CHAPTER 6

Saddlepoint approximation

6.1 Introduction

One difficulty that arises as a result of approximating the density function or log density function is that the density is not invariant under affine transformation of X. Any approximation, therefore, ought to have similar non-invariant properties and this requirement raises difficulties when we work exclusively with tensors. For example, if $Y^r = a^r + a^r_i X^i$ is an affine transformation of X, then the density function of Y at y is

$$f_Y(y) = |A|^{-1} f_X(x)$$

where $x^i = b^i_r(y^r - a^r)$, $b^i_r a^r_j = \delta^i_j$ and $|A|$ is the determinant of a^r_i, assumed to be non-zero. In the terminology of Thomas (1965), the density is said to be an invariant of *weight* 1: ordinary invariants have weight zero.

One way to exploit the advantages of working with invariants, at least under affine transformation, is to work with probabilities of sets rather than with probability densities. The difficulty then is to specify the sets in an invariant manner, for example as functions of the invariant polynomials

$$(x^i - \kappa^i)(x^j - \kappa^j)\kappa_{i,j}, \quad (x^i - \kappa^i)(x^j - \kappa^j)(x^k - \kappa^k)\kappa_{i,j,k},$$
$$(x^i - \kappa^i)\kappa^{j,k}\kappa_{i,j,k}, \quad (x^i - \kappa^i)(x^j - \kappa^j)\kappa^{k,l}\kappa_{i,j,k,l}$$

and so on. Another way, more convenient in the present circumstances, is to specify the probability density with respect to a so-called carrier measure on the sample space. This can always be done in such a way that the approximating density is invariant and the carrier measure transforms in such a way as to absorb the Jacobian of the transformation.

To be more explicit, suppose that the density function of X is written as the product

$$f_X(x) = |\kappa^{i,j}|^{-1/2} g(x).$$

Now let Y be an affine transformation of X as before. The covariance matrix, being a contravariant tensor, transforms to

$$\bar{\kappa}^{r,s} = a^r_i a^s_j \kappa^{i,j}.$$

Then the density of Y at y is simply

$$f_Y(y) = |\bar{\kappa}^{r,s}|^{-1/2} g(x).$$

Thus, with the inverse square root of the determinant of the covariance matrix playing the role of carrier measure, the density $g(x)$ is invariant under affine transformation of coordinates.

From this viewpoint, the usual normal-theory approximation uses the constant carrier measure $(2\pi)^{-p/2}|\kappa^{i,j}|^{-1/2}$ together with the invariant quadratic approximation

$$(x^i - \kappa^i)(x^j - \kappa^j)\kappa_{i,j}/2$$

for the negative log density. The Edgeworth approximation retains the carrier measure but augments the approximation for the negative log density by the addition of further invariant polynomial terms, namely

$$-\kappa^{i,j,k}h_{ijk}/3! - \kappa^{i,j,k,l}h_{ijkl}/4! - \kappa^{i,j,k}\kappa^{l,m,n}h_{ijk,lmn}[10]/6! - \cdots$$

In the Edgeworth system of approximation, the carrier measure is taken as constant throughout the sample space. Thus the whole burden of approximation lies on the invariant series approximation. Furthermore, this property of constancy of the carrier measure is preserved only under transformations for which the Jacobian is constant, i.e. under affine transformation alone. It is therefore appropriate to investigate the possibility of using alternative systems of approximation using non-polynomial invariants together with carrier measures that are not constant throughout the sample space. In all cases considered, the carrier measure is the square root of the determinant of a covariant tensor or, equivalently so far as transformation properties are concerned, the inverse square root of the determinant of a contravariant tensor.

6.2 Legendre transformation of $K(\xi)$

6.2.1 *Definition*

In what follows, it is convenient to consider the set of ξ-values, denoted by Ξ, for which $K(\xi) < \infty$ as being, in a sense, complementary or dual to the sample space of possible averages of identically distributed Xs. Thus, $\mathcal{X}$ is the interior of the convex hull of the sample space appropriate to a single X. In many cases, the two sample spaces are identical, but, particularly in the case of discrete random variables there is an important technical distinction. Both $\mathcal{X}$ and Ξ are subsets of p-dimensional space. It is essential, however, to think of the spaces as distinct and qualitatively different: if we are contemplating the effect of linear transformation on X, then vectors in $\mathcal{X}$ are contravariant whereas vectors in Ξ are covariant and inner products are invariant. To keep the distinction clear, it is sometimes helpful to think of Ξ as a parameter space even though we have not yet introduced any parametric models in this context.

Corresponding to the cumulant generating function $K(\xi)$ defined on Ξ, there is a dual function $K^*(x)$ defined on $\mathcal{X}$ such that the derivatives $K^r(\xi)$ and $K^*_r(x)$ are functional inverses. In other words, the solution in ξ to the p equations

$$K^r(\xi) = x^r \tag{6.1}$$

is

$$\xi_i = K^*_i(x), \tag{6.2}$$

where $K^*_i(x)$ is the derivative of $K^*(x)$. The existence of a solution to (6.1) has been demonstrated by Daniels (1954); see also Barndorff-Nielsen (1978, Chapter 5), where $K^*(x)$ is called the *conjugate function* of $K(\xi)$. Uniqueness follows from the observation that $K(\xi)$ is a strictly convex function (Exercise 6.2).

The function $K^*(x)$ is known as the Legendre or Legendre-Fenchel transformation of $K(\xi)$: it occurs in the theory of large deviations (Ellis, 1985, p.220), where $-K^*(x)$ is also called the *entropy* or *point entropy* of the distribution $f_X(x)$.

In the literature on convex analysis, the term convex conjugate is also used (Fenchel, 1949; Rockafellar, 1970, Sections 12, 26).

An alternative, and in some ways preferable definition of $K^*(x)$ is

$$K^*(x) = \sup_{\xi} \{\xi_i x^i - K(\xi)\}. \tag{6.3}$$

To see that the two definitions are mutually consistent, we note that (6.1) or (6.2) determines the stationary points of the function in (6.3). Now write $h(x)$ for the maximum value, namely

$$h(x) = x^i K^*_i(x) - K(K^*_i(x))$$

Differentiation with respect to x^i gives $h_i(x) = K^*_i(x)$, showing that $K^*(x) = h(x)$+const. The constant is identified by (6.3) but not by the previous definition. The turning point gives a

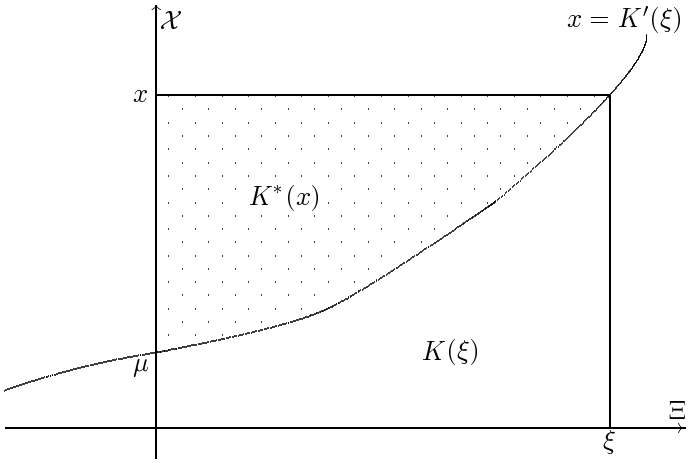

Figure 6.1: *Graphical construction of Legendre transformation in the univariate case.*

maximum because $K(\xi)$ is convex on Ξ. Note also that $K^*(x)$ is convex on $\mathcal{X}$ and achieves its minimum value of zero at $x^i = \kappa^i$.

Figure 6.1 illustrates graphically the construction of the Legendre transformation in the univariate case. The solid line gives the graph of $K'(\xi)$ against ξ and has positive gradient since $K''(\xi) > 0$. In the illustration, the intercept, which is equal to $E(X)$ is taken to be positive. The area under the curve from the origin to ξ is equal to the cumulant generating function $K(\xi) = \int_0^\xi K'(t)dt$. For any given value of x, ξx is the area of the rectangle whose opposite corners are at the origin and (ξ, x). Evidently, from the geometry of the diagram, $\xi x - K(\xi)$ is maximized at the value ξ satisfying $K'(\xi) = x$. The shaded area above the curve is equal to $\xi K'(\xi) - K(\xi)$. Regarded as a function of x, this is the Legendre transformation of $K(\xi)$. Equivalently, the graph reflected about the 45° line gives the inverse function of $K'(\xi)$. Integration from μ to x gives the area shaded above the line, showing that the two definitions given earlier are equivalent.

A similar geometrical description applies in the p-dimensional case with the two axes in Figure 6.1 replaced by the appropriate p-dimensional dual spaces. The graph is replaced by a mapping from Ξ to $\mathcal{X}$, but unfortunately, this is awkward to visualize even for $p = 2$ because four dimensions are involved.

Equations (6.1) or (6.2) identify corresponding points in Ξ and $\mathcal{X}$. In fact, every point in Ξ has a unique image in $\mathcal{X}$ given by (6.1) and conversely, every point in $\mathcal{X}$ has a unique image in Ξ given by (6.2). For example, the point $x^i = \kappa^i$ in $\mathcal{X}$ is identified with the point $\xi = 0$ in Ξ. The values of the two functions at these points are $K(0) = 0$ and $K^*(\kappa^i) = 0$ respectively.

Schematically, the relation between the first few derivatives of $K(\xi)$ and the dual function, $K^*(x)$ is as follows.

$$\begin{array}{ccc} K(\xi) & & K^*(x) \\ \downarrow \partial/\partial\xi_i & & \downarrow \partial/\partial x^i \\ K^i(\xi) & \longleftrightarrow \text{ inverse function } \longleftrightarrow & K^*_i(x) \\ \downarrow \partial/\partial\xi_j & & \downarrow \partial/\partial x^j \\ K^{ij}(\xi) & \longleftrightarrow \text{ matrix inverse } \longleftrightarrow & K^*_{ij}(x) \end{array}$$

Apart, therefore, from the choice of constant at the integration step, it is evident that the procedure can be reversed and that the Legendre transformation of $K^*(x)$ is just $K(\xi)$, i.e. $K^{**} = K$.

Equivalently, we may show algebraically that

$$K(\xi) = \sup_x \{x^i \xi_i - K^*(x)\},$$

which is obvious from the diagram in Figure 6.1.

6.2.2 *Applications*

The following is a brief informal description mostly without proofs of the role of the Legendre transformation in approximating distributions. For a more formal and rigorous account of the theory in the context of large deviations, see Ellis (1985). For the connection with exponential family models, see Barndorff-Nielsen (1978).

Large deviations: The following inequality, which is central to much of the theory of large deviations in the univariate case, helps to explain the role played by the Legendre transformation. Let X be a real-valued random variable with mean μ and let $x > \mu$ be a given number. We show that

$$\text{pr}(X \geq x) \leq \exp\{-K^*(x)\}. \tag{6.4}$$

To derive this inequality, we first write the required probability in the form

$$\text{pr}(X - x \geq 0) = \text{pr}\{\exp(\xi(X - x)) \geq 1\},$$

which is valid for all $\xi > 0$. Thus, since $\exp(\xi x) \geq H(x)$, where $H(.)$ is the Heaviside function, it follows that

$$\begin{aligned} \text{pr}(X \geq x) &\leq \inf_{\xi > 0} \exp\{-\xi x + K(\xi)\} \\ &= \exp\{-K^*(x)\} \end{aligned}$$

and the inequality is proved.

More generally, for vector-valued X, it may be shown that, for any set A in $\mathcal{X}$,

$$\text{pr}(X \in A) \leq \exp\{I(A)\}$$

where

$$I(A) = \sup_{x \in A} \{-K^*(x)\}$$

is called the entropy of the set A. Evidently, this is a generalization of the univariate inequality. The multivariate inequality follows from the univariate inequality together with the observation that $A \subset B$ implies $I(A) \leq I(B)$.

To see the relevance of the above inequalities to the theory of large deviations, let $X_1, \ldots, X_n$ be independent and identically distributed with cumulant generating function $K(\xi)$ and let $\bar{X}_n$ be the sample average. Large deviation theory is concerned mainly with approximations for the probability of the event $\bar{X}_n \geq x$ where $x > \mu$ is a fixed value independent of n. It is not difficult to see from the law of large numbers that the event in question has negligible probability for large n: in fact, for fixed x, the probability decreases exponentially fast as $n \to \infty$. The central result due to Cramér (1938) and Chernoff (1952), which determines the exponential rate of decrease, is that

$$n^{-1} \log \text{pr}(\bar{X}_n \geq x) \to -K^*(x) \tag{6.5}$$

as $n \to \infty$. Note that $nK^*(x)$ is the Legendre transformation of $\bar{X}_n$, implying, in effect, that the inequality (6.4) becomes increasingly sharp as $n \to \infty$. As a numerical device for approximating tail probabilities however, the above limit is rarely of sufficient accuracy for statistical purposes. Better approximations are described in Section 6.2.6.

The Cramér-Chernoff key limit theorem may be extended to vector-valued random variables in the following way. Let $\bar{X}_n \in R^p$ and $A \subset R^p$ be open and convex. The condition that A be open can often be dropped. Then the required limit may be written

$$n^{-1} \log \mathrm{pr}(\bar{X}_n \in A) \to I(A) \tag{6.6}$$

as $n \to \infty$. Note that if $E(X)$ lies in the interior of A then $I(A)$ is zero, which is consistent with the law of large numbers, $\mathrm{pr}(\bar{X}_n \in A) \to 1$. Similarly, if $E(X)$ lies on the boundary of A, $I(A)$ is zero and $\mathrm{pr}(\bar{X}_n \in A) \to \mathrm{const} \leq \frac{1}{2}$. In other words, if $\mathrm{pr}(\bar{X}_n \in A)$ tends to a finite limit, the limiting value cannot be deduced from the entropy limit (6.6).

Proofs of the limits (6.5) and (6.6) may be found in the books by Bahadur (1971) and Ellis (1985). Bahadur & Zabell (1979) give a number of generalizations.

Likelihood ratio statistics: Consider now the exponential family of distributions parameterized by θ in the form

$$f_X(x;\theta) = \exp\{\theta_i x^i - K(\theta)\} f_X(x). \tag{6.7}$$

First note that since $K(.)$ is the cumulant generating function of the distribution $f_X(x)$, it follows that $f_X(x;\theta)$ is a probability distribution on $\mathcal{X}$ for all θ in Ξ. The cumulant generating function of $f_X(x;\theta)$ is $K(\xi+\theta) - K(\theta)$.

Suppose we wish to test the hypothesis that $\theta = 0$ based on an observed value x on X. The log likelihood ratio is

$$\log f_X(x;\theta) - \log f_X(x) = \theta_i x^i - K(\theta).$$

The maximized log likelihood ratio statistic, maximized over θ, gives

$$\sup_\theta \{\theta_i x^i - K(\theta)\} = K^*(x)$$

where the maximum occurs at $\hat{\theta}_i = K_i^*(x)$. In this setting, the Legendre transformation is none other than the maximized log likelihood ratio statistic. It follows then, by the usual asymptotic property of likelihood ratio statistics, that

$$2nK^*(\bar{X}_n) \sim \chi_p^2 + o(1)$$

and typically the error of approximation is $O(n^{-1})$. For a derivation of this result including an explicit expression for the $O(n^{-1})$ term, see Section 6.2.4.

Saddlepoint approximation: One reason for considering the Legendre transformation of $K(\xi)$ is that it is a function on $\mathcal{X}$, invariant under affine transformation, that is useful for approximating the density of X, and hence, the density of any one-to-one function of X. In fact, the saddlepoint approximation, derived in Section 6.3, for the density of X may be written in the form

$$(2\pi)^{-p/2} |K_{rs}^*(x)|^{1/2} \exp\{-K^*(x)\}. \tag{6.8}$$

This approximation arises from applying the inversion formula to $K(\xi)$, namely

$$f_X(x) = (2\pi i)^{-p} \int_{c-i\infty}^{c+i\infty} \exp\{K(\xi) - \xi_i x^i\}\, d\xi.$$

The saddlepoint of the integrand is given by (6.1) and occurs at a point for which ξ_i is real. Approximation of the exponent in the neighbourhood of the saddlepoint by a quadratic function and integrating gives (6.8). The approximation can be justified as an asymptotic approximation if X is an average or standardized sum of n independent random variables where n is assumed to be large.

From the point of view discussed in Section 6.1, approximation (6.8) uses the carrier measure $(2\pi)^{-p/2}|K^*_{rs}(x)|^{1/2}$ on $\mathcal{X}$, together with the invariant approximation $K^*(x)$ for the negative log density. Note that the carrier measure in this case is not constant on $\mathcal{X}$. This property permits us to consider non-linear transformations of X incorporating the non-constant Jacobian into the determinant and leaving the exponent in (6.8) unaffected. Of particular importance is the distribution of the random variable $Y_i = K^*_i(X)$. Direct substitution using (6.8) gives as the required approximation

$$(2\pi)^{-p/2}|K^{rs}(y)|^{1/2}\exp\{-K^*(x)\}.$$

In this case, $Y_i = \hat{\theta}_i$, the maximum likelihood estimate of the parameter in the exponential family derived from $f_X(x)$.

Greater accuracy in the asymptotic sense can be achieved by retaining the carrier measure and replacing the exponent in (6.8) by

$$-K^*(x) - (3\rho^*_4(x) - 4\rho^{*2}_{23}(x))/4! \tag{6.9}$$

where

$$\rho^*_4 = K^*_{ijkl}K^{*ij}K^{*kl}$$
$$\rho^{*2}_{23} = K^*_{ijk}K^*_{rst}K^{*ir}K^{*js}K^{*kt}$$

and K^{*ij}, the matrix inverse of K^*_{ij}, is identical to $K^{ij}(\xi)$ with ξ satisfying (6.2). The additional correction terms in (6.9) arise from expanding the exponent, $K(\xi) - \xi_i x^i$, about the saddlepoint as far as terms of degree four in ξ. These correction terms may alternatively be expressed in terms of the derivatives of $K(\xi)$ evaluated at the saddlepoint. Thus (6.9) becomes

$$-K^*(x) + (3\rho_4(\xi) - 3\rho^2_{13}(\xi) - 2\rho^2_{23}(\xi))/4!$$

where, for example,

$$\rho^2_{13}(\xi) = K_{ijk}K_{rst}K^{ij}K^{kr}K^{st}$$

and the functions are evaluated at the saddlepoint. Evidently, $\rho^2_{13}(\xi)$ evaluated at $\xi = 0$ is identical to ρ^2_{13}, and similarly for the remaining invariants.

6.2.3 *Some examples*

In general, it is not easy to compute the Legendre transformation of an arbitrary cumulant generating function in terms of known functions. The following examples show that the calculation is feasible in a number of cases.

Multivariate normal density: The cumulant generating function for the normal distribution is

$$K(\xi) = \kappa^i\xi_i + \kappa^{i,j}\xi_i\xi_j/2!.$$

The derivative, which is linear in ξ, may be inverted giving the Legendre transformation, which is

$$K^*(x) = (x^i - \kappa^i)(x^j - \kappa^j)\kappa_{i,j}/2!.$$

Evidently, since $K^*(x)$ is quadratic in x, K^*_{rs} is a constant equal to $\kappa_{r,s}$ and the saddlepoint approximation (6.8) is exact.

Gamma distribution: Suppose that X has the univariate gamma distribution with mean μ and variance μ^2/ν, where ν is the index or precision parameter. The cumulant generating function is

$$K(\xi) = -\nu\log(1 - \mu\xi/\nu).$$

Taylor expansion about $\xi = 0$ yields the higher-order cumulants $\kappa_r = (r-1)!\,\mu^r/\nu^{r-1}$. On solving the equation $K'(\xi) = x$, we find

$$\xi(x) = \frac{\nu}{\mu} - \frac{\nu}{x}.$$

Integration gives

$$K^*(x) = \nu \left\{ \frac{x-\mu}{\mu} - \log\left(\frac{x}{\mu}\right) \right\},$$

which is exactly one half of the deviance contribution for gamma models (McCullagh & Nelder, 1983, p.153).

Since the second derivative of $K^*(x)$ is ν/x^2, it follows that the saddlepoint approximation for the density is

$$\frac{\left(\frac{\nu x}{\mu}\right)^{\nu} \exp\left(\frac{-\nu x}{\mu}\right) \frac{1}{x}\, dx}{(2\pi)^{1/2} \nu^{\nu-1/2} \exp(-\nu).}$$

This approximation differs from the exact density only to the extent that Stirling's approximation is used in the denominator in place of $\Gamma(\nu)$. In this example, this kind of defect can be corrected by choosing a multiplicative constant in such a way that the total integral is exactly one. This re-normalized saddlepoint approximation is exact for all gamma distributions.

Poisson distribution: The cumulant generating function is $\mu(\exp(\xi) - 1)$ showing that all cumulants are equal to μ. Following the procedure described above, we find

$$K^*(x) = x \log(x/\mu) - (x - \mu).$$

In the literature on log linear models, $2K^*(x)$ is more familiar as the deviance contribution or contribution to the likelihood ratio test statistic. The second derivative of K^* is just x^{-1}. Thus, the saddlepoint approximation gives

$$\frac{\exp(-\mu)\mu^x}{(2\pi)^{1/2} x^{x+1/2} \exp(-x).}$$

Again, this differs from the exact distribution in that Stirling's approximation has been used in place of $x!$. Unlike the previous example, however, re-normalization cannot be used to correct for this defect because x is not a constant and the *relative* weights given to the various values of x by the approximation are not exact.

6.2.4 *Transformation properties*

The Legendre transformation possesses a number of invariance properties that help to explain its role in approximating densities. First, let $Y^r = a^r + a^r_i X^i$ be an affine transformation of X. Any point in $\mathcal{X}$ can be identified either by its x-coordinates or by its y-coordinates, and the two are, in a sense, equivalent. The inverse transformation may be written $X^i = b^i_r(Y^r - a^r)$. It is easy to show that the Legendre transformation of $K_Y(\xi)$ is identical to the Legendre transformation of $K_X(\xi)$. To demonstrate this fact, we note first that

$$K_Y(\xi) = K_X(\xi_i a^i_j) + \xi_i a^i$$

Differentiation with respect to ξ_i and equating the derivative to y^i gives

$$K^i_X(\xi_i a^i_j) = b^i_j(y^j - a^j) = x^i.$$

Hence

$$\xi_i = b^j_i K^*_j(x),$$

which is the derivative with respect to y^i of $K^*(x)$.

For an alternative proof using definition (6.3) directly, see Exercise 6.9.

More generally, if a_i^r does not have full rank, it is easily shown that

$$K_X^*(x) \geq K_Y^*(y)$$

for all $x \in \mathcal{X}$ and y in the image set. To derive this inequality, we work directly from definition (6.3), giving

$$\begin{aligned} K_Y(y) &= \sup_\zeta \{\zeta_r y^r - K_Y(\zeta)\} \\ &= \sup_\zeta \{(\zeta_r a_i^r)x^i - K_X(\zeta_r a_i^r)\} \\ &\leq \sup_\xi \{\xi_i x^i - K_X(\xi)\} = K^*(x). \end{aligned}$$

This inequality is intuitively obvious from the interpretation of $K^*(x)$ as a log likelihood ratio statistic or *total deviance* in the sense of McCullagh & Nelder (1983).

A second important property of the Legendre transformation concerns its behaviour under exponential tilting of the density $f_X(x)$. The exponentially tilted density is

$$f_X(x;\theta) = \exp\{\theta_i x^i - K(\theta)\} f_X(x)$$

where θ is the tilt parameter, otherwise known as the canonical parameter of the exponential family. The effect of exponential tilting on the negative log density is to transform from $-\log f_X(x)$ to

$$-\log f_X(x) - \theta_i x^i + K(\theta).$$

To see the effect on the Legendre transform, we write

$$K^*(x;\theta) = \sup_\xi \{\xi_i x^i - K(\xi+\theta) + K(\theta)\} \tag{6.10}$$

where $K(\xi+\theta)-K(\theta)$ is the cumulant generating function of the tilted density (6.7). An elementary calculation gives

$$K^*(x;\theta) = K^*(x) - \theta_i x^i + K(\theta) \tag{6.11}$$

so that, under this operation, the Legendre transformation behaves exactly like the negative log density.

Table 6.1 *Transformation properties of the Legendre transform*

Transformation	*Cumulant generating function*	*Legendre transform*
Identity	$K(\xi)$	$K^*(y)$
Convolution (sum)	$nK(\xi)$	$nK^*(y/n)$
Average	$nK(\xi/n)$	$nK^*(y)$
Location shift	$K(\xi) + \xi_i a^i$	$K^*(y-a)$
Exponential tilt	$K(\xi+\theta) - K(\theta)$	$K^*(y) - \theta_i y^i + K(\theta)$
Affine	$K(a_j^i \xi_i) + \xi_i a^i$	$K^*(b_j^i(y^j - a^j))$

The above transformation properties have important consequences when the Legendre transform is used to approximate the negative log density function. In particular, whatever the error incurred in using the saddlepoint approximation to $f_X(x)$, the same error occurs uniformly for all θ in the saddlepoint approximation to $f_X(x;\theta)$, the exponentially tilted density.

Table 6.1 provides a list of some of the more important transformation properties of the Legendre transformation.

6.2.5 *Expansion of the Legendre transformation*

In this section, we derive the Taylor expansion of $K^*(x)$ about $x^i = \kappa^i$. The expansion is useful for finding the approximate distribution of the random variables $K^*(X)$ and $K_i^*(X)$. As shown in Section 6.2.2, the first of these is a likelihood ratio statistic: the second is the maximum likelihood estimate of the canonical parameter in the exponential family generated by $f_X(x)$.

The first step is to expand the derivative of the cumulant generating function in a Taylor series about the origin giving

$$K^r(\xi) = \kappa^r + \kappa^{r,i}\xi_i + \kappa^{r,i,j}\xi_i\xi_j/2! + \kappa^{r,i,j,k}\xi_i\xi_j\xi_k/3! + \cdots .$$

On solving the equation $K^r(\xi) = x^r$ by reversal of series and substituting $z^i = x^i - \kappa^i$, we find

$$\begin{aligned}\xi_i = \kappa_{i,r}z^r &- \kappa_{i,r,s}z^rz^s/2! \\ &- \{\kappa_{i,r,s,t} - \kappa_{i,r,u}\kappa_{s,t,v}\kappa^{u,v}[3]\}z^rz^sz^t/3! + \cdots .\end{aligned}$$

After integrating term by term and after applying the boundary condition $K^*(\kappa^r) = 0$, we find

$$\begin{aligned}K^*(x) = \kappa_{i,j}z^iz^j/2! &- \kappa_{i,j,k}z^iz^jz^k/3! \\ &- \{\kappa_{i,j,k,l} - \kappa_{i,j,r}\kappa_{k,l,s}\kappa^{r,s}[3]\}z^iz^jz^kz^l/4! \\ &- \{\kappa_{i,j,k,l,m} - \kappa_{i,j,k,r}\kappa_{l,m,s}\kappa^{r,s}[10] \\ &\qquad + \kappa_{i,j,r}\kappa_{k,l,s}\kappa_{m,t,u}\kappa^{r,t}\kappa^{s,u}[15]\}z^i\cdots z^m/5! \\ &- \cdots .\end{aligned} \tag{6.12}$$

Note that the invariance of $K^*(x)$ follows immediately from the above expansion.

The close formal similarity between (6.12) and expansion (5.5) for the negative log density is remarkable: in fact, the two expansions are identical except that certain polynomial terms in (6.12) are replaced by generalized Hermite tensors in (5.5). For example, in (5.5) the coefficient of $\kappa_{i,j,k}$ is $-h^{ijk}/3!$ while the coefficient of $\kappa_{i,j,r}\kappa_{k,l,s}$ is $-h^{ijr,kls}[10]/6!$, which is quartic in x.

If $\bar{X}_n$ is the mean of n identically distributed Xs, then $Z^r = \bar{X}_n^r - \kappa^r$ is $O_p(n^{-1/2})$. A straightforward calculation using (6.12) reveals that the likelihood ratio statistic has expectation

$$\begin{aligned}E(2nK^*(\bar{X})) &= p\{1 + (3\bar{\rho}_{13}^2 + 2\bar{\rho}_{23}^2 - 3\bar{\rho}_4)/(12n)\} + O(n^{-2}) \\ &= p\{1 + b/n\} + O(n^{-2}),\end{aligned} \tag{6.13}$$

where the ρs are the invariant cumulants of X_1. It is evident from expansion (6.12) that $2nK^*(\bar{X}_n)$ has a limiting χ_p^2 distribution because the terms beyond the first are negligible. With a little extra effort, it can be shown that all cumulants of $\{1 + b/n\}^{-1}2nK^*(\bar{X}_n)$ are the same as those of χ_p^2 when terms of order $O(n^{-2})$ are neglected. This adjustment to the likelihood ratio statistic is known as the Bartlett factor.

The key idea in the proof is to write the likelihood ratio statistic as a quadratic form in derived variables Y. Thus

$$2nK^*(\bar{X}_n) = Y^rY^s\kappa_{r,s}$$

where Y is a polynomial in Z. Rather intricate, but straightforward calculations then show that Y has third and fourth cumulants of orders $O(n^{-3/2})$ and $O(n^{-2})$ instead of the usual $O(n^{-1/2})$ and $O(n^{-1})$. Higher-order cumulants are $O(n^{-3/2})$ or smaller. To this order of approximation, therefore, Y has a normal distribution and $2nK^*(\bar{X}_n)$ has a non-central χ_p^2 distribution for which the rth cumulant is

$$\kappa_r = \{1 + b/n\}^r2^{r-1}p(r-1)! + O(n^{-2}).$$

Thus the correction factor $1 + b/n$, derived as a correction for the mean, corrects all the cumulants simultaneously to the same order of approximation. See also Exercises 6.16 and 6.17.

Details of the proof are not very interesting but can be found in McCullagh (1984b, Section 7.1).

6.2.6 *Tail probabilities in the univariate case*

All of the calculations of the previous section apply equally to the univariate case, particularly (6.12) and (6.13). In order to find a suitably accurate approximation to the distribution function $\mathrm{pr}(X \le x)$, or to the tail probability $\mathrm{pr}(X \ge x)$, there are several possible lines of attack. The first and most obvious is to attempt to integrate the saddlepoint approximation directly. This method has the advantage of preserving the excellent properties of the saddlepoint approximation but it is cumbersome because the integration must be carried out numerically. An alternative method that is less cumbersome but retains high accuracy is to transform from X to a new scale defined by

$$T(x) = \pm\{2K^*(x)\}^{1/2}$$

where the sign of $T(x)$ is the same as that of $x - \mu$. The random variable $T(X)$ is sometimes called the *signed likelihood ratio statistic* although the term is appropriate only in the context of the exponential family of densities (6.7).

From the discussion in the previous section and from Exercises 6.16, 6.17, it may be seen that $T(X)$ is nearly normally distributed with mean and variance

$$\begin{aligned} E(T(X)) &\simeq -\rho_3/6 \\ \mathrm{var}(T(X)) &\simeq 1 + (14\rho_3^2 - 9\rho_4)/36, \end{aligned}$$

where $\rho_3 = \kappa_3/\kappa_2^{3/2}$ is the usual univariate standardized measure of skewness of X. In fact, the cumulants of $T(X)$ differ from those of the normal distribution having the above mean and variance, by $O(n^{-3/2})$ when X is a mean or total of n independent observations. Since $T(x)$ is increasing in x, it follows that

$$\begin{aligned} \mathrm{pr}(X \ge x) &= \mathrm{pr}\{T(X) \ge T(x)\} \\ &\simeq 1 - \Phi\left(\frac{T(x) + \rho_3/6}{1 + (14\rho_3^2 - 9\rho_4)/72}\right). \end{aligned} \tag{6.14}$$

A similar, but not identical, formula was previously given by Lugannani & Rice (1980). In fact, Daniels's (1987) version of the Lugannani-Rice formula may be written

$$\mathrm{pr}(X \ge x) \simeq 1 - \Phi(T(x)) + \phi(T(x))\left(\frac{1}{S(x)} - \frac{1}{T(x)}\right), \tag{6.15}$$

where $K'(\hat\xi) = x$ defines the saddlepoint and $S(x) = \hat\xi\{K''(\hat\xi)\}^{1/2}$ is a kind of Wald statistic. Note that the standard error is calculated under the supposition that the mean of X is at x rather than at μ.

Admittedly, approximation (6.14) has been derived in a rather dubious fashion. Neither the nature of the approximation nor the magnitude of the error have been indicated. In fact, the approximation may be justified as an asymptotic expansion if X is a sum or average of n independent random variables, in which case ρ_3 is $O(n^{-1/2})$ and ρ_4 is $O(n^{-1})$. The error incurred in using (6.14) for normal deviations is typically $O(n^{-3/2})$. By normal deviations, we mean values of x for which $K^*(x)$ is $O(1)$, or equivalently, values of x that deviate from $E(X)$ by a bounded multiple of the standard deviation. This range includes the bulk of the probability. Further adjustments to (6.14) are required in the case of discrete random variables: it often helps, for example, to make a correction for continuity.

For large deviations, the approximation of Lugannani & Rice has relative error of order $O(n^{-1})$. On the other hand, the relative error of (6.14) is $O(1)$, but despite this substantial asymptotic inferiority, (6.14) is surprisingly accurate even for moderately extreme tail probability calculations of the kind that occur in significance testing.

For further discussion of the above and related approximations, the reader is referred to Daniels (1987).

6.3 Derivation of the saddlepoint approximation

The most direct derivation of the saddlepoint approximation, by inversion of the cumulant generating function, was described in Section 6.2.2, if only briefly. A simpler method of derivation is to apply the Edgeworth approximation not to the density $f_X(x)$ directly, but to an appropriately chosen member of the conjugate family or exponential family

$$f_X(x;\theta) = \exp\{\theta_i x^i - K(\theta)\} f_X(x). \tag{6.16}$$

We aim then, for each value of x, to chose the most advantageous value of θ in order to make the Edgeworth approximation to $f_X(x;\theta)$ as accurate as possible. This is achieved by choosing $\theta = \hat{\theta}(x)$ in such a way that x is at the mean of the conjugate density under $\hat{\theta}$. In other words, we chose $\hat{\theta}$ such that

$$K^r(\hat{\theta}) = x^r \quad \text{or} \quad \hat{\theta}_r = K^*_r(x).$$

As the notation suggests, $\hat{\theta}$ is the maximum likelihood estimate of θ based on x in the family (6.16).

From (5.5) or (5.10), the Edgeworth approximation for the log density at the mean may be written

$$-\tfrac{1}{2}p\log(2\pi) - \tfrac{1}{2}\log|\kappa^{r,s}| + (3\rho_4 - 3\rho_{13}^2 - 2\rho_{23}^2)/4! + \cdots .$$

Taking logs in (6.16) gives

$$\log f_X(x) = -K^*(x) + \log f_X(x;\hat{\theta}).$$

Applying the Edgeworth expansion to the second term on the right gives

$$\begin{aligned}\log f_X(x) &= -\tfrac{1}{2}p\log(2\pi) - \tfrac{1}{2}\log|K^{rs}(\hat{\theta})| \\ &\qquad - K^*(x) + (3\rho_4(\hat{\theta}) - 3\rho_{13}^2(\hat{\theta}) - 2\rho_{23}^2(\hat{\theta}))/4! + \cdots \\ &= -\tfrac{1}{2}p\log(2\pi) + \tfrac{1}{2}\log|K^*_{rs}(x)| \\ &\qquad - K^*(x) - (3\rho_4^*(x) - 4\rho_{23}^{*2}(x))/4! + \cdots .\end{aligned}$$

The first three terms above constitute the saddlepoint approximation for the log density: the fourth term is a correction term that is $O(n^{-1})$ in large samples. Terms that are ignored are $O(n^{-2})$.

Alternatively, and sometimes preferably, we may write the approximate density function as

$$(2\pi c)^{-p/2}|K^*_{rs}|\exp\{-K^*(x)\} \tag{6.17}$$

where c is a constant chosen to make the integral equal to one. To a first order of approximation, we may write

$$\log(c) = (3\bar{\rho}_{13}^2 + 2\bar{\rho}_{23}^2 - 3\bar{\rho}_4)/12.$$

Thus, $\log(c)$ is just the Bartlett adjustment term defined in (6.13). The error of approximation is $O(n^{-3/2})$ when the above approximation is used for the constant of integration.

The advantages of the saddlepoint approximation over the Edgeworth series are mainly connected with accuracy. Although both approximations are asymptotic, the saddlepoint approximation is often sufficiently accurate for statistical purposes even when n is small, less than 10, say. In addition, the saddlepoint approximation retains high relative accuracy over the whole range of possible values of x. The Edgeworth approximation, on the other hand, is valid only for values of $\bar{X}$ that deviate from $E(\bar{X})$ by $O(n^{-1/2})$. The implication of this restriction is that the Edgeworth series may not be of adequate accuracy to judge the probability of unusual events.

On the negative side, the saddlepoint approximation applies to the density, and if tail probability calculations are required, integration is necessary. Unfortunately, the saddlepoint approximation, unlike the Edgeworth series, cannot usually be integrated analytically. Numerical integration is one

answer, but this is often cumbersome. An alternative and more convenient solution is described in Section 6.2.6. A second argument against the saddlepoint approximation in favour of the Edgeworth series is that in order to compute the saddlepoint, it is necessary to have an explicit formula for the cumulant generating function. To use the Edgeworth series, on the other hand, it is necessary only to know the first few cumulants, and these can often be computed without knowing the generating function. There may, in fact, be no closed form expression for the generating function: see, for example, Exercise 2.30. In short, the Edgeworth series is often easier to use in practice but is usually inferior in terms of accuracy, particularly in the far tails of the distribution.

6.4 Approximation to conditional distributions

6.4.1 *Conditional density*

Suppose that we require an approximation for the conditional distribution of a statistic X_2 given that $X_1 = x_1$. Both components may be vector valued. Calculations of this kind arise in a number of important areas of application. The following are a few examples.

(i) Elimination of nuisance parameters by conditioning, particularly where matched retrospective designs are used to study factors that influence the incidence of rare diseases (Breslow & Day, 1980, Chapter 7).
(ii) Conditioning to take account of the observed value of an ancillary statistic (Cox, 1958).
(iii) Testing for goodness of fit when the model to be tested contains unknown parameters (McCullagh, 1985).

The simplest and most natural way to proceed from the cumulant generating functions $K_{X_1X_2}(.)$ and $K_{X_1}(.)$ is to compute the corresponding Legendre transformations, $K^*_{X_1X_2}(x_1, x_2)$ and $K^*_{X_1}(x_1)$. The saddlepoint approximation is then used twice, once for the joint density and once for the marginal density of X_1. Thus,

$$f_{X_1X_2}(x_1, x_2) \simeq c_{12}|K^*_{X_1X_2;rs}|^{1/2} \exp\{-K^*_{X_1X_2}(x_1, x_2)\}$$
$$f_{X_1}(x_1) \simeq c_1|K^*_{X_1;rs}|^{1/2} \exp\{-K^*_{X_1}(x_1)\},$$

where c_{12} and c_1 are normalizing constants. On subtracting the approximate log densities, we find

$$\begin{aligned}\log f_{X_2|X_1}(x_2|x_1) \simeq{}& \log c_{12} - \log c_1 \\ &+ \tfrac{1}{2}\log|K^*_{X_1X_2;rs}| - \tfrac{1}{2}\log|K^*_{X_1;rs}| \\ &- K^*_{X_1X_2}(x_1, x_2) + K^*_{X_1}(x_1). \end{aligned} \tag{6.18}$$

In large samples, the error of approximation is $O(n^{-3/2})$ provided that the constants of integration c_1 and c_{12} are appropriately chosen.

Approximation (6.18) is sometimes called the *double saddlepoint approximation*. It is not the same as applying the saddlepoint approximation directly to the conditional cumulant generating function of X_2 given $X_1 = x_1$. For an example illustrating the differences, see Exercises 6.18–6.20.

6.4.2 *Conditional tail probability*

Suppose now that X_2 is a scalar and that we require an approximation to the conditional tail probability

$$\mathrm{pr}(X_2 \geq x_2|X_1 = x_1).$$

Expression (6.15) gives the required unconditional tail probability: the surprising fact is that the same expression, suitably re-interpreted, applies equally to conditional tail probabilities.

In the double saddlepoint approximation, there are, of course, two saddlepoints, one for the joint distribution of (X_1, X_2) and one for the marginal distribution of X_1. These are defined by

$$K^r(\hat{\xi}_1, \hat{\xi}_2) = x_1^r, \quad r = 1, \ldots, p-1; \quad K^p(\hat{\xi}_1, \hat{\xi}_2) = x_2$$

for the joint distribution, and

$$K^r(\tilde{\xi}_1, 0) = x_1^r, \quad r = 1, \ldots, p-1$$

for the marginal distribution of the $p-1$ components of X_1. In the above expressions, ξ_1 has $p-1$ components and ξ_2 is a scalar, corresponding to the partition of X.

The signed likelihood ratio statistic $T = T(x_2|x_1)$ is most conveniently expressed in terms of the two Legendre transformations, giving

$$T = \text{sign}(\hat{\xi}_2)\{K^*_{X_1X_2}(x_1, x_2) - K^*_{X_1}(x_1)\}.$$

Further, define the generalized conditional variance, $V = V(x_2|x_1)$, by the determinant ratio

$$V = \frac{|K^{rs}(\hat{\xi}_1, \hat{\xi}_2)|}{|K^{rs}(\tilde{\xi}_1, 0)|} = \frac{|K^*_{X_1;rs}|}{|K^*_{X_1X_2;rs}|}.$$

Using these expressions, the double saddlepoint approximation becomes

$$cV^{-1/2} \exp(-T^2/2).$$

In the conditional sample space, $V^{-1/2}$ plays the role of carrier measure and the exponent is invariant. The conditional tail probability is given by (6.15) using the Wald statistic

$$S(x) = \hat{\xi}_2 V^{1/2}.$$

This important result is due to I. Skovgaard (1986, personal communication) and is given here without proof.

6.5 Bibliographic notes

The following is a very brief description of a few key references. Further references can be found cited in these papers.

The Cramér-Chernoff large deviation result is discussed in greater detail by Bahadur (1971) and by Ellis (1985).

The derivation of the saddlepoint approximation by using an Edgeworth expansion for the exponentially tilted density goes back to the work of Esscher (1932), Cramér (1938), Chernoff (1952) and Bahadur & Ranga-Rao (1960). In a series of important papers, Daniels (1954, 1980, 1983) develops the saddlepoint method, derives the conditions under which the re-normalized approximation is exact and finds approximations for the density of a ratio and the solution of an estimating equation. Barndorff-Nielsen & Cox (1979) discuss double saddlepoint approximation as a device for approximating to conditional likelihoods.

The Legendre transformation plays an important role in the literature on large deviations, and is emphasized by Ellis (1985), Bahadur & Zabell (1979) and also to an extent, Daniels (1960).

The relationship between the Bartlett adjustment factor and the normalization factor in the saddlepoint formula is discussed by Barndorff-Nielsen & Cox (1984).

Tail probability calculations are discussed by Lugannani & Rice (1980), Robinson (1982) and by Daniels (1987).

6.6 Further results and exercises 6

6.1 Show that the array

$$M^{ij}(\xi) = E\{X^i X^j \exp(\xi_r x^r)\}$$

is positive definite for each ξ. Hence deduce that the function $M(\xi)$ is convex. Under what conditions is the inequality strict?

6.2 By using Hölder's inequality, show for any $0 \leq \lambda \leq 1$, that

$$K(\lambda\xi_1 + (1-\lambda)\xi_2) \leq \lambda K(\xi_1) + (1-\lambda)K(\xi_2)$$

proving that $K(\xi)$ is a convex function.

6.3 Prove directly that $K^{rs}(\xi)$ is positive definite for each ξ in Ξ. Hence deduce that $K^*(x)$ is a convex function on $\mathcal{X}$.

6.4 Prove that $K^*(x) \geq 0$, with equality only if $x^i = \kappa^i$.

6.5 Prove the following extension of inequality (6.4) for vector-valued X

$$\mathrm{pr}(X \in A) \leq \exp\{I(A)\}$$

where $I(A) = \sup_{x \in A}\{-K^*(x)\}$.

6.6 From the entropy limit (6.6), deduce the law of large numbers.

6.7 Show that the Legendre transformation of $Y = X_1 + \cdots + X_n$ is $nK^*(y/n)$, where the Xs are *i.i.d.* with Legendre transformation $K^*(x)$.

6.8 Show that, for each θ in Ξ,

$$\exp(\theta_i x^i - K(\theta)) f_X(x)$$

is a distribution on $\mathcal{X}$. Find its cumulant generating function and the Legendre transformation.

6.9 From the definition

$$K_Y^*(y) = \sup_{\xi}\{\xi_i y^i - K_Y(\xi)\}$$

show that the Legendre transformation is invariant under affine transformation of coordinates on $\mathcal{X}$.

6.10 By writing ξ_i as a polynomial in z

$$\xi_i = a_{ir} z^r + a_{irs} z^r z^s/2! + a_{irst} z^r z^s z^t/3! + \cdots,$$

solve the equation

$$\kappa^{r,i}\xi_i + \kappa^{r,i,j}\xi_i\xi_j/2! + \kappa^{r,i,j,k}\xi_i\xi_j\xi_k/3! + \cdots \quad = z^r$$

by series reversal. Hence derive expansion (6.12) for $K^*(x)$.

6.11 Using expansion (6.12), find the mean of $2nK^*(\bar{X}_n)$ up to and including terms that are of order $O(n^{-1})$.

6.12 Show that the matrix inverse of $K^{rs}(\xi)$ is $K^*_{rs}(x)$, where $x^r = K^r(\xi)$ corresponds to the saddlepoint.

6.13 Using (6.12) or otherwise, show that, for each x in $\mathcal{X}$,

$$K^*_{ijk}(x) = -K^{rst}K_{ri}K_{sj}K_{tk}$$
$$K^*_{ijkl}(x) = -\{K^{rstu} - K^{rsv}K^{tuw}K_{vw}[3]\}K_{ri}K_{sj}K_{tk}K_{ul}$$

where all functions on the right are evaluated at $\xi_r = K^*_r(x)$, the saddlepoint image of x.

6.14 Using the results given in the previous exercise, show, using the notation of Section 6.3, that

$$\rho_{13}^2(\hat{\theta}) = \rho_{13}^{*2}(x) \qquad \rho_{23}^2(\hat{\theta}) = \rho_{23}^{*2}(x)$$
$$\rho_4(\hat{\theta}) = -\rho_4^*(x) + \rho_{13}^{*2}(x) + 2\rho_{23}^{*2}(x).$$

6.15 By using the expansion for ξ_i given in Section 6.2.4, show that the maximum likelihood estimate of θ based on $\bar{X}_n$ in the exponential family (6.14) has bias

$$E(n^{1/2}\hat{\theta}_r) = -\tfrac{1}{2}n^{-1/2}\kappa^{i,j,k}\kappa_{i,r}\kappa_{j,k} + O(n^{-3/2}).$$

6.16 Show that if $Z^r = \bar{X}_n^r - \kappa^r$ and

$$n^{-1/2}Y^r = Z^r - \kappa^{r,s,t}\kappa_{s,i}\kappa_{t,j}Z^iZ^j/6$$
$$+ \{8\kappa^{r,s,t}\kappa^{u,v,w}\kappa_{s,i}\kappa_{t,u}\kappa_{v,j}\kappa_{w,k} - 3\kappa^{r,s,t,u}\kappa_{s,i}\kappa_{t,j}\kappa_{u,k}\}Z^iZ^jZ^k/72$$

then $Y = O_p(1)$ and

$$2nK^*(\bar{X}_n) = Y^rY^s\kappa_{r,s} + O(n^{-2}).$$

6.17 Show that Y^r defined in the previous exercise has third cumulant of order $O(n^{-3/2})$ and fourth cumulant of order $O(n^{-1})$. Hence show that $2nK^*(\bar{X}_n)$ has a non-central χ_p^2 distribution for which the rth cumulant is

$$\{1 + b/n\}^r 2^{r-1}(r-1)!p + O(n^{-2}).$$

Find an expression for b in terms of the invariant cumulants of X.

6.18 Show that, in the case of the binomial distribution with index m and parameter π, the Legendre transformation is

$$y\log\left(\frac{y}{\mu}\right) + (m-y)\log\left(\frac{m-y}{m-\mu}\right)$$

where $\mu = m\pi$. Hence show that the saddlepoint approximation is

$$\frac{\pi^y(1-\pi)^{m-y}m^{m+1/2}}{(2\pi)^{1/2}y^{y+1/2}(m-y)^{m-y+1/2}}.$$

In what circumstances is the saddlepoint approximation accurate? Derive the above as a double saddlepoint approximation to the conditional distribution of Y_1 given $Y_1 + Y_2 = m$, where the Ys are independent Poisson random variables.

6.19 Let X_1, X_2 be independent exponential random variables with common mean μ. Show that the Legendre transformation of the joint cumulant generating function is

$$K^*(x_1, x_2; \mu) = \frac{x_1 + x_2 - 2\mu}{\mu} - \log\left(\frac{x_1}{\mu}\right) - \log\left(\frac{x_2}{\mu}\right).$$

Show also that the Legendre transformation of the cumulant generating transformation of $\bar{X}$ is

$$K^*(\bar{x}; \mu) = 2\left(\frac{\bar{x} - \mu}{\mu}\right) - 2\log\left(\frac{\bar{x}}{\mu}\right).$$

Hence derive the double saddlepoint approximation for the conditional distribution of X_1 given that $X_1 + X_2 = 1$. Show that the re-normalized double saddlepoint approximation is exact.

6.20 Extend the results described in the previous exercise to gamma random variables having mean μ and indices ν_1, ν_2. Replace $\bar{X}$ by an appropriately weighted mean.

6.21 In the notation of Exercise 6.18, show that the second derivative of $K^*(x, 1-x; 1/2)$ at $x = 1/2$ is 8, whereas the conditional variance of X_1 given that $X_1 + X_2 = 1$ is $1/12$. Hence deduce that the double saddlepoint approximation to the conditional density of X_1 is not the same as applying the ordinary saddlepoint approximation directly to the conditional cumulant generating function of X_1.

6.22 Using the asymptotic expansion for the normal tail probability

$$1 - \Phi(x) \simeq \frac{\phi(x)}{x} \qquad\qquad x \to \infty$$

and taking $x > E(X)$, show, using (6.14), that

$$n^{-1} \log \mathrm{pr}\{\bar{X}_n > x\} \to -K^*(x)$$

as $n \to \infty$, where $\bar{X}_n$ is the average of n independent and identically distributed random variables. By retaining further terms in the expansion, find the rate of convergence to the entropy limit (6.5).

6.23 Using (6.11), show that the Legendre transform $K^*(x; \theta)$ of the exponentially tilted density satisfies the partial differential equations

$$\frac{\partial K^*(x;\theta)}{\partial x^r} = \hat{\theta}_r(x) - \theta_r$$
$$-\frac{\partial K^*(x;\theta)}{\partial \theta_i} = x^i - K^i(\theta).$$

Hence show that in the univariate case,

$$K^*(x;\theta) = \int_\mu^x \frac{x-t}{v(t)}\, dt,$$

where $\mu = K'(\theta)$ and $v(\mu) = K''(\theta)$, (Wedderburn, 1974; Nelder & Pregibon, 1986).

6.24 By using Taylor expansions for $S(x)$ and $T(x)$ in (6.15), show that, for normal deviations, the tail probability (6.15) reduces to

$$1 - \Phi(T) + \phi(T)\left(-\frac{\rho_3}{6} + \frac{5\rho_3^2 - 3\rho_4}{24} T\right) + O(n^{-3/2}).$$

Hence deduce (6.14).

6.25 Let X be a random variable with density function

$$f_X(x;\theta) = \exp\{\theta_i x^i - K(\theta)\} f_0(x)$$

depending on the unknown parameter θ. Let θ have Jeffreys's prior density

$$\pi(\theta) = |K^{rs}(\theta)|^{1/2}.$$

Using Bayes's theorem, show that the posterior density for θ given x is approximately

$$\pi(\theta|x) \simeq c \exp\{\theta_i x^i - K(\theta) - K^*(x)\} |K^{rs}(\theta)|^{1/2},$$

whereas the density of the random variable $\hat{\theta}$ is approximately

$$p(\hat{\theta}|\theta) \simeq c \exp\{\theta_i x^i - K(\theta) - K^*(x)\} |K^{rs}(\hat{\theta})|^{1/2}.$$

In the latter expression x is considered to be a function of $\hat{\theta}$.

Find expressions for the constants in both cases.

6.26 Consider the conjugate density $f_X(x;\theta)$ as given in the previous exercise, where $K(\theta)$ is the cumulant generating function for $f_0(x)$ and $K^*(x)$ is its Legendre transform. Show that

$$E_\theta\left\{\log\left(\frac{f_X(X;\theta)}{f_0(X)}\right)\right\} = K^*(E_\theta(X))$$

where $E_\theta(.)$ denotes expectation under the conjugate density. [In this context, $K^*(E_\theta(X))$ is sometimes called the *Kullback-Leibler distance* between the conjugate density and the original density.]