Exponential Families

November 7, 2024

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 [Bernardo and Smith, 2009]. [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 $\boldsymbol{\eta} \in H$,whose probability density functions with respect to measure μ on $\mathcal{X}$ have the following form

$$p_{\boldsymbol{\eta}}(\boldsymbol{x}) = h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^\top \boldsymbol{s}(\boldsymbol{x}) - a(\boldsymbol{\eta})\} \tag{1.1.1}$$

where $\boldsymbol{s} : \mathcal{X} \to \mathbb{R}^p$ is the *sufficient statistics function*, $h : \mathcal{X} \to \mathbb{R}$ is the *carrier density*, $\boldsymbol{\eta} \in H \subset \mathbb{R}^p$ is the *natural parameter*, and $a : H \to \mathbb{R}$ is a strictly convex and C^∞ differentiable real-valued function known as the *log normalizer* or *log partition function*; that is $a(\boldsymbol{\eta}) = \log Z(\boldsymbol{\eta})$ where

$$Z(\boldsymbol{\eta}) := \int h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^\top \boldsymbol{s}(\boldsymbol{x})\} \, \mu(d\boldsymbol{x}) \tag{1.1.2}$$

We refer to μ as the *base measure*.[2] Typically μ is the Lesbesgue measure or counting measure. If μ is the Lesbesgue measure, then the density $p_{\boldsymbol{\eta}}$ is a probability density function. If μ is the counting measure, then the density $p_{\boldsymbol{\eta}}$ is a probability mass function.[3] △

Remark 1.1.1. (*Alternate constructions for the exponential family.*) Observe [Chua, 2019] that without loss of generality, we can

a) *(Ignorability of the carrier density.)* Set $h(\boldsymbol{x}) \equiv 1$ if we change μ to $\widetilde{\mu}$, chosen so that its Radon-Nikodym derivative with respect to μ is h. That is, we can always absorb h into μ and write (1.1.1) as $p_{\boldsymbol{\eta}}(\boldsymbol{x}) = \exp\{\boldsymbol{\eta}^\top \boldsymbol{s}(\boldsymbol{x}) - a(\boldsymbol{\eta})\}$, in which case (1.1.2) becomes $Z(\boldsymbol{\eta}) := \int \exp\{\boldsymbol{\eta}^\top \boldsymbol{s}(\boldsymbol{x})\} \, \widetilde{\mu}(d\boldsymbol{x})$ where $\widetilde{\mu}(A) := \int_A h d\mu$ for any measurable set A. This observation streamlines most computations with integrals, e.g. see the proof of Prop. 1.3.3.

b) *(Non-uniqueness of natural parameter.)* Assume $M \in \mathbb{R}^{p \times p}$ is invertible, and define $\widetilde{\boldsymbol{s}} = M\boldsymbol{s}(\boldsymbol{x})$. Then defining $\widetilde{\boldsymbol{\eta}} = M^{-1}\boldsymbol{\eta}$ results in an equivalent family. As a special case, $\boldsymbol{\eta}$ could be multiplied by a non-zero constant c if $\boldsymbol{s}(\boldsymbol{x})$ is divided by c. Thus, a natural parameterization is not unique; we should speak of *a* natural parameter, rather than *the* natural parameter.

[1] TODO: Get clearer on the relationship. There is a brief discussion on this in [Bernardo and Smith, 2009]. An answer from mathstackexchange at https://stats.stackexchange.com/questions/176668/can-anyone-explain-conjugate-priors-in-simplest-possible-terms gives a promising quote, although I do not know the source: "Outside this exponential family setting, there is no non-trivial family of distributions with a fixed support that allows for conjugate priors. This is a consequence of the Darmois-Pitman-Koopman lemma."

[2] Some presentations use the term "base measure" to refer to $h(x)$, but that is an abuse of notation, as h is not a measure.

[3] If measure theory is off-putting, just take μ to be the Lesbesgue measure for continuous random variables, in which case one can remove it from the equation, and simply write $Z(\boldsymbol{\eta}) := \int h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^\top \boldsymbol{s}(\boldsymbol{x})\} \, d\boldsymbol{x}$, and take μ to be the counting measure for discrete random variables, in which case one can write $Z(\boldsymbol{\eta}) := \sum_{x \in \mathcal{X}} h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^\top \boldsymbol{s}(\boldsymbol{x})\}$. Indeed, we will make precisely these assumptions throughout the document, unless otherwise noted.

c) Change h to $\widetilde{h}(\boldsymbol{x}) = p_{\mathbf{0}}(\boldsymbol{x})$, assuming that $\mathbf{0} \in H$. Then, changing $a(\boldsymbol{\eta})$ to $\widetilde{a}(\boldsymbol{\eta}) = a(\boldsymbol{\eta}) - a(\mathbf{0})$ results in an equivalent family.

△

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

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 $\boldsymbol{\theta}$. If one writes the natural parameter as a continuous function $\boldsymbol{\eta}(\boldsymbol{\theta})$, then the density (1.1.1) becomes

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \boldsymbol{s}(\boldsymbol{x}) - a(\boldsymbol{\eta}(\boldsymbol{\theta}))\} \tag{1.1.3}$$

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

Remark 1.1.2. Exponential family members can have intractable normalization constants. Consider, for example, the Ising model. See pp. 3 of [Taylor, 2013]. △

Definition 1.1.4. An exponential family is said to be **minimal** if the components of the sufficient statistics $\boldsymbol{s}(\boldsymbol{x})$ are linearly independent (μ-a.e.).[4] That is, there must be no $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ such that $\boldsymbol{\eta}(\boldsymbol{\theta})^\top \boldsymbol{s}(\boldsymbol{x}) = 0$ (μ-a.e.).[5] △

An example of a non-minimal exponential family is the categorical distribution (Example 1.2.1).[6]

TODO: I believe a non-minimal EF can always be reparametrized into a minimal EF. This makes sense based on the definition, based on the categorical distribution, and based on an implication from Jordan [2010a, Sec. 8.4]. Justify this.

1.2 Examples

Here we give examples of exponential families, showing how to derive the exponential family forms. Note that the result of such computations are readily available for a wide variety of exponential family members (see e.g. [Nielsen and Garcia, 2009] or the Wikipedia page on 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 $\boldsymbol{x} \in \{0, 1\}^K$ and simplex-valued parameter $\pi \in \Delta_{K-1}$, we can write

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

with natural parameter, $\boldsymbol{\eta}(\pi) = \log \pi$, the sufficient statistics $\boldsymbol{s}(\boldsymbol{x}) = \boldsymbol{x}$, carrier density $h(\boldsymbol{x}) = 1$ and log normalizer 0. △

[4] CHECK: This statement, when given by David Blei, made no mention of almost everywhere. The next statement, however, which came from [Johnson et al., 2016], does. I attempted to align them by adding "almost everywhere" to the linear independence claim. Hopefully this is valid.

[5] Is it strictly speaking necessary to assume that the parameters are real-valued? If so, why?

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

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 $\boldsymbol{\eta}(\alpha) = [\alpha_1 - 1, ..., \alpha_K - 1]^\top$, sufficient statistics $\boldsymbol{s}(\pi) = \log \pi = [\log \pi_1, ..., \log \pi_K]^\top$, carrier density $h(\pi) = 1$, 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 B.3.

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

In those notes, we see

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

where (1) uses a well-known exponential family property (see Proposition 1.3.1) 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 $\Omega \in \mathbb{R}$. [7] 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.

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

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

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

[7] For more information on the truncated normal, see e.g. [Burkardt, 2014] or http://parker.ad.siu.edu/Olive/ch4.pdf.

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 $\boldsymbol{\eta} = (\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2})\top$, sufficient statistics function $\boldsymbol{s}(x) = (-\frac{1}{2}x^2, x)^\top$, carrier density $h(x) = \frac{1}{\sqrt{2\pi}}\mathbb{I}_{a\le x\le b}$, and log normalizer $a(\boldsymbol{\theta}) = \log K + \frac{\mu^2}{\sigma^2} - \log\sigma$.

△

Remark 1.2.2. The truncated normal distribution differs from the normal distribution only in its carrier density $h(x)$ and therefore log normalizer $a(\boldsymbol{\theta})$. The natural parameter $\boldsymbol{\eta}$ and sufficient statistics function $\boldsymbol{s}(x)$ are identical. Thus, knowing $\boldsymbol{\eta}$ and $\boldsymbol{s}(x)$ 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.[8] We can write the density of the Inverse Gamma $\mathcal{IG}(\alpha, \beta)$ distribution in exponential family form:

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

with natural parameter $\boldsymbol{\eta}(\alpha) = [-\alpha - 1, -\beta]\top$, sufficient statistics $\boldsymbol{s}(x) = [\log x, \frac{1}{x}]^\top$, carrier density $h(x) = 1$, 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

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

with natural parameter $\boldsymbol{\eta}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(-\frac{1}{2}\mathrm{vec}(\boldsymbol{\Sigma}^{-1}),\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$, sufficient statistics $\boldsymbol{s}(\boldsymbol{x}) = \left(\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^\top), \boldsymbol{x}\right)$, carrier density $h(\boldsymbol{x}) = (2\pi)^{-d/2}$ 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 $\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = \mathrm{tr}(\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}) = \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\boldsymbol{x}^\top) = \mathrm{vec}(\boldsymbol{\Sigma}^{-1})^\top\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^\top)$.[9] △

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 $\exp\{-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^\top\boldsymbol{b}\}$ for some matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$, then $\boldsymbol{x}$ must be multivariate normal with covariance $\boldsymbol{A}^{-1}$ and mean $\boldsymbol{A}^{-1}\boldsymbol{b}$. △

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

[9] Recall $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{vec}(\boldsymbol{A})^\top\mathrm{vec}(\boldsymbol{B})$.

1.2.6 Inverse Wishart distribution

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

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

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 (A.2.1).

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

△

1.2.7 Gaussian Mixture Models

For a detailed explanation of how to represent Gaussian mixture models in the exponential family, see https://www.youtube.com/playlist?list=PLD0F06AA0D2E8FFBA, videos 16.6-16.8. More generally, it can be shown that mixtures of any exponential family distributions are still in the exponential family.

1.2.8 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.[10]

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

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

where we have defined

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

- $y_{1:T} = (y_1, ..., y_T)$ 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, $P(y_t \mid x_t = k) = P(y_t \mid \phi_k)$.
- $\boldsymbol{\theta} = (\pi, A, \phi)$: model parameters

We can write the complete data likelihood (1.2.4) as

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

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, the sufficient statistics for $\log A_{kk'}$ are $\sum_{t=2}^{T} x_{t-1}^k x_t^{k'}$, and the sufficient statistics for the natural parameters of the emissions distributions $p(y_t \mid \phi_k)$ are $\sum_{t=1}^{T} x_t^k \cdot \{\texttt{sufficient statistics of emissions}\}$.

△

1.2.9 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. [The gradient of the log normalizer equals the expected sufficient statistics.] *Let X have an exponential family distribution with natural parameter $\boldsymbol{\eta}$, sufficient statistics function $\boldsymbol{s}$, and log normalizer function a. Then*

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

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

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{\eta}^\top \boldsymbol{s}(\boldsymbol{x}) - a(\boldsymbol{\eta}) + \log h(\boldsymbol{x})\}$$

where

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

Thus

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

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 $\exp a(\boldsymbol{\eta})$. □

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:

$$\mathrm{Cov}[\boldsymbol{s}(X)] = \nabla^2 a(\boldsymbol{\eta})$$

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. △

For more information on Proposition 1.3.1, see Jordan [2010a], Jordan [2010b], Nielsen and Nock [2010], or Nielsen and Garcia [2009].

1.3.2 Entropy, cross-entropy, and KL divergence

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 $\boldsymbol{\eta}$ and $\widetilde{\boldsymbol{\eta}}$, respectively. Then the KL-divergence (i.e. relative entropy) is given by*

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

Proof. We assume for simplicity of notation (but without loss of generality) that μ in Definition 1.1.1 is the Lesbesgue measure.

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

□

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

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

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 carrier density satisfies $h(x) \equiv 1$, as is the case, for example, with the Gaussian, Dirichlet, and inverse gamma distributions.[11]. In that case, we have

$$\mathbb{H}[\boldsymbol{\eta}] = a(\boldsymbol{\eta}) - \big\langle \boldsymbol{\eta},\ \nabla a(\boldsymbol{\eta}) \big\rangle \tag{1.3.4a}$$
$$\mathbb{H}[\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}] = a(\widetilde{\boldsymbol{\eta}}) - \big\langle \widetilde{\boldsymbol{\eta}},\ \nabla a(\boldsymbol{\eta}) \big\rangle \tag{1.3.4b}$$

Example 1.3.1. Let us apply (1.3.3a) to compute the entropy of a centered Laplace distribution, which has pdf

$$f(x \mid \sigma) = -\frac{1}{2\sigma} \exp\left\{ -\frac{|x|}{\sigma} \right\}$$

The natural parameter is $\boldsymbol{\eta} = -\frac{1}{\sigma}$, the log normalizer is $a(\boldsymbol{\eta}) = \log(-\frac{2}{\boldsymbol{\eta}})$ (and so $\nabla a(\boldsymbol{\eta}) = -\frac{1}{\boldsymbol{\eta}}$), and the carrier density is $h(x) = 1$ (which allows us to use (1.3.4a) instead of (1.3.3a)). Using this, we can easily compute

$$\mathbb{H}[\boldsymbol{\eta}] = a(\boldsymbol{\eta}) - \big\langle \boldsymbol{\eta},\ \nabla a(\boldsymbol{\eta}) \big\rangle = \log\left(-\frac{2}{\boldsymbol{\eta}} \right) - \big\langle \boldsymbol{\eta},\ -\frac{1}{\boldsymbol{\eta}} \big\rangle = \log\left(-\frac{2}{\boldsymbol{\eta}} \right) + 1$$
$$\implies \mathbb{H}[\sigma] = \log 2\sigma + 1$$

△

Example 1.3.2. Let us apply (1.3.3a) to compute the entropy of a univariate Gaussian. The necessary exponential family quantities can be found in many tables (e.g. see [Nielsen and Garcia, 2009]); they are:

$$\boldsymbol{\eta} = \left(\frac{\mu}{\sigma^2},\ -\frac{1}{2\sigma^2} \right)^\top, \quad a(\boldsymbol{\eta}) = \frac{-\boldsymbol{\eta}_1^2}{4\boldsymbol{\eta}_2} + \frac{1}{2}\log(-\frac{\pi}{\boldsymbol{\eta}_2}), \quad \nabla a(\boldsymbol{\eta}) = \left(-\frac{\boldsymbol{\eta}_1}{2\boldsymbol{\eta}_2},\ -\frac{1}{2\boldsymbol{\eta}_2} + \frac{\boldsymbol{\eta}_1^2}{4\boldsymbol{\eta}_2^2} \right)^\top$$

and $h(x) = 1$ (which allows us to use (1.3.4a) instead of (1.3.3a)).

Using this, we can easily compute

$$\begin{aligned} \mathbb{H}[\boldsymbol{\eta}] &= a(\boldsymbol{\eta}) - \big\langle \boldsymbol{\eta},\ \nabla a(\boldsymbol{\eta}) \big\rangle \\ &= \cancel{\frac{-\boldsymbol{\eta}_1^2}{4\boldsymbol{\eta}_2}} + \frac{1}{2}\log(-\frac{\pi}{\boldsymbol{\eta}_2}) - \Big[\cancel{\frac{-\boldsymbol{\eta}_1^2}{2\boldsymbol{\eta}_2}} - \frac{1}{2} + \cancel{\frac{\boldsymbol{\eta}_1^2}{4\boldsymbol{\eta}_2}} \Big] \\ \implies \mathbb{H}[\sigma] &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2} \end{aligned}$$

Consider how much easier this computation is than that of, say, https://proofwiki.org/wiki/Differential_Entropy_of_Gaussian_Distribution. △

[11] Some presentations give $h(x) = C$ for some constant C, but note that the constant can simply be absorbed into the log-normalizer $a(\boldsymbol{\eta})$.

Remark 1.3.2. (*Cross-entropy between two distributions which belong to different exponential families*)

What does (1.3.3b) look like in the case that we we want to compute the cross-entropy between two distributions which belong to *different* exponential families (with the same support)?

Let

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{\eta}^\top \boldsymbol{s}(\boldsymbol{x}) - a(\boldsymbol{\eta}) + \log h(\boldsymbol{x})\}$$
$$\widetilde{p}(\boldsymbol{x} \mid \widetilde{\boldsymbol{\eta}}) = \exp\{\widetilde{\boldsymbol{\eta}}^\top \widetilde{\boldsymbol{s}}(\boldsymbol{x}) - \widetilde{a}(\widetilde{\boldsymbol{\eta}}) + \log \widetilde{h}(\boldsymbol{x})\}$$

Then the cross entropy from $\widetilde{p}$ to p is given by

$$\begin{aligned}
\mathbb{H}(p, \widetilde{p}) &= -\mathbb{E}_p[\log \widetilde{p}(\boldsymbol{x})] \\
&= -\int p(\boldsymbol{x} \mid \boldsymbol{\eta}) \ \log \widetilde{p}(\boldsymbol{x} \mid \widetilde{\boldsymbol{\eta}}) \, d\boldsymbol{x} \\
&= -\int p(\boldsymbol{x} \mid \boldsymbol{\eta}) \Big[\langle \widetilde{\boldsymbol{s}}(\boldsymbol{x}), \ \widetilde{\boldsymbol{\eta}} \rangle - \widetilde{a}(\widetilde{\boldsymbol{\eta}}) + \log \widetilde{h}(\boldsymbol{x}) \Big] d\boldsymbol{x} \\
&= \widetilde{a}(\widetilde{\boldsymbol{\eta}}) - \langle \mathbb{E}_p[\widetilde{\boldsymbol{s}}(\boldsymbol{x})], \ \widetilde{\boldsymbol{\eta}} \rangle - \mathbb{E}_p[\log \widetilde{h}(\boldsymbol{x})]
\end{aligned}$$

So there are now two potentially problematic terms, although in the case where $\widetilde{h}(\boldsymbol{x}) \equiv 1$, only one potentially problematic term remains. △

For more information on information theoretical quantities in exponential families, including connection to Bregman divergences, see [Nielsen and Nock, 2010] or [Nielsen and Garcia, 2009].

1.3.3 Convexity

Proposition 1.3.3. [Convexity properties of the exponential family [Jordan, 2010a].] *The natural parameter space H is convex (as a set) and the log normalizer function $a(\boldsymbol{\eta})$ is convex (as a function). If the family is minimal then $a(\boldsymbol{\eta})$ is strictly convex.*

Proof. The proofs of both convexity results follow from an application of Hölder's inequality. Consider distinct parameters $\boldsymbol{\eta}_1 \in H$ and $\boldsymbol{\eta}_2 \in H$ and let $\boldsymbol{\eta} = \lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2$, for $0 < \lambda < 1$.

Then

$$\begin{aligned}
\exp a(\boldsymbol{\eta}) &\overset{1}{=} \int e^{(\lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2)^\top \boldsymbol{s}(\boldsymbol{x})} \nu(d\boldsymbol{x}) \\
&\overset{2}{\leq} \Bigg(\underbrace{\int e^{\boldsymbol{\eta}_1^\top \boldsymbol{s}(\boldsymbol{x})} \nu(d\boldsymbol{x})}_{\text{assumed finite}} \Bigg)^{\lambda} \Bigg(\underbrace{\int e^{\boldsymbol{\eta}_2^\top \boldsymbol{s}(\boldsymbol{x})} \nu(d\boldsymbol{x})}_{\text{assumed finite}} \Bigg)^{1-\lambda} \\
&< \infty
\end{aligned}$$

Hence $a(\boldsymbol{\eta})$ is finite, and so $\boldsymbol{\eta} \in H$. Moreover, taking logarithms of the above yields

$$a(\boldsymbol{\eta}) \leq \lambda a(\boldsymbol{\eta}_1) + (1-\alpha) a(\boldsymbol{\eta}_2),$$

and so a is a convex function. For an argument that minimality implies strict convexity, see Jordan [2010a].

For details:

- In (1) we used the definition of the log normalizer function for an exponential family. For simplicity, we also applied Item a) of Remark 1.1.1, changing the ambient measure to $\nu(d\boldsymbol{x}) = h(\boldsymbol{x})\mu(d\boldsymbol{x})$, so that we can ignore the carrier density (the carrier density $\widetilde{h}$ with respect to ν satisfies $\widetilde{h}(\boldsymbol{x}) \equiv 1$).
- In (2), we applied Hölder's inequality. Recall that Hölder's inequality is that $||fg||_1 \leq ||f||_p||g||_q$ whenever $p, q \in [1, \infty]$ and $\frac{1}{p} + \frac{1}{q} = 1$. Recall also the definition of norm $||f||_p = (\int f^p \ d\mu \ )^{1/p}$ in some underlying measure space (with measure μ). To apply Hölder's inequality, we take the functions to be $f(\boldsymbol{x}) = e^{\lambda \boldsymbol{\eta}_1^\top \boldsymbol{s}(\boldsymbol{x})}$ and $g(\boldsymbol{x}) = e^{(1-\lambda)\boldsymbol{\eta}_2^\top \boldsymbol{s}(\boldsymbol{x})}$, the underlying measure to be $\mu = \nu$, and the Hölder exponents to be $p = \frac{1}{\lambda}$ and $q = \frac{1}{1-\lambda}$.

□

1.3.4 Identifiability

A model is called *globally identifiable* if $p(\cdot \mid \boldsymbol{\theta}_1) = p(\cdot \mid \boldsymbol{\theta}_2)$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ [Cole, 2020]. For exponential families as defined in Def. 1.1.1,[12] if the Fisher information matrix is non-singular, then the model is globally identifiable. For exponential families, we may compute the Fisher information matrix as[13]

$$I(\boldsymbol{\theta}) = \mathbb{E}[-\nabla^2 \log p(\boldsymbol{x} \mid \boldsymbol{\theta})] \tag{1.3.5}$$

For more information on identifiability in exponential families, see [Cole, 2020] pp.62.

1.3.5 i.i.d samples from an exponential family

If $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ are n independent samples from the same exponential family, then

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} h(\boldsymbol{x}_i) \exp\left\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \sum_{i=1}^{n} \boldsymbol{s}(\boldsymbol{x}_i) - n\, a(\boldsymbol{\eta}(\boldsymbol{\theta}))\right\} \tag{1.3.6}$$

1.3.6 Mean parametrization

In Prop. 1.3.1 (see Sec. 1.3.1), we showed that the expected sufficient statistics $\mu \triangleq \mathbb{E}[\boldsymbol{s}(X)]$ can be obtained as a function of the natural parameter $\boldsymbol{\eta}$; namely

$$\nabla a(\boldsymbol{\eta}) = \mathbb{E}[\boldsymbol{s}(X)]$$

For minimal families (Def. 1.1.4), it turns out that this relationship is invertible [Jordan, 2010a, Sec. 8.4]. TODO: Prove this. This implies that a distribution in the exponential family can be parameterized not only by $\boldsymbol{\eta}$ – the canonical parametrization – but also by μ – the *mean parametrization*. TODO: Prove that we can always reparametrize a non-minimal EF into a minimal EF.

2 Frequentist Inference

2.1 Maximum Likelihood Estimation

The goal for maximum likelihood is to determine the parameter

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

Let us assume that $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ are i.i.d observations from a fixed exponential family, so that the likelihood has form (1.3.6). Let us compute the gradient with respect to the natural parameter $\boldsymbol{\eta}$ of $\ell(\boldsymbol{\eta}) := \log p(\boldsymbol{x} \mid \boldsymbol{\eta})$

[12] Note that the log normalizer function is assumed to be continuously differentiable by definition.
[13] For justification, see https://www2.stat.duke.edu/courses/Spring05/sta215/lec/wk06a.pdf.

$$\nabla_{\boldsymbol{\eta}} \ell(\boldsymbol{\eta}) = \sum_{i=1}^{n} \boldsymbol{s}(\boldsymbol{x}_i) - n \, \nabla_{\boldsymbol{\eta}} a(\boldsymbol{\eta})$$

Setting the gradient to zero, we obtain

$$\nabla_{\boldsymbol{\eta}} a(\boldsymbol{\eta}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{s}(\boldsymbol{x}_i)$$

But $\nabla_{\boldsymbol{\eta}} a(\boldsymbol{\eta}) = \mathbb{E}[\boldsymbol{s}(X)]$ (see Proposition 1.3.1). Thus, we should set $\boldsymbol{\theta}_{ML}$ such that

$$\mu(\boldsymbol{\theta}_{ML}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{s}(\boldsymbol{x}_i) \tag{2.1.2}$$

where $\mu := \mathbb{E}[\boldsymbol{s}(X)]$ refers to the mean parametrization (Sec. 1.3.6) of the likelihood. [14]

2.2 Latent variable models

Some models have latent variables associated with each observation. Let us assume that $(\boldsymbol{x}, \boldsymbol{z}) = ((\boldsymbol{x}_1, \boldsymbol{z}_1), ..., (\boldsymbol{x}_n, \boldsymbol{z}_n))$ where $\boldsymbol{x}$ is observed data and $\boldsymbol{z}$ is unobserved data. Then we say that the *complete data likelihood* is an exponential family if we can write

$$p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) = h(\boldsymbol{x}, \boldsymbol{z}) \exp\left\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z}) - a(\boldsymbol{\eta}(\boldsymbol{\theta}))\right\}. \tag{2.2.1}$$

for some functions $h, \boldsymbol{\eta}, \boldsymbol{s}, a$.

2.3 Expectation Maximization

When a model contains latent variables as in Eq. (2.3.26), it is not possible to use the maximum likelihood strategy of Sec. 2.1 directly. Indeed, the maximum likelihood solution of Eq. (2.1.2) cannot be applied because it is a function of unobserved variables $\boldsymbol{z}$. However, when complete data likelihood is an exponential family as in Eq. (2.3.26), we can perform frequentist inference via Expectation Maximization (EM). We see how this plays out in exponential families by following the logic of Section 2.1; see Salakhutdinov et al. [2002, Sec. 3] and Miller [2011] for useful references. TODO: Rewrite to clarify that EM applies more broadly than just to EF models.

2.3.1 The EM algorithm

The **Expectation Maximization (EM)** algorithm is to recursively update parameters $\boldsymbol{\theta}$ by solving at iteration $t = 0, 1, \ldots$

$$\boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}} \; Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) \tag{2.3.1a}$$

where

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) \triangleq \mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})}\Big[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big] \tag{2.3.1b}$$

Notation 2.3.1. (*Notation for expected values.*) In the Q-function of Eq. (2.3.1b) above, our notation means

$$\mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})}\Big[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big] \triangleq \int \ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) \, p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)}) \, \mu(d\boldsymbol{z}).$$

[14] TODO: This switching of parameterization should be handled more constructively.

This Q function is perhaps more clearly denoted by Miller [2011] as

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\Big[\ln p_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Z}) \mid \boldsymbol{X} = \boldsymbol{x}\Big].$$

DISCUSS: However, I believe both notations are somewhat misleading. They only make sense formally, because the density of a random variable $p(\boldsymbol{Y})$ is not a function of the random variable $\boldsymbol{Y}$. △

In the EM algorithm of Eq. (2.3.1), we refer to Eq. (2.3.1b) as the *E-step* and Eq. (2.3.1a) as the *M-step*.

What justifies this particular algorithm? One way to justify EM is as a special case of Coordinate Ascent Variational Inference (CAVI) [Wojnowicz, XXXX].[15] In particular, suppose we have a frequentist latent variable model (with observed data $\boldsymbol{x}$, random latent variables $\boldsymbol{z}$, and fixed but unknown parameters $\boldsymbol{\theta}$) for which we can compute the exact posterior. In this setting, an unrestricted variational family is $\mathcal{Q} = \{q : q = q_{\boldsymbol{z}} q_{\boldsymbol{\theta}} \text{ where } q_{\boldsymbol{\theta}} = \delta_{\boldsymbol{\theta}}\}$. Now at the t-th iteration of coordinate ascent, we do:

- *Update to* $q_{\boldsymbol{z}}$. We have:

$$\texttt{VLBO}\left(q_{\boldsymbol{z}}, q_{\boldsymbol{\theta}}^{(t)}\right) = \log p(\boldsymbol{x} \mid \boldsymbol{\theta}^{(t)}) \quad \text{if} \quad q_{\boldsymbol{z}} = p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)}) \tag{2.3.2}$$

  That is, by setting $q_{\boldsymbol{z}}$ to the exact posterior, the variational lower bound (VLBO) becomes tight, exactly equaling the log marginal likelihood. Hence, it is the optimal update when there are no restrictions on $\mathcal{Q}_z \ni q_{\boldsymbol{z}}$. This is precisely the E-step of the EM algorithm.

- *Update to* $q_{\boldsymbol{\theta}}$. To update $q_{\boldsymbol{\theta}}$, we note that when we evaluate the variational lower bound at the exact posterior over latent variables, we obtain

$$\texttt{VLBO}\left(q_{\boldsymbol{\theta}}, q_{\boldsymbol{z}}^{(t)}\right) = \underbrace{Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})}_{\text{energy}} + \underbrace{H(\boldsymbol{\theta}^{(t)})}_{\text{entropy}} \tag{2.3.3}$$

  And so the optimal coordinate ascent step for $q_{\boldsymbol{\theta}}$ amounts to maximizing the Q-function

$$\max_{q_{\boldsymbol{\theta}}} \texttt{VLBO}\left(q_{\boldsymbol{\theta}}, q_{\boldsymbol{z}}^{(t)}\right) = \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}). \tag{2.3.4}$$

  This is precisely the M-step of the EM algorithm.

TODO: Justify the EM algorithm WITHOUT going through CAVI. How did the original EM paper do it? How did Neal's "perspective" paper do it? DISCUSS: Does Prop 2.3.3 give a/the justification?

2.3.2 Gradient of the Q-function

Here, we give the gradient of the Q-function in Prop. 2.3.1. Using that, we can provide an interesting representation of the EM step Eq. (2.3.1) in Prop. 2.3.2.

Proposition 2.3.1. [The gradient of the Q-function [Salakhutdinov et al., 2002].] *The gradient of the Q-function of Eq.* (2.3.1b) *is given by*

$$\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}} = \mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})}[\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z})] - \mathbb{E}_{p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})}[\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z})] \tag{2.3.5}$$

[15] My notes in Wojnowicz [XXXX] cite Beal [2003] for this insight. However, per Salakhutdinov et al. [2003, pp. 2], I believe this perspective on EM may trace all the way back to Neal and Hinton [1998].

Proof of Prop. 2.3.1. No proof was provided in Salakhutdinov et al. [2002]. We provide a proof here.

We have

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) &\overset{1}{=} p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) \Big(\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z}) - \mathbb{E}_{p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})}[\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z})] \Big) \\
\implies \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) &\overset{2}{=} \frac{\frac{\partial p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}{p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})} \overset{3}{=} \boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z}) - \mathbb{E}_{p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})}[\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z})] \\
\implies \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) &\overset{4}{=} \int \frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)}) \, \mu(d\boldsymbol{z}) \\
&\overset{5}{=} \left(\int \boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z}) p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)}) \, \mu(d\boldsymbol{z}) \right) - \mathbb{E}_{p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})}[\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z})] \\
&\overset{6}{=} \mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})}[\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z})] - \mathbb{E}_{p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})}[\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z})]
\end{aligned}$$

Here, in (1), we used the definition in Eq. (2.2.1) of exponential family for latent variable models, as well as the chain rule for derivatives. Note that here $\mathbb{E}_{p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})}[\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z})] = \int \boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z}) p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) \mu(d\boldsymbol{x}, d\boldsymbol{z})$. In (2), we used the derivative of the logarithm along with chain rule for derivatives. In (3) we substituted the result of the first line. In (4), we used the definition in Eq. (2.3.1b) of the Q-function, along with an (assumed) interchange of the integral and the derivative. In (5), we used linearity of the integral; in the second summand, we have a constant with respect to the density $p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})$, which integrates to 1. In (6), we simply applied notation. □

Corollary 2.3.1. [The gradient of the Q function as the difference in expected sufficient statistics with clamped and unclamped data.] *By Prop. 2.3.1, the gradient of the Q-function evaluated at the current estimate of the parameter*

$$\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = \mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})}[\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z})] - \mathbb{E}_{p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}^{(t)})}[\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z})] \tag{2.3.6}$$

can be interpreted as the difference in the expected sufficient statistic vector when the observed data is clamped and unclamped.

Notation 2.3.2. The interpretation of the RHS of Eq. (2.3.5) was given in Notation 2.3.1. Using the alternative notation of Miller [2011] that we mentioned in Notation 2.3.1, we could also write the gradient in Eq. (2.3.5) as

$$\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\boldsymbol{s}(\boldsymbol{X}, \boldsymbol{Z}) \mid \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{s}(\boldsymbol{X}, \boldsymbol{Z})] \tag{2.3.7}$$

Similarly, we could write Eq. (2.3.6) as

$$\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\boldsymbol{s}(\boldsymbol{X}, \boldsymbol{Z}) \mid \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\boldsymbol{s}(\boldsymbol{X}, \boldsymbol{Z})] \tag{2.3.8}$$

△

Proposition 2.3.2. [EM update to parameters.] *The solution to the EM objective in Eq. (2.3.1a) for obtaining the next parameter value $\boldsymbol{\theta}^{(t+1)}$ given the previous parameter value $\boldsymbol{\theta}^{(t)}$ is given by finding the $\boldsymbol{\theta}^{(t+1)}$ that solves*

$$\mathbb{E}_{\boldsymbol{\theta}^{(t+1)}}[\boldsymbol{s}(\boldsymbol{X}, \boldsymbol{Z})] = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\boldsymbol{s}(\boldsymbol{X}, \boldsymbol{Z}) \mid \boldsymbol{X} = \boldsymbol{x}] \tag{2.3.9}$$

That is, we must find the parameter value which makes the expected sufficient statistic vector with unclamped data equal to that with clamped data.

Proof. We find a local optimum of the Q-function by finding its critical point, that is by solving $\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}} = 0$. TODO: How do we know that we don't need to look at boundary conditions? How do we know that the optimum is a maximum and not a minimum? By applying the representation of $\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}}$ in Eq. (2.3.7), we see that this is equivalent to finding $\boldsymbol{\theta} \in \boldsymbol{\theta}$ which solves the equation

$$\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{s}(\boldsymbol{X}, \boldsymbol{Z})] = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\boldsymbol{s}(\boldsymbol{X}, \boldsymbol{Z}) \mid \boldsymbol{X} = \boldsymbol{x}]$$

Setting the solution $\boldsymbol{\theta}^*$ as the next parameter value $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$ gives the result. TODO: How do we know the solution exists? □

2.3.3 Relating the EM step and the ordinary gradient

Proposition 2.3.3. [The Q-function and log marginal likelihood have equal gradients [Salakhutdinov et al., 2002].] *Consider a latent variable model whose complete data likelihood $p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})$ is in the exponential family. Define its log marginal likelihood as*

$$\ell(\boldsymbol{\theta}) \triangleq \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \log \int p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) \mu(d\boldsymbol{z}). \tag{2.3.10}$$

Then

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}. \tag{2.3.11}$$

Proof. No proof was provided in Salakhutdinov et al. [2002]. We provide a proof here.

On the one hand, we have

$$\begin{aligned}
Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) &= \int p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)}) \log p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) \, \mu(d\boldsymbol{z}) && \text{Def. Q function} \\
&= \int \frac{p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}^{(t)})}{p(\boldsymbol{x} \mid \boldsymbol{\theta}^{(t)})} \log p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) \, \mu(d\boldsymbol{z}) && \text{Bayes rule} \\
\implies \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} &= \int \frac{p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}^{(t)})}{p(\boldsymbol{x} \mid \boldsymbol{\theta}^{(t)})} \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \mu(d\boldsymbol{z}) && \text{Assume deriv interchanges with int.} \\
&= \int \frac{\cancel{p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}^{(t)})}}{p(\boldsymbol{x} \mid \boldsymbol{\theta}^{(t)})} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}}{\cancel{p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}^{(t)})}} \mu(d\boldsymbol{z}) && \text{Deriv. of logarithms} \\
&= \frac{1}{p(\boldsymbol{x} \mid \boldsymbol{\theta}^{(t)})} \int \frac{\partial}{\partial \boldsymbol{\theta}} p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \mu(d\boldsymbol{z}) && \text{Pull out constant}
\end{aligned} \tag{2.3.12}$$

On the other hand, we have

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} &= \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{x} \mid \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} && \text{Definition of } \ell \\
&= \frac{\frac{\partial}{\partial \boldsymbol{\theta}} p(\boldsymbol{x} \mid \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}}{p(\boldsymbol{x} \mid \boldsymbol{\theta}^{(t)})} && \text{Deriv of logarithm} \\
&= \frac{1}{p(\boldsymbol{x} \mid \boldsymbol{\theta}^{(t)})} \frac{\partial}{\partial \boldsymbol{\theta}} \left[\int p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) \, \mu(d\boldsymbol{z}) \right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} && \text{Law of total probability} \\
&= \frac{1}{p(\boldsymbol{x} \mid \boldsymbol{\theta}^{(t)})} \int \frac{\partial}{\partial \boldsymbol{\theta}} p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \mu(d\boldsymbol{z}) && \text{Assume deriv. interchanges with int.}
\end{aligned} \tag{2.3.13}$$

Together, Eq. (2.3.12) and Eq. (2.3.13) imply that

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \left(= \frac{1}{p(\boldsymbol{x} \mid \boldsymbol{\theta}^{(t)})} \int \frac{\partial}{\partial \boldsymbol{\theta}} p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \mu(d\boldsymbol{z}) \right)$$

as claimed.

□

TODO: It might be possible to give a much faster proof using the idea of EM as exact CAVI inference.

Corollary 2.3.2. [The ordinary gradient of the log marginal likelihood of an exponential family latent variable model.]

We have

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\boldsymbol{s}(\boldsymbol{X}, \boldsymbol{Z}) \mid \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\boldsymbol{s}(\boldsymbol{X}, \boldsymbol{Z})] \tag{2.3.14}$$

Proof. This follows immediately from the gradient of the Q-function (see Eq. (2.3.8)) along with the fact that the Q-function and the log marginal likelihood have equal gradients when evaluated at the current parameter value (Prop. 2.3.3). □

Proposition 2.3.4. [Relating the EM step to the ordinary gradient [Salakhutdinov et al., 2002].] *There exists a symmetric positive define "transformation matrix" $\boldsymbol{P}(\boldsymbol{\theta}^{(t)})$ such that the EM step $\boldsymbol{\theta}^{(t-1)} - \boldsymbol{\theta}^{(t)}$ can be expressed as*

$$\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)} = \boldsymbol{P}(\boldsymbol{\theta}^{(t)}) \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \tag{2.3.15}$$

where ℓ the the log marginal likelihood as in Eq. (2.3.10).

Proof. Define

$$\overline{T}(\boldsymbol{\theta}) \triangleq \int p(\boldsymbol{z}, \boldsymbol{x} \mid \boldsymbol{\theta}) \boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z}) \mu_{\boldsymbol{x},\boldsymbol{z}}(d\boldsymbol{x}, d\boldsymbol{z}) \tag{2.3.16}$$

$$\overline{T}_z(\boldsymbol{\theta}) \triangleq \int p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}) \boldsymbol{s}(\boldsymbol{x}, \boldsymbol{z}) \mu_{\boldsymbol{z}}(d\boldsymbol{z}) \tag{2.3.17}$$

Using these definitions, we can write the gradient of the log marginal likelihood from Eq. (2.3.14) as

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = \overline{T}_z(\boldsymbol{\theta}^{(t)}) - \overline{T}(\boldsymbol{\theta}^{(t)}) \tag{2.3.18}$$

Using these same definitions, the EM update of Eq. (2.3.9) is given by $\boldsymbol{\theta}^{(t+1)}$ which solves

$$\overline{T}(\boldsymbol{\theta}^{(t+1)}) = \overline{T}_z(\boldsymbol{\theta}^{(t)})$$

Since $\overline{T}$ is an invertible function (see Sec. 1.3.6), the EM update is given by

$$\boldsymbol{\theta}^{(t+1)} = \overline{T}^{-1} \overline{T}_z(\boldsymbol{\theta}^{(t)}) \tag{2.3.19}$$

Substituting Eq. (2.3.19) and Eq. (2.3.18) into the claim of Eq. (2.3.15) to be justified, we obtain the question of whether

$$\left[\overline{T}^{-1} \overline{T}_z(\boldsymbol{\theta}^{(t)}) - \boldsymbol{\theta}^{(t)}\right] \stackrel{?}{=} \boldsymbol{P}(\boldsymbol{\theta}^{(t)}) \left[\overline{T}_z(\boldsymbol{\theta}^{(t)}) - \overline{T}(\boldsymbol{\theta}^{(t)})\right] \tag{2.3.20}$$

for some symmetric positive definite matrix $\boldsymbol{P}$.

By defining notation for the LHS and RHS of Eq. (2.3.20)

$$\boldsymbol{v}(\boldsymbol{\theta}^{(t)}) \triangleq \overline{T}^{-1}\overline{T}_z(\boldsymbol{\theta}^{(t)}) - \boldsymbol{\theta}^{(t)} \qquad \text{LHS} \tag{2.3.21}$$

$$\boldsymbol{u}(\boldsymbol{\theta}^{(t)}) \triangleq \overline{T}_z(\boldsymbol{\theta}^{(t)}) - \overline{T}(\boldsymbol{\theta}^{(t)}), \qquad \text{RHS} \tag{2.3.22}$$

we can further simplify Eq. (2.3.20). In particular, we must show that there exists symmetric positive definite matrix $\boldsymbol{P}$ such that

$$\boldsymbol{v}(\boldsymbol{\theta}^{(t)}) \stackrel{?}{=} \boldsymbol{P}(\boldsymbol{\theta}^{(t)})\boldsymbol{u}(\boldsymbol{\theta}^{(t)}) \tag{2.3.23}$$

Eq. (2.3.23) clearly holds by defining

$$\boldsymbol{P}(\boldsymbol{\theta}^{(t)}) \triangleq \frac{\boldsymbol{v}(\boldsymbol{\theta}^{(t)})\boldsymbol{v}(\boldsymbol{\theta}^{(t)})^\top}{\boldsymbol{v}(\boldsymbol{\theta}^{(t)})^\top \boldsymbol{u}(\boldsymbol{\theta}^{(t)})}. \tag{2.3.24}$$

In particular, by substituting Eq. (2.3.24) into Eq. (2.3.23), we find

$$\boldsymbol{P}\boldsymbol{u} = \frac{\boldsymbol{v}\cancel{\boldsymbol{v}^\top}}{\cancel{\boldsymbol{v}^\top}\boldsymbol{u}}\cancel{\boldsymbol{u}} = \boldsymbol{v}, \tag{2.3.25}$$

as desired.

Now all that remains is to show that $\boldsymbol{P}$ is symmetric positive definite. It is clear from Eq. (2.3.24) that $\boldsymbol{P}$ is symmetric. We also see from Eq. (2.3.24) that $\boldsymbol{P}$ is positive definite if $\boldsymbol{v}^\top \boldsymbol{u} > 0$, and that is shown in Salakhutdinov et al. [2002, Eqn. 26]. □

Corollary 2.3.3. *The EM step gives an ascent direction for the log marginal likelihood.*

Proof. Since the matrix $\boldsymbol{P}$ in Prop. 2.3.4 is symmetric positive definite, the EM step $\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}$ is an ascent direction by Nocedal and Wright [2006, pp.31, Eqn. (3.2)]. Namely, using shorthand notation $\nabla \ell_t \triangleq \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$, we compute the directional derivative as

$$\nabla \ell_t^\top (\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}) \stackrel{(Prop.\ 2.3.4)}{=} \nabla \ell_t^\top \boldsymbol{P} \nabla \ell_t \stackrel{p.d.}{>} 0,$$

and hence the EM step $\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}$ is an ascent direction. □

TODO:

Consider WHY/HOW gradient chasing could ever be faster than EM.

- Both are ascent directions. How/when is one better than another?
- Is there some way to improve EM using ordinary gradient but without bringing in the 2nd order info? I.e., is there some *other* $\boldsymbol{P}$ that would be better? If so, why?
- Qin et al. [2000, abstract] mentions that (at least in the context of HMMs), EM is especially slow (relative to direct optimization) in the low STN regime. Why would this be?
- Discuss with Jeff: Given this analysis, aren't both EM and direct optimization just different first-order methods, in contrast with bike/car/racecar analogy?
- Salakhutdinov et al. [2002] says the EM step has "positive projection" onto the true gradient of the likelihood function. How is this a projection? What are the implications for comparing EM and direct optimization?

Also, how does the $\boldsymbol{P}$ matrix magically keep update in param bounds?

2.3.4 Convergence behavior of the EM algorithm

Here, we discuss the convergence behavior of the EM algorithm, following Salakhutdinov et al. [2002] and Salakhutdinov et al. [2003].

The EM algorithm implicitly defines a mapping $\boldsymbol{M} : \Theta \to \Theta$ where $\boldsymbol{M}(\boldsymbol{\theta}^{(t)}) = \boldsymbol{\theta}^{(t+1)}$.

TODO: Use this material to address some of the questions at the end of the section on relating the EM step and the ordinary gradient.

2.3.5 The EM algorithm for i.i.d data

In the special case where $(\boldsymbol{x}, \boldsymbol{z}) = ((\boldsymbol{x}_1, \boldsymbol{z}_1), ..., (\boldsymbol{x}_n, \boldsymbol{z}_n))$ are n i.i.d samples from some exponential family, we can write the complete data likelihood of Eq. (2.2.1) as

$$p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) = \left(\prod_{i=1}^{n} \widetilde{h}(\boldsymbol{x}_i, \boldsymbol{z}_i) \right) \exp \left\{ \boldsymbol{\eta}(\boldsymbol{\theta})^\top \sum_{i=1}^{n} \widetilde{\boldsymbol{s}}(\boldsymbol{x}_i, \boldsymbol{z}_i) - n\, \widetilde{a}(\boldsymbol{\eta}(\boldsymbol{\theta})) \right\} \tag{2.3.26}$$

for some functions $\boldsymbol{\eta}, \widetilde{\boldsymbol{s}}, \widetilde{h}, \widetilde{a}$ describing the exponential family for a single observation (the so-called *unit complete data likelihood*).

Following the logic of Section 2.1, we expect that in the i.i.d case, we should select $\boldsymbol{\theta}^{(t+1)}$ such that

$$\mu(\boldsymbol{\theta}^{(t+1)}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})} \boldsymbol{s}(\boldsymbol{x}_i, \boldsymbol{z}_i) \tag{2.3.27}$$

where $\mu := \mathbb{E}[\boldsymbol{s}(\boldsymbol{x}_i, \boldsymbol{z}_i)]$ refers to the mean parametrization (Sec. 1.3.6) of the *unit complete data likelihood*. This is true. TODO: Justify. TODO: I believe that I read Eq. (2.3.27) in some paper – perhaps notes by Jordan. Find the source so that I can cite it and perhaps provide additional context.

Eq. (2.3.27) reveals 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.

Note that the update in Eq. (2.3.27) also corroborates with the more general update (not assuming i.i.d models) given in Eq. (2.3.9). TODO: Prove directly that the latter implies the former in the i.i.d case.

TODO: Check this section, especially with respect to the fact that I am dealing with three parametrizations here - $\mu, \boldsymbol{\theta}, \nu$; 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.

TODO (Notation): Fix notation here and throughout. For instance: (1) As usual, the boldfacing is giving me headaches. For instance, if we use boldface $\boldsymbol{\theta}$ to represent arbitrary params, then we should presumably use boldface for the natural and mean parameterizations. (2) Be more consistently careful about when we're referring to random variables, and when to realizations.

3 Bayesian inference for conjugate and semi-conjugate models

Conjugacy can be defined as follows [Gelman et al., 2013]. If $\mathcal{F}$ is a class of sampling distributions and $\mathcal{P}$ is a class of prior distributions for $\boldsymbol{\theta}$, then the class $\mathcal{P}$ is *conjugate* for $\mathcal{F}$ if

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

Conditional conjugacy (sometimes called semi-conjugacy) can be defined similarly [Gelman et al., 2013]. If $\mathcal{F}$ is a class of sampling distributions and $\mathcal{P}$ is a class of prior distributions for $\boldsymbol{\theta} \mid \phi$, then

the class $\mathcal{P}$ is *conditionally conjugate* for $\mathcal{F}$ if

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

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 [Gelman, 2006]. [16] 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 [Minka, 2001]. △

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.

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*

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

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

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

[16] 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.

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 [Hoff, 2009] (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.[17]. In other words, this reparametization gives us an interpretation of the the prior in terms of "equivalent data" [Box and Tiao, 2011].[18]

△

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

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

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

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

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

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

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

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

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

[18] I like this description. I first saw it referenced on pp. 517 of [Gelman, 2006].

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

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

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 $\boldsymbol{y} := (\boldsymbol{y}_1, ..., \boldsymbol{y}_n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

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

where $\boldsymbol{S}_\mu := \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^\top$ 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.

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

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

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

where S_μ was defined above. Thus, we have

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

And so the conjugate updates are given by

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

$$\boldsymbol{\Psi}' \leftarrow \boldsymbol{\Psi} + \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^\top \tag{3.2.7}$$

△

For interpretation of the parameters of the Inverse Wishart, see Remark C.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 $\sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{A} \boldsymbol{x}_i = \operatorname{tr}(\boldsymbol{A} \sum \boldsymbol{x}_i \boldsymbol{x}_i^\top)$ for $\boldsymbol{x} \in \mathbb{R}^d$, and $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ symmetric.

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

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

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

△

3.2.3 Example: Bayesian normal model with conditionally conjugate prior

Consider the following model with a normal sampling distribution and conditionally conjugate prior[19]:

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

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. [20] In particular

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

[19] 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.)

[20] 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 (D.2.2), noting that the natural parameters for a multivariate Gaussian are its precision and precision-weighted mean.

where

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

and

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

where

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

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. [21]

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.

Remark 3.2.2. (*Purpose of a prior on* $\boldsymbol{\Sigma}$) As mentioned by [Imai and Van Dyk, 2005] (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 $\boldsymbol{y} = (y_1, ..., y_n)^\top$ for such a model is also Gaussian.

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

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

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

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

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

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

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) \propto \exp \left\{ -\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right\} \tag{3.3.3a}$$

$$= \exp \left\{ -\frac{1}{2\sigma^2} (-2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\beta}) \right\} \tag{3.3.3b}$$

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

$$p(\boldsymbol{\beta}) \propto \exp \left\{ -\frac{1}{2} (-2\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_0 \boldsymbol{\mu}_0 + \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\beta}) \right\} \tag{3.3.4}$$

Thus, by Bayes rule

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

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 [Grosse et al., 2019] or [Bishop, 2006]. 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. (*Posterior parameters in terms of maximum likelihood estimates*) Now recall

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\text{ML}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \\ \text{Var}(\widehat{\boldsymbol{\beta}}_{\text{ML}}) &= \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \end{aligned}$$

So the posterior parameters in (3.3.2) have the interpretation

$$\underbrace{\widetilde{\boldsymbol{\Sigma}}^{-1}}_{\text{posterior precision}} = \underbrace{\boldsymbol{\Sigma}_0^{-1}}_{\text{prior precision}} + \underbrace{\frac{1}{\sigma^2} \boldsymbol{X}^\top \boldsymbol{X}}_{\text{“data" precision, } \text{Var}(\widehat{\boldsymbol{\beta}}_{\text{ML}})^{-1}}$$

$$\underbrace{\widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{\mu}}}_{\text{posterior precision-weighted mean}} = \underbrace{\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0}_{\text{prior precision-weighted mean}} + \underbrace{\frac{1}{\sigma^2} \boldsymbol{X}^\top \boldsymbol{y}}_{\text{“data" pwm } \text{Var}(\widehat{\boldsymbol{\beta}}_{\text{ML}})^{-1} \widehat{\boldsymbol{\beta}}_{\text{ML}}}$$

△

Remark 3.3.3. (*Bayesian linear regression as a compromise between the prior and maximum likelihood value.*) 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 [Hoff, 2009] (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 $\boldsymbol{\mu} \approx (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$, 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

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

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.2.1.

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 $\boldsymbol{y} = (y_1, ..., y_n)^\top$ for such a model is also Inverse Gamma.

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

$$\sigma^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right) \tag{3.3.5a}$$

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) \tag{3.3.5b}$$

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 (3.3.5) is given by

$$\sigma^2 \mid \boldsymbol{y}, \boldsymbol{\beta} \sim \mathcal{IG}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + SSR(\boldsymbol{\beta})}{2}\right) \tag{3.3.6a}$$

where the sum of squared residuals is

$$SSR(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{3.3.6b}$$

Proof. [22]

[22] For corroboration, see [Hoff, 2009, pp. 155], who obtains the same result.

$$p(\sigma^2 \mid \text{rest}) \propto \underbrace{\left(\frac{1}{\sigma}\right)^n \exp\left(-\frac{1}{2}\frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right)}_{\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)} \underbrace{\left(\sigma^2\right)^{-\frac{\nu_0}{2}-1} \exp\left(-\frac{\frac{1}{2}\nu_0\sigma_0^2}{\sigma^2}\right)}_{\sigma^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)}$$
$$= (\sigma^2)^{-\frac{n}{2}-\frac{\nu_0}{2}-1} \exp\left(-\frac{\frac{1}{2}(\text{SSR}(\boldsymbol{\beta})+\nu_0\sigma_0^2)}{\sigma^2}\right)$$

□

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 [Hoff, 2009].

A simple version of this model is[23] [24]:

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

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 [Hoff, 2009]) are given by

$$\begin{aligned}
\boldsymbol{\beta}_j \mid \boldsymbol{\Sigma}, \boldsymbol{\mu}, \sigma^2, \boldsymbol{y} &\sim \mathcal{N}(\boldsymbol{\mu}_j', \boldsymbol{\Sigma}_j') \\
\boldsymbol{\Sigma}_j' &= \left(\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}_j^\top \boldsymbol{X}_j\right)^{-1} \\
\boldsymbol{\mu}_j' &= \boldsymbol{\Sigma}_j'\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{\sigma^2}\boldsymbol{X}_j^\top \boldsymbol{y}_j\right)
\end{aligned}$$

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

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

[24] 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

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

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

$$\begin{aligned}
\boldsymbol{\Sigma} \mid \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \boldsymbol{\mu} &\sim \mathcal{W}^{-1}(\boldsymbol{\eta}', \boldsymbol{\Psi}') \\
\boldsymbol{\eta}' &= \boldsymbol{\eta}_0 + J \\
\boldsymbol{\Psi}' &= \boldsymbol{\Psi}_0 + \sum_{j=1}^{J} (\boldsymbol{\beta}_j - \boldsymbol{\mu})(\boldsymbol{\beta}_j - \boldsymbol{\mu})^\top
\end{aligned} \tag{3.4.2}$$

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 [Hoff, 2009] for ideas, although that discussion takes the prior on the mean to depend on the sampling variance σ^2. △

3.5 General formalism

Let us write the data model's density $p(x \mid \boldsymbol{\eta})$ in exponential family form, and assume that the prior $p(\boldsymbol{\eta} \mid \phi)$ has an exponential family construction as well.

$$
\begin{aligned}
p(x \mid \boldsymbol{\eta}) &= h(x) \exp\{\boldsymbol{\eta}^\top \boldsymbol{s}(x) - a(\boldsymbol{\eta})\} \\
p(\boldsymbol{\eta} \mid \phi) &= \widetilde{h}(\boldsymbol{\eta}) \exp\{\phi^\top \ \widetilde{\boldsymbol{s}}(\boldsymbol{\eta}) - \widetilde{a}(\phi)\}
\end{aligned}
$$

We use the tilde to denote that the carrier density h, sufficient statistics function $\boldsymbol{s}$, and log normalizer a can differ for these densities.

Our goal is to find the form of a conjugate prior, given the form of the data model. To do this, we work with the posterior

$$
\begin{aligned}
p(\boldsymbol{\eta} \mid x, \phi) &\propto p(x \mid \boldsymbol{\eta}) \quad p(\boldsymbol{\eta} \mid \phi) && \text{bayes law} \\
&\propto \cancel{h(x)} \exp\{\boldsymbol{\eta}^\top \boldsymbol{s}(x) - a(\boldsymbol{\eta})\} \quad \widetilde{h}(\boldsymbol{\eta}) \exp\{\phi^\top \ \widetilde{\boldsymbol{s}}(\boldsymbol{\eta}) - \widetilde{a}(\phi)\} && \text{e.f. form; constant of prop.} \\
&\overset{1}{=} \exp\left\{ \begin{bmatrix} \boldsymbol{s}(x) \\ 1 \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\eta} \\ -a(\boldsymbol{\eta}) \end{bmatrix} \right\} \quad \widetilde{h}(\boldsymbol{\eta}) \exp\left\{ \phi^\top \ \begin{bmatrix} \boldsymbol{\eta} \\ -a(\boldsymbol{\eta}) \end{bmatrix} - \widetilde{a}(\phi) \right\} && \text{explained below} \\
&= \widetilde{h}(\boldsymbol{\eta}) \exp\left\{ \left(\phi + \begin{bmatrix} \boldsymbol{s}(x) \\ 1 \end{bmatrix} \right)^\top \ \begin{bmatrix} \boldsymbol{\eta} \\ -a(\boldsymbol{\eta}) \end{bmatrix} - \widetilde{a}(\phi) \right\} && (3.5.1)
\end{aligned}
$$

where in (1) we reorganize the likelihood to isolate terms in $\boldsymbol{\eta}$, and then take $\widetilde{\boldsymbol{s}}(\boldsymbol{\eta})$ to have this same form in order to get conjugacy.

From this, we conclude

1. The vector $\widetilde{\boldsymbol{s}}(\boldsymbol{\eta}) = \begin{bmatrix} \boldsymbol{\eta} \\ -a(\boldsymbol{\eta}) \end{bmatrix}$ gives sufficient statistics of the conjugate prior, which can be used to identify its (or identify its non-existence).
2. We can decompose the prior hyperparameter as $\phi = \begin{bmatrix} \tau \\ n_0 \end{bmatrix}$, where τ has the dimensionality of the canonical parameter $\boldsymbol{\eta}$ and n_0 is a scalar. Moreover, τ can be interpreted as the prior sufficient statistics and n_0 can be interpreted as the prior sample size. (The latter becomes more clear below in the case of i.i.d samples.)
3. The prior-to-posterior conversion can be summarized with the following update rules

$$
\begin{aligned}
\tau &\to \tau + \boldsymbol{s}(x) \\
n_0 &\to n_0 + 1
\end{aligned}
\qquad (3.5.2)
$$

Remark 3.5.1. As this derivation shows, there are multiple possible conjugate priors, depending on the choice of $\widetilde{h}$. For instance, whenever the normal distribution is conjugate, so is the truncated normal, since those distributions differ only in their carrier density (and therefore their log normalizer; see Remark 1.2.2.) More generally, the prior density can be defined with respect to an arbitrarily-chosen dominating measure ν on Φ; for a discussion, see here . △

Remark 3.5.2. Note that 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. See Section D.2. △

3.5.1 General formalism for multiple i.i.d samples

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

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

as the likelihood (see Section 1.3.5). In this case, the posterior under the conjugate prior (3.5.1) becomes

$$\begin{aligned} p(\boldsymbol{\eta} \mid \boldsymbol{x}, \phi) &\propto \sum_{i=1}^{\infty} p(x_i \mid \boldsymbol{\eta})\; p(\boldsymbol{\eta} \mid \phi) \\ &\propto \exp\left\{ \left(\phi + \begin{bmatrix} \sum_{i=1}^{n} \boldsymbol{s}(x_i) \\ n \end{bmatrix} \right)^\top \begin{bmatrix} \boldsymbol{\eta} \\ -a(\boldsymbol{\eta}) \end{bmatrix} - \widetilde{a}(\phi) \right\} \end{aligned} \tag{3.5.3}$$

From this, we see that under multiple i.i.d samples, the prior-to-posterior conversion can be summarized with the following update rules

$$\tau \to \tau + \sum_{i=1}^{n} \boldsymbol{s}(x_i) \tag{3.5.4a}$$

$$n_0 \to n_0 + n \tag{3.5.4b}$$

3.5.2 Application: Finding conjugate priors (and identifying when there isn't one)

In application, we don't often use $\begin{bmatrix} \boldsymbol{\eta} \\ -a(\boldsymbol{\eta}) \end{bmatrix}$ to identify the sufficient statistics, but the set of basis functions necessary to construct $\begin{bmatrix} \boldsymbol{\eta} \\ -a(\boldsymbol{\eta}) \end{bmatrix}$. Said differently, below we show how a (semi-)conjugate prior can be determined by simply rewriting the likelihood as a function of the parameter for which we want to set the (semi-)conjugate prior.

Example 3.5.1. Let $X \sim \mathcal{N}(\mu, \beta^{-1})$, where we parameterize in terms of the precision parameter β. The log pdf under the natural parameterization is given by

$$\log p(x \mid \mu, \beta) = \begin{bmatrix} \beta\mu \\ -\frac{1}{2}\beta \end{bmatrix}^\top \begin{bmatrix} x \\ x^2 \end{bmatrix} + \frac{1}{2}\big(\log \beta - \beta\mu^2 - \log(2\pi) \big) \tag{3.5.5}$$

Suppose we want to find a (semi-)conjugate prior for the precision parameter β. Considering (3.5.5) as a function of β, the basis functions needed are

$$\widetilde{\boldsymbol{s}}(\beta) := \begin{bmatrix} \beta \\ \log \beta \end{bmatrix}$$

which is the sufficient statistics function of the gamma distribution. Thus, a conjugate prior for the precision parameter is the gamma distribution. (Note that a conjugate prior can have any desired carrier density, so we can freely set $h \equiv 1$; see Remark 3.5.1.)[25] △

[25] Well, technically, the carrier density must be chosen such that the ambient measure dominates the prior distribution. I think that will always hold regardless of choice of h (since h basically is a Radon-Nikodym derivative), but I should double check this.

Example 3.5.2. Let $X \sim \text{Gamma}(\alpha, \beta)$, where $\text{Gamma}(\alpha, \beta)$ refers to the shape-rate parametrization. The log pdf is given by

$$\log p(x \mid \alpha, \beta) = \alpha \log \beta - \ln \Gamma(\alpha) - (\alpha - 1)x - \beta x \tag{3.5.6}$$

Suppose we want to find a (semi-)conjugate prior for the shape parameter α. Considering (3.5.5) as a function of α, the basis functions needed are

$$\widetilde{\boldsymbol{s}}(\alpha) := \begin{bmatrix} \alpha \\ \Gamma(\alpha) \end{bmatrix}$$

There is no exponential family distribution with these sufficient statistics.[26] TODO: Explain. △

3.6 Compound distributions

3.6.1 The negative binomial distribution

In Prop. 3.6.1, we show that the negative binomial distribution can be represented as a compound Poisson-Gamma distribution. This fact is useful, e.g. for easily computing its mean and variance. This representation fits into the conjugate Bayes section since the Gamma is the conjugate prior for the Poisson.

Proposition 3.6.1. *Let*

$$X \sim \textit{NegativeBinomial}\big(r, p\big)$$

We can write this as the marginal distribution of a Poisson observation model with a Gamma prior

$$\begin{aligned} X &\sim \textit{Poisson}(\lambda) \\ \lambda &\sim \textit{Gamma}\left(r, \frac{p}{1-p}\right). \end{aligned}$$

Proof. Let $f_{(r,p)}$ be the density of a Negative Binomial random variable, g_λ be the density of a Poisson random variable and $h_{(\alpha,\beta)}$ be the density of a Gamma random variable. Then we must show that

$$f_{(r,p)}(k) = \int_0^\infty g_\lambda(k)\; h_{(r,p/1-p)}(\lambda)\; d\lambda$$

for any non-negative integer $k = 0, 1, 2, \ldots$.

[26] This is stated by [Winn et al., 2005], although I am not currently sure why this is true.

We have

$$
\begin{aligned}
f_{(r,p)}(k) &\stackrel{?}{=} \int_0^\infty g_\lambda(k) h_{(r,p/1-p)}(\lambda)\, d\lambda \\
&= \int_0^\infty \underbrace{\left(\frac{\lambda^k e^{-\lambda}}{k!}\right)}_{\text{Poisson}} \underbrace{\frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}}_{\text{Gamma}}\, d\lambda && \text{Substitute pdfs} \\
&= \frac{\beta^\alpha}{k!\Gamma(\alpha)} \underbrace{\int_0^\infty \lambda^{\alpha+k-1} e^{-(\beta+1)\lambda}\, d\lambda}_{\text{Unnormalized } Gamma(\alpha+k,\beta+1)} && \text{Combine like terms, pull out constant} \\
&= \frac{\beta^\alpha}{k!\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{(\beta+1)^{\alpha+k}} && \text{Normalizing constant of Gamma dist'n} \\
&\stackrel{(1)}{=} \binom{k+\alpha-1}{k} \frac{\beta^\alpha}{(\beta+1)^{\alpha+k}} && \text{See below} \\
&\stackrel{(2)}{=} \binom{k+r-1}{k} \left(\frac{p}{1-p}\right)^r (1-p)^{r+k} && \text{See below}
\end{aligned}
$$

which is indeed the density of NegativeBinomial(r,p).

Equation (1) follows from the definition of binomial coefficients with real numbers. In detail, the definition of binomial coefficients

$$\binom{x}{y} \triangleq \frac{x!}{(x-y)!y!}$$

when x, y are non-negative numbers such that $y \leq x$ is well-known. The definition can be generalized to real-numbers x, y by

$$\binom{x}{y} \triangleq \frac{\Gamma(x+1)}{\Gamma(x-y+1)\Gamma(y+1)}$$

since $\Gamma(x+1) = x!$.

Equation (2) follows by substituting the choices shape $\alpha = r$ and rate $\beta = \frac{p}{1-p}$ for the parameters of the Gamma distribution. Note that $\beta = \frac{p}{1-p} \implies \beta + 1 = \frac{1}{1-p}$. □

In Prop. 3.6.2 below, we use the compound representation of the Negative Binomial (Prop. 3.6.1) to easily obtain its mean and variance. These expressions may be hard to prove directly TODO: Confirm/justify.

Proposition 3.6.2. *Let*

$$X \sim \textit{NegativeBinomial}(r,p).$$

Then

$$
\begin{aligned}
\mathbb{E}[X] &= \frac{r(1-p)}{p} \\
\text{Var}[X] &= \frac{r(1-p)}{p^2}
\end{aligned}
$$

Proof. Representing the NegativeBinomial(r,p) distribution as the marginal of a $Poisson(\lambda)$ with

Gamma($\alpha = r, \beta = \frac{p}{1-p}$) prior (Prop. 3.6.1), we obtain

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}\mathbb{E}[X \mid \lambda] && \text{Law of iterated expectation} \\
&= \mathbb{E}\lambda && \text{Poisson mean} \\
&= \frac{\alpha}{\beta} && \text{Gamma mean} \\
&= \frac{r(1-p)}{p} && \text{Reparametrize}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\text{Var}[X] &= \mathbb{E}\text{Var}[X \mid \lambda] + \text{Var}\mathbb{E}[X \mid \lambda] && \text{Law of total variance} \\
&= \mathbb{E}\lambda + \text{Var}\lambda && \text{Poisson mean, var} \\
&= \frac{\alpha}{\beta} + \frac{\alpha}{\beta^2} && \text{Gamma mean, var} \\
&= \frac{\alpha(\beta+1)}{\beta^2} && \text{algebra} \\
&= \frac{r(1-p)}{p^2} && \text{Reparametrize, algebra}
\end{aligned}$$

□

Remark 3.6.1. (*Negative binomial as an overdispersed count model.*)

Note from Prop. 3.6.2 that if $X \sim \text{NegativeBinomial}(r, p)$, we have

$$\text{Var}[X] = \frac{\mathbb{E}[X]}{p},$$

where $p \in [0, 1]$. Hence, a Negative Binomial random variable has greater variance than a Poisson, since if $Y \sim \text{Poisson}(\lambda)$, we have

$$\text{Var}[Y] = \mathbb{E}[Y].$$

△

TODO: Expand the above remark to show that a negative binomial is specifically an overdispersed *Poisson*.

Proposition 3.6.3. [Poisson as the limiting distribution of Negative Binomials.] *Consider a sequence $X_n \sim$ NegativeBinomial(r_n, p_n) such that $r_n \to \infty$ and $p_n \to 1$ but where we hold the mean constant $\mathbb{E}[X_n] = \frac{r_n(1-p_n)}{p_n} = \lambda$. Then the limiting distribution of X_n is Poisson(λ).*

TODO: Give some intuition on what these limit conditions mean, in terms of the "successes" and "failures" of a negative binomial. Note that this will require giving some intuition up front about the construction of the negative binomial density.

Proof. The density of the random variables in the sequence is given by

$$\begin{aligned}
f(k; r_n, p_n) &= \frac{\Gamma(k+r_n)}{k!\Gamma(r_n)}(1-p_n)^k p_n^{r_n} && \text{Substitute pmf} \\
&= \frac{\Gamma(k+r_n)}{k!\Gamma(r_n)}\left(\frac{\lambda}{\lambda+r_n}\right)^k\left(\frac{r_n}{\lambda+r_n}\right)^{r_n} && \text{Substitute constant mean} \\
&= \frac{\lambda^k}{k!}\underbrace{\frac{\Gamma(k+r_n)}{\Gamma(r_n)(r_n+\lambda)^k}}_{\to 1}\underbrace{\left(\frac{1}{1+\frac{\lambda}{r_n}}\right)^{r_n}}_{\to e^{-\lambda}} && \text{Algebra} \\
&\to \frac{\lambda^k}{k!}e^{-\lambda}
\end{aligned}$$

which is the density of a Poisson(λ) random variable. □

References

Alvarez, I., Niemi, J., and Simpson, M. (2014). Bayesian inference for a covariance matrix. *arXiv preprint arXiv:1408.4050*.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. University of London, University College London (United Kingdom).

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, volume 405. John Wiley & Sons.

Bickson, D. (2008). Gaussian belief propagation: Theory and aplication. *arXiv preprint arXiv:0811.2518*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Box, G. E. and Tiao, G. C. (2011). *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons.

Burkardt, J. (2014). The truncated normal distribution. *Department of Scientific Computing Website, Florida State University*, pages 1–35.

Chua, K. (2019). Stats 210a lecture notes. https://kchua.github.io/notes/Stats210A.pdf.

Cole, D. (2020). *Parameter redundancy and identifiability*. CRC Press.

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

Englehardt, B. (2013). Gaussian models. https://www.cs.princeton.edu/~bee/courses/scribe/lec_09_09_2013.pdf.

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Grosse, R. et al. (2019). Bayesian linear regression. https://www.cs.toronto.edu/~rgrosse/courses/csc411_f18/slides/lec19-slides.pdf.

Gundersen, G. (2019). Completing the square. http://gregorygundersen.com/blog/2019/09/18/completing-the-square/.

Gupta, M. and Srivastava, S. (2010). Parametric bayesian estimation of differential entropy and relative entropy. *Entropy*, 12(4):818–843.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*, volume 580. Springer.

Hughes, M. (2012). Inverse wishart distribution. https://www.michaelchughes.com/blog/probability-basics/inverse-wishart-distribution/.

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

Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P., and Datta, S. R. (2016). Composing graphical models with neural networks for structured representations and fast inference. *Advances in neural information processing systems*, 29:2946–2954.

Jordan, M. (2010a). The exponential family: Basics. https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/other-readings/chapter8.pdf.

Jordan, M. (2010b). The exponential family: Conjugate priors. https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/other-readings/chapter9.pdf.

Krishnan, R. G., Shalit, U., and Sontag, D. (2016). Structured inference networks for nonlinear state space models. *arXiv preprint arXiv:1609.09869*.

Miller, J. (2011). Why em makes sense (video lecture from course on machine learning). https://www.youtube.com/watch?v=6JZ-PKpx5Kc&list=PLD0F06AA0D2E8FFBA&index=117.

Minka, T. (2001). Bayesian inference of a uniform distribution. https://tminka.github.io/papers/minka-uniform.pdf.

Neal, R. M. and Hinton, G. E. (1998). A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer.

Nielsen, F. and Garcia, V. (2009). Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*.

Nielsen, F. and Nock, R. (2010). Entropies and cross-entropies of exponential families. In *2010 IEEE International Conference on Image Processing*, pages 3621–3624. IEEE.

Nocedal, J. and Wright, S. J. (2006). *Numerical optimization*. Springer.

Qin, F., Auerbach, A., and Sachs, F. (2000). A direct optimization approach to hidden markov modeling for single channel kinetics. *Biophysical journal*, 79(4):1915–1927.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear gaussian models. *Neural computation*, 11(2):305–345.

Salakhutdinov, R., Roweis, S., and Ghahramani, Z. (2002). Relationship between gradient and em steps in latent variable models.

Salakhutdinov, R., Roweis, S. T., and Ghahramani, Z. (2003). Optimization with em and expectation-conjugate-gradient. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 672–679.

Taylor, J. (2013). Multiparameter exponential families, part ii. http://statweb.stanford.edu/~jtaylo/courses/stats306b/restricted/notebooks/multiparameter_partII.pdf.

Winn, J., Bishop, C. M., and Jaakkola, T. (2005). Variational message passing. *Journal of Machine Learning Research*, 6(4).

Wojnowicz, M. (2022). Variational inference for categorical models. https://github.com/mikewojnowicz/fall_2020/blob/master/reports/categorical_models_with_vi/categorical_models_with_vi.pdf.

Wojnowicz, M. (XXXX). *Variational Inference: Foundations*. Available (with permission) at https://github.com/mikewojnowicz/vi_foundations.

Wolpert, R. (2011). Change of variables. https://www2.stat.duke.edu/courses/Spring11/sta114/lec/114mvnorm.pdf.

A Matrix Facts

A.1 Multivariate completing the square

A nice overview of multivariate completing the square is given by [Gundersen, 2019]. See also [Bishop, 2006, pp. 86] for application to the Gaussian case.

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

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

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

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

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

B Gaussian Facts

B.1 Entropy facts about Multivariate Gaussian

If p, q are the densities of two different d-variate Gaussian distributions with parameters $\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p$ and $\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q$, respectively, then the entropy is given by

$$\mathbb{H}[q] = \frac{1}{2} \log \left[(2\pi e)^d |\boldsymbol{\Sigma}_q| \right] \tag{B.1.1}$$

The KL divergence is given by

$$\begin{aligned} \mathtt{KL}\big[q \,||\, p\big] = \frac{1}{2}\Big[&\log \frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_q|} - d \\ &+ (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^\top \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p) + \operatorname{tr}(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q) \Big] \end{aligned} \tag{B.1.2}$$

The cross-entropy of two multivariate Gaussians can then be determined from (B.1.1) and (B.1.2) via the relation

$$\mathbb{H}[q, p] = \mathbb{H}[q] + \mathtt{KL}\big[q \,\big|\big|\, p\big] \tag{B.1.3}$$

B.2 The simplest linear Gaussian model

Proposition B.2.1. *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

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

Proof. See Section 2.3.3 of [Bishop, 2006]. The necessary computation is for the density

$$p(y) = \int p(y \mid x)\, p(x)\, dx \tag{B.2.1}$$

□

Remark B.2.1. The necesary computation in (B.2.1) can be seen as the convolution of two Gaussians. △

Remark B.2.2. 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 [Bishop, 2006], or [Roweis and Ghahramani, 1999]. △

B.3 Exponential family representation of Multivariate Gaussian in message passing

In a dissertation on Gaussian Belief Propagation [Bickson, 2008], referred to in [Krishnan et al., 2016], a multivariate Gaussian is considered as a Markov Random Field.

In particular, consider the Markov Random field

$$p(x) = \frac{1}{Z}\left(\prod_{i=1}^{n} \phi(x_i) \prod_{i,j} \psi(x_i, x_j)\right) \tag{B.3.1}$$

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

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

as this is just the exponential family form of a Gaussian (e.g., see [Englehardt, 2013]), 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 (B.3.1) 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 [Krishnan et al., 2016], 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 [Krishnan et al., 2016]. 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 [Bickson, 2008]

C The Inverse Wishart Distribution

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

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

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

Remark C.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) $\mathbf{\Psi}$.

△

Remark C.0.1 also provides inuition on the expected value. For a visualization of how samples are affected by the parameters, see [Hughes, 2012].

Remark C.0.2. (*Peter Hoff's notation for the Inverse Wishart: A warning*) Note that some authors (e.g. [Hoff, 2009], pp.257) use the notation $\mathcal{W}^{-1}(\nu, \boldsymbol{M})$ to refer to the density under reparametrization

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

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

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

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

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

△

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

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

△

C.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 [Wolpert, 2011].[27,28]

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

The inverse Wishart can be seen as a generalization of the inverse gamma distribution to multiple dimensions.[29]

C.2 Entropy and relative entropy

Let $\mathbf{\Sigma}$ have an Inverse Wishart distribution (parametrized to have density (C.0.1)). Then its entropy is given by [Gupta and Srivastava, 2010]:

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

where ψ denotes the digamma function, $\psi(x) = \frac{d}{dx}\Gamma(x)$.[30]

The relative entropy between two Inverse Wishart distributions p_1, p_2 with parameters $\nu_1, \mathbf{\Psi}_1$ and $\nu_2, \mathbf{\Psi}_2$ is given by [Gupta and Srivastava, 2010]:

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

C.3 Sampling

A sample $\mathbf{\Sigma}$ from the $\mathcal{W}^{-1}(\nu, \mathbf{\Psi})$ distribution (using the natural parametrization of (C.0.1)) can be obtained by the following scheme[31] [Hoff, 2009]:

1. Sample $\boldsymbol{z}_1, ..., \boldsymbol{z}_\nu \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Psi}^{-1})$

[27] It is claimed in Wikipedia that if $\boldsymbol{X} \sim \mathcal{W}(\nu, \mathbf{\Psi})$ then $\boldsymbol{X}^{-1} \sim \mathcal{W}^{-1}(\nu, \mathbf{\Psi}^{-1})$. I almost was able to show this using the multivariate change of variables [Wolpert, 2011] along with (15.15) of [Dwyer, 1967], but I was off by a negative when attempting to combine the two terms with $|\boldsymbol{X}^{-1}|$ raised to an exponent.

[28] TO DO: Provide derivation.

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

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

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

2. Calculate $\boldsymbol{Z}^\top \boldsymbol{Z} = \sum_{i=1}^{\nu} \boldsymbol{z}_i \boldsymbol{z}_i^\top$.
3. Set $\boldsymbol{\Sigma} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}$.

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

C.4 Evaluation as a model for covariance matrices

The Inverse Wishart is a popular choice for modeling covariance matrices (e.g. see [Hoff, 2009]), 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 [Alvarez et al., 2014]. 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 [Wojnowicz, 2022].

D General Conjugacy Formalism: Alternate Approaches

D.1 General Conjugacy Formalism: Alternate Approach 1[32]

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

We can write the prior as

$$p(\boldsymbol{\theta}) = \exp\left\{ \left\langle \boldsymbol{\eta}^o_{\boldsymbol{\theta}},\ \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \right\rangle - \log Z_{\boldsymbol{\theta}}(\boldsymbol{\eta}^o_{\boldsymbol{\theta}}) \right\}$$

And the likelihood for a single observation y_i as

$$\begin{aligned} p(y_i \mid \boldsymbol{\theta}) &\overset{1}{=} \exp\left\{ \left\langle \boldsymbol{\eta}_y(\boldsymbol{\theta}),\ \boldsymbol{s}_y(y_i) \right\rangle - \log Z_y(\boldsymbol{\eta}_y(\boldsymbol{\theta})) \right\} \\ &\overset{2}{=} \exp\left\{ \left\langle \big(\boldsymbol{\eta}_y(\boldsymbol{\theta}),\ -\log Z_y(\boldsymbol{\eta}_y(\boldsymbol{\theta}))\big),\ \big(\boldsymbol{s}_y(y_i),\ 1\big) \right\rangle \right\} \\ &\overset{3}{=} \exp\left\{ \left\langle \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{\theta}),\ \big(\boldsymbol{s}_y(y_i), 1\big) \right\rangle \right\} \end{aligned}$$

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

[32] This argument follows the argument (and notation) of [Johnson et al., 2016], Appendix B. As of now, I find it more intuitive then the argument given in the main body.

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

$$\begin{aligned} p(\boldsymbol{\theta} \mid y_i) &\propto p(\boldsymbol{\theta}, y_i) \\ &= \exp\left\{ \left\langle \boldsymbol{\eta}_{\boldsymbol{\theta}}(y_i),\ \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \right\rangle - \log Z_{\boldsymbol{\theta}}(\boldsymbol{\eta}^o_{\boldsymbol{\theta}}) \right\} \end{aligned}$$

where $\boldsymbol{\eta}_{\boldsymbol{\theta}}(y_i) = \boldsymbol{\eta}^o_{\boldsymbol{\theta}} + \big(\boldsymbol{s}_y(y_i), 1\big)$, 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

$$p(\boldsymbol{\theta} \mid y_i) = \exp\left\{ \left\langle \boldsymbol{\eta}_{\boldsymbol{\theta}}(y_i),\ \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \right\rangle - \log Z_{\boldsymbol{\theta}}\big(\boldsymbol{\eta}_{\boldsymbol{\theta}}(y_i)\big) \right\} \tag{D.1.1}$$

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

$$p(\boldsymbol{\theta} \mid y) = \exp\left\{ \left\langle \boldsymbol{\eta}_{\boldsymbol{\theta}}(y),\ \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \right\rangle - \log Z_{\boldsymbol{\theta}}\big(\boldsymbol{\eta}_{\boldsymbol{\theta}}(y)\big) \right\} \tag{D.1.2}$$

where $\boldsymbol{\eta}_{\boldsymbol{\theta}}(y) = \boldsymbol{\eta}^o_{\boldsymbol{\theta}} + \big(\sum_{i=1}^n \boldsymbol{s}_y(y_i), n\big)$.

This motivates interpreting the prior parameter as $\boldsymbol{\eta}_{\boldsymbol{\theta}} = (\tau_0, n_0)$, where $\tau_0 \in \mathbb{R}^{\dim(\boldsymbol{\eta}_{\boldsymbol{\theta}})-1}$ 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 (D.2.2).

D.2 Alternate approach 2

Following [Jordan, 2010b], here we provide a general formalism for conjugate priors for exponential family data models.

Writing the exponential family density in canonical form, we have

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

where $\boldsymbol{\eta}$ is the canonical parameter, $T(x)$ are the sufficient statistics, $h(x)$ is the carrier density, and $A(\boldsymbol{\eta})$ is the log normalizer (and so is *not* a degree of freedom).

The natural parameter space is

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

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

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

as the likelihood function.

A conjugate prior can be obtained by mimicking the likelihood

$$p(\boldsymbol{\eta} \mid \tau, n_0) = H(\tau, n_0) \exp\{\tau^\top \boldsymbol{\eta} - n_0 A(\boldsymbol{\eta})\} \tag{D.2.1}$$

where now $H(\tau, n_0)$ is the normalizing factor. (For conditions on normalizability, see [Jordan, 2010b]). Note that τ has the dimensionality of the canonical parameter $\boldsymbol{\eta}$ and n_0 is a scalar.

To verify conjugacy, we compute the posterior density

$$p(\boldsymbol{\eta} \mid \boldsymbol{x}, \tau, \boldsymbol{\eta}_0) \propto \exp\left\{ \left(\tau + \sum_{n=1}^{N} T(x_n) \right)^\top \boldsymbol{\eta} - (n_0 + N) A(\boldsymbol{\eta}) \right\}$$

which retains the form of (D.2.1). For an argument that is perhaps more intuitive, see Section D.1

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

$$\begin{aligned} \tau &\to \tau + \sum_{n=1}^{N} T(x_n) \\ n_0 &\to n_0 + N \end{aligned} \tag{D.2.2}$$

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

$$p(x_{\text{new}} \mid x) = \frac{H(\tau_{\text{post}}, n_0 + N)}{H(\tau_{\text{post}} + T(x_{\text{new}}), n_0 + N + 1)} \tag{D.2.3}$$

For a proof of this, see Sec. E.

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.

E Posterior Predictives

Definition E.0.1. A **Bayesian i.i.d. model** with likelihood $F_{\boldsymbol{\theta}}$ (with parameter $\boldsymbol{\theta}$ and density $f_{\boldsymbol{\theta}}$) and prior π is given by

$$\begin{aligned} \boldsymbol{\theta} &\sim \pi \\ x_i &\overset{\text{i.i.d.}}{\sim} F_{\boldsymbol{\theta}} \qquad\qquad \forall i = 1, \ldots, n \end{aligned}$$

△

TODO: Give PGM

The proposition below shows that the posterior predictive density for a Bayesian i.i.d. model is the ratio of successive marginal densities (the "new" marginal to the "old" marginal).

Proposition E.0.1. [The posterior predictive density for a Bayesian i.i.d. model is the ratio of the complete marginal density to the marginal density of the conditioning set.]

Given a Bayesian i.i.d. model (Def. E.0.1), define the posterior predictive density by

$$PP(x_n; \boldsymbol{x}_{1:n-1}) \triangleq \int f(x_n \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x}_{1:n-1})\, d\boldsymbol{\theta}$$

Then

$$PP(x_n; x_{1:n-1}) = \frac{m(\boldsymbol{x}_{1:n})}{m(\boldsymbol{x}_{1:n-1})}$$

where m is the marginal density of the observations (marginalizing over the parameters $\boldsymbol{\theta}$).

Proof.

$$
\begin{aligned}
\texttt{PP}(x_n; x_{1:n-1}) &\triangleq \int f(x_n \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x}_{1:n-1})\, d\boldsymbol{\theta} && \text{def} \\
&= \int f(x_n \mid \boldsymbol{\theta}) \left[\frac{\prod_{i=1}^{n-1} f(x_i \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\int \prod_{i=1}^{n-1} f(x_i \mid \boldsymbol{\theta}') \pi(\boldsymbol{\theta}')\, d\boldsymbol{\theta}'} \right] d\boldsymbol{\theta} && \text{Bayes law} \\
&= \frac{\int \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int \prod_{i=1}^{n-1} f(x_i \mid \boldsymbol{\theta}') \pi(\boldsymbol{\theta}')\, d\boldsymbol{\theta}'} && \text{pull out constant multiple} \\
&= \frac{m(\boldsymbol{x}_{1:n})}{m(\boldsymbol{x}_{1:n-1})} && \text{notation}
\end{aligned}
$$

□

Remark E.0.1. By Prop E.0.1, we can notate the posterior predictive density as a conditional marginal density, since

$$
\texttt{PP}(x_n; x_{1:n-1}) \overset{\text{Prop E.0.1}}{=} \frac{m(\boldsymbol{x}_{1:n})}{m(\boldsymbol{x}_{1:n-1})} \overset{\text{def. cond. prob.}}{=} m(x_n \mid \boldsymbol{x}_{1:n-1})
$$

△

F Bayesian networks

F.1 Overview

Definition F.1.1. A *Bayesian network* (also known as a belief network) is a representation of a joint probability distribution that specifies the conditional dependencies among random variables via a directed acyclic graph (DAG). △

In a Bayesian network, a joint density is represented as follows:

$$
p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i \mid \mathbf{pa}_i) \tag{F.1.1}
$$

where $\mathbf{pa}_i$ refers to random variable i's conditioning set, which is called the *parents* of node i. The reason for this terminology is as follows: once we have specified our desired factorization via (F.1.1), we can identify it with a directed acyclic graph $\mathcal{G} = (E, V)$ by identifying each random variable with a node, and drawing a directed arc from A to B if A is a parent of B [**?**].

Equation (F.1.1) of course simplifies the factorization which is *always* true, by the chain rule of probability:

$$
p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i \mid x_1, .., x_{i-1})
$$

In other words, (F.1.1) restricts our consideration to a certain subset of joint probability distributions.

The problem is that it is not clear in advance that the arbitrary collection of conditionals given in (F.1.1) is internally consistent, and are indeed the conditionals for the provided joint. Luckily, there is a theorem (not proven here) which guarantees this to be true.[33]. As a result, *valid joint distributions can be constructed via specifying arbitrary collections of conditional distributions*.

[33] This may be demonstrated in Chapter 3 of [**?**]; I'm not sure.

Example F.1.1. An example of a Bayesian network is given in Figure 2, corresponding to the factorization

$$p(x) = p(x_1)\,p(x_2 \mid x_1)\,p(x_3 \mid x_1)\,p(x_4 \mid x_2)\,p(x_5 \mid x_3)\,p(x_6 \mid x_5, x_2)$$

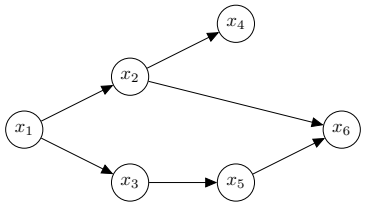

Figure 2: An example Bayesian network

△

Definition F.1.2. Random variable Y is called a *child* of random variable X if X is a parent of Y. The set of children of random variable X is denoted $\mathbf{ch}_X$. △

For example, in Example F.1.1, x_4 is a child of x_2, and x_2 is a child of x_1.

Definition F.1.3. Random variables X and Z are called *coparents* if there is some random variable Y that is a child of each. △

For example, in Example F.1.1, x_2 and x_5 are coparents of x_6.

F.2 Markov blankets

Definition F.2.1. The Markov blanket of a random variable is its children, parents, co-parents, △

Proposition F.2.1. *The complete conditional of a random variable in a Bayesian network depends only on its Markov blanket.*

Proof. Let $p(x_1, ..., x_n)$ represent the joint density over all random variables in the network. Then

$$\begin{aligned}
p(x_j \mid x_{-j}) &\propto p(x_1, ..., x_n) && \text{def. conditional} \\
&\overset{\text{(F.1.1)}}{=} \prod_{v \in V} p(x_v \mid \mathbf{pa}_v) && \text{Bayesian network} \\
&\overset{1}{\propto} p(x_j \mid \mathbf{pa}_j) \prod_{i \in \mathbf{ch}_j} p(x_i \mid \mathbf{pa}_i) && \text{explained below} \\
&= p(x_j \mid \mathbf{pa}_j) \prod_{i \in \mathbf{ch}_j} p(x_i \mid x_j, \mathbf{cp}_j) && \text{A child of } x_j \text{ has parents: } x_j, \mathbf{cp}_j.
\end{aligned}$$

In Equation 1, all terms are absorbed into the constant of proportionality except for terms of the form

1. $p(x_j \mid \mathbf{pa}_j)$, i.e. terms where x_j is the child (on the left of the conditional)
2. $\{p(x_i \mid \mathbf{pa}_i) : i \in \mathbf{ch}_j\}$, i.e. terms where x_j is one of the parents (on the right of the conditional bar).

□

G Bayesian multivariate linear regression: Alternate derivations

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

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

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

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

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

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

Proof. By Bayes rule,

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

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

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

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 $\boldsymbol{X}$. [34]

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 $(\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X})^{-1}$. 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. [35] Using this, we obtain

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

$$\begin{aligned} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) &= (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^\top \boldsymbol{\Omega} (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}) \\ &= \Big(\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{X}^+ \boldsymbol{y}) \Big)^\top \boldsymbol{\Omega} \Big(\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{X}^+ \boldsymbol{y}) \Big) \\ &= (\boldsymbol{\beta} - \boldsymbol{X}^+ \boldsymbol{y})^\top \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X} (\boldsymbol{\beta} - \boldsymbol{X}^+ \boldsymbol{y}) \end{aligned}$$

.

[35] See, for reference, Section B.3.

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

where

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

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 G.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 G.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 G.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.[36].

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

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

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

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

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

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

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

where

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

Equality (1) is obtained by noting $\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, 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 $\boldsymbol{M} = \frac{1}{\sigma^2} \boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{V}^{-1}$ and $\boldsymbol{b}^\top = \frac{1}{\sigma^2} \boldsymbol{y}^\top \boldsymbol{X}^\top$. □