Chapter 8

The exponential family: Basics

In this chapter we extend the scope of our modeling toolbox to accommodate a variety of additional data types, including counts, time intervals and rates. We introduce the exponential family of distributions, a family that includes the Gaussian, binomial, multinomial, Poisson, gamma, von Mises and beta distributions, as well as many others. In this chapter we focus on unconditional models and in the following chapter we show how these ideas can be carried over to the setting of conditional models.

At first blush this chapter may appear to involve a large dose of mathematical detail, but appearances shouldn't deceive—most of the detail involves working out examples that show how the exponential family formalism relates to more familiar material. The real message of this chapter is the simplicity and elegance of exponential family. Once the new ideas are mastered, it is often easier to work within the general exponential family framework than with specific instances.

8.1 The exponential family

Given a measure η, we define an *exponential family* of probability distributions as those distributions whose density (relative to η) have the following general form:

$$p(x \mid \eta) = h(x) \exp\{\eta^T T(x) - A(\eta)\} \tag{8.1}$$

for a parameter vector η, often referred to as the *canonical parameter*, and for given functions T and h. The statistic $T(X)$ is referred to as a *sufficient statistic*; the reasons for this nomenclature are discussed below. The function $A(\eta)$ is known as the *cumulant function*. Integrating Eq. (**??**) with respect to the measure ν, we have:

$$A(\eta) = \log \int h(x) \exp\{\eta^T T(x)\} \nu(dx) \tag{8.2}$$

where we see that the cumulant function can be viewed as the logarithm of a normalization factor.[1] This shows that $A(\eta)$ is not a degree of freedom in the specification of an exponential family density; it is determined once ν, $T(x)$ and $h(x)$ are determined.[2]

The set of parameters η for which the integral in Eq. (**??**) is finite is referred to as the *natural parameter space*:

$$\mathcal{N} = \{\eta : \int h(x) \exp\{\eta^T T(x)\} \nu(dx) < \infty\}. \tag{8.3}$$

We will restrict ourselves to exponential families for which the natural parameter space is a nonempty open set. Such families are referred to as *regular*.

In many cases we are interested in representations that are in a certain sense non-redundant. In particular, an exponential family is referred to as *minimal* if there are no linear constraints among the components of the parameter vector nor are there linear constraints among the components of the sufficient statistic (in the latter case, with probability one under the measure ν). Non-minimal families can always be reduced to minimal families via a suitable transformation and reparameterization.[3]

Even if we restrict ourselves to minimal representations, however, the same probability distribution can be represented using many different parameterizations, and indeed much of the power of the exponential family formalism derives from the insights that are obtained from considering different parameterizations for a given family. In general, given a set Θ and a mapping $\Phi : \Theta \to \mathcal{N}$, we consider densities obtained from Eq. (**??**) by replacing η with $\Phi(\theta)$:

$$p(x \,|\, \theta) = h(x) \exp\{\Phi(\theta)^T T(x) - A(\Phi(\theta))\}. \tag{8.4}$$

where Φ is a one-to-one mapping whose image is all of $\mathcal{N}$.

We are also interested in cases in which the image of Φ is a strict subset of $\mathcal{N}$. If this subset is a linear subset, then it is possible to transform the representation into an exponential family on that subset. When the representation is not reducible in this way, we refer to the exponential family as a *curved exponential family*.

---

[1] The integral in this equation is a Lebesgue integral, reflecting the fact that in general we wish to deal with arbitrary η. Actually, let us take the opportunity to be more precise and note that η is required to be a σ-finite measure. But let us also reassure those readers without a background in measure theory and Lebesgue integration that standard calculus will suffice for an understanding of this chapter. In particular, in all of the examples that we will treat, ν will either be Lebesgue measure, in which case "$\nu(dx)$" reduces to "dx" and the integral in Eq. (**??**) can be handled using standard multivariable calculus, or counting measure, in which case the integral reduces to a summation.

[2] It is also worth noting that ν and $h(x)$ are not really independent degrees of freedom. We are always free to absorb $h(x)$ in the measure ν. Doing so yields measures that are variations on Lebesgue measure and counting measure, and thus begins to indicate the elegance of the formulation in terms of general measures.

[3] For a formal proof of this fact, see Chapter 1 of **?**.

8.1.1 Examples

The Bernoulli distribution

A Bernoulli random variable X assigns probability measure π to the point $x = 1$ and probability measure $1 - \pi$ to $x = 0$. More formally, define ν to be counting measure on $\{0, 1\}$, and define the following density function with respect to ν:

$$\begin{aligned} p(x \,|\, \pi) &= \pi^x (1-\pi)^{1-x} && (8.5) \\ &= \exp\left\{ \log\left(\frac{\pi}{1-\pi}\right) x + \log(1-\pi) \right\}. && (8.6) \end{aligned}$$

Our trick for revealing the canonical exponential family form, here and throughout the chapter, is to take the exponential of the logarithm of the "usual" form of the density. Thus we see that the Bernoulli distribution is an exponential family distribution with:

$$\begin{aligned} \eta &= \frac{\pi}{1-\pi} && (8.7) \\ T(x) &= x && (8.8) \\ A(\eta) &= -\log(1-\pi) = \log(1+e^{\eta}) && (8.9) \\ h(x) &= 1. && (8.10) \end{aligned}$$

Note moreover that the relationship between η and π is invertible. Solving Eq. (**??**) for π, we have:

$$\pi = \frac{1}{1+e^{-\eta}}, \tag{8.11}$$

which is the logistic function.

The reader can verify that the natural parameter space is the real line in this case.

The Poisson distribution

The probability mass function (i.e., the density respect to counting measure) of a Poisson random variable is given as follows:

$$p(x \,|\, \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}. \tag{8.12}$$

Rewriting this expression we obtain:

$$p(x \,|\, \lambda) = \frac{1}{x!} \exp\{x \log \lambda - \lambda\}. \tag{8.13}$$

Thus the Poisson distribution is an exponential family distribution, with:

$$\begin{aligned} \eta &= \log \lambda && (8.14)\\ T(x) &= x && (8.15)\\ A(\eta) &= \lambda = e^{\eta} && (8.16)\\ h(x) &= \frac{1}{x!}. && (8.17) \end{aligned}$$

Moreover, we can obviously invert the relationship between η and λ:

$$\lambda = e^{\eta}. \tag{8.18}$$

The Gaussian distribution

The (univariate) Gaussian density can be written as follows (where the underlying measure is Lebesgue measure):

$$\begin{aligned} p(x \,|\, \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} && (8.19)\\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{ \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}\mu^2 - \log\sigma \right\}. && (8.20) \end{aligned}$$

This is in the exponential family form, with:

$$\begin{aligned} \eta &= \begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix} && (8.21)\\ T(x) &= \begin{bmatrix} x \\ x^2 \end{bmatrix} && (8.22)\\ A(\eta) &= \frac{\mu^2}{2\sigma^2} + \log\sigma = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2) && (8.23)\\ h(x) &= \frac{1}{\sqrt{2\pi}}. && (8.24) \end{aligned}$$

Note in particular that the univariate Gaussian distribution is a two-parameter distribution and that its sufficient statistic is a vector.

The multivariate Gaussian distribution can also be written in the exponential family form; we leave the details to Exercise **??** and Chapter 13.

The von Mises distribution

Suppose that we wish to place a distribution on an angle x, where $x \in (0, 2\pi)$. This is readily accomplished within the exponential family framework:

$$p(x \,|\, \kappa, \mu) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \cos(x-\mu)\} \tag{8.25}$$

where μ is a location parameter, κ is a scale parameter and $I_0(\kappa)$ is the modified Bessel function of order 0. This is the *von Mises distribution*.

The von Mises distribution can be viewed as an analog of the Gaussian distribution on a circle. Expand the cosine function in a Taylor series: $\cos(z) \approx 1 - 1/2z^2$. Plugging this into Eq. (**??**), we obtain a Gaussian distribution. Thus, locally around μ, the von Mises distribution acts like a Gaussian distribution as a function of the angular variable x, with mean μ and inverse variance κ.

This example can be generalized to higher dimensions, where the sufficient statistics are cosines of general spherical coordinates. The resulting exponential family distribution is known as the *Fisher-von Mises distribution*.

The multinomial distribution

As a final example, let us consider the multinomial distribution. Let $X = (X_1, X_2, \ldots, X_K)$ be a collection of integer-valued random variables representing event counts, where X_k represents the count of the number of times the kth event occurs in a set of M independent trials. Let π_k represent the probability of the ith event occuring in any given trial. We have:

$$p(x \,|\, \pi) = \frac{M!}{x_1! x_2! \cdots x_K!} \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_K^{x_K}, \tag{8.26}$$

as the probability mass function for such a collection, where the underlying measure is counting measure on the set of K-tuples of nonnegative integers for which $\sum_{k=1}^K x_k = M$.

Following the strategy of our previous examples, we rewrite the multinomial distribution as follows:

$$p(x \,|\, \pi) = \frac{M!}{x_1! x_2! \cdots x_m!} \exp\left\{ \sum_{k=1}^K x_k \log \pi_k \right\}. \tag{8.27}$$

While this suggests that the multinomial distribution is in the exponential family, there are some troubling aspects to this expression. In particular it appears that the cumulant function is equal to zero. As we will be seeing (in Section **??**), one of the principal virtues of the exponential family form is that means and variances can be calculated by taking derivatives of the cumulant function; thus, the seeming disappearance of this term is unsettling.

In fact the cumulant function is not equal to zero. We must recall that the cumulant function is defined on the natural parameter space, and the natural parameter space in this case is all of $\mathbb{R}^K$; it is not restricted to parameters η such that $\sum_{k=1}^K e^{\eta_k} = 1$. The cumulant function is not equal to zero on the larger space. However, it is inconvenient to work in the larger space, because it is a redundant representation—it yields no additional probability distributions beyond those obtained from the constrained parameter space.

Another way to state the issue is to note that the representation in Eq. (**??**) is not minimal. In particular, the constraint $\sum_{k=1}^K X_k = 1$ is a linear constraint on the components

of theh sufficient statistic. To achieve a minimal representation for the multinomial, we parameterize the distribution using the first $K-1$ components of π:

$$\begin{aligned} p(x \mid \pi) &= \exp\left\{\sum_{k=1}^{K} x_k \log \pi_k\right\} && (8.28)\\ &= \exp\left\{\sum_{k=1}^{K-1} x_k \log \pi_k + \left(1 - \sum_{k=1}^{K-1} x_k\right) \log\left(1 - \sum_{k=1}^{K-1} \pi_k\right)\right\} && (8.29)\\ &= \exp\left\{\sum_{k=1}^{K-1} \log\left(\frac{\pi_k}{1-\sum_{k=1}^{K-1}\pi_k}\right) x_k + \log\left(1 - \sum_{k=1}^{K-1} \pi_k\right)\right\}. && (8.30) \end{aligned}$$

where we have used the fact that $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$.

From this representation we obtain:

$$\eta_k = \log\left(\frac{\pi_k}{1-\sum_{k=1}^{K-1}\pi_k}\right) = \log\left(\frac{\pi_k}{\pi_K}\right) \tag{8.31}$$

for $i = 1, \ldots, K-1$. For convenience we also can define η_K; Eq. (**??**) implies that if we do so we must take $\eta_K = 0$.

As in the other examples of exponential family distributions, we can invert Eq. (**??**) to obtain a mapping that expresses π_k in terms of η_k. Taking the exponential of Eq. (**??**) and summing we obtain:

$$\pi_k = \frac{e^{\eta_k}}{\sum_{j=1}^{K} e^{\eta_j}}, \tag{8.32}$$

which is known as the *multinomial logit* or *softmax* function.

Finally, from Eq. (**??**) we obtain:

$$A(\eta) = -\log\left(1 - \sum_{k=1}^{K-1} \pi_k\right) = \log\left(\sum_{k=1}^{K} e^{\eta_k}\right) \tag{8.33}$$

as the cumulant function for the multinomial.

8.2 Convexity

We now turn to a more general treatment of the exponential family. As we will see, the exponential family has several appealing statistical and computational properties. Many of these properties derive from the two fundamental results regarding convexity that are summarized in the following theorem.

Theorem 1. *The natural parameter space $\mathcal{N}$ is convex (as a set) and the cumulant function $A(\eta)$ is convex (as a function). If the family is minimal then $A(\eta)$ is strictly convex.*

Proof. The proofs of both convexity results follow from an application of Hölder's inequality. Consider distinct parameters $\eta_1 \in \mathcal{N}$ and $\eta_2 \in \mathcal{N}$ and let $\eta = \lambda\eta_1 + (1-\lambda)\eta_2$, for $0 < \lambda < 1$. We have:

$$\begin{aligned}
\exp(A(\eta)) &= \int e^{(\lambda\eta_1+(1-\lambda)\eta_2)^T T(x)} h(x)\nu(dx) && (8.34)\\
&\leq \left(\int \left(e^{\lambda\eta_1^T T(x)}\right) \frac{1}{\lambda} h(x)\nu(dx)\right)^{\lambda} \left(\int \left(e^{((1-\lambda)\eta_2^T T(x)}\right) \frac{1}{(1-\lambda)} h(x)\nu(dx)\right)^{1-\lambda} && (8.35)\\
&= \left(\int \left(e^{\eta_1^T T(x)}\right) \frac{1}{\lambda} h(x)\nu(dx)\right)^{\lambda} \left(\int \left(e^{\eta_2^T T(x)}\right) \frac{1}{(1-\lambda)} h(x)\nu(dx)\right)^{1-\lambda}. && (8.36)
\end{aligned}$$

This establishes that $\mathcal{N}$ is convex, because it shows that the integral on the left is finite if both integrals on the right are finite. Also, taking logarithms yields:

$$A(\lambda\eta_1 + (1-\lambda)\eta_2) \leq \lambda A(\eta_1) + (1-\lambda)A(\eta_2), \tag{8.37}$$

which establishes the convexity of $A(\eta)$.

Hölder's inequality is strict unless $e^{\eta_1^T T(X)}$ is proportional to $e^{\eta_2^T T(X)}$ (with probability one). But this would imply that $(\eta_1 - \eta_2)^T T(X)$ is equal to a constant with probability one, which is not possible in a minimal family.

8.3 Means, variances and other cumulants

The mean and variance of probability distributions are defined as integrals with respect to the distribution. Thus it may come as a surprise that the mean and variance of distributions in the exponential family can be obtained by calculating derivatives. Moreover, this should be a pleasant surprise because derivatives are generally easier to compute than integrals.

In this section we show that the mean can be obtained by computing a first derivative of the cumulant function $A(\eta)$ and the variance can be obtained by computing second derivatives of $A(\eta)$.

The mean and variance are the first and second *cumulants* of a distribution, respectively, and the results of this section explain why $A(\eta)$ is referred to as a cumulant function. We will define cumulants below in general, but for now we will be interested only in the mean and variance.

Let us begin with an example. Recall that in the case of the Bernoulli distribution we

have $A(\eta) = \log(1 + e^\eta)$. Taking a first derivative yields:

$$\begin{aligned} \frac{dA}{d\eta} &= \frac{e^\eta}{1 + e^\eta} && (8.38) \\ &= \frac{1}{1 + e^{-\eta}} && (8.39) \\ &= \mu, && (8.40) \end{aligned}$$

which is the mean of a Bernoulli variable.

Taking a second derivative yields:

$$\begin{aligned} \frac{d^2a}{d\eta^2} &= \frac{d\mu}{d\eta} && (8.41) \\ &= \mu(1 - \mu), && (8.42) \end{aligned}$$

which is the variance of a Bernoulli variable.

Let us now consider computing the first derivative of $A(\eta)$ for a general exponential family distribution. The computation begins as follows:

$$\frac{\partial A}{\partial \eta^T} = \frac{\partial}{\partial \eta^T} \left\{ \log \int \exp\{\eta^T T(x)\} h(x) \nu(dx) \right\}. \quad (8.43)$$

To proceed we need to move the gradient past the integral sign. In general derivatives cannot be moved past integral signs (both are certain kinds of limits, and sequences of limits can differ depending on the order in which the limits are taken). However, it turns out that the move is justified in this case. The justification is obtained by an appeal to the dominated convergence theorem; see, e.g., **?** for details. Thus we continue our computation:

$$\begin{aligned} \frac{\partial A}{\partial \eta^T} &= \frac{\int T(x) \exp\{\eta^T T(x)\} h(x) \nu(dx)}{\int \exp\{\eta T(x)\} h(x) \nu(dx)} && (8.44) \\ &= \int T(x) \exp\{\eta^T T(x) - A(\eta)\} h(x) \nu(dx) && (8.45) \\ &= \mathbb{E}[T(X)]. && (8.46) \end{aligned}$$

We see that the first derivative of $A(\eta)$ is equal to the mean of the sufficient statistic.

Let us now take a second derivative:

$$\begin{aligned} \frac{\partial^2 A}{\partial \eta \partial \eta^T} &= \int T(x)(T(x) - \frac{\partial}{\partial \eta^T} A(\eta))^T \exp\{\eta^T T(x) - A(\eta)\} h(x) \nu(dx) && (8.47) \\ &= \int T(x)(T(x) - \mathbb{E}[T(X)])^T \exp\{\eta^T T(x) - A(\eta)\} h(x) \nu(dx) && (8.48) \\ &= \mathbb{E}[T(X)T(X)^T] - \mathbb{E}[T(X)]\mathbb{E}[T(X)])^T && (8.49) \\ &= \mathrm{Var}[T(X)], && (8.50) \end{aligned}$$

and thus we see that the second derivative of $A(\eta)$ is equal to the variance (i.e., the covariance matrix) of the sufficient statistic.

In general cumulants are defined as coefficients in the Taylor series expansion of an object known as the *cumulant generating function*:

$$C(\mathbf{s}) = \log \mathbb{E}[e^{\mathbf{s}^T X}]. \tag{8.51}$$

In Exercise **??** we ask the reader to compute the cumulant generating function for exponential family distributions. The result turns out to be:

$$C(\mathbf{s}) = A(\mathbf{s} + \eta) - A(\eta). \tag{8.52}$$

This is *not* the same function as the cumulant function $A(\eta)$. But if we take derivatives of $C(\mathbf{s})$ with respect to $\mathbf{s}$ and then set $\mathbf{s}$ equal to zero (to obtain the coefficients of the Taylor series expansion; i.e., the cumulants), we get the same answer as is obtained by taking derivatives of $A(\eta)$ with respect to η.

This result shows that higher-order cumulants can be obtained by taking higher-order derivatives of A.[4]

Example

Let us calculate the mean and variance of the univariate Gaussian distribution. Recall the form taken by $A(\eta)$:

$$A(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2), \tag{8.53}$$

where $\eta_1 = \mu/\sigma^2$ and $\eta_2 = -1/2\sigma^2$.

Taking the derivative with respect to η_1 yields:

$$\begin{aligned} \frac{\partial A}{\partial \eta_1} &= \frac{\eta_1}{2\eta_2} && (8.54)\\ &= \frac{\mu/\sigma^2}{1/\sigma^2} && (8.55)\\ &= \mu, && (8.56) \end{aligned}$$

which is the mean of X, the first component of the sufficient statistic.

Taking a second derivative with respect to η_1 yields:

$$\begin{aligned} \frac{\partial^2 A}{\partial \eta_1^2} &= -\frac{1}{2\eta_2} && (8.57)\\ &= \sigma^2, && (8.58) \end{aligned}$$

[4]The mean and variance are also referred to as *central moments*. While cumulants and central moments are the same up to second order, third and higher-order central moments are not the same as cumulants. But in general central moments can be obtained from cumulants and vice versa.

which is the variance of X.

Given that X^2 is the second component of the sufficient statistic, we can also compute the variance by calculating the partial of A with respect to η_2. Moreover, we can calculate third cumulants by computing the mixed partial, and fourth cumulants by taking the second partial with respect to η_2 (see Exercise **??**).

8.4 The mean parameterization

In the previous section we showed that the $\mu := \mathbb{E}[T(X)]$, can be obtained as a function of the canonical parameter η:

$$\mu = \frac{\partial A}{\partial \eta^T}. \tag{8.59}$$

For minimal families it turns out this relationship is invertible. To see this, recall that we demonstrated in Theorem **??** that $A(\eta)$ is strictly convex in the case of minimal families (see also Exercise **??**). Strictly convex functions obey Fenchel's inequality (see Appendix XXX):

$$(\eta_1 - \eta_2)^T(\mu_1 - \mu_2) > 0, \tag{8.60}$$

where η_1 and η_2 are assumed distinct, and where μ_1 and μ_2 are the derivatives $\partial A/\partial \eta^T$ evaluated at η_1 and η_2, respectively. This implies that $\mu_1 \neq \mu_2$ whenever $\eta_1 \neq \eta_2$, and hence that the mapping from η to μ is invertible.

This argument implies that a distribution in the exponential family can be parameterized not only by η—the canonical parameterization—but also by μ—the *mean parameterization*. Many distributions are traditionally parameterized using the mean parameterization; indeed, in Section **??** our starting point was the mean parameterization for most of the examples. We subsequently reparameterized these distribution using the canonical parameterization. We also computed the mapping from η to μ in each case, recovering some familiar functions, including the logistic function. This topic will return in Chapter **??** when we discuss generalized linear models.

8.5 Sufficiency

In this section we discuss the important concept of *sufficiency*. Sufficiency characterizes what is essential in a data set, or, alternatively, what is inessential and can therefore be thrown away. While the notion of sufficiency is broader than the exponential family, the ties to the exponential family are close, and it is natural to introduce the concept here.

A *statistic* is any function on the sample space that is not a function of the parameter. In particular, let X be a random variable and let $T(X)$ be a statistic. Suppose that the distribution of X depends on a parameter θ. The intuitive notion of sufficiency is that $T(X)$

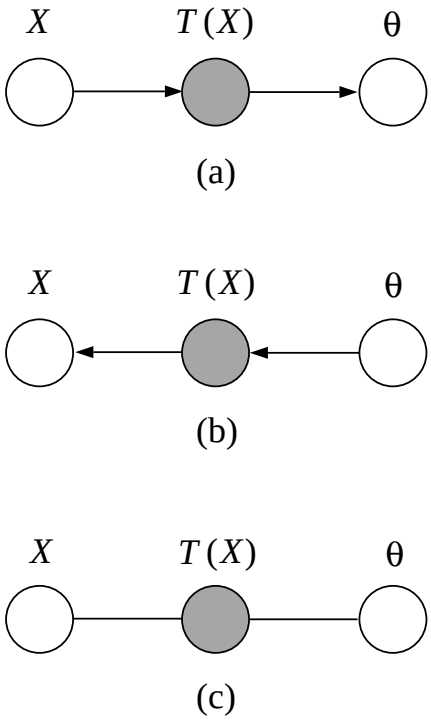

Figure 8.1: Graphical models whose conditional independence properties capture the notion of sufficiency in three equivalent ways.

is sufficient for θ if there is no information in X regarding θ beyond that in $T(X)$. That is, having observed $T(X)$, we can throw away X for the purposes of inference with respect to θ. Let us make this notion more precise.

Sufficiency is defined in somewhat different ways in the Bayesian and frequentist frameworks. Let us begin with the Bayesian approach (which is arguably more natural). In the Bayesian approach, we treat θ as a random variable, and are therefore licensed to consider conditional independence relationships involving θ. We say that $T(X)$ is sufficient for θ if the following conditional independence statement holds:

$$\theta \perp\!\!\!\perp X \,|\, T(X). \tag{8.61}$$

We can also write this in terms of probability distributions:

$$p(\theta \,|\, T(x), x) = p(\theta \,|\, T(x)). \tag{8.62}$$

Thus, as shown graphically in Figure **??**(a), sufficiency means that θ is independent of X, when we condition on $T(X)$. This captures the intuitive notion that $T(X)$ contains all of the essential information in X regarding θ.

To obtain a frequentist definition of sufficiency, let us consider the graphical model in Figure **??**(b). This model expresses the same conditional independence semantics as Figure **??**(a), asserting that θ is independent of X conditional on $T(X)$, but the model is parameterized in a different way. From the factorized form of the joint probability we obtain:

$$p(x \,|\, T(x), \theta) = p(x \,|\, T(x)). \tag{8.63}$$

This expression suggests a frequentist definition of sufficiency. In particular, treating θ as a label rather than a random variable, we define $T(X)$ to be sufficient for θ if the conditional distribution of X given $T(X)$ is not a function of θ.

Both the Bayesian and frequentist definitions of sufficiency imply a factorization of $p(x \,|\, \theta)$, and it is this factorization which is generally easiest to work with in practice. To obtain the factorization we use the undirected graphical model formalism. Note in particular that Figure **??**(c) expresses the same conditional independence semantics as Figure **??**(a) and Figure **??**(b). Moreover, from Figure **??**(c), we know that we can express the joint probability as a product of potential functions ψ_1 and ψ_2:

$$p(x, T(x), \theta) = \psi_1(T(x), \theta)\psi_2(x, T(x)), \tag{8.64}$$

where we have absorbed the constant of proportionality Z in one of the potential functions. Now $T(x)$ is a deterministic function of x, which implies that we can drop $T(x)$ on the left-hand side of the equation. Dividing by $p(\theta)$ we therefore obtain:

$$p(x \,|\, \theta) = g(T(x), \theta)h(x, T(x)), \tag{8.65}$$

for given functions g and h. Although we have motivated this result by using a Bayesian calculation, the result can be utilized within either the Bayesian or frequentist framework. Its equivalence to the frequentist definition of sufficiency is known as the Neyman factorization theorem.

8.5.1 Sufficiency and the exponential family

An important feature of the exponential family is that it one can obtain sufficient statistics by inspection, once the distribution is expressed in the standard form. Recall the definition:

$$p(x \,|\, \eta) = h(x) \exp\{\eta^T T(x) - A(\eta)\}. \tag{8.66}$$

From Eq. (**??**) we see immediately that $T(X)$ is a sufficient statistic for η.

8.5.2 Random samples

The reduction obtainable by using a sufficient statistic is particularly notable in the case of i.i.d. sampling. Suppose that we have a collection of N independent random variables,

$X = (X_1, X_2, \ldots, X_N)$, sampled from the same exponential family distribution. Taking the product, we obtain the joint density (with respect to product measure):

$$p(x \mid \eta) = \prod_{n=1}^{N} h(x_n) \exp\{\eta^T T(x_n) - A(\eta)\} = \left(\prod_{n=1}^{N} h(x_n)\right) \exp\left\{\eta^T \sum_{n=1}^{N} T(x_n) - NA(\eta)\right\}. \tag{8.67}$$

From this result we see that X is itself an exponential distribution, with sufficient statistic $\sum_{n=1}^{N} T(X_n)$.

For several of the examples we discussed earlier (in Section **??**), including the Bernouilli, the Poisson, and the multinomial distributions, the sufficient statistic $T(X)$ is equal to the random variable X. For a set of N i.i.d. observations from such distributions, the sufficient statistic is equal to $\sum_{n=1}^{N} x_n$. Thus in this case, it suffices to maintain a single scalar or vector, the sum of the observations. The individual data points can be thrown away.

For the univariate Gaussian the sufficient statistic is the pair $T(X) = (X, X^2)$, and thus for N i.i.d. Gaussians it suffices to maintain two numbers: the sum $\sum_{n=1}^{N} x_n$, and the sum of squares $\sum_{n=1}^{N} x_n^2$.

8.6 Maximum likelihood estimates

In this section we show how to obtain maximum likelihood estimates of the mean prameter in exponential family distributions. We obtain a generic formula that generalizes our earlier work on density estimation in Chapter **??**.

Consider an i.i.d. data set, $\mathcal{D} = (x_1, x_2, \ldots, x_N)$. From Eq. (**??**) we obtain the following log likelihood:

$$l(\eta \mid \mathcal{D}) = \log\left(\prod_{n=1}^{N} h(x_n)\right) + \eta^T \left(\sum_{n=1}^{N} T(x_n)\right) - NA(\eta). \tag{8.68}$$

Taking the gradient with respect to η yields:

$$\nabla_\eta l = \sum_{n=1}^{N} T(x_n) - N\nabla_\eta A(\eta), \tag{8.69}$$

and setting to zero gives:

$$\nabla_\eta A(\hat{\eta}) = \frac{1}{N} \sum_{n=1}^{N} T(x_n). \tag{8.70}$$

Finally, defining $\mu := E[T(X)]$, and recalling Eq. (**??**), we obtain:

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} T(x_n) \tag{8.71}$$

as the general formula for maximum likelihood estimation of the mean parameter in the exponential family.

It should not be surprising that our formula involves the data only via the sufficient statistic $\sum_{n=1}^{N} T(X_n)$. This gives operational meaning to sufficiency—for the purpose of estimating parameters we retain only the sufficient statistic.

For distributions in which $T(X) = X$, which include the the Bernoulli distribution, the Poisson distribution, and the the multinomial distribution, our result shows that the sample mean is the maximum likelihood estimate of the mean.

For the univariate Gaussian distribution, we see that the sample mean is the maximum likelihood estimate of the mean and the sample variance is the maximum likelihood estimate of the variance. For the multivariate Gaussian we obtain the same result, where by "variance" we mean the covariance matrix.

Let us also provide a frequentist evaluation of the quality of the maximum likelihood estimate $\hat{\mu}_{ML}$. To simplify the notation we consider univariate distributions.

Note first that $\hat{\mu}_{ML}$ is an unbiased estimator:

$$\begin{aligned} \mathbb{E}[\hat{\mu}_{ML}] &= \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[T(X_n)] && (8.72) \\ &= \frac{1}{N} N\mu && (8.73) \\ &= \mu. && (8.74) \end{aligned}$$

Recall that the variance of any estimator of μ is lower bounded by the inverse Fisher information (the Cramér-Rao lower bound). We first compute the Fisher information for the canonical parameter:

$$\begin{aligned} I(\eta) &= -\mathbb{E}\left[\frac{d^2 \log p(X \mid \eta)}{d\eta^2}\right] && (8.75) \\ &= -\mathbb{E}\left[\frac{d^2 A(\eta)}{d\eta^2}\right] && (8.76) \\ &= \mathrm{Var}[T(X)]. && (8.77) \end{aligned}$$

It is now an exercise with the chain rule to compute the Fisher information for the mean parameter (see Exercise **??**):

$$I(\eta) = \frac{1}{\mathrm{Var}[T(X)]}. \tag{8.78}$$

But $\mathrm{Var}[T(X)]$ is the variance of $\hat{\mu}_{ML}$. We see that $\hat{\mu}_{ML}$ attains the Cramér-Rao lower bound and in this sense it is an optimal estimator.

8.7 Kullback-Leibler divergence

The *Kullback-Leibler (KL) divergence* is a measure of "distance" between probability distributions, where "distance" is in quotes because the KL divergence does not possess all of the properties of a distance (i.e., a metric). We discuss the KL divergence from the point of view of information theory in Appendix **??** and we will see several statistical roles for the KL divergence throughout the book. In this section we simply show how to compute the KL divergence for exponential family distributions.

Recall the definition of the KL divergence:

$$D(p(x \mid \eta_1) \parallel p(x \mid \eta_2)) = \int p(x \mid \eta_1) \log \frac{p(x \mid \eta_1)}{p(x \mid \eta_2)} \nu(dx) \tag{8.79}$$

$$= \mathbb{E}_{\theta_1} \log \frac{p(X \mid \eta_1)}{p(X \mid \eta_2)} \tag{8.80}$$

We will also abuse notation and write $D(\eta_1 \parallel \eta_2)$. For exponential families we have:

$$D(\eta_1 \parallel \eta_2) = \mathbb{E}_{\eta_1} \log \frac{p(X \mid \eta_1)}{p(X \mid \eta_2)} \tag{8.81}$$

$$= \mathbb{E}_{\eta_1} (\eta_1 - \eta_2)^T T(X) - A(\eta_1) + A(\eta_2) \tag{8.82}$$

$$= (\eta_1 - \eta_2)^T \mu_1 - A(\eta_1) + A(\eta_2), \tag{8.83}$$

where $\mu_1 = \mathbb{E}_{\eta_1}[T(X)]$ is the mean parameter.

Recalling that the mean parameter is the gradient of the cumulant function, we obtain from Eq. (**??**) the geometrical interpretation of the KL divergence shown in Figure **??**.

8.8 Maximum likelihood and the KL divergence

In this section we point out a simple relationship between the maximum likelihood problem and the KL divergence. This relationship is general; it has nothing to do specifically with the exponential family. We discuss it in the current chapter, however, because we have a hidden agenda. Our agenda, to be gradually revealed in Chapters **??**, **??** and **??**, involves building a number of very interesting and important relationships between the exponential family and the KL divergence. By introducing a statistical interpretation of the KL divergence in the current chapter, we hope to hint at deeper connections to come.

To link the KL divergence and maximum likelihood, let us first define the *empirical distribution*, $\tilde{p}(x)$. This is a distribution which places a point mass at each data point x_n in our data set $\mathcal{D}$. We have:

$$\tilde{p}(x) := \frac{1}{N} \sum_{n=1}^{N} \delta(x, x_n), \tag{8.84}$$

where $\delta(x, x_n)$ is a Kronecker delta function in the continuous case. In the discrete case, $\delta(x, x_n)$ is simply a function that is equal to one if its arguments agree and equal to zero otherwise.

If we integrate (in the continuous case) or sum (in the discrete case) $\tilde{p}(x)$ against a function of x, we evaluate that function at each point x_n. In particular, the log likelihood can be written this way. In the discrete case we have:

$$\begin{aligned}
\sum_x \tilde{p}(x) \log p(x \mid \theta) &= \sum_x \frac{1}{N} \sum_{n=1}^{N} \delta(x, x_n) \log p(x \mid \theta) && (8.85)\\
&= \frac{1}{N} \sum_{n=1}^{N} \sum_x \delta(x, x_n) \log p(x \mid \theta) && (8.86)\\
&= \frac{1}{N} \sum_{n=1}^{N} \log p(x_n \mid \theta) && (8.87)\\
&= \frac{1}{N} l(\theta \mid \mathcal{D}). && (8.88)
\end{aligned}$$

Thus by computing a cross-entropy between the empirical distribution and the model, we obtain the log likelihood, scaled by the constant $1/N$. We obtain an identical result in the continuous case by integrating.

Let us now calculate the KL divergence between the empirical distribution and the model $p(x \mid \theta)$. We have:

$$\begin{aligned}
D(\tilde{p}(x) \parallel p(x \mid \theta)) &= \sum_x \tilde{p}(x) \log \frac{\tilde{p}(x)}{p(x \mid \theta)} && (8.89)\\
&= \sum_x \tilde{p}(x) \log \tilde{p}(x) - \sum_x \tilde{p}(x) \log p(x \mid \theta) && (8.90)\\
&= +\sum_x \tilde{p}(x) \log \tilde{p}(x) - \frac{1}{N} l(\theta \mid \mathcal{D}). && (8.91)
\end{aligned}$$

The first term, $\sum_x \tilde{p}(x) \log \tilde{p}(x)$, is independent of θ. Thus, the minimizing value of θ on the left-hand side is equal to the maximizing value of θ on the right-hand side.

In other words: *minimizing the KL divergence to the empirical distribution is equivalent to maximizing the likelihood.* This simple result will prove to be very useful in our later work.

Bibliography