Chapter 3

The Exponential Family

3.1 The exponential family of distributions

See also Section 5.2, Davison (2002).

It is possible to derive the properties (eg. mean, variance and maximum likelihood estimators - to be defined properly later on) for every distribution of interest. However, this can be cumbersome, the algebra can be tedious and we may not see the 'big picture'. Instead, we now consider an 'umbrella' family of distributions which include several well known distributions. We will derive a general expression for the mean and variance of such distributions (which will be useful when we consider Generalised Linear models later in this course), and use these results to show that the maximum likelihood estimator is a function of the sufficient statistic :- thus is the best unbiased estimator (under the assumption of completeness). In other words, we that for this family of distributions the maximum likelihood estimator (which we have encountered many times previously) is indeed the best parameter estimator (in terms of minimum variance).

Suppose that the distribution of the random variable X_t can be written in the form

$$f(y;\omega) = \exp\big(s(y)\eta(\omega) - b(\omega) + c(y)\big). \tag{3.1}$$

If the distribution of X_t (both the probability distribution function for discrete random variables and probability density function for continuous random variables) has the above representation, then X_t is said to belong to the exponential family of distributions. A large number of well known distribution functions belong to this family. Hence by understanding the properties the exponential family, we can draw conclusions on a large number of distribution functions.

Example 3.1.1 *(a) The exponential distribution $X \sim Exp(\lambda)$, hence the pdf is $f(y;\lambda) = \lambda\exp(-\lambda y)$, which can be written as*

$$\log f(y;\lambda) = (-y\lambda + \log\lambda).$$

Therefore $s(y) = -y$ *and* $\eta(\lambda) = \lambda$.

(b) *The binomial distribution* $P(X = y) = \binom{n}{y}\pi(1-\pi)^{n-y}$ *can be rewritten as*

$$\log P(y;\lambda) = y\log(\frac{\pi}{1-\pi}) + n\log(1-\pi) + \log\binom{n}{y}.$$

Therefore $s(y) = y$, $\eta(\pi) = \log(\frac{\pi}{1-\pi})$, $b(\pi) = n\log(1-\pi)^{-1}$ *and* $c(y) = \log\binom{n}{y}$.

It should be mentioned that it is straightforward to generalise the exponential family to the case that θ is a vector of dimension greater than one. Suppose that θ is a p-dimensional vector. The order p exponential family is defined as distributions which satisfy

$$f(y;\omega) = \exp(s(y)'\theta(\omega) - b(\omega) + c(y)),$$

where $s(y) = (s_1(y),\ldots,s_p(y))$ (with $\{s_i\}$ linearly independent) and $\theta(\omega) = (\theta_1(\omega),\ldots,\theta_p(\omega))$.

3.1.1 The natural exponential family

If we let $\theta = \eta(\omega)$ and η is an invertible function (hence there is a one-to-one correspondence between the space containing ω and the space containing θ), then we can rewrite (3.1) we

$$f(y;\theta) = \exp(s(y)\theta - \kappa(\theta) + c(y)),$$

where $\kappa(\theta) = b(\eta^{-1}(\theta))$. The *natural exponential family* is when $\underline{s(y) = y}$.

Now by transformation we give example of distributions which have natural form.

(i) The exponential distribution is already in natural exponential form.

(ii) For the binomial distribution we let $\theta = \log(\frac{\pi}{1-\pi})$, since $\log(\frac{\pi}{1-\pi})$ is invertible this gives the log distribution as

$$\log f(y;\theta) = \log f(y;\log\frac{\pi}{1-\pi}) = \big(y\theta - n\log\big(\frac{1}{1+\exp(\theta)}\big) + \log\binom{n}{y}\big).$$

Hence the parameter of interest, π, has been transformed, and often we fit a model (later in the course) to θ, and transform back to obtain an estimator of π.

Some properties of the natural exponental

Distributions which have a natural exponential have interesting properties which we now discuss.

Lemma 3.1.1 *Suppose that* X *is a random variable which has the natural exponential representation. Then the moment generating function of* X *is* $\mathbb{E}(\exp(Xt)) = \exp(\kappa(t+\theta) - \kappa(\theta))$. *Furthermore,* $\mathbb{E}(X) = \kappa'(\theta)$ *and* $var(X) = \kappa''(\theta)$.

PROOF. Let us suppose that t is sufficiently small such that $f(y;(\theta+t))$ is a distribution. The mfg is

$$\begin{aligned} M_X(t) = \mathbb{E}(\exp(tY)) &= \int \exp(ty)\exp(\theta y - \kappa(\theta) + c(y))dy \\ &= \exp(\kappa(\theta+t) - \kappa(\theta)) \int \exp((\theta+t)y - \kappa(\theta+t) + c(y))dy \\ &= \exp(\kappa(\theta+t) - \kappa(\theta)), \end{aligned}$$

since $\int \exp((\theta+t)y - \kappa(\theta+t) + c(y))dy = \int f(y;(\theta+t))dy = 1$. To obtain the moments we recall that $M_X'(0) = \mathbb{E}(X)$ and $\text{var}(X) = M_X''(0) - (M_X'(0))^2$. Therefore

$$\begin{aligned} M_X'(t) &= \kappa'(\theta+t)\exp(\kappa(\theta+t) - \kappa(\theta)) \\ M_X'(t) &= \left(\kappa''(\theta+t) + (\kappa'(\theta+t))^2\right)\exp(\kappa(\theta+t) - \kappa(\theta)). \end{aligned}$$

Hence $M_X'(0) = \kappa'(\theta)$ and $M_X''(0) = \kappa''(\theta) + \kappa'(\theta)^2$, which gives the result.

Remark 3.1.1 *The mean and variance of the natural exponential family make obtaining the mle estimators quite simple. We derive this later but we first observe that since $\mathbb{E}(X) = \kappa'(\theta)$, therefore the mean of X is a function of θ, hence we can write $\mu(\theta) = \kappa'(\theta)$. Moreover, since var$(X) = \kappa''(\theta)$, then the derivative of μ, $\mu'(\theta)$, is strictly positive. In other words, $\mu(\theta)(= \kappa'(\theta))$ is an increasing function in θ. Thus $\mu(\theta)$ is an invertible function, therefore given $\mu(\theta)$, we can uniquely determine θ. This observation will prove useful later when obtaining the mle estimators of θ.*

3.1.2 Maximum likelihood estimation for the exponential family

Suppose that $\{X_t\}$ are iid random variables which have a natural exponential distribution representation. Then the log likelihood function is

$$\mathcal{L}_T(\underline{X};\theta) = \theta\sum_{t=1}^{T} X_t - T\kappa(\theta) + T\sum_{t=1}^{T} c(X_t).$$

Hence by using the factorisation theorem we see that the sufficient statistic for θ is $s(\underline{X}) = \sum_{t=1}^{T} X_t$. Hence, supposing that Assumption 1.1.1 is satisfied, then the minimum variance unbiased estimator of θ should be a function of $s(\underline{X})$. We now obtain the maximum likelihood estimator of θ, and derive conditions under which the mle is a function of $s(\underline{X})$ (hence, by Rao-Blackwell theorem and the Lehmann-Scheffe lemma it is the best estimator).

The mle of θ is $\hat{\theta}_T$ where

$$\hat{\theta}_T = \arg\max_{\theta\in\Theta} \big\{\theta\sum_{t=1}^{T} X_t - T\kappa(\theta) + \sum_{t=1}^{T} c(X_t)\big\}.$$

The natural way to obtain $\hat{\theta}_T$ is to find the solution of $\frac{\partial \mathcal{L}_T(\underline{X};\theta)}{\partial \theta} = 0$. However for $\frac{\partial \mathcal{L}_T(\underline{X};\theta)}{\partial \theta}\rfloor_{\theta=\hat{\theta}_T} = 0$, depends on a few conditions. Before we derive these conditions we first consider the solution of the derivative of $\mathcal{L}_T(\underline{X};\theta)$. Differentiating $\mathcal{L}_T(\underline{X};\theta)$ gives

$$\frac{\partial \mathcal{L}_T(\underline{X};\theta)}{\partial \theta} = \sum_{t=1}^{T} X_t - T\kappa'(\theta).$$

Therefore, since $\mu(\theta) = \kappa^{prime}(\theta)$ is an invertible function, then $\frac{\partial \mathcal{L}_T(\underline{X};\theta)}{\partial \theta} = 0$ when

$$\hat{\theta}_T = \mu^{-1}\big(\frac{1}{T}\sum_{t=1}^{T} X_t\big).$$

Of course, we need to know under what conditions

$$\mu^{-1}\big(\frac{1}{T}\sum_{t=1}^{T} X_t\big) = \arg\max_{\theta\in\Theta}\big\{\theta\sum_{t=1}^{T} X_t - T\kappa(\theta) + T\sum_{t=1}^{T} c(X_t)\big\}.$$

The above really depends on the parameter space Θ.

Definition 3.1.1 *Let Θ be the parameter space of θ and the space of outcomes of the random variable X, $\mathcal{Y}$. Let $\mathcal{M} = \{\mu = \mu(\theta); \theta \in \Theta\}$ denote the man space. Let $\bar{\mathcal{Y}}_T = \{y = \frac{1}{T}\sum_{t=1}^{T} x_t; x_t \in \mathcal{Y}\}$ the sample mean space.*

Lemma 3.1.2 *Suppose that $\{X_t\}$ are iid random variables which have a natural exponential representation. If $\mathcal{Y}_T \subset \mathcal{M}$ then*

$$\mu^{-1}\big(\frac{1}{T}\sum_{t=1}^{T} X_t\big) = \arg\max_{\theta\in\Theta}\big\{\theta\sum_{t=1}^{T} X_t - T\kappa(\theta) + T\sum_{t=1}^{T} c(X_t)\big\}.$$

PROOF. The proof is straightforward, since the first derivative is zero when $\theta = \mu^{-1}\big(\frac{1}{T}\sum_{t=1}^{T} X_t\big)$. Then this is the maximum of $\ell(\underline{X};\theta)$ in the sample mean space $\bar{\mathcal{Y}}_T$. Hence in order for it the minimum over the mean space $\mathcal{M}$, then either $\mathcal{M} = \bar{\mathcal{Y}}_T$ or $\bar{\mathcal{Y}}_T \subset \mathcal{M}$. $\square$

Remark 3.1.2 (Minimum variance unbiased estimators) *Suppose X_t has a distribution in the natural exponential family, the conditions of the above lemma are satisfied and $s(\underline{X})$ is the complete statistic of θ. Moreover if $\mu^{-1}(\frac{1}{T}\sum_{t=1}^{T} X_t)$ is an unbiased estimator of θ, then $\mu^{-1}(\frac{1}{T}\sum_{t=1}^{T} X_t)$ is the minumum variance unbiased estimator of θ. However, in general, this will not be case. But by using Slutsky's theorem it can be shown that $\mu^{-1}(\frac{1}{T}\sum_{t=1}^{T} X_t) \xrightarrow{\mathcal{P}} \theta$.*

Remark 3.1.3 (Estimating ω) *Often we are interested in estimating ω, where $\theta = \theta(\omega)$. However, since*

$$\frac{\partial \ell(\underline{X};\theta)}{\partial \omega} = \frac{\partial \theta}{\partial \omega} \times \frac{\partial \ell(\underline{X};\theta)}{\partial \theta} = \theta'(\omega)\big(\sum_{t=1}^{T} X_t - T\kappa'(\theta)\big).$$

Then if all conditions regarding parameter and sample mean spaces are satisfied then the mle of ω is

$$\hat{\omega}_T = \eta^{-1}\bigg(\mu^{-1}\big(\frac{1}{T}\sum_{t=1}^{T} X_t\big)\bigg).$$

It should be noted that one great advantage of the exponential family of distributions is that the mle is easy to obtain (with explicit expressions!).

Many of the results above can be generalised to the setting that $\{X_t\}$ are independent but not necessarily identically distributed and there exists regressors z which are known to influence the mean of X_t. We will revisit this problem when we consider generalised linear models.