Exponential Family

August 20, 2021

Contents

1 Exponential Families

We are interested in exponential families primarily because they 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 distributions with conjugate priors are often also exponential families [1]. [1] More complicated models may not be exponential families, 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 Definitions

Definition 1.1.1. We can define an *exponential family* as a set of probability distributions, indexed by natural parameter η,[2] whose probability density functions have the following form

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

where $a(\eta)$ is a C^∞ differentiable real-valued convex function. We refer to h as the base measure[3], η as the natural parameter, t as the sufficient statistics, and a as the log normalizer or log partition function; that is $a(\eta) = \log Z(\eta)$ where

$$Z(\eta) := \int h(x) \exp\{\eta^T t(x)\} \, \nu_{\mathcal{X}}(dx) \tag{1.1.2}$$

where $\nu_{\mathcal{X}}$ is a measure on $\mathcal{X}$.[4] [5] [6] [7] △

Definition 1.1.2. The *natural parameter space* is the set of parameters η for which the integral (1.1.2) is finite; i.e., it is $H := \{\eta : Z(\eta) < \infty\}$ △

Definition 1.1.3. An exponential family is said to be *regular* if the natural parameter space is an open set. △

One can *reparameterize* a regular exponential family with some other coordinates θ. If one writes the natural parameter as a continuous function $\eta(\theta)$, then the density (1.1.1) becomes

$$p(x \mid \theta) = h(x) \exp\{\eta(\theta)^T t(x) - a(\eta(\theta))\} \tag{1.1.3}$$

The reparameterized family is regular as well, since $\Theta := \eta^{-1}(H)$ is open.

[1] TODO: Get clearer on the relationship. There is a brief discussion on this in [1]

[2] TODO: Add restriction on parameter space so that the density is normalizable

[3] This an abuse of notation, as explained in the immediately following footnotes.

[4] If measure theory is off-putting, just take ν to be the familiar Lesbesgue measure, in which case one can remove it from the equation, and simply write $Z(\eta) := \int h(x) \exp\{\eta^T t(x)\} \, dx$

[5] Some presentations assume Lebesgue measure on $\mathcal{X}$, and write (1.1.2) more simply as $Z(\eta) := \int h(x) \exp\{\eta^T t(x)\} \, dx$. In contrast, presentations which allow for general measures $\nu_{\mathcal{X}}$ (e.g. [2], or this one) can simply absorb $h(x)$ into the measure $\nu_{\mathcal{X}}$ and write (1.1.2) as $Z(\eta) := \int \exp\{\eta^T t(x)\} \, \nu_{\mathcal{X}}(dx)$. In the former case, the term *base measure* refers to $h(x)$ – although this is an abuse of notation. In the latter case, the term *base measure* refers to $\nu_{\mathcal{X}}$.

[6] TODO: Align more closely with Jordan, who uses integration against probability measure here. He remarks on this somewhere in his exponential family lecture notes. What also may be helpful is this beautiful excerpt from pp.38 of [3]: " [...] 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 $\mathbb{R}$."

[7] TODO: The $\mathcal{X}$ appears out of nowhere. It needs to be defined.

Remark 1.1.1. (*Non-uniqueness of natural parameter*) Note from (1.1.3) 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. [8] Thus, we should speak of *a* natural parameter, rather than *the* natural parameter. △

Remark 1.1.2. Exponential family members can have intractable normalization constants. Consider, for example, the Ising model. See pp. 3 of [4]. △

Definition 1.1.4. An exponential family is said to be *minimal* if the components of the sufficient statistics $t(x)$ are linearly independent ($\nu_{\mathcal{X}}$-a.e.).[9] That is, there must be no $\eta(\theta) \in \mathbb{R}^n \backslash \{0\}$ such that $\eta(\theta)^T t(x) = 0$ ($\nu_{\mathcal{X}}$-a.e.).[10] △

An example of a non-minimal exponential family is the categorical distribution (Example 1.2.1).[11]

1.2 Examples

Here we give examples of exponential families.

1.2.1 Categorical distribution

Example 1.2.1. (Categorical Distribution) We can write the density of the categorical distribution in exponential family form. Given one-hot encoded observations $x \in [0,1]^K$ and simplex-valued parameter $\pi \in \Delta_{K-1}$, we can write

$$p(x \mid \pi) = \prod_{k=1}^{K} \pi_k^{x_k} = \exp\{\sum_{k=1}^{K} x_k \log \pi_k\}$$

with natural parameter, $\eta(\pi) = \log \pi$, the sufficient statistics $t(x) = x$, base measure $h(x) = 1$ and log normalizer 0.[12] △

1.2.2 Dirichlet distribution

Example 1.2.2. (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 F.

Remark 1.2.1. The exponential family representation of the Dirichlet, as given in Example 1.2.2, is useful when we want to compute the expectation of a log probability from a Dirichlet distributed

[8] Are they unique up to scalar multiplication?

[9] CHECK: This statement, when given by David Blei, made no mention of almost everywhere. The next statement, however, which came from [2], does. I attempted to align them by adding "almost everywhere" to the linear independence claim. Hopefully this is valid.

[10] TODO: Until now, I haven't yet assumed that the parameters are real valued. Is this necessary?

[11] Justify, and state how to rectify. Also demonstrate the importance of this.

[12] TODO: Write this up as its own example.

probability vector (as happens in the derivation of LDA with variational inference; see my notes on variational inference).

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*.

△

1.2.3 Truncated normal distribution

Example 1.2.3. (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 $\Omega \in \mathbb{R}$. [13] 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 (1.1.3) where, in this case, we have natural parameter $\eta = (-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2})^T$, sufficient statistics function $t(x) = (x^2, x)^T$, base measure $h(x) = \frac{1}{\sqrt{2\pi}} 1_{a \leq x \leq b}$, and log normalizer $a(\theta) = \log K + \frac{\mu^2}{\sigma^2} - \log\sigma$.

△

Remark 1.2.2. The truncated normal distribution differs from the normal distribution only in its base measure $h(x)$ and log normalizer $a(\theta)$. The natural parameter η and sufficient statistics function $T(x)$ are identical. Thus, knowing η and $T(x)$ is not sufficient to determine the form of the probability distribution. △

1.2.4 Inverse Gamma distribution

Example 1.2.4. (Inverse Gamma Distribution) The Inverse Gamma distribution is the distribution of the reciprocal of a Gamma random variable.[14] We can write the density of the Inverse Gamma

[13] For more information on the truncated normal, see e.g. [5] or `http://parker.ad.siu.edu/Olive/ch4.pdf`.

[14] 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, $f_Y(y) = f_X(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)|$.

$\mathcal{IG}(\alpha, \beta)$ distribution in exponential family form:

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

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 $a(\alpha, \beta) = \log \frac{\beta^\alpha}{\Gamma(\alpha)}$.

△

1.2.5 Multivariate normal

Example 1.2.5. (Multivariate normal) We can write the density of a multivariate normal $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution in exponential form

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\ &\overset{1}{=} (2\pi)^{-d/2} \exp\left\{-\frac{1}{2}\underbrace{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}_{-\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}^{-1})^T\text{vec}(\boldsymbol{x}\boldsymbol{x}^T)} + \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\log|\boldsymbol{\Sigma}^{-1}|\right\} \end{aligned} \tag{1.2.3}$$

with natural parameter $\eta(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(-\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}^{-1}),\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$, sufficient statistics $t(\boldsymbol{x}) = \left(\text{vec}(\boldsymbol{x}\boldsymbol{x}^T), \boldsymbol{x}\right)$, base measure $h(x) = (2\pi)^{-d/2}$ and log normalizing $a(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}\boldsymbol{\mu}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\log|\boldsymbol{\Sigma}^{-1}|$. △

Remark 1.2.3. From Example 1.2.5, we see that the natural parameters of the MVN are the *precision* $\boldsymbol{\Sigma}^{-1}$ and *precision-weighted mean* $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. △

Remark 1.2.4. The underbrace representation in Equation (1) is given by $\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = \text{tr}(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}) = \text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\boldsymbol{x}^T) = \text{vec}(\boldsymbol{\Sigma}^{-1})^T\text{vec}(\boldsymbol{x}\boldsymbol{x}^T)$.[15] △

Remark 1.2.5. In Section 3.2.3, we use the exponential family representation to derive the updates to the mean for a Bayesian normal model with conditionally conjugate prior. △

Remark 1.2.6. Equation (1.2.3) also says that if a random vector $\boldsymbol{x}$ has a density on $\mathbb{R}^d$ that is proportional to $\exp\{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{b}\}$ for some matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$, then $\boldsymbol{x}$ must be multivariate normal with covariance $\boldsymbol{A}^{-1}$ and mean $\boldsymbol{A}^{-1}\boldsymbol{b}$. △

1.2.6 Inverse Wishart distribution

Example 1.2.6. (Inverse Wishart distribution) The Inverse Wishart distribution (Section E) 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:

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

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

[15] Recall $\text{tr}(\boldsymbol{A}\boldsymbol{B}) = \text{vec}(\boldsymbol{A})^T\text{vec}(\boldsymbol{B})$.

$\nu > p - 1$. Equation (2) uses the fact that the trace of a matrix product behaves like a dot product (A.2.1).

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

△

1.2.7 Hidden Markov Models

Example 1.2.7. (Hidden Markov Models) A hidden Markov model (HMM) is a tool for representing probability distributions over sequences of observations. The HMM assumes that the observation at time t was generated by some process whose state x_t is hidden from the observer. Moreover, it assumes that the sequence of states satisfies the *Markov property*: conditional on the current state x_t, its future and past hidden states are independent. Finally, there is a Markov property on outputs: conditional on the current state x_t, the output y_t is independent of all other hidden states and outputs.[16]

The the *complete data likelihood* for the HMM is given by

$$\begin{aligned} p(x_{1:T}, y_{1:T} \mid \theta) &= p(x_1 \mid \theta)p(y_1 \mid x_1, \theta)\prod_{t=2}^{T} p(x_t \mid x_{t-1}, \theta)p(y_t \mid x_t, \theta) \\ &= p(x_1 \mid \pi)p(y_1 \mid x_1, \phi)\prod_{t=2}^{T} p(x_t \mid x_{t-1}, A)p(y_t \mid x_t, \phi) \\ &= \pi_{x_1}\prod_{t=2}^{T} A_{x_{t-1},x_t}\prod_{t=1}^{T} p(y_1 \mid \phi_{x_t}) \end{aligned} \tag{1.2.4}$$

where we have defined

- $y_{1:T} = (y_1, ..., y_T)$ observed sequence
- $x_{1:T} = (x_1, ..., x_T)$: hidden state sequence ($x_t \in \{1, ..., K\}$)
- $\pi = \{\pi_k\}, \pi_k = P(x_1 = k)$: initial state distribution
- $A = \{A_{kk'}\}, A_{kk'} = P(x_t = k' \mid x_{t-1} = k)$: state transition probability matrix
- $\phi = (\phi_k)_{k=1}^K$ a set of parameters, each governing an output distribution (also called emissions distribution) associated to each hidden state; that is, $P(y_t \mid x_t = k) = P(y_t \mid \phi_k)$.
- $\theta = (\pi, A, \phi)$: model parameters

We can write the complete data likelihood (1.2.4) as

$$\begin{aligned} p(x_{1:T}, y_{1:T} \mid \theta) &= \exp\left\{\log p(x_1 \mid \pi) + \sum_{t=2}^{T}\log p(x_t \mid x_{t-1}, A) + \sum_{t=1}^{T}\log p(y_t \mid x_t, \phi)\right\} \\ &= \exp\left\{\log \pi_{x_1} + \sum_{t=2}^{T}\log A_{x_{t-1},x_t} + \sum_{t=1}^{T}\log p(y_1 \mid \phi_{x_t})\right\} \\ &= \exp\left\{\sum_{k=1}^{K} x_1^k \log \pi_k + \sum_{t=2}^{T}\sum_{k,k'=1}^{K} x_{t-1}^k x_t^{k'} \log A_{kk'} + \sum_{t=1}^{T}\sum_{k=1}^{K} x_t^k \log p(y_t \mid \phi_k)\right\} \end{aligned} \tag{1.2.5}$$

[16] I might have lifted this paragraph overviewing HMM's from somewhere; check into that.

where we have defined

$$x_t^k = \begin{cases} 1, & \text{if the latent state at time } t \text{ is } k \\ 0, & \text{otherwise} \end{cases}$$

and (1.2.5) shows that the HMM is an exponential family, so long as the emissions distributions are. The sufficient statistics for $\log \pi_k$ are x_1^k, and the sufficient statistics for $\log A_{kk'}$ are $\sum_{t=2}^T x_{t-1}^k x_t^{k'}$.

△

1.2.8 Non-examples

Some non-examples include

- The Cauchy distribution (since, as we will see in Remark 1.3.1, any exponential family must have finite moments)
- The uniform distribution, whose density cannot be written in the form (1.1.3).

1.3 Properties

1.3.1 Relationship between moments and the normalizer function

Proposition 1.3.1. *Let X have an exponential family distribution with sufficient statistics function t and log normalizer function a. Then*

$$\nabla a(\eta) = \mathbb{E}[t(X)] \tag{1.3.1}$$

Proof. Since X is in the exponential family, its density can be written in the form[17]

$$p(x \mid \eta) = \exp\{\eta^T t(x) - a(\eta) + k(x)\}$$

where

$$a(\eta) = \log \int_{\mathcal{X}} \exp\{< t(x), \eta > + k(x)\} \, d\nu_{\mathcal{X}}$$

Thus

$$\begin{aligned} \nabla a(\eta) &\overset{1}{=} \frac{\int_{\mathcal{X}} t(x) \exp\{< t(x), \eta > + k(x)\} \, d\nu_{\mathcal{X}}}{\int_{\mathcal{X}} \exp\{< t(x), \eta > + k(x)\} \, d\nu_{\mathcal{X}}} \\ &\overset{2}{=} \int_{\mathcal{X}} t(x) \exp\{< t(x), \eta > - a(\eta) + k(x)\} \, d\nu_{\mathcal{X}} \\ &= \int_{\mathcal{X}} t(x) \, p(x \mid \eta) \, d\nu_{\mathcal{X}} \\ &= \mathbb{E}[t(X)] \end{aligned}$$

where in Equation (1) we take the derivative of a logarithm (interchanging the gradient and the integral), and in Equation (2) we recognize the denominator as $\exp a(\eta)$. □

Task 1.3.1. Justify formally the interchange of gradient and integral in Proposition 1.3.1. △

Remark 1.3.1. In a manner similar to that of Proposition 1.3.1, we can show that the covariance matrix of the sufficient statistics is the Hessian of the log-normalizer calculated at its natural parameter:

$$\text{Cov}[t(X)] = \nabla^2 a(\eta)$$

[17] Note that here we let $k(x) = \log h(x)$.

In fact, all moments of an exponential family are finite (recall from Definition 1.1.1 that exponential family membership requires a to be a C^∞ function). This explains why the Cauchy distribution (of undefined mean) is not an exponential family. $\triangle$

For more information on Proposition 1.3.1, see [6], [7], or [8].

1.3.2 Entropy, cross-entropy, and KL divergence between members

We can provide a closed-form expression for the KL divergence between two members of the same exponential family.

Proposition 1.3.2. *Consider two probability distributions from the same exponential family with density p, and and let their natural parameters denoted η and $\widetilde{\eta}$, respectively. Then the KL-divergence (i.e. relative entropy) is given by*

$$KL(\widetilde{\eta}||\eta) = \left\langle \nabla a(\widetilde{\eta}), \widetilde{\eta} - \eta \right\rangle + a(\eta) - a(\widetilde{\eta}) \tag{1.3.2}$$

Proof. We assume for simplicity of notation (but without loss of generality) that $\nu_{\mathcal{X}}$ in Definition 1.1.1 is the Lesbesgue measure.

$$\begin{aligned}
\mathtt{KL}(\widetilde{\eta}||\eta) &= \int p(x \mid \widetilde{\eta}) \log\left(\frac{p(x \mid \widetilde{\eta})}{p(x \mid \eta)}\right) dx \\
&= \int p(x \mid \widetilde{\eta}) \left[\left\langle t(x), \widetilde{\eta} - \eta \right\rangle + a(\eta) - a(\widetilde{\eta}) \right] dx \\
&= \left\langle \mathbb{E}_{\widetilde{\eta}}[t(X)], \widetilde{\eta} - \eta \right\rangle + a(\eta) - a(\widetilde{\eta}) \\
&\overset{(1.3.1)}{=} \left\langle \nabla a(\widetilde{\eta}), \widetilde{\eta} - \eta \right\rangle + a(\eta) - a(\widetilde{\eta})
\end{aligned}$$

□

By reasoning in a similar way as the proof of Proposition 1.3.2, expressions for the entropy $\mathbb{H}[\eta] = -\mathbb{E}_{\eta}[\log p(\eta)]$ and cross-entropy $\mathbb{H}[\widetilde{\eta}, \eta] = -\mathbb{E}_{\widetilde{\eta}}[\log p(\eta)]$ can also be provided:

$$\mathbb{H}[\widetilde{\eta}] = a(\widetilde{\eta}) - \left\langle \widetilde{\eta}, \nabla a(\widetilde{\eta}) \right\rangle - \mathbb{E}_{\widetilde{\eta}}[\log h(X)] \tag{1.3.3a}$$

$$\mathbb{H}[\widetilde{\eta}, \eta] = -a(\eta) + \left\langle \eta, \nabla a(\widetilde{\eta}) \right\rangle + \mathbb{E}_{\widetilde{\eta}}[\log h(X)] \tag{1.3.3b}$$

Note that unlike with KL divergence (1.3.2), the expressions for entropy and cross entropy (1.3.3) may not have a closed form solution. However, note that these expressions will always automatically have closed form solution when the base measure satisfies $h(x) \equiv 1$, as is the case, for example, with the Gaussian, Dirichlet, and inverse gamma distributions.

For more information on information theoretical quantities in exponential families, including connection to Bregman divergences, see [7] or [8].

1.3.3 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

$$p(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} h(x_i) \exp\left\{\eta(\theta)^T \sum_{i=1}^{n} t(x_i) - n\, a(\eta(\theta))\right\} \tag{1.3.4}$$

2 Frequentist Inference

2.1 Maximum Likelihood Estimation

The goal for maximum likelihood is to determine the parameter

$$\theta_{ML} = \operatorname*{argmax}_{\theta} \log p(\boldsymbol{x} \mid \theta) \tag{2.1.1}$$

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 (1.3.4). 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 $\nabla_\eta a(\eta) = \mathbb{E}[t(X)]$ (see Proposition 1.3.1). 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. [18]

2.2 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 an exponential family.

The expectation maximization algorithm is

$$\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})} \Big[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) \Big] \tag{2.2.1}$$

We see how this plays out in exponential families by following the logic of Section 2.1. 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 an exponential family

$$p(\boldsymbol{x}, \boldsymbol{z} \mid \theta) = \prod_{i=1}^{n} h(x_i, z_i) \exp\left\{ \eta(\theta)^T \sum_{i=1}^{n} t(x_i, z_i) - n \, a(\eta(\theta)) \right\} \tag{2.2.2}$$

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

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

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

Following the logic of Section 2.1, 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.

3 Conjugate and semi-conjugate models

Conjugacy can be defined as follows [9]. If $\mathcal{F}$ is a class of sampling distributions and $\mathcal{P}$ is a class of prior distributions for θ, then the class $\mathcal{P}$ is *conjugate* for $\mathcal{F}$ if

$$p(\theta \mid y) \in \mathcal{P} \text{ for all } p(\cdot \mid \theta) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P}$$

Conditional conjugacy (sometimes called semi-conjugacy) can be defined similarly [9]. If $\mathcal{F}$ is a class of sampling distributions and $\mathcal{P}$ is a class of prior distributions for $\theta \mid \phi$, then the class $\mathcal{P}$ is *conditionally conjugate* for $\mathcal{F}$ if

$$p(\theta \mid \phi, y) \in \mathcal{P} \text{ for all } p(\cdot \mid \theta, \phi) \in \mathcal{F} \text{ and } p(\cdot \mid \phi) \in \mathcal{P}$$

Remark 3.0.1. (*On conditional conjugacy*) In other words, 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 [10]. [19] In Section 3.2.3, we give perhaps the simplest example of a conditionally conjugate model.

△

Why are conjugate and conditionally conjugate models of interest? The posterior distributions for conditionally conjugate models are 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.

Remark 3.0.2. Although most distributions with conjugate priors are exponential families, EF membership is not a *necessary* condition for admitting a conjugate prior. For instance, the uniform distribution on $[0, a]$ is not an exponential family (the distributions don't all have the same support), but the Pareto distribution is a conjugate prior for the parameter a [11]. △

3.1 Univariate normal model

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

TODO: Fill in. Note also that we can obtain this as a special case of the multivariate case, which is handled in Section 3.2.1.

[19] Add some notes, or refer back to notes from regular conjugacy (once they're created), pointing out how this definition can be vapid, and also how conjugate priors are not unique.

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

Proposition 3.1.1. *Consider the following Bayesian univariate normal model with known mean μ and random variance σ^2*

$$\begin{aligned} \sigma^2 &\sim \mathcal{IG}(\alpha_0, \beta_0) \\ y_i \mid \mu, \sigma^2 &\overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad i = 1, ..., n \end{aligned} \tag{3.1.1}$$

where $\mathcal{IG}$ denotes the Inverse Gamma distribution. The posterior distribution is given by

$$\sigma^2 \mid \boldsymbol{y}, \mu \sim \mathcal{IG}\left(\alpha_0 + \frac{n}{2}\ ,\ \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \tag{3.1.2}$$

Proof. We have

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

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 3.1.1. (*Reparametrizing the inverse gamma prior for greater interpretability*) As observed by Peter Hoff [12] (pp.74), the form of (3.1.2) suggests parametrizing the prior as

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

for greater intepretability. In this case, we find that the posterior is given by

$$\sigma^2 \mid \boldsymbol{y}, \mu \sim \mathcal{IG}\left(\frac{\nu_0 + n}{2},\quad \frac{\nu_0\, \sigma_0^2 + n\, \widehat{\sigma}^2_{\text{MLE}}}{2}\right)$$

where the maximum likelihood estimator of the variance $\widehat{\sigma}^2_{\text{MLE}}$ is defined by

$$\widehat{\sigma}^2_{\text{MLE}} := \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$$

So ν_0 plays the role of a prior sample size and σ_0^2 plays the role of the variance within that prior sample.[20]

△

[20] Note that the maximum likelihood estimator of the variance, $\widehat{\sigma}^2_{\text{MLE}}$, could also be expressed as the mean squared error, MSE.

3.2 Multivariate normal model

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

Here we provide the posterior for the mean of a multivariate Gaussian in the case where the covariance is known.

Given data $\boldsymbol{y} := (\boldsymbol{y}_1, ..., \boldsymbol{y}_n)$, consider the model

$$\begin{aligned} \boldsymbol{\mu} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ \boldsymbol{y}_i \mid \boldsymbol{\mu} &\overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad\qquad i = 1, ..., n \end{aligned}$$

We use the exponential family representation of the MVN (Example 1.2.5) to represent the prior in terms of its natural parameters

$$p(\boldsymbol{\mu}) \propto \exp\left\{ -\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 \right\} \tag{3.2.1}$$

And similarly, we write the likelihood $L(\boldsymbol{\mu}) = p(\boldsymbol{y} \mid \boldsymbol{\mu}) = \prod_{i=1}^n p(\boldsymbol{y}_i \mid \boldsymbol{\mu})$ as

$$\begin{aligned} L(\boldsymbol{\mu}) &\propto \exp\left\{ -\frac{1}{2}\sum_{i=1}^n (\boldsymbol{y}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}) \right\} \\ &= \exp\left\{ -\frac{1}{2}\boldsymbol{\mu}^T n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}n\bar{\boldsymbol{y}} \right\} \end{aligned} \tag{3.2.2}$$

So by Bayes' law, combining the like terms in $\boldsymbol{\mu}$ of (3.2.1) and (3.2.2), we find

$$\begin{aligned} p(\boldsymbol{\mu} \mid \boldsymbol{y}) &\propto \underbrace{p(\boldsymbol{\mu})}_{\text{prior}}\underbrace{p(\boldsymbol{y} \mid \boldsymbol{\mu})}_{\text{likelihood}} \\ &= \exp\left\{ -\frac{1}{2}\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\mu} + \boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}_0\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}n\bar{\boldsymbol{y}}\right) \right\} \end{aligned}$$

which reveals that the posterior is normal (Remark 1.2.6), along with the particular forms for its natural parameters (precision and precision-weighted mean). In particular, we have, $\boldsymbol{\mu} \mid \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, where

$$\begin{aligned} \boldsymbol{\Sigma}_n &= \left(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1} \\ \boldsymbol{\mu}_n &= \boldsymbol{\Sigma}_n\left(\boldsymbol{\Sigma}_0\boldsymbol{\mu}_0 + n\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{y}}\right) \end{aligned}$$

On the precision scale, $\boldsymbol{\Sigma}_n$ is the sum of the prior precision matrix $\boldsymbol{\Sigma}_0^{-1}$ and n copies of the precision for each observation, $\boldsymbol{\Sigma}^{-1}$. Similarly, $\boldsymbol{\mu}_n$ is the precision-weighted convex combination of $\boldsymbol{\mu}_0$, the prior mean, and the empirical average, $\bar{\boldsymbol{y}}$.

3.2.2 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 3.2.1. (*Inverse Wishart prior on the covariance of a Multivariate Normal sampling model with known mean*)

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

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) &\propto |\boldsymbol{\Sigma}|^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})\right] \\ &= |\boldsymbol{\Sigma}|^{-n/2} \exp\left[-\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_\mu)\right] \end{aligned} \tag{3.2.3}$$

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 (3.2.3) is justified in Remark 3.2.1.

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.

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+d+1)/2} \exp\left[-\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi})\right] \tag{3.2.4}$$

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 (3.2.3) and prior (3.2.4) that the Inverse Wishart is a conjugate prior in this context. In particular

$$p(\boldsymbol{\Sigma} \mid \boldsymbol{\mu}, \boldsymbol{y}) \propto |\boldsymbol{\Sigma}|^{-(\nu+n+d+1)/2} \exp\left[-\frac{1}{2}\text{tr}\big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Psi} + \boldsymbol{S}_\mu)\big)\right] \tag{3.2.5}$$

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

$$\nu' \leftarrow \nu + n \tag{3.2.6}$$

$$\boldsymbol{\Psi}' \leftarrow \boldsymbol{\Psi} + \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^T \tag{3.2.7}$$

$\triangle$

For interpretation of the parameters of the Inverse Wishart, see Remark E.0.1.

Remark 3.2.1. (*Expressing the Multivariate Gaussian density in a nice form for the Inverse Wishart prior on the Covariance Matrix*)

Here we justify the equality of (3.2.3).

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.

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

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

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

△

3.2.3 Example: Bayesian normal model with conditionally conjugate prior

Consider the following model with a normal sampling distribution and conditionally conjugate prior[21]:

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

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, and have in fact already been provided by Sections 3.2.1 and 3.2.2. [22] In particular

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \boldsymbol{x} \sim \mathcal{N}_d(\boldsymbol{m}, \boldsymbol{V}) \tag{3.2.8}$$

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

$$\boldsymbol{\Sigma} \mid \boldsymbol{\mu}, \boldsymbol{x} \sim \mathcal{W}^{-1}(\nu, \boldsymbol{\Psi}) \tag{3.2.9}$$

where

$$\begin{aligned} \nu &= \nu_0 + N \\ \boldsymbol{\Psi} &= \Psi_0 + \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T \end{aligned} \tag{3.2.10}$$

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. [23]

These conjugate posterior updates have nice interpretations:

[21] 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.)

[22] We still need to add a derivation for (3.2.8) TODO , but the birds' eye view for one approach is to use the general formalism for conjugacy updates in the exponential family (3.5.2), noting that the natural parameters for a multivariate Gaussian are its precision and precision-weighted mean.

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

- **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.

Remark 3.2.2. (*Purpose of a prior on* $\boldsymbol{\Sigma}$) As mentioned by [13] (pp. 315) a prior distribution on $\boldsymbol{\Sigma}$ is generally not meant to convey substantive information, but rather to be weakly informative, and provide some shrinkage of the eigenvalues (i.e., the variances along the principal directions) and correlations. △

3.3 Bayesian linear regression

3.3.1 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 3.3.1. *Consider the Bayesian linear multiple regression model with known observation noise* σ^2

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ y_i \mid \boldsymbol{\beta}, \sigma^2 &\overset{ind}{\sim} \mathcal{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2), \quad i = 1, ..., n \end{aligned} \tag{3.3.1}$$

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 (3.3.1) *is given by*

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

where

$$\begin{aligned} \boldsymbol{\Sigma} &= \left(\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \\ \boldsymbol{\mu} &= \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{y} \right) \end{aligned} \tag{3.3.2}$$

Proof. First, we consider the likelihood $L(\boldsymbol{\beta}) := p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2)$, dropping terms proportional to $\boldsymbol{\beta}$.

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

Doing the same for the prior $p(\boldsymbol{\beta})$, we have

$$p(\boldsymbol{\beta}) \propto \exp\left\{ -\frac{1}{2} (-2\boldsymbol{\beta}^T \boldsymbol{\Sigma}_0 \boldsymbol{\mu}_0 + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\beta}) \right\}$$

Thus, by Bayes rule

$$\begin{aligned} p(\boldsymbol{\beta} \mid \boldsymbol{y}, \sigma^2) &\propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}) \\ &\propto \exp\left\{ \boldsymbol{\beta}^T \underbrace{\left(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{y} \right)}_{:= \boldsymbol{b}} - \frac{1}{2} \boldsymbol{\beta}^T \underbrace{\left(\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X} \right)}_{:= \boldsymbol{A}} \boldsymbol{\beta} \right\} \end{aligned}$$

which reveals that the posterior is normal (Remark 1.2.6), along with the particular form of its parameter (covariance $\boldsymbol{A}^{-1}$ and mean $\boldsymbol{A}^{-1}\boldsymbol{b}$).

□

Remark 3.3.1. For a nice conceptual overview of Bayesian linear regression., see [14] or [15]. 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 3.3.2. (*Intuition about posterior of Bayesian linear regression*) Equation (3.3.2) gives the posterior for Bayesian linear multiple regression in the case where the observation noise is known. As pointed out by [12] (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. △

Posterior predictive The posterior predictive distribution for Bayesian linear regression with known observation noise (3.3.1), after observing n observations $\boldsymbol{y} = (y_1, ..., y_n)$, has density

$$\begin{aligned} p(y_{\text{new}} \mid \boldsymbol{y}) &= \int p(y_{\text{new}} \mid \boldsymbol{\beta})\, p(\boldsymbol{\beta} \mid \boldsymbol{y})\, d\boldsymbol{\beta} \\ &= \int f_{\mathcal{N}}(\boldsymbol{x}_{\text{new}}^T \boldsymbol{\beta}, \sigma^2)\, f_{\mathcal{N}}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)\, d\boldsymbol{\beta} \\ &\overset{1}{=} f_{\mathcal{N}}\left(\boldsymbol{x}_{\text{new}}^T \boldsymbol{\mu}_n,\ \sigma^2 + \boldsymbol{x}_{\text{new}}^T \boldsymbol{\Sigma}_n \boldsymbol{x}_{\text{new}} \right) \end{aligned}$$

where $f_{\mathcal{N}}(m, v)$ refers to the density of a univariate Gaussian with mean m and variance v, and where $(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ are the posterior parameters given by (3.3.2), and where Equality (1) holds by Proposition B.1.1.

3.3.2 Example: Bayesian linear regression with inverse gamma prior on observation noise and known regression weights

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

Proposition 3.3.2. *Consider the Bayesian linear multiple regression model with known regression weights $\boldsymbol{\beta}$.*

$$\begin{aligned} \sigma^2 &\sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2} \right) \\ y_i \mid \boldsymbol{\beta}, \sigma^2 &\overset{ind}{\sim} \mathcal{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2), \quad i = 1, ..., n \end{aligned} \tag{3.3.3}$$

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 (3.3.3) *is given by*

$$\sigma^2 \mid \boldsymbol{y}, \boldsymbol{\beta} \sim \mathcal{IG}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + SSR(\boldsymbol{\beta})}{2}\right)$$

where the sum of squared residuals is

$$SSR(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{3.3.4}$$

Proof. See [12] pp. 155. □

3.4 Hierarchical Bayesian linear regression

Consider a Bayesian hierarchical linear regression. We take the regression to be hierarchical in the sense that we take the regression weights $\boldsymbol{\beta}_j$ to be distinct for each of $j = 1, ..., J$ groups, but we assume that the $\boldsymbol{\beta}_j$'s are drawn from some distribution. The model allows for "sharing statistical strength" in the sense that uncertainty about the jth group's regression parameters, to the extent that it exists, can be reduced by borrowing information from the other groups $k \neq j$. In other words, for grouped data, we allow the information from the other groups to play the role that is played by the prior in Bayesian linear regression.To further motivate this model, see [12].

A simple version of this model is[24] [25]:

$$\begin{aligned}
\boldsymbol{\mu} &\sim \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{V}_0) \\
\boldsymbol{\Sigma} &\sim \mathcal{W}^{-1}(\eta_0, \boldsymbol{\Psi}_0) \\
\boldsymbol{\beta}_j &\overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\sigma^2 &\sim \mathcal{IG}(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2) \\
\epsilon_{ij} &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \\
y_{ij} &= \boldsymbol{\beta}_j^T \boldsymbol{x}_{ij} + \epsilon_{ij}
\end{aligned} \tag{3.4.1}$$

This model can be seen as a Bayesian linear regression to model within-group data, put beneath a Bayesian normal sampling model to handle between-group heterogeneity in the regression weights.

The complete conditionals (e.g. see Section 11.2 of [12]) are given by

$$\begin{aligned}
\boldsymbol{\beta}_j \mid \boldsymbol{\Sigma}, \boldsymbol{\mu}, \sigma^2, \boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{\mu}_j', \boldsymbol{\Sigma}_j') \\
\boldsymbol{\Sigma}_j' &= \left(\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}_j^T\boldsymbol{X}_j\right)^{-1} \\
\boldsymbol{\mu}_j' &= \boldsymbol{\Sigma}_j'\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{\sigma^2}\boldsymbol{X}_j^T\boldsymbol{y}_j\right)
\end{aligned}$$

$$\sigma^2 \mid \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \boldsymbol{y} \sim \mathcal{IG}\left(\frac{1}{2}(\nu_0 + N), \frac{1}{2}(\nu_0\sigma_0^2 + \text{SSR}(\boldsymbol{\beta}))\right)$$

[24] The version is simple because, for example, we ignore problems with the Inverse Wishart for modeling covariance matrices (see Section E), we are not imagining that the regression coefficients are sparse, etc.

[25] Recall that the inverse gamma distribution is parametrized in a convenient way for interpretability, where ν_0 is a prior sample size from which a prior sample variance of σ_0^2 has been obtained. This parametrization, and corresponding interpretation, falls out of the use of the inverse gamma as a prior on the variance in a univariate normal model (see Remark 3.1.1).

where

$$
\begin{aligned}
N &:= \sum_{j=1}^{J} n_j \\
\text{SSR}(\boldsymbol{\beta}) &:= \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \boldsymbol{\beta}_j^T \boldsymbol{x}_{ij})^2
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\mu} \mid \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \boldsymbol{\Sigma} &\sim \mathcal{N}(\boldsymbol{m}', \boldsymbol{V}') \\
\boldsymbol{m}' &= \boldsymbol{V}' \left(\boldsymbol{V}_0^{-1} \boldsymbol{m}_0 + J \boldsymbol{\Sigma}^{-1} \overline{\boldsymbol{\beta}} \right), \qquad \overline{\boldsymbol{\beta}} := \frac{1}{J} \sum_{j=1}^{J} \boldsymbol{\beta}_j \\
\boldsymbol{V}' &= \left(\boldsymbol{V}_0^{-1} + J \boldsymbol{\Sigma}^{-1} \right)^{-1}
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\Sigma} \mid \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \boldsymbol{\mu} &\sim \mathcal{W}^{-1}(\eta', \boldsymbol{\Psi}') \\
\eta' &= \eta_0 + J \\
\boldsymbol{\Psi}' &= \boldsymbol{\Psi}_0 + \sum_{j=1}^{J} (\boldsymbol{\beta}_j - \boldsymbol{\mu})(\boldsymbol{\beta}_j - \boldsymbol{\mu})^T
\end{aligned}
\tag{3.4.2}
$$

Where note that we have defined, as shorthands,

$$
\begin{aligned}
\boldsymbol{\mu}_j' &:= \mathbb{E}[\boldsymbol{\beta}_j \mid \boldsymbol{\Sigma}, \boldsymbol{\mu}, \sigma^2, \boldsymbol{y}] \\
\boldsymbol{\Sigma}_j' &:= \text{Var}[\boldsymbol{\beta}_j \mid \boldsymbol{\Sigma}, \boldsymbol{\mu}, \sigma^2, \boldsymbol{y}]
\end{aligned}
$$

and likewise throughout (3.4.2).

Following are some thoughts on these complete conditionals; in particular, on their relationship to the complete conditionals from the simpler models (Bayesian linear regression, multivariate normal sampling model) from which the Bayesian hierarchical linear regression is composed:

- The complete conditional for the expected regression weights across groups, $\boldsymbol{\mu}$, is just the conditional distribution for the mean of a multivariate normal sampling model (3.2.8), but where the "data" are the (latent) regression weights, $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J$.
- The complete conditional for the variance in regression weights across groups, $\boldsymbol{\Sigma}$, is just the conditional distribution for the variance of a multivariate normal sampling model (3.2.9), but where the "data" are the (latent) regression weights, $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J$.
- The complete conditionals for the group-specific regression weights, $\boldsymbol{\beta}_j$, are just the conditional distributions for the regression coefficients from Bayesian linear regression (3.3.1), but where we use group j's data alone, and where the prior on these regression weights is not an external prior, but a normal distribution with mean equal to $\boldsymbol{\mu}$, the expected regression weights across groups, and variance equal to $\boldsymbol{\Sigma}$, the variance in regression weights across groups.

Question 3.4.1. The hierarchical linear regression model (3.4.1) parametrizes the prior on the variance σ^2 such that its hyperparameters (σ_0^2, ν_0) can be interpreted as the sample variance and sample size of prior observations. However, there is a weird asymmetry because we don't construct the prior on the mean in this manner. It would be nice to provide the option of doing so. My guess

is that in the simplest form, this would just mean parameterizing the top-level prior on $\boldsymbol{\mu}$ to have a variance given by $\frac{1}{\kappa_0}V_0$, where κ_0 is the number of psuedo observations relevant to estimating the mean. See pp.74-75 of [12] for ideas, although that discussion takes the prior on the mean to depend on the sampling variance σ^2. $\triangle$

3.5 General formalism

Here we provide some notes, following [6], 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).

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

$$p(\eta \mid \tau, n_0) = H(\tau, n_0) \exp\{\tau^T \eta - n_0 A(\eta)\} \tag{3.5.1}$$

where now $H(\tau, n_0)$ is the normalizing factor. (For conditions on normalizability, see [6]). 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 (3.5.1). For an argument that is perhaps more intuitive, see Section C

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

$$\begin{aligned} \tau &\to \tau + \sum_{n=1}^{N} T(x_n) \\ n_0 &\to n_0 + N \end{aligned} \tag{3.5.2}$$

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

$$p(x_{\text{new}} \mid x) = \frac{H(\tau_{\text{post}}, n_0 + N)}{H(\tau_{\text{post}} + T(x_{\text{new}}), n_0 + N + 1)} \tag{3.5.3}$$

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.

A Matrix Facts

A.1 Multivariate completing the square

A nice overview of multivariate completing the square is given by [16].

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

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

A.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

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

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

B Gaussian Facts

B.1 The simplest linear Gaussian model

Proposition B.1.1. *Let*

$$\begin{aligned} X &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ Y \mid X &\sim \mathcal{N}(\boldsymbol{A}X + \boldsymbol{b}, \boldsymbol{V}) \end{aligned}$$

Then

$$Y \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{V} + \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$$

Proof. See Section 2.3.3 of [15]. The necessary computation is for the density

$$p(y) = \int p(y \mid x)\, p(x)\, dx \tag{B.1.1}$$

□

Remark B.1.1. The necesary computation in (B.1.1) can be seen as the convolution of two Gaussians. △

Remark B.1.2. This scheme can be generalized to *linear Gaussian models*. These are probabilistic graphical models whose conditional distributions are all Gaussian, with a mean that is a linear function of parents, and a covariance that is independent of parents. See Sec 8.1.4 of [15], or [17]. △

C General Conjugacy Formalism: Alternate Approach[26]

Let $p(y \mid \theta)$ be an exponential family likelihood, and let $p(\theta)$ be its conjugate prior.

We can write the prior as

$$p(\theta) = \exp\left\{ \left\langle \eta_\theta^o, t_\theta(\theta) \right\rangle - \log Z_\theta(\eta_\theta^o) \right\}$$

And the likelihood for a single observation y_i as

$$\begin{aligned} p(y_i \mid \theta) &\overset{1}{=} \exp\left\{ \left\langle \eta_y(\theta),\ t_y(y_i) \right\rangle - \log Z_y(\eta_y(\theta)) \right\} \\ &\overset{2}{=} \exp\left\{ \left\langle \big(\eta_y(\theta),\ -\log Z_y(\eta_y(\theta))\big),\ \big(t_y(y_i),\ 1\big) \right\rangle \right\} \\ &\overset{3}{=} \exp\left\{ \left\langle t_\theta(\theta),\ \big(t_y(y_i), 1\big) \right\rangle \right\} \end{aligned}$$

where (1) is true by the exponential family assumption, (2) regroups terms to make conjugacy clearer and (3) must be true given conjugacy.

By Bayes law, the posterior after a single observation y_i is given by

$$\begin{aligned} p(\theta \mid y_i) &\propto p(\theta, y_i) \\ &= \exp\left\{ \left\langle \eta_\theta(y_i),\ t_\theta(\theta) \right\rangle - \log Z_\theta(\eta_\theta^o) \right\} \end{aligned}$$

where $\eta_\theta(y_i) = \eta_\theta^o + \big(t_y(y_i), 1\big)$, i.e. the posterior natural parameter is the sum of the prior natural parameter and the sufficient statistics concatenated with the number of samples.

And so after re-normalizing

$$p(\theta \mid y_i) = \exp\left\{ \left\langle \eta_\theta(y_i),\ t_\theta(\theta) \right\rangle - \log Z_\theta\big(\eta_\theta(y_i)\big) \right\} \tag{C.0.1}$$

After seeing multiple i.i.d observations $y = (y_1, ..., y_n)$ from the likelihood, the posterior is given by

$$p(\theta \mid y) = \exp\left\{ \left\langle \eta_\theta(y),\ t_\theta(\theta) \right\rangle - \log Z_\theta\big(\eta_\theta(y)\big) \right\} \tag{C.0.2}$$

where $\eta_\theta(y) = \eta_\theta^o + \big(\sum_{i=1}^n t_y(y_i), n\big)$.

This motivates interpreting the prior parameter as $\eta_\theta = (\tau_0, n_0)$, where $\tau_0 \in \mathbb{R}^{\dim(\eta_\theta)-1}$ is interpreted as sufficient statistics and $n_0 \in \mathbb{R}$ is interpreted as a the sample size of a prior psuedo-dataset.

Note that this argument yields the same parameter updating scheme of (3.5.2).

D More on Bayesian multivariate linear regression

Below we give two alternate proofs for the posterior of the regression weights in multivariate linear regression, compared to what was given in Proposition D.0.1.

We begin by restating the proposition here, then we provide the alternate proof.

[26] This argument follows the argument (and notation) of [2], Appendix B. As of now, I find it more intuitive then the argument given in the main body.

Proposition D.0.1. *Consider the Bayesian linear multiple regression model with known observation noise σ^2*

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ y_i \mid \boldsymbol{\beta} &\overset{ind}{\sim} \mathcal{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2), \quad i = 1, ..., n \end{aligned} \tag{D.0.1}$$

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 (3.3.1) *is given by*

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

where

$$\begin{aligned} \boldsymbol{\Sigma} &= \left(\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \\ \boldsymbol{\mu} &= \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{y} \right) \end{aligned} \tag{D.0.2}$$

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 $\boldsymbol{\Omega} \in \mathbb{R}^{n \times n} : \boldsymbol{\Omega} = \text{diag}(\frac{1}{\sigma^2}, ..., \frac{1}{\sigma^2})$, 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 $\boldsymbol{X}$. [27]

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. [28] Using this, we obtain

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

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

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

.

[28] See, for reference, Section F.

where

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

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

Now we provide a third proof, which may be of interest. Whereas both proofs of proposition D.0.1 that we have seen so far (both the proof given in Section 3.3.1, as well as the one given immediately above) refer to exponential family properties, the proof below does not, and instead uses multivariate completing the square to do the heavy lifting.

Note that proposition D.0.2 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 D.0.1 exactly, and the proof could be adjusted accordingly to match the additional generality. The difference in statements is just an unnecessary presentational blemish.[29].

Proposition D.0.2. *Consider the Bayesian linear multiple regression model*

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

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

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

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

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

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

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 (A.1.1), and where in that notation we have $\boldsymbol{M} = \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{V}^{-1}$ and $\boldsymbol{b}^T = \frac{1}{\sigma^2}\boldsymbol{y}^T\boldsymbol{X}^T$ □

E The Inverse Wishart Distribution

The Inverse Wishart is a distribution on symmetric, positive definite matrices. The Inverse Wishart distribution, denoted $\mathcal{W}^{-1}(\nu, \boldsymbol{\Psi})$, has density

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+d+1)/2} \exp\left[-\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi})\right] \tag{E.0.1}$$

where $\boldsymbol{\Sigma} \succ 0$ and $\nu > d-1$ to have a proper prior. The expected value of an Inverse Wishart random variable parametrized as in (E.0.1) is given by $\mathbb{E}[\boldsymbol{\Sigma}] = \frac{\boldsymbol{\Psi}}{\nu-d-1}$.

Remark E.0.1. (*Interpreting the parameters of the Inverse Wishart*) Note that the parameters of the Inverse Wishart can be interpreted (as per conjugacy; see (3.2.10)) in the following way: the covariance was estimated from ν observations with a residual sum of squares (a.k.a. sum of pairwise deviation products) $\boldsymbol{\Psi}$.

△

Remark E.0.1 also provides inuition on the expected value. For a visualization of how samples are affected by the parameters, see [18].

Remark E.0.2. (*Peter Hoff's notation for the Inverse Wishart: A warning*) Note that some authors (e.g. [12], pp.257) use the notation $\mathcal{W}^{-1}(\nu, \boldsymbol{M})$ to refer to the density under reparametrization

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+d+1)/2} \exp\left[-\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-1})\right] \tag{E.0.2}$$

and therefore appropriately altered normalization constant. The expected value of an Inverse Wishart random variable parametrized as in (E.0.2), is given by $\mathbb{E}[\boldsymbol{\Sigma}] = \frac{\boldsymbol{M}^{-1}}{\nu-d-1}$.

However, Hoff later introduces the reparametrization $\boldsymbol{S} := \boldsymbol{M}^{-1}$, and so writes $\mathcal{W}^{-1}(\nu, \boldsymbol{S}^{-1})$ to mean

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+d+1)/2} \exp\left[-\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S})\right] \tag{E.0.3}$$

which is (E.0.1), and which we would write as $\mathcal{W}^{-1}(\nu, \boldsymbol{S})$.

△

Remark E.0.3. (*Comparing parametrizations*) Regarding Remark E.0.2, prefer our notation because

- It is the natural parameterization (see Example 1.2.6).
- It lets $\boldsymbol{\Psi}$ be interpreted directly as a prior residual sum of squares (see Remark E.0.1).
- It matches the parametrization used throughout Wikipedia, e.g. in its conjugacy tables.

△

E.1 Relation to other distributions

The Wishart distribution has the same support as the Inverse Wishart; however, the Wishart does not give a conditionally conjugate prior on the covariance of a normal distribution. The Inverse Wishart density can be derived from the Wishart via the multivariate change of variables [19]. [30][31]

In particular, we have the relation $\boldsymbol{\Sigma} \sim \mathcal{W}^{-1}(\nu, \boldsymbol{\Psi}) \implies \boldsymbol{\Sigma}^{-1} \sim \mathcal{W}(\nu, \boldsymbol{\Psi}^{-1})$. Thus, if covariance matrix $\boldsymbol{\Sigma}$ has this Inverse Wishart distribution, then we obtain the expected value of the precision matrix as $\mathbb{E}[\boldsymbol{\Sigma}^{-1}] = \nu\boldsymbol{\Psi}^{-1}$.

The inverse Wishart can be seen as a generalization of the inverse gamma distribution to multiple dimensions.[32]

E.2 Entropy and relative entropy

Let $\boldsymbol{\Sigma}$ have an Inverse Wishart distribution (parametrized to have density (E.0.1)). Then its entropy is given by [21]:

$$\mathbb{H}(\boldsymbol{\Sigma}) = \ln \Gamma_d\left(\frac{\nu}{2}\right) + \frac{\nu d}{2} + \frac{d+1}{2} \ln\left|\frac{\boldsymbol{\Psi}}{2}\right| - \frac{\nu + d + 1}{2} \sum_{i=1}^{d} \psi\left(\frac{\nu - d + i}{2}\right)$$

where ψ denotes the digamma function, $\psi(x) = \frac{d}{dx}\Gamma(x)$.[33]

The relative entropy between two Inverse Wishart distributions p_1, p_2 with parameters $\nu_1, \boldsymbol{\Psi}_1$ and $\nu_2, \boldsymbol{\Psi}_2$ is given by [21]:

$$\mathtt{KL}\big[p_1 \,\big|\big|\, p_2\big] = \ln\left(\frac{\Gamma_d(\frac{\nu_2}{2})}{\Gamma_d(\frac{\nu_1}{2})}\right) + \frac{\nu_1}{2}\mathrm{tr}(\boldsymbol{\Psi}_1^{-1}\boldsymbol{\Psi}_2) - \frac{\nu_1 d}{2} - \frac{\nu_2}{2}\ln\left|\boldsymbol{\Psi}_1^{-1}\boldsymbol{\Psi}_2\right| - \frac{\nu_2 - \nu_1}{2}\sum_{i=1}^{d}\psi\left(\frac{\nu_1 - d + i}{2}\right)$$

E.3 Sampling

A sample $\boldsymbol{\Sigma}$ from the $\mathcal{W}^{-1}(\nu, \boldsymbol{\Psi})$ distribution (using the natural parametrization of (E.0.1)) can be obtained by the following scheme[34] [12]:

1. Sample $\boldsymbol{z}_1, ..., \boldsymbol{z}_\nu \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}^{-1})$
2. Calculate $\boldsymbol{Z}^T\boldsymbol{Z} = \sum_{i=1}^{\nu} \boldsymbol{z}_i\boldsymbol{z}_i^T$.
3. Set $\boldsymbol{\Sigma} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$.

[30] It is claimed in Wikipedia that if $\boldsymbol{X} \sim \mathcal{W}(\nu, \boldsymbol{\Psi})$ then $\boldsymbol{X}^{-1} \sim \mathcal{W}^{-1}(\nu, \boldsymbol{\Psi}^{-1})$. I almost was able to show this using the multivariate change of variables [19] along with (15.15) of [20], but I was off by a negative when attempting to combine the two terms with $|\boldsymbol{X}^{-1}|$ raised to an exponent.

[31] TO DO: Provide derivation.

[32] TODO: fill in. Make explicit how it is a generalization.

[33] It is unfortunate that in our notation, $\boldsymbol{\Psi}$ and ψ mean completely different things; fix this.

[34] TODO: Provide derivation of this scheme. See perhaps https://www.math.wustl.edu/~sawyer/hmhandouts/Wishart.pdf.

The intuition is that the Inverse Wishart models covariance matrices as an inverse sum of squares (again, see Remark E.0.1).

E.4 Evaluation as a model for covariance matrices

The Inverse Wishart is a popular choice for modeling covariance matrices (e.g. see [12]), due to at least the fact that is a conditionally conjugate prior on the covariance of a normal distribution. (See Section 3.2.3.) It seems to me that a weakly informative prior could be constructed by setting $\nu = d + 2$ (the smallest integer for which ν is in the parameter space) and $\boldsymbol{\Psi} = (\nu - d - 1)\boldsymbol{I} = \boldsymbol{I}$. This would presumably be reasonable at least if one expected unit variances and wanted to make a prior assumption of independence across dimensions.

Some problems with the Inverse Wishart as a model for covariance matrices is summarized in [22]. We highlight that:

1. When $\nu > 1$, the implied scaled inv-χ^2 distribution on the individual variances has extremely low density in the region near zero.
2. The prior imposes a dependency between the correlations and the variances. In particular, larger variances are associated with absolute values of the correlations near 1 while small variances are associated with correlations near zero.

For additional discussion on the problems with Inverse Wishart, especially when used in hierarchical models, and for a remedy using a half-t distribution that also has a conditionally conjugate construction, see [23].

F EF representation of Multivariate Gaussian in message passing

In a dissertation on Gaussian Belief Propagation [24], referred to in [25], a multivariate Gaussian is considered as a Markov Random Field.

In particular, consider the Markov Random field

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

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

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

as this is just the exponential family form of a Gaussian (e.g., see [26]), where the natural parameters are given in terms of the *precision* Σ^{-1}

$$\begin{aligned} A &= \Sigma^{-1} \\ b &= \Sigma^{-1}\mu \end{aligned}$$

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

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

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 [25], 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 [25]. 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 [24]

References

[1] José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

[2] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. *Advances in neural information processing systems*, 29:2946–2954, 2016.

[3] Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

[4] Johnathan Taylor. *Multiparameter exponential families, Part II*. Available at `http://statweb.stanford.edu/~jtaylo/courses/stats306b/restricted/notebooks/multiparameter_partII.pdf`.

[5] John Burkardt. The truncated normal distribution. *Department of Scientific Computing Website, Florida State University*, pages 1–35, 2014.

[6] Michael Jordan. *The exponential family: Conjugate priors*, (accessed September 11, 2020).

[7] Frank Nielsen and Richard Nock. Entropies and cross-entropies of exponential families. In *2010 IEEE International Conference on Image Processing*, pages 3621–3624. IEEE, 2010.

[8] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2009.

[9] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[10] Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

[11] Tom Minka. *Bayesian inference of a uniform distribution*. Available at https://tminka.github.io/papers/minka-uniform.pdf.

[12] Peter D Hoff. *A first course in Bayesian statistical methods*, volume 580. Springer, 2009.

[13] Kosuke Imai and David A Van Dyk. A bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of econometrics*, 124(2):311–334, 2005.

[14] Roger Grosse et al. *Bayesian Linear Regression*. Available at https://www.cs.toronto.edu/~rgrosse/courses/csc411_f18/slides/lec19-slides.pdf.

[15] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[16] Gregory Gundersen. *Completing the square*. Available at http://gregorygundersen.com/blog/2019/09/18/completing-the-square/.

[17] Sam Roweis and Zoubin Ghahramani. A unifying review of linear gaussian models. *Neural computation*, 11(2):305–345, 1999.

[18] Michael Hughes. *Inverse Wishart Distribution*. Available at https://www.michaelchughes.com/blog/probability-basics/inverse-wishart-distribution/.

[19] Robert Wolpert. *Change of variables*. Available at https://www2.stat.duke.edu/courses/Spring11/sta114/lec/114mvnorm.pdf.

[20] Paul S Dwyer. Some applications of matrix derivatives in multivariate analysis. *Journal of the American Statistical Association*, 62(318):607–625, 1967.

[21] Maya Gupta and Santosh Srivastava. Parametric bayesian estimation of differential entropy and relative entropy. *Entropy*, 12(4):818–843, 2010.

[22] Ignacio Alvarez, Jarad Niemi, and Matt Simpson. Bayesian inference for a covariance matrix. *arXiv preprint arXiv:1408.4050*, 2014.

[23] Michael Wojnowicz. *Variational Inference for categorical models*. Available at https://github.com/mikewojnowicz/fall_2020/blob/master/reports/categorical_models_with_vi/categorical_models_with_vi.pdf with permission.

[24] Danny Bickson. Gaussian belief propagation: Theory and aplication. *arXiv preprint arXiv:0811.2518*, 2008.

[25] Rahul G Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. *arXiv preprint arXiv:1609.09869*, 2016.

[26] Barbara Englehardt. *Gaussian Models*, (accessed November 22, 2020).