

TODO: Fill in

[6] TODO: This switching of parameterization should be handled much more explicitly.

4.1.2 Example: Inverse gamma prior on the variance of a univariate Gaussian with known mean

Proposition 4.1.1. *Consider the Bayesian univariate normal model with known mean μ and random variance σ^2*

$$\begin{aligned} \sigma^2 &\sim \mathcal{IG}(\alpha_0, \beta_0) \\ y_i \mid \mu, \sigma^2 &\overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad i = 1, ..., n \end{aligned} \tag{4.1.1}$$

where $\mathcal{IG}$ denotes the Inverse Gamma distribution.

The posterior distribution for (4.1.1) *is given by*

$$\sigma^2 \mid \boldsymbol{y}, \mu \sim \mathcal{N}(\alpha_n, \beta_n)$$

where

$$\begin{aligned} \alpha_n &= \alpha_0 + \frac{1}{2} n \\ \beta_n &= \beta_0 + \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned} \tag{4.1.2}$$

Proof. We have

$$\begin{aligned} p(\sigma^2 \mid \boldsymbol{y}, \mu) &\overset{1}{\propto} \underbrace{p(\sigma^2)}_{\text{prior}} \underbrace{\prod_{i=1}^{n} p(y_i \mid \mu, \sigma^2)}_{\text{likelihood}} \\ &\overset{2}{\propto} \underbrace{(\sigma^2)^{-\alpha_0 - 1} \exp\left\{ -\frac{\beta_0}{\sigma^2} \right\}}_{\text{prior}} \underbrace{(\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}}_{\text{likelihood}} \\ &\overset{3}{\propto} (\sigma^2)^{-(\alpha_0 + n/2) - 1} \exp\left\{ -\frac{\beta_0 + \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2}{\sigma^2} \right\} \end{aligned}$$

where (1) is by Bayes rule (and conditional independence of the observation model), (2) fills in the pdfs, and (3) combines like terms so as to look like an Inverse Gamma density. □

Remark 4.1.1. (*Reparametrizing the inverse gamma prior for greater interpretability*) Peter Hoff [**?**] (pp.74) suggests parametrizing the prior as

$$\sigma^2 \sim \mathcal{IG}(\nu_0, \, \nu_0 \, \sigma_0^2 / 2)$$

for greater intepretability. In this case, we find that the posterior IG parameters are given by

$$\begin{aligned} \alpha_n &= \frac{1}{2}(\nu_0 + n) \\ \beta_n &= \frac{1}{2}(\nu_0 \, \sigma_0^2 + n \, \text{MSE}) \end{aligned}$$

where the mean squared error MSE is defined by

$$\text{MSE} := \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^2$$

So η_0 plays the role of a prior sample size and σ_0^2 plays the role of the variance within that prior sample. □

4.2 Multivariate normal model

4.2.1 Example: Normal prior on mean of multivariate Gaussian with known covariance

TODO. Note that in Section 4.3.1, I've already written up a self-enclosed argument for Bayesian linear regression; some of that argument can likely be factored out to here.

4.2.2 Example: Inverse Wishart prior on covariance matrix of multivariate Gaussian with known mean

Here we will show that the Inverse Wishart is a conjugate prior for the covariance of a multivariate normally distributed random variable with known mean.

This situation comes up

Example 4.2.1. (*Inverse Wishart prior on the covariance of a Multivariate Normal sampling model with known mean*)

Consider the sampling model for $\boldsymbol{y} := (\boldsymbol{y}_1, ..., \boldsymbol{y}_n) \overset{\text{iid}}{\sim} \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\begin{aligned} p(y \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) &\propto |\boldsymbol{\Sigma}|^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})\right] \\ &= |\boldsymbol{\Sigma}|^{-n/2} \exp\left[-\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_\mu)\right] \end{aligned} \tag{4.2.1}$$

where $\boldsymbol{S}_\mu := \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^T$ is the sum of pairwise deviation products, and where the equality in (4.2.1) is justified in Remark 4.2.2.

Let us take the mean $\boldsymbol{\mu}$ to be known, and let us take the prior on the covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ to be given by $\boldsymbol{\Sigma} \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$, i.e.

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+d+1)/2} \exp\left[-\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi})\right] \tag{4.2.2}$$

where $\boldsymbol{\Sigma} \succ 0$ and $\nu > d-1$ to have a proper prior. Note that $\mathbb{E}[\boldsymbol{\Sigma}] = \frac{\boldsymbol{\Psi}}{\nu-d-1}$.

It is easy to see from the forms of the likelihood (4.2.1) and prior (4.2.2) that the Inverse Wishart is a conjugate prior in this context. In particular

$$p(\boldsymbol{\Sigma} \mid \boldsymbol{\mu}, \boldsymbol{y}) \propto |\boldsymbol{\Sigma}|^{-(\nu+n+d+1)/2} \exp\left[-\frac{1}{2}\text{tr}\big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Psi} + \boldsymbol{S}_\mu)\big)\right] \tag{4.2.3}$$

where S_μ was defined above. Thus, we have

$$\boldsymbol{\Sigma} \mid \boldsymbol{\mu}, \boldsymbol{y} \sim \mathcal{W}^{-1}\left(\boldsymbol{\Psi} + \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^T, \nu + n\right)$$

And so the conjugate updates are given by

$$\nu' \leftarrow \nu + n \tag{4.2.4}$$

$$\boldsymbol{\Psi}' \leftarrow \boldsymbol{\Psi} + \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^T \tag{4.2.5}$$

Remark 4.2.1. (*Interpreting the hyperparameters of the Inverse Wishart*) Note that the hyperparameters of the Inverse Wishart can be interpreted (as per conjugacy) in the following way: the covariance was estimated from ν observations with a sum of pairwise deviation products $\boldsymbol{\Psi}$. [7]

Remark 4.2.2. (*Expressing the Multivariate Gaussian density in a nice form for the Inverse Wishart prior on the Covariance Matrix*)

Here we justify the equality of (4.2.1).

We will show that $\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{A} \boldsymbol{x}_i = \text{tr}(\boldsymbol{A}\sum \boldsymbol{x}_i \boldsymbol{x}_i^T)$ for $\boldsymbol{x} \in \mathbb{R}^d$, and $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ symmetric.

[7] This interpretation also makes the formula for $\mathbb{E}[\boldsymbol{\Sigma}]$ more intuitive.

$$\begin{aligned}\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{A} \boldsymbol{x}_i &= \sum_{i=1}^{n} \sum_{j,k=1}^{n} a_{jk} x_{ij} x_{ik} \\ &= \sum_{j,k=1}^{n} \left(\boldsymbol{A} \circ \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \right)_{jk} \\ &\overset{(*)}{=} \operatorname{tr}(\boldsymbol{A} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T)\end{aligned}$$

where $\circ$ is the Hadamard, also called the elementwise, operator, and where (*) holds by properties of the tr operator

$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) = \sum_{i,j} (\boldsymbol{A}^T \circ \boldsymbol{B})_{ij} \overset{\boldsymbol{A} \text{ symmetric}}{=} \sum_{i,j} (\boldsymbol{A} \circ \boldsymbol{B})_{ij}$$

4.3 Bayesian linear regression

4.3.1 Example: Bayesian linear regression with normal prior on regression weights and known observation noise

In this section, we will show that the normal prior on $\boldsymbol{\beta}$ is a conjugate prior for the regression weights $\boldsymbol{\beta}$ of a Bayesian multiple regression model with known observation noise σ^2. That is, the posterior on $\boldsymbol{\beta}$ given $\boldsymbol{y} = (y_1, ..., y_n)^T$ for such a model is also Gaussian.

Proposition 4.3.1. *Consider the Bayesian linear multiple regression model with known observation noise σ^2*

$$\begin{aligned}\boldsymbol{\beta} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ y_i \mid \boldsymbol{\beta} &\overset{ind}{\sim} \mathcal{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2), \quad i = 1, ..., n\end{aligned} \tag{4.3.1}$$

where $\boldsymbol{x}_i$ designates the i-th row of the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$.

The posterior distribution for (4.3.1) *is given by*

$$\boldsymbol{\beta} \mid \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where

$$\begin{aligned}\boldsymbol{\Sigma} &= \left(\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \\ \boldsymbol{\mu} &= \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{y} \right)\end{aligned} \tag{4.3.2}$$

Proof. By Bayes rule,

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto p(\boldsymbol{\beta}) \exp \left\{ \sum_{i=1}^{N} -\frac{1}{2\sigma^2} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} \right)^2 \right\}$$

and defining $\boldsymbol{\Omega} \in \mathbb{R}^{n \times n} : \boldsymbol{\Omega} = \operatorname{diag}(\frac{1}{\sigma^2}, ..., \frac{1}{\sigma^2})$, we have

$$\begin{aligned}&\overset{1}{\propto} p(\boldsymbol{\beta}) \exp \left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Omega} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right\} \\ &\overset{2}{\propto} p(\boldsymbol{\beta}) \exp \left\{ -\frac{1}{2} (\boldsymbol{X}^+ \boldsymbol{y} - \boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{\Omega} \boldsymbol{X} (\boldsymbol{X}^+ \boldsymbol{y} - \boldsymbol{\beta}) \right\}\end{aligned}$$

where (1) writes the weighted sum of squares in matrix notation, and (2) isolates $\boldsymbol{\beta}$, using $\boldsymbol{X}^+$, the Moore-Penrose psuedo-inverse of $\boldsymbol{X}$. [8]

Thus, we see that $p(\boldsymbol{\beta} \mid \boldsymbol{y})$ is proportional to the product of two multivariate Gaussians: $p(\boldsymbol{\beta})$, which has mean $\boldsymbol{\mu}_0$ and covariance $\boldsymbol{\Sigma}_0$, and another Gaussian, which has mean $\boldsymbol{X}^+\boldsymbol{y}$ and covariance $(\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X})^{-1}$. We know from the exponential family representation of the Gaussian that the resulting distribution can be obtained by summing at the scale of natural parameters – which for the Gaussian are the precision and precision-weighted mean. [9] Using this, we obtain

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where

$$\begin{aligned}
\boldsymbol{\Sigma} &= \left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X}\right)^{-1} \\
\boldsymbol{\mu} &= \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X}\boldsymbol{X}^{+}\boldsymbol{y}\right) \\
&= \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{y}\right)
\end{aligned}$$

recalling that we defined $\boldsymbol{\Omega} = \text{diag}(\frac{1}{\sigma^2}, ..., \frac{1}{\sigma^2})$ completes the proof. □

Remark 4.3.1. For a nice conceptual overview of Bayesian linear regression., see [**?**] or [**?**]. Among other things, these resources demonstrate how Bayesian regression makes predictions using an infinite collection of regression models (whose contributions are weighted by their posterior probabilities). They also show how the linear model is less restrictive than it might first seem; it can be used to model nonlinear functional relationships by using nonlinear basis functions.

Remark 4.3.2. (*Intuition about posterior of Bayesian linear regression*) Equation (4.3.2) gives the posterior for Bayesian linear multiple regression in the case where the observation noise is known. As pointed out by [**?**] (pp. 155), intuition can be obtained by considering the limiting cases. When the prior on the regression coefficients $\boldsymbol{\beta}$ is diffuse, the elements of the prior precision matrix $\boldsymbol{\Sigma}_0^{-1}$ will be small, and so the posterior mean satisfies $\boldsymbol{\mu} \approx (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, i.e. it approximately equals the standard least squares estimate. On the other hand, when the observation variance σ^2 is large, then the measurement precision is small, and the posterior mean satisfies $\boldsymbol{\mu} \approx \boldsymbol{\mu}_0$, i.e. it approximately equals the prior mean.

TODO: Add Bayesian linear regression where variance is also unknown.

4.4 General formalism

Here we provide some notes, following [**?**], about conjugate priors for exponential family data models.

Writing the exponential family density in canonical form, we have

$$p(x \mid \eta) = h(x)\exp\{\eta^T T(x) - A(\eta)\}$$

where η is the canonical parameter, $T(x)$ are the sufficient statistics, $h(x)$ is the base measure, and $A(\eta)$ is the log normalizer (and so is *not* a degree of freedom).

[8] Specifically, since $\boldsymbol{X}\boldsymbol{X}^+ = \boldsymbol{I}$, we use

$$\begin{aligned}
(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Omega}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) &= (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T\boldsymbol{\Omega}(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}) \\
&= \left(\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{X}^+\boldsymbol{y})\right)^T\boldsymbol{\Omega}\left(\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{X}^+\boldsymbol{y})\right) \\
&= (\boldsymbol{\beta} - \boldsymbol{X}^+\boldsymbol{y})^T\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{X}^+\boldsymbol{y})
\end{aligned}$$

.

[9] See, for reference, Section A.

The natural parameter space is

$$\left\{\eta : \int h(x) \exp\{\eta^T T(x) - A(\eta)\} < \infty\right\}$$

Given a random sample, $\boldsymbol{x} = (x_1, x_2, .., x_N)$, we obtain:

$$p(\boldsymbol{x} \mid \eta) = \left(\prod_{i=1}^{n} h(x_i)\right) \exp\left\{\eta^T \sum_{i=1}^{n} T(x_i) - NA(\eta)\right\}$$

as the likelihood function.

A conjugate prior can be obtained by mimicking the likelihood

$$p(\eta \mid \tau, \eta_0) = H(\tau, \eta_0) \exp\{\tau^T \eta - \eta_o A(\eta)\} \tag{4.4.1}$$

where now $H(\tau, \eta_0)$ is the normalizing factor. (For conditions on normalizability, see [**?**]). Note that τ has the dimensionality of the canonical parameter η and n_0 is a scalar.

To verify conjugacy, we compute the posterior density

$$p(\eta \mid \boldsymbol{x}, \tau, \eta_0) \propto \exp\left\{\left(\tau + \sum_{n=1}^{N} T(x_n)\right)^T \eta - (n_0 + N)A(\eta)\right\}$$

which retains the form of (4.4.1)

Thus, the prior-to-posterior conversion can be summarized with the following update rules

$$\begin{aligned} \tau &\to \tau + \sum_{n=1}^{N} T(x_n) \\ n_0 &\to n_0 + N \end{aligned} \tag{4.4.2}$$

For conjugate Bayesian models, the predictive posterior distribution, $p(x_{\text{new}} \mid x)$ is always tractable, because it has the same form (integrating a likelihood against the parameter distribution) as does the evidence term in Bayes law. For exponential family models, the predictive posterior takes the form of a ratio of normalizing factors

$$p(x_{\text{new}} \mid x) = \frac{H(\tau_{\text{post}}), n_0 + N}{H(\tau_{\text{post}} + T(x_{\text{new}}), n_0 + N + 1)} \tag{4.4.3}$$

TODO: Redo some of the examples using the exponential family conjugate prior formalism. A possibly useful resource in the giant table at https://en.wikipedia.org/wiki/Exponential_family.

5 Conditional Conjugacy

A family of prior distributions for a parameter is called conditionally conjugate if the conditional posterior distribution (often called the *complete conditional*), given the data and all other parameters in the model, is also in that class [**?**]. The posterior distribution for conditionally conjugate models is easily approximated with Gibbs sampling or Mean Field Variational Inference – the former samples from the complete conditional, whereas the latter takes variational expectations with respect to the natural parameter of the complete conditional.

Below we give perhaps the simplest example.

5.1 Example: Bayesian normal model with conditionally conjugate prior

Consider the following model with a normal sampling distribution and conditionally conjugate prior[10]:

$$\begin{aligned}\boldsymbol{\mu} &\sim \mathcal{N}_d(\boldsymbol{m}_0, \boldsymbol{V}_0)\\ \boldsymbol{\Sigma} &\sim \mathcal{W}^{-1}(\nu_0, \boldsymbol{\Psi}_0)\\ \boldsymbol{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} &\overset{\text{iid}}{\sim} \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1, ..., N\end{aligned}$$

We define $\boldsymbol{x} := (\boldsymbol{x}_1, \dots \boldsymbol{x}_N)$, where each $\boldsymbol{x}_i \in \mathbb{R}^d$.

The complete conditionals are well-known. In particular

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \boldsymbol{x} \sim \mathcal{N}_d(\boldsymbol{m}, \boldsymbol{V}) \tag{5.1.1}$$

where

$$\begin{aligned}\boldsymbol{m} &= \left(\boldsymbol{V}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{V}_0^{-1}\boldsymbol{m}_0 + N\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}}\right)\\ \boldsymbol{V} &= \left(\boldsymbol{V}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)^{-1}\end{aligned}$$

and

$$\boldsymbol{\Sigma} \mid \boldsymbol{\mu}, \boldsymbol{x} \sim \mathcal{W}^{-1}(\nu, \boldsymbol{\Psi}) \tag{5.1.2}$$

where

$$\begin{aligned}\nu &= \nu_0 + N\\ \boldsymbol{\Psi} &= \Psi_0 + \sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T\end{aligned}$$

Indeed, we derived (5.1.2) in Section 4.2.2. [11]

Note that the model is different than the model fully conjugate (Normal-Inverse-Wishart) prior on the pair $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The conditionally conjugate prior lacks closed-form posterior updating, but is also more expressive. [12]

These conjugate posterior updates have nice interpretations:

- **Hyperparameter updates for** $(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \boldsymbol{x})$: On the precision scale, $\boldsymbol{V}$ is the sum of the prior precision matrix $\boldsymbol{V}_0^{-1}$ and N copies of the precision for each observation, $\boldsymbol{\Sigma}^{-1}$. Similarly, $\boldsymbol{m}$ is the precision-weighted convex combination of $\boldsymbol{m}_0$, the prior mean and the empirical average $\bar{\boldsymbol{x}}$.
- **Hyperparameter updates for** $(\boldsymbol{\Sigma} \mid \boldsymbol{\mu}, \boldsymbol{x})$: The covariance was estimated from ν observations with a sum of pairwise deviation products Ψ.

[10] TODO: Prove that the prior, although conditionally conjugate, is not conjugate. (I *believe* this is true, based on context clues from experience, but I am not currently certain about it.)

[11] We still need to add a derivation for (5.1.1) TODO , but the birds' eye view for one approach is to use the general formalism for conjugacy updates in the exponential family (4.4.2), noting that the natural parameters for a multivariate Gaussian are its precision and precision-weighted mean.

[12] Is it also more expressive once we move to a variational approximation? i.e., can we get more expressive marginals this way?

A EF representation of Multivariate Gaussian in message passing

In a dissertation on Gaussian Belief Propagation [**?**], referred to in [**?**], a multivariate Gaussian is considered as a Markov Random Field.

In particular, consider the Markov Random field

$$p(x) = \frac{1}{Z}\left(\prod_{i=1}^{n}\phi(x_i)\prod_{i,j}\psi(x_i,x_j)\right) \tag{A.0.1}$$

Now note that a multivariate Gaussian has a joint distribution which can be expressed as

$$p(x) \propto \exp\left\{-\frac{1}{2}x^T Ax + b^T x\right\}$$

as this is just the exponential family form of a Gaussian (e.g., see [**?**]), where the natural parameters are given in terms of the *precision* Σ^{-1}

$$A = \Sigma^{-1}$$
$$b = \Sigma^{-1}\mu$$

Thus, the multivariate Gaussian is a MRF where the potentials in (A.0.1) are given by

$$\psi_{ij}(x_i,x_j) := \exp\left\{-\frac{1}{2}x_i A_{ij} x_j\right\}$$
$$\phi_i(x_i) := \exp\left\{-\frac{1}{2}A_{ii}x_i^2 + b_i x_i\right\}$$

This seems to be useful in inference for state space models, where one multiplies multiple "messages" that are different Gaussian densities *over the same variable*. For example, see the equations for μ_t and σ_t^2 in Section 4 of [**?**], where messages from the past and the future of a time series model are combined to get a posterior distribution on the state z_t. The combined parameters have an expression which may at first be puzzling:

$$\mu_t = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}, \quad \sigma_t^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$$

However, these messages have an intuitive form when considered in terms of the natural parameterizations: the combined mean is a weighted combination of the original means, with the weights given by the precisions. The combined precision (inverse covariance) is given simply by the sum of the original precisions. Very nice!

See Figure 1 for the general expression, which explains the formula in [**?**]. This is an example of where the natural parametrization provides more insight than the standard parametrization.

Lemma 12. *Let $f_1(x)$ and $f_2(x)$ be the probability density functions of a Gaussian random variable with two possible densities $\mathcal{N}(\mu_1, P_1^{-1})$ and $\mathcal{N}(\mu_2, P_2^{-1})$, respectively. Then their product, $f(x) = f_1(x)f_2(x)$ is, up to a constant factor, the probability density function of a Gaussian random variable with distribution $\mathcal{N}(\mu, P^{-1})$, where*

$$\mu = P^{-1}(P_1\mu_1 + P_2\mu_2), \tag{2.9}$$
$$P^{-1} = (P_1+P_2)^{-1}. \tag{2.10}$$

Figure 1: Lemma 12 of [**?**]

B More on Bayesian multivariate linear regression

Below we give an alternate proof for the posterior of the regression weights in multivariate linear regression, compared to what was given in Proposition 4.3.1. This alternate proof may be of interest. Whereas the proof of proposition 4.3.1 given in Section 4.3.1 refers to exponential family properties, the proof below does not, and instead uses multivariate completing the square to do the heavy lifting.

Note that proposition B.0.1 below, as stated, is slightly more restrictive in that it assumes the prior mean is zero. This additional restriction is not necessary; the proposition could be rewritten to match Proposition 4.3.1 exactly, and the proof could be adjusted accordingly to match the additional generality. The difference in statements is just an unnecessary presentational blemish.[13].

Proposition B.0.1. *Consider the Bayesian linear multiple regression model*

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{V})$$
$$y_i \mid \boldsymbol{\beta} \overset{ind}{\sim} \mathcal{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2), \quad i = 1, ..., n$$

where $\boldsymbol{x}_i$ designates the i-th row of the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$.

The posterior distribution for this model is given by

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\mu} = \frac{1}{\sigma^2} \boldsymbol{\Sigma} \boldsymbol{X}^T \boldsymbol{y}$$
$$\boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{V}^{-1} \right)^{-1}$$

Proof. The posterior on $\boldsymbol{\beta}$ given $\boldsymbol{y} = (y_1, ..., y_N)^T$ is Gaussian, since

$$\begin{aligned}
\ln p(\boldsymbol{\beta} \mid \boldsymbol{y}) &= \sum_{i=1}^{n} \ln p(y_i \mid \boldsymbol{\beta}) + \ln p(\boldsymbol{\beta}) \\
&= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 - \frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{V}^{-1} \boldsymbol{\beta} + \text{constant} \\
&\overset{1}{=} -\frac{1}{2\sigma^2} \left(\boldsymbol{y}^T \boldsymbol{y} - 2 \boldsymbol{y}^T \boldsymbol{X}^T \boldsymbol{\beta} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} \right) - \frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{V}^{-1} \boldsymbol{\beta} + \text{constant} \\
&\overset{2}{=} -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}) + \text{constant}
\end{aligned}$$

where

$$\boldsymbol{\mu} = \frac{1}{\sigma^2} \boldsymbol{\Sigma} \boldsymbol{X}^T \boldsymbol{y}$$
$$\boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{V}^{-1} \right)^{-1}$$

Equality (1) is obtained by noting $\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, FOIL-ing, and observing that the cross-products are scalars. Equality (2) is obtained by completing the square, where $\boldsymbol{\beta}$ plays the role of $\boldsymbol{x}$ in (C.1.1), and where in that notation we have $\boldsymbol{M} = \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{V}^{-1}$ and $\boldsymbol{b}^T = \frac{1}{\sigma^2} \boldsymbol{y}^T \boldsymbol{X}^T$ □

[13] TODO: fix up the unnecessary presentational blemish – assuming that we don't end up sacrifing too much pedagogical clarity for the sake of generality

C Matrix Facts

C.1 Multivariate completing the square

A nice overview of multivariate completing the square is given by [**?**].

Let $\boldsymbol{x}, \boldsymbol{b}$ be d-dimensional vectors, and let $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ be a symmetric invertible matrix. Then

$$\boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x} - 2\boldsymbol{b}^T \boldsymbol{x} = (\boldsymbol{x} - \boldsymbol{M}^{-1}\boldsymbol{b})^T \boldsymbol{M} (\boldsymbol{x} - \boldsymbol{M}^{-1}\boldsymbol{b}) - \boldsymbol{b}^T \boldsymbol{M}^{-1} \boldsymbol{b} \tag{C.1.1}$$

C.2 The trace of a matrix product

The trace of a matrix product behaves like a dot product.

Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$. Then

$$\mathrm{tr}(\boldsymbol{A}^T \boldsymbol{B}) = \sum_{i=1}^{n} (\boldsymbol{A}^T \boldsymbol{B})_i = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} b_{ij} \tag{C.2.1}$$

i.e., the trace of the matrix product is obtained by summing up the element-wise products.