CUMULANTS AND PSEUDO-CUMULANTS FOR ASYMPOTOTIC EXPANSIONS

Nancy Reid
University of Toronto

1. Introduction
For a density with a cumulant generating function, a standard method of proof of the asymptotic normality of a sample mean from that density shows that the cumulant generating function converges to $t^2/2$. In contrast, in a location model the limiting normal distribution for the mean, conditional on the configuration statistic, is established by showing that the log density converges to $b - x^2/2$, where $e^b$ is the norming constant for the density. Both these asymptotic expansion limits can be extended to include terms to fourth order, and there is a simple correspondence between the coefficients of these expansion, neglecting terms of order $O(n^{-3/2})$. This correspondence is presented for both the univariate and multivariate cases. The results are established in Fraser and Reid (1992), and used there to obtain various approximations to tail areas.

2. The univariate case
Suppose that $Y$ is a sample mean from a distribution with a moment generating function that exists in an open interval including zero. Let $X$ be the standardized version of $Y$; i.e. $X = \sqrt{n}(\bar{Y} - \mu)/\sigma$, and let $c(t) = \log E\{\exp(tX)\}$ be the cumulant generating function of $X$. The usual proof of the central limit theorem shows that

$$c(t) = \frac{1}{2}t^2 + o(1),$$

and thus that $c(t)$ converges to the cumulant generating function for a standard normal distribution, as $n \to \infty$. A refinement of this result can be obtained (with additional regularity conditions) by examining the form of the $o(1)$ term in (1) above. Including the next two terms in the expansion we have

$$c(t) = \frac{1}{2}t^2 + \frac{\alpha_3}{6\sqrt{n}}t^3 + \frac{\alpha_4}{24n}t^4 + O(n^{-3/2}).$$
For models where the cumulant generating function is either not available, or not convenient to work with, it is often possible to establish asymptotic normality for a suitably standardized variable working with the likelihood function or log density function directly. For example, if we have a sample from a location model, we can conclude that the sample mean $\bar{Y}$ is asymptotically normally distributed, conditionally on the ancillary statistic $d = (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})$. Fraser and McDunnough (1984) show that for the standardized version of $\bar{Y}$, $X = (\bar{Y} - m)/s$,

$$\log f(x|d) = b - \frac{1}{2} x^2 + o(1),$$

where $m$ is the mode of $f(\bar{Y}|d)$, $s^2$ is the Euclidean length of $d$, and $b = -(1/2) \log(2\pi) + o(1)$. Again, it is possible to refine this argument by including terms of higher order. DiCiccio, Field and Fraser (1990) show that if the log-density for $X$ has the expansion

$$l(x) = b - \frac{1}{2} x^2 + \frac{a_3}{6\sqrt{n}} x^3 + \frac{a_4}{24n} x^4 + O(n^{-3/2}) \quad (3)$$

as is typically the case for the conditional density in a location model, then the asymptotic normality result $F(x|d) \doteq \Phi(x)$ can be improved to

$$F(x|d) = \Phi(r) + \phi(r) \left( \frac{1}{r} - \frac{1}{u} \right) + O(n^{-3/2}) \quad (4)$$

where

$$r = \pm 2 \{l(0) - l(x)\}^{1/2}$$

$$u = l'(x) \{-l''(0)\}^{-1/2}. \quad (5)$$

An analogous result can be obtained from the cumulant generating function expansion (2), and the resulting tail area approximation is that due to Lugannani and Rice (1980) and is described in the present context at (9).

There is a direct connection between the expansion of the cumulant generating function and the expansion of the log density, that can be formally obtained simply by working with the asymptotic expansions of these two functions. The details are outlined in Fraser and Reid (1992), and the result is the following.
If the log-density for \( X \) has the expansion (3), then the standard-
ized variable \( Z = (X - \mu) / \sigma \) has a cumulant generating function \( c(t) \) of the form (2), where

\[
\mu = \frac{a_3}{2\sqrt{n}}, \quad \sigma^2 = 1 + \frac{a_4 + 2a_3^2}{2n}, \quad \alpha_3 = a_3, \quad \text{and} \quad \alpha_4 = a_4 + 3a_3^2.
\] (6)

The result in the other direction is that if \( Z \) has a cumulant gen-
erating function given by (2), then \( X = (Z - m) / s \) has an expansion for the log-density given by (3), where

\[
m = -\frac{\alpha_3}{2\sqrt{n}}, \quad s^2 = 1 - \frac{a_4 + 2a_3^2}{2n}, \quad \alpha_3 = a_3, \quad \text{and} \quad \alpha_4 = a_4 - 3a_3^2.
\] (7)

To establish the equivalence we first integrate the expansion of \( \exp\{l(x)\} \) given in (3) to verify that 
\[b = -(1/2) \log(2\pi) - (3a_4 + 5a_3^2)/(24n) + O(n^{-3/2}),\] and then obtain an expansion to third order of \( \tilde{c}(t) = \log \int \exp\{tx + l(x)\} dx \), which turns out to be

\[
\tilde{c}(t) = \frac{a_3}{2\sqrt{n}} t + \frac{1}{2} \left(1 + \frac{a_4 + 2a_3^2}{2n}\right) + \frac{1}{6} t^3 \frac{a_3}{\sqrt{n}} + \frac{1}{24} t^4 \left(\frac{a_4 + 3a_3^2}{n}\right)
\] (8)

from which the form (2) of \( c(t) \) is obtained from the mean and variance adjustment indicated in (7).

The Lugannani and Rice approximation referred to above can easily be obtained from this result, by embedding a density of the form (3) in an exponential model, using the expansion here. That is, letting

\[
f(x; \theta) = \exp\{x\theta - \tilde{c}(\theta)\} \exp\{l(x)\}
\]
and substituting the expansions for \( \tilde{c}(\theta) \) and \( l(x) \), we can verify the saddlepoint approximation to the density of \( X \) (Daniels, 1954):

\[
f(x) = c_n \{\tilde{c}''(\tilde{\theta})\}^{-1/2} \exp\{l(x) - \tilde{\theta} x + \tilde{c}(\tilde{\theta})\} \{1 + O(n^{-3/2})\}
\]

where \( \tilde{c}(\tilde{\theta}) = x \). This expression can be integrated following the tech-
nique of DiCiccio, Field and Fraser (1990) to give the following ap-
proximation to the distribution function for \( X \) (Lugannani and Rice, 1980; Daniels, 1987):

\[
F(x) = \Phi(r) + \phi(r) \left(\frac{1}{r} - \frac{1}{q}\right) + O(n^{-3/2}),
\] (9)
where
\[
\begin{align*}
    r &= \pm [2\{l(0) - l(\hat{\theta})\}]^{1/2} \\
    q &= \hat{\theta}\{-l''(\hat{\theta})\}^{1/2}.
\end{align*}
\] (10)

Fraser and Reid (1992) also use the correspondence between log-density and cumulant generating function described above to construct a ‘third order asymptotic model’ for the univariate case. Any suitably smooth density \( f(x; \theta) \) that depends on a one-dimensional variable and a one-dimensional parameter can be approximated by a series expansion in \( x \) and \( \theta \):
\[
f(x; \theta) = f(x_0; \theta_0) + (x - x_0)f_x(x_0; \theta_0) + (\theta - \theta_0)f_\theta(x_0; \theta_0) + \cdots
\]

It turns out that by rescaling and recentering with respect to the parameter and the variable separately, the expansion can be expressed in the form
\[
\log f(x; \theta) = P(x) + A(\theta) + x\theta + \frac{1}{4n}x^2\theta^2 + O(n^{-3/2})
\]

where \( P(x) \) and \( A(\theta) \) are fourth degree polynomials in \( x - x_0 \) and \( \theta - \theta_0 \) respectively, and the correspondence described at (6) and (7) determines the relationship between the coefficients of \( P(x) \) and the coefficients of \( A(\theta) \). The fact that it differs from a canonical exponential family only by a single term of order \( O(n^{-1}) \) is useful for establishing general asymptotic results.

3. The multivariate case
The correspondence between log-density and cumulant generating function can also be established in the multivariate case, using the same techniques; the only difficulty is essentially notational. The log density expansion corresponding to (2) is
\[
l(x) = b - \frac{1}{2} I^{ij} x_i x_j + \frac{1}{6} A^{ijk} x_i x_j x_k + \frac{1}{24} A^{ijkl} x_i x_j x_k x_l + O(n^{-3/2})
\] (11)

where \( x \) represents a centered sum, standardized with respect to \( n \), \( I^{ij} \) is the asymptotic covariance matrix for this standardized variable, and \( I^{ij} \) is its inverse. The standardization of the sum is required in order that the covariance matrix be \( O(1) \); then the third and fourth derivative arrays \( A^{ijk} \) and \( A^{ijkl} \) are \( O(n^{-1/2}) \) and \( O(n^{-1}) \) respectively. It is more convenient in this case not to rescale \( X \) to have the identity
covariance matrix.) To compute the cumulant generating function for this density, we first obtain an expansion for the normalizing constant as

\[
\exp(b) = (2\pi)^{-p/2}|I|^{-1/2}\{1 + O(n^{-1})\}
\]

where the leading term of the remainder term is given in Fraser and Reid (1992, eq. (4.3)) in the present notation, and in DiCiccio, Field and Fraser (1990, eq. (8)) and references therein. The cumulant generating function

\[
\tilde{c}(t) = \int \exp\{t^ax_a + l(x)\}dx
\]

is computed after several changes of variable to have the expansion

\[
\tilde{c}(t) = \mu_a t^a + \frac{1}{2}\sigma_{ab} t^a t^b + \frac{1}{6}\alpha_{abc} t^a t^b t^c + \frac{1}{24}\alpha_{abcd} t^a t^b t^c t^d + O(n^{-3/2}),
\]

where

\[
\begin{align*}
\mu_a &= \frac{1}{2}A^i_i I_{ij}, & \sigma_{ab} &= I_{ab} + \frac{1}{2}\Delta_{ab} \\
\alpha_{abc} &= A^{ijk}I_{ia}I_{jb}I_{kc} = A_{abc} \\
\alpha_{abcd} &= A_{abcd} + 3A^i_{ab}A^j_{cd}I_{ij} \\
\Delta_{ab} &= A^{ij}_{ab}I_{ij} + A_{ab}A^{ij}I_{ij}I_{kl} + A^{ijk}A^{kl}_{ab}I_{ij}I_{kl}
\end{align*}
\]

which is the multivariate version of (6). Similarly there is a correspondence in the other direction, and the multivariate version of (7) is given in Fraser and Reid (1992, eq. (4.14)).

As in the univariate case, an exponential model can be constructed by tilting a density of the form (10):

\[
f(x; \theta) = \exp\{\theta^t x_i - \tilde{c}(\theta) + l(x)\}
\]

and various asymptotic results for the multivariate setting, such as the \(O(n^{-3/2})\) accuracy of the saddlepoint approximation to the density of \(X\) can be obtained. As another example, the Bartlett factor for the likelihood ratio statistic is readily obtained by writing

\[
w = 2\{(\hat{\theta}^t - \theta^t)x_i - \tilde{c}(\hat{\theta}) + \tilde{c}(\theta)\},
\]

using the expansion for \(\hat{\theta}^t\) obtained from \(\tilde{c}(\hat{\theta}) = x_i\), and the expansion of \(\tilde{c}\) given in (11) to verify that

\[
Ew = \sigma^{ij}\sigma_{ij} + O(n^{-1}) = p + O(n^{-1})
\]
where the leading term in the remainder is given by

\[-\frac{1}{4} A^{ijkl} I_{ij} I_{kl} - \frac{1}{3} A^{ijk} A^{lmn} I_{il} I_{jm} I_{kn}.\]

More general expressions for the Bartlett factor are given in McCullagh (1987, Ch. 7).

Another use for the expansions outlined here is to obtain approximations for conditional or marginal densities of components of $x$. DiCiccio, Field and Fraser (1990) use the expansion (10) to derive an approximation to the cumulative distribution function for $x_1$, marginalizing over the other components, which is of the form of (4), but involves computing the restricted maximum of $l(x)$, over $(x_2, \ldots, x_n)$ for fixed $x_1$. Fraser and Reid (1992) use the embedded exponential model described above to find an expansion for the conditional density of one component, given the remainder.

References


