DIFFERENTIAL AND INTEGRAL GEOMETRY IN STATISTICAL INFERENCE

O. E. Barndorff-Nielsen*

*Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus, Aarhus, Denmark

1. INTRODUCTION

This paper gives an account of some of the recent developments in statistical inference in which concepts and results from integral and differential geometry have been instrumental.

A great many important contributions to the field of integral and differential geometry in statistics are not discussed or even referred to here, but a rather comprehensive overview of the field can be obtained from the material compiled in the present volume and from the survey paper by Barndorff-Nielsen, Cox and Reid (1986).

Section 2 reviews pertinent parts of statistics and of integral and differential geometry, and introduces some of the terminology and notation that will be used in the rest of the paper.

A considerable part of the material in sections 3, 4, 5 and 8 and in the appendices, which are mainly concerned with the systematic theory of transformation models and exponential transformation models, has not been published elsewhere.

Sections 6 and 7 describe a theory of "observed geometries" and its relation to an asymptotic expansion of the formula $c|\hat{j}|^{\frac{1}{2}}\bar{L}$ for the conditional distribution of the maximum likelihood estimator; the results there are mostly taken from Barndorff-Nielsen (1986a). Briefly speaking, the observed geometries on the parameter space of a statistical model consist of a Riemannian metric and an associated one-parameter family of affine connections, constructed from the observed information matrix and from an auxiliary statistic a chosen such that $(\hat{\omega},a)$, where $\hat{\omega}$ denotes the maximum likelihood estimator of the parameter of the model, is minimal sufficient. The observed geometries and the closely related expansion of $c|\hat{j}|^{\frac{1}{2}}\bar{L}$ form a parallel to the "expected geometries" and the associated conditional Edgeworth expansions for curved exponential families studied primarily by Amari (cf., in particular, Amari 1985, 1986), but with some essential differences. In particular, the developments in sections 6 and 7 are, in a sense, closer to the actual data and they do not require integrations over the sample space; instead they employ "mixed derivatives of the log model function." Furthermore, whereas the studies of expected geometries have been largely concerned with curved exponential families the approach taken here makes it equally natural to consider other parametric models, and in particular transformation models. The viewpoint of conditional inference has been instrumental for the constructions in question. However, the observed geometrical calculus, as discussed in section 6, does not require the employment of exact or approximate ancillaries.

The observed geometries provide examples of the concept of statistical manifolds discussed by Lauritzen (1986).

Throughout the paper examples are given to illustrate the general results.

2. REVIEW AND PRELIMINARIES

We shall consider parametrized statistical models $\underline{M}$ specified by $(\underline{X},p(x;\omega),\Omega)$ where $\underline{X}$ is the sample space, Ω is the parameter space and $p(x;\omega)$ is the model function, i.e. $p(x;\omega) = dP_\omega/d\mu$ for some dominating measure μ. The dimension of the parameter ω will usually be denoted by d and we write ω on coordinate form as $(\omega^1,\ldots,\omega^d)$. Generic coordinates of ω will be indicated as ω^r, ω^s, ω^t, etc.

The present section is organized in a number of subsections and it serves two purposes: to provide a survey of previous results and to set the stage for the developments in the following sections.

Combinants. It is useful to have a term for functions which depend on both the observation x and the parameter ω and we shall call any such function a combinant.

Jacobians. Our vectors are row vectors and we denote transposition of a matrix by an asterix *. If f is a differentiable transformation of a space $\underline{Y}$ then the Jacobian $\partial f/\partial y^*$ of f at $y \in \underline{Y}$ is also denoted by $\underline{J}_f(y)$, while we write $J_f(y)$ for the Jacobian determinant, i.e. $J_f = |\underline{J}_f|$. When appropriate we interpret $J_f(y)$ as an absolute value, without explicitly stating this. We shall repeatedly use the fact that for differentiable transformations f and g we have

$$\underline{J}_{f \circ g}(y) = \underline{J}_g(y)\underline{J}_f(g(y)) \tag{2.1}$$

and hence

$$J_{f \circ g}(y) = J_f(g(y))J_g(y). \tag{2.2}$$

Foliations. A partition of a manifold of dimension k into submanifolds all of dimension $m<k$ is called a foliation and the submanifolds are said to be the leaves of the foliation.

A dimension-reducing statistical hypothesis may often, in a natural way, be viewed as a leaf of an associated foliation of the parameter space Ω.

Likelihood. We let $L = L(\omega) = L(\omega;x)$ denote an arbitrary version of the likelihood function for ω and we set $l = \log L$. Furthermore, we write $\partial_r = \partial/\partial\omega^r$, and $l_r = \partial_r l$, $l_{rs} = \partial_r\partial_s l$, etc. The observed information is the matrix

$$j(\omega) = -[l_{rs}] \tag{2.3}$$

and the expected information is

$$i(\omega) = E_\omega j(\omega). \tag{2.4}$$

The inverse matrices of j and i are referred to as observed and expected formation, respectively.

Suppose the minimal sufficient statistic t for $\underline{M}$ is of dimension k. We then speak of $\underline{M}$ as a (k,d)-model (d being the dimension of the parameter ω). Let $(\hat{\omega},a)$ be a one-to-one transformation of t, where $\hat{\omega}$ is the maximum likelihood estimator of ω and a, of dimension k-d, is an auxiliary statistic.

In most applications it will be essential to choose a so as to be distribution constant either exactly or to the relevant asymptotic order. Then a is ancillary and according to the conditionality principle the conditional model for $\hat{\omega}$ given a is considered the appropriate basis for inference on ω.

However, unless explicitly stated, distribution constancy of a is not assumed in the following.

There will be no loss of generality in viewing the log likelihood $l = l(\omega)$ in its dependence on the observation x as being a function of the minimal sufficient $(\hat{\omega},a)$ only. Henceforth we shall think of l in this manner and we will indicate this by writing

$$l = l(\omega,\hat{\omega},a).$$

Similarly, in the case of observed information we write

$$j = j(\omega;\hat{\omega},a)$$

etc. It turns out to be of interest to consider the function

$$\ell(\omega) = \ell(\omega;a) = l(\omega;\omega,a), \tag{2.5}$$

obtained from $l(\omega;\hat{\omega},a)$ by substituting ω for $\hat{\omega}$. Similarly we write

$$\jmath(\omega) = \jmath(\omega;a) = j(\omega;\omega,a). \tag{2.6}$$

For a general parametric model $p(x;\omega)$ and for a general auxiliary a a conditional probability function $p^*(\hat{\omega};\omega|a)$ for $\hat{\omega}$ given a may be defined by

$$p^*(\hat{\omega};\omega|a) = c|\hat{j}|^{\frac{1}{2}}\bar{L} \tag{2.7}$$

where $\bar{L}$ is the normed likelihood function, i.e.

$$\bar{L} = p(x;\omega)/p(x;\hat{\omega}),$$

and where $c = c(\omega,a)$ is a norming constant determined so as to make the integral of (2.7) with respect to $\hat{\omega}$ equal to 1.

Suppose now that a is approximately or exactly distribution constant. Then the probability function $p^*(\hat{\omega};\omega|a)$, given by (2.7), is to be considered as an approximation to the conditional probability function $p(\hat{\omega};\omega|a)$ of the maximum likelihood estimator $\hat{\omega}$ given a, cf. Barndorff-Nielsen (1980, 1983). In general, $p^*(\hat{\omega};\omega|a)$ is simple to calculate since it only requires knowledge of standard likelihood quantities plus an integration over the sample space to determine the norming constant c. Moreover, to sufficient accuracy this norming constant can often be approximated by $(2\pi)^{-d/2}$, where d is the dimension of ω; and a more refined approximation to c solely in terms of mixed derivatives of the log model function is also available, cf. the next subsection and section 7. In a great number of cases, including virtually all transformation models, $p^*(\hat{\omega};\omega|a)$ is, in fact, equal to $p(\hat{\omega};\omega|a)$. Furthermore, outside these exactness cases one often has an asymptotic relation of the form

$$p(\hat{\omega};\omega|a) = p^*(\hat{\omega};\omega|a)\{1 + O(n^{-3/2})\} \tag{2.8}$$

uniformly in $\hat{\omega}$ for $\sqrt{n}(\hat{\omega}-\omega)$ bounded, where n denotes sample size. This holds, in particular, for (k,d) exponential models. For more details and further discussion, see Barndorff-Nielsen (1980, 1983, 1984, 1985, 1986a,b) and Barndorff-Nielsen and Blaesild (1984).

Expansion of $c|\hat{j}|^{\frac{1}{2}}\bar{L}$ in the single-parameter case. Suppose ω is one-dimensional. From formulas (4.2) and (4.5) of Barndorff-Nielsen and Cox (1984) we have

$$c\hat{j}^{\frac{1}{2}}\bar{L} = \phi(\hat{\omega}-\omega;\hat{\jmath})\{1 + C_1\}\{1 + A_1(\hat{\jmath}^{\frac{1}{2}}(\hat{\omega}-\omega)) + A_2(\hat{\jmath}^{\frac{1}{2}}(\hat{\omega}-\omega))\} \cdot\{1 + O(n^{-3/2})\}. \tag{2.9}$$

Here $\phi(w;\gamma)$ denotes the probability density function of the normal distribution with mean 0 and variance γ^{-1}. Furthermore, C_1, A_1, and A_2 are given by

$$C_1 = \frac{1}{24}\{-3U_4 + 12U_{3,1} - 5U_3^2 + 24U_{2,1}U_3 - 24U_{2,1}^2 - 12U_{2,2}\} \tag{2.10}$$

and

$$A_1(u) = P_1(u)U_{2,1} + P_2(u)U_3$$

$$A_2(u) = P_3(u)U_{2,2} + P_4(u)U_{2,1}^2 + P_5(u)U_4 + P_6(u)U_{3,1} + P_7(u)U_3^2 + P_8(u)U_{2,1}U_3$$

where $P_i(u)$, $i = 1,\dots,8$, are polynomials, the explicit forms of which are given in Barndorff-Nielsen (1985), and where $U_\nu = U_{\nu,0}$ and $U_{\nu,s}$ are defined as

$$U_{\nu,s}(\omega) = \frac{\partial^s\{l^{(\nu)}(\omega;\omega,a)\}}{\hat{\jmath}^{(\nu+s)/2}}, \quad \begin{array}{l}\nu = 1,2,3,\dots \\ s = 0,1,2\dots\end{array},$$

$l^{(\nu)}$ denoting the ν-th order derivative of $l = l(\omega;\hat{\omega},a)$ with respect to ω and ∂^s indicating differentiation s times with respect to ω. Note that, in the repeated sampling situation, $U_{\nu,s}$ is of order $O(n^{-(\nu+s-2)/2})$. Hence the quantities C_1, A_1 and A_2 are of order $O(n^{-1})$, $O(n^{-\frac{1}{2}})$ and $O(n^{-1})$, respectively.

Integration of (2.7) yields an approximation to the conditional distribution of the likelihood ratio statistic

$$w = 2\{l(\hat{\omega}) - l(\hat{\omega}_0)\} \tag{2.11}$$

for testing a dimension reducing hypothesis Ω_0 of Ω. In particular, if Ω_0 is a point hypothesis, $\Omega_0 = \{\omega_0\}$, we have

$$p^*(w;\omega_0|a) = ce^{-\frac{1}{2}w} \int_{\hat{\omega}|w,a} |\hat{j}|^{\frac{1}{2}} d\hat{\omega} \tag{2.12}$$

as an approximation to $p(w;\omega_0|a)$. (The leading term of (2.9) together with (2.12) yields the usual χ^2 approximation for w. For a connection to Bartlett adjustment factors see Barndorff-Nielsen and Cox (1984)).

Furthermore, (2.9) may be integrated termwise to obtain expansions for the conditional distribution function for $\hat{\omega}$ and, by inversion, for confidence limits for ω, correct to order $O(n^{-3/2})$, conditionally as well as unconditionally, cf. Barndorff-Nielsen (1985). The resulting expressions allow one to carry out "conditional inference without conditioning and without integration."

For extensions to the case of multidimensional parameters see section 7.

<u>Reparametrization</u>. A basic form of invariance is <u>parametrization invariance</u> of statistical procedures (though parametrization equivariance might be a more proper term). If we think of an <u>inference frame</u> as consisting of the data in conjunction with the model and a particular parametrization of the model, and of a <u>statistical procedure</u> π as a method which leads from the inference frame to a conclusion formulated in terms of the parametrization of the inference frame then parametrization invariance may be formally specified as commutativity of the diagram

$$\begin{tikzcd}
\text{inference frame} \arrow[r, "\text{reparametrization}"] \arrow[d, "\text{procedure } \pi"'] & \text{inference frame} \arrow[d, "\text{procedure } \pi"] \\
\text{conclusion} \arrow[r, "\text{reparametrization}"'] & \text{conclusion}
\end{tikzcd}$$

In words, the procedure π is parametrization invariant if changing the inference base by shifting to another parametrization and then applying π yields the same conclusion as first applying π and then translating the conclusion so as to be expressed in terms of the new parametrization. (We might describe a parametrization invariant procedure as a 0-th order generalized tensor.) Maximum likelihood estimation and likelihood ratio testing are instances of parametrization invariant procedures.

Example 2.1. Consider any log-likelihood function $l(\omega)$, of a one-dimensional parameter ω. Define the functions $r^{[\nu]} = r^{[\nu]}(\omega)$, $\nu = 1,2,\ldots$, recursively by

$$r^{[1]}(\omega) = l^{(1)}(\omega)/i(\omega)^{\frac{1}{2}}$$

$$r^{[\nu]} = \frac{dr^{[\nu-1]}(\omega)}{d\omega}/i(\omega)^{\frac{1}{2}}, \qquad \nu = 2,3,\ldots,$$

and set $\hat{r}^{[\nu]} = r^{[\nu]}(\hat{\omega})$. The derivatives $\hat{r}^{[\nu]}$ are parametrization invariant, i.e. $\hat{r}^{[\nu]}$ takes the same value whatever the parametrization employed.

While parametrization invariance is clearly a desirable property, there are a number of useful, and virtually indispensable, statistical methods which do not have this property. Thus procecures which rely on the asymptotic normality of the maximum likelihood estimator, such as the Wald test or standard ways of setting confidence intervals in non-linear regression problems, are mostly not parametrization invariant. However, in cases of non parametrization invariance particular caution must be exercised, as demonstrated for instance for the Wald test by Hauck and Donner (1977) and Vaeth (1985).

We shall be interested in how various quantities behave under reparametrizations of the model $\underline{M}$. Let ψ, of dimension d, be the parameter of some parametrization of $\underline{M}$, alternative to that indicated by ω. Coordinates of ψ will be denoted by ψ^ρ, ψ^σ, etc. and we write ∂_ρ for $\partial/\partial\psi^\rho$ and

$$\omega^r_{/\rho} = \partial\omega^r/\partial\psi^\rho, \qquad \omega^r_{/\rho\sigma} = \partial^2\omega^r/\partial\psi^\rho\partial\psi^\sigma,$$

etc. Furthermore, we write $l(\psi)$ for the log likelihood under the parametrization by ψ, though formally this is in conflict with the notation $l(\omega)$, and correspondingly we let $l_\rho = \partial_\rho l = \partial_\rho l(\psi)$, etc.; similarly for other parameter dependent quantities. Finally, the symbol ^ over such a quantity indicates that the maximum likelihood estimate has been substituted for the parameter.

Using this notation and adopting the summation convention that if a suffix occurs repeatedly in a single expression then summation over that suffix is understood, we have

$$l_\rho = l_r \omega^r_{/\rho}$$

$$l_{\rho\sigma} = l_{rs}\omega^r_{/\rho}\omega^s_{/\sigma} + l_r\omega^r_{/\rho\sigma} \tag{2.13}$$

$$l_{\rho\sigma\tau} = l_{rst}\omega^r_{/\rho}\omega^s_{/\sigma}\omega^t_{/\tau} + l_{rs}\omega^r_{/\rho\sigma}\omega^s_{/\tau}[3] + l_r\omega^r_{/\rho\sigma\tau} \tag{2.14}$$

etc., where [3] signifies a sum of three similar terms determined by permutation of the indices ρ,σ,τ. On substituting $\hat{\omega}$ for ω in (2.13) we obtain the well-known relation

$$\hat{j}_{\rho\sigma} = \hat{j}_{rs}\hat{\omega}^r_{/\rho}\hat{\omega}^s_{/\sigma}$$

which, now by substitution of ω for $\hat{\omega}$, may be reexpressed as

$$\mathfrak{j}_{\rho\sigma} = \mathfrak{j}_{rs}\omega^r_{/\rho}\omega^s_{/\sigma} \tag{2.15}$$

or, written more explicitly,

$$\mathfrak{j}_{\rho\sigma}(\psi;a) = \mathfrak{j}_{rs}(\omega;a)\,\frac{\partial\omega^r}{\partial\psi^\rho}\,\frac{\partial\omega^s}{\partial\psi^\sigma}\,.$$

Equation (2.15) shows that $\mathfrak{j}$ is a metric tensor on $\underline{M}$, for any given value of the auxiliary statistic a. Moreover, in wide generality $\mathfrak{j}$ will be positive definite on $\underline{M}$, and we assume henceforth that this is the case. In fact, for any $\hat{\omega}\varepsilon\Omega$ we have $\hat{\mathfrak{j}} = \hat{j}$, i.e. observed information at the maximum likelihood point, which is generally positive definite (though counterexamples do exist).

Let $A(\omega) = [A^{r_1 \ldots r_p}_{s_1 \ldots s_q}(\omega)]$ be an array, depending on ω and where each of the p + q indices runs from 1 to d. Then A is said to be a <u>(p,q) tensor</u>, or a tensor of <u>contravariant rank p</u> and <u>covariant rank q</u>, if under reparametrization from ω to ψ A obeys the transformation law

$$A^{\rho_1\ldots\rho_p}_{\sigma_1\ldots\sigma_q}(\psi) = \omega^{s_1}_{/\sigma_1}\cdots\omega^{s_q}_{/\sigma_q}\,\psi^{\rho_1}_{/r_1}\cdots\psi^{\rho_p}_{/r_p}\,A^{r_1\ldots r_p}_{s_1\ldots s_q}(\omega).$$

Example 2.2. A covariant tensor of rank q is given by

$$E_\omega\left\{\frac{\partial l}{\partial\omega^{r_1}}\cdots\frac{\partial l}{\partial\omega^{r_q}}\right\}.$$

In particular, the expected information i is a (0,2) tensor.

The inverse $[i^{rs}]$ of $i = [i_{rs}]$ is a contravariant second order tensor.

The (outer) product of two tensors $A^{r_1 r_2\ldots}_{s_1 s_2\ldots}$ and $B^{t_1 t_2\ldots}_{u_1 u_2\ldots}$ is defined as the array C given by

$$C^{r_1 r_2\ldots t_1 t_2\ldots}_{s_1 s_2\ldots u_1 u_2\ldots} = A^{r_1 r_2\ldots}_{s_1 s_2\ldots}\,B^{t_1 t_2\ldots}_{u_1 u_2\ldots}.$$

This product is again a tensor, of rank (p' + p", q' + q") if (p',q') and (p",q") are the ranks of A and B.

Lower rank tensors may be derived from higher rank tensors by contraction, i.e. by pairwise identification of upper and lower indices (which implies a summation).

The parameter space as a manifold. The parameter space Ω may be viewed as a (pseudo-) Riemannian manifold with (pseudo-) metric determined by a metric tensor ϕ, i.e. ϕ is a rank 2 covariant, regular and symmetric tensor. The associated Riemannian connection $\overset{0}{\nabla}$ is determined by the Christoffel symbols $\overset{0}{\Gamma}{}^t_{rs}$ where

$$\overset{0}{\Gamma}{}^t_{rs} = \phi^{tu}\,\overset{0}{\Gamma}_{rsu}$$

and

$$\overset{0}{\Gamma}_{rst} = \tfrac{1}{2}(\partial_r\phi_{st} - \partial_t\phi_{rs} + \partial_s\phi_{rt}). \tag{2.16}$$

If ∇ is any affine connection with connection symbols Γ^t_{rs} then these symbols satisfy

$$\nabla_{\partial_r}\partial_s = \Gamma^t_{rs}\partial_t \tag{2.17}$$

and the transformation law

$$\Gamma^\tau_{\rho\sigma}(\psi) = [\Gamma^t_{rs}(\omega)\omega^r_{/\rho}\omega^s_{/\sigma} + \omega^t_{/\rho\sigma}]\psi^\tau_{/t} . \tag{2.18}$$

On the other hand, any set of functions $[\Gamma^t_{rs}]$ which satisfy the law (2.18) constitute the connection symbols of an affine connection on Ω. It follows that all affine connections on Ω are of the form

$$\Gamma^t_{rs} = \overset{o}{\Gamma}{}^t_{rs} + S^t_{rs} \tag{2.19}$$

where the S^t_{rs} are characterized by the transformation law

$$S^\tau_{\rho\sigma}(\psi) = S^t_{rs}(\omega)\omega^r_{/\rho}\omega^s_{/\sigma}\psi^\tau_{/t} . \tag{2.20}$$

If, for a given metric tensor ϕ, we define Γ_{rst} and S_{rst} by

$$\Gamma_{rst} = \Gamma^u_{rs}\phi_{tu} \text{ and } S_{rst} = S^u_{rs}\phi_{tu}$$

then (2.18), (2.19) and (2.20) are equivalent to, respectively,

$$\Gamma_{\rho\sigma\tau}(\psi) = \Gamma_{rst}(\omega)\omega^r_{/\rho}\omega^s_{/\sigma}\omega^t_{/\tau} + \phi_{tu}(\omega)\omega^t_{/\rho\sigma}\omega^u_{/\tau} \tag{2.21}$$

$$\Gamma_{rst} = \overset{o}{\Gamma}_{rst} + S_{rst} \tag{2.22}$$

and

$$S_{\rho\sigma\tau} = S_{rst}\omega^r_{/\rho}\omega^s_{/\sigma}\omega^t_{/\tau}. \tag{2.23}$$

Thus, in particular, $[S_{rst}]$ is a tensor.

Suppose $\psi:\beta \to \omega$ is a mapping of full rank from an open subset B of a Euclidean space of dimension $d_0 < d$ into Ω. Then ψ is said to be an immersion of B in Ω. We denote coordinates of β by β^a, β^b, etc. If ϕ is a metric tensor on Ω then the metric tensor on B induced from Ω by ψ is defined by

$$\phi_{ab}(\beta) = \phi_{rs}(\omega)\omega^r_{/a}\omega^s_{/b} . \tag{2.24}$$

If $\Gamma^t_{rs}(\omega)$ is a connection on Ω and if $\Gamma_{rst} = \Gamma^u_{rs}\phi_{tu}$ then the <u>induced connection</u>

on B is defined by $\Gamma^{c}_{ab}(\beta) = \Gamma_{abd}(\beta)\phi^{cd}(\beta)$ and by

$$\Gamma_{abc}(\beta) = \Gamma_{rst}(\omega)\omega^{r}_{/a}\omega^{s}_{/b}\omega^{t}_{/c} + \phi_{tu}\omega^{t}_{/ab}\omega^{u}_{/c} . \tag{2.25}$$

Let G be a group acting smoothly on the parameter space. A metric tensor ϕ is said to be (G-) invariant if

$$\phi_{rs}(\omega) = \frac{\partial(g\omega)^{r'}}{\partial\omega^{r}} \phi_{r's'}(g\omega) \frac{\partial(g\omega)^{s'}}{\partial\omega^{s}}, \quad g\varepsilon G. \tag{2.26}$$

For a given g let a new parametrization be introduced by $\psi = g\omega$. From the transformation law for tensors it follows that ϕ is invariant if and only if

$$\phi_{rs}(\psi) = \phi_{rs}(g\omega), \quad g\varepsilon G. \tag{2.27}$$

(On the left hand side the tensor is expressed in ψ coordinates, on the right hand side in ω coordinates.) Similarly, a connection Γ is said to be invariant if

$$\Gamma^{t}_{rs}(\psi) = \Gamma^{t}_{rs}(g\omega), \quad g\varepsilon G. \tag{2.28}$$

The pseudo-Riemannian connection derived from an invariant metric tensor is invariant.

In generalization of (2.27) an arbitrary covariant tensor $A_{r_1 \ldots r_q}$ is said to be (G-) invariant if

$$A_{r_1 \ldots r_q}(\psi) = A_{r_1 \ldots r_q}(g\omega), \quad g\varepsilon G.$$

If Γ^{t}_{rs} is a G-invariant connection and if ϕ_{rs} and S_{rst} are G-invariant tensors, with ϕ_{rs} being a metric tensor, then $\tilde{\Gamma}^{t}_{rs}$ defined by

$$\tilde{\Gamma}^{t}_{rs} = \Gamma^{t}_{rs} + \phi^{tu}S_{rsu}$$

is a G-invariant connection.

Now, let ϕ be the information tensor i on Ω. Then (2.16) takes the form

$$\overset{0}{\Gamma}_{rst} = E\{l_{rs}l_{t}\} + \tfrac{1}{2}E\{l_{r}l_{s}l_{t}\}.$$

Obviously,

$$T_{rst} = E\{l_{r}l_{s}l_{t}\} \tag{2.29}$$

satisfies (2.23) and hence, for any real α an affine connection is defined by

$$\overset{\alpha}{\Gamma}_{rst} = E\{l_{rs}l_t\} + \frac{1-\alpha}{2} E\{l_r l_s l_t\}. \tag{2.30}$$

These are the <u>α-connections</u> introduced and studied by Chentsov (1972) and Amari (1982a,b, 1985, 1986).

However, we shall be mainly concerned with another type of connection, determined from observed information, more specifically from the metric tensor $\mathfrak{j}$, see sections 6-8. We refer to i and $\mathfrak{j}$ as expected and observed information metric on $\underline{M}$, respectively.

Suppose, as above, that $\psi:\beta \rightarrow \omega$ is an immersion of B in Ω. The submodel $\underline{M}_0$ of $\underline{M}$ obtained by restricting ω to lie in $\Omega_0 = \psi(B)$ has expected information

$$i(\beta) = \frac{\partial\omega}{\partial\beta^*} i(\omega) \frac{\partial\omega^*}{\partial\beta} . \tag{2.31}$$

Thus $i(\beta)$ equals the Riemannian metric induced from the metric $i(\omega)$ on Ω to the imbedded submanifold Ω_0. Furthermore, the α-connection of the model $\underline{M}_0$ equals the connection on Ω_0 induced from the α-connection on Ω, by the general construction (2.25).

The measures on Ω defined by

$$|i|^{\frac{1}{2}}d\omega \tag{2.32}$$

and

$$|\mathfrak{j}|^{\frac{1}{2}}d\omega \tag{2.33}$$

are both geometric measures, relative to expected and observed information metric, respectively. Note that (2.33) depends on the value of the auxiliary statistic a. We shall speak of (2.32) and (2.33) as expected and observed information measure, respectively. It is an important property of these measures that they are parametrization invariant. This property follows from the fact that i and $\mathfrak{j}$ are covariant tensors of rank 2. As a consequence we have that $c|\hat{j}|^{\frac{1}{2}}\bar{L}$ (of (2.7)) is parametrization invariant.

<u>Invariant measures</u>. A measure μ on $\underline{X}$ is said to be <u>invariant</u> with respect to a group G acting on $\underline{X}$ if $g\mu = \mu$ for all $g\varepsilon G$.

Invariant measures, when they exist, may often be constructed from a quasi-invariant measure, as follows.

A measure μ on $\underline{X}$ is called quasi-invariant with multiplier $\chi = \chi(g,x)$ if $g\mu$ and μ are mutually absolutely continuous for every $g \varepsilon G$ and if

$$d(g^{-1}\mu)(x) = \chi(g,x)d\mu(x).$$

Furthermore, define a function m on $\underline{X}$ to be a modulator with associated multiplier $\chi(g,x)$ if m is positive and

$$m(gx) = \chi(g,x)m(x).$$

Then, if μ^X is quasi-invariant with multiplier $\chi(g,x)$ and if m is a modulator with the same multiplier we have that

$$\mu = m^{-1}\mu^X$$

is an invariant measure on $\underline{X}$.

As quasi-invariance is clearly a very weak property the problem in constructing invariant measures lies mainly in finding appropriate modulators. It is usually possible to specify the modulators in terms of Jacobians.

In particular, in applications it is often the case that $\underline{X}$ is an open subset of a Euclidean space. By the standard theorem on transformation of integrals, Lebesgue measure λ on $\underline{X}$ is then quasi-invariant with multiplier $J_{\gamma(g)}(x)$. Under mild conditions an invariant measure on $\underline{X}$ is then given by

$$d\mu(x) = J_{\gamma(z)}(u)^{-1}d\lambda(x). \tag{2.34}$$

Here $J_{\gamma(g)}$ denotes the Jacobian determinant of the mapping $\gamma(g)$ of $\underline{X}$ onto itself determined by $g \varepsilon G$ and (z,u) constitutes an orbital decomposition of x, i.e. (z,u) is a one-to-one transformation of x such that $u \varepsilon \underline{X}$ and u is maximal invariant while $z \varepsilon G$ and $x=zu$. For a more detailed discussion see section 3 and appendix 1.

Transformation models. Let G be a group acting on the sample space $\underline{X}$. If the class $\underline{P}$ of probability measures given by the statistical model is invariant under the induced action of G on the set of all probability measures on $\underline{X}$ then the model is called a composite transformation model and if $\underline{P}$

consists of a single orbit we use the term transformation model. For a composite transformation model, G acts on $\underline{P}$ and we may, of course, equally think of G as acting on the parameter space Ω. A parameter (function) λ which is maximal invariant under this action is said to be an index parameter. Virtually all composite transformation models of interest have the property that after minimal sufficient reduction (and possibly after deletion of a null set from $\underline{X}$) there exists a sub-group K of G such that K is the isotropy group for a point on every one of the orbits of $\underline{X}$ and of Ω. Each of these orbits is then isomorphic to the homogeneous space $G/K = \{gK:g\varepsilon G\}$ of left cosets of K.

For a transformation model the information measures (2.32) and (2.33) are invariant measures relative to the action of G on Ω induced from the action of G on $\underline{X}$ via the maximum likelihood estimator $\hat{\omega}$, which is an equivariant mapping from $\underline{X}$ to Ω. This action is the same as the above-mentioned action of G on $\underline{P} \equiv \Omega$ and also the same as the natural action of G on $G/K \equiv \Omega$.

It follows that relative to information measure on Ω the formula (2.7) for the conditional distribution of $\hat{\omega}$ is simply $c\bar{L}$. From this it may be shown that, with the auxiliary a as the maximal invariant statistic, $p^*(\hat{\omega},\omega|a)$ is exactly equal to $p(\hat{\omega};\omega|a)$.

These results are shown in outline in Barndorff-Nielsen (1983). A more general statement will be derived in section 5.

Exponential models. A (k,d) exponential model has model function of the form

$$p(x;\omega) = \exp\{\theta(\omega)\cdot t(x) - \kappa(\theta(\omega)) - h(x)\}. \tag{2.35}$$

Here k is the order of the model (2.35) and is equal to the common dimension of the vectors $\theta(\omega)$ and $t(x)$, while d denotes the dimension of the parameter ω. The full exponential model generated by (2.35) has model function

$$p(x;\theta) = \exp\{\theta\cdot t(x) - \kappa(\theta) - h(x)\} \tag{2.36}$$

and $\kappa(\theta)$ is the cumulant transform of the canonical statistic $t = t(x)$. From the viewpoint of inference on ω there is no restriction in assuming $x = t$, since t is minimal sufficient, and we shall often do so. We set $\tau = \tau(\theta) = E_\theta t$, i.e. τ is the mean value parameter of (2.36), and we write T for $\tau(\mathrm{int}\Theta)$ where Θ denotes the canonical parameter domain of the full model (2.36).

Let f be a real differentiable function defined on an open subset of R^k. The <u>Legendre transform</u> $f^\dagger$ of f is defined by

$$f^\dagger(y) = x \cdot y - f(x)$$

where

$$y = (Df)(x) = \frac{\partial f}{\partial x}(x) .$$

The Legendre transform is a useful tool in studying various, dualistic aspects of exponential models (cf. Barndorff-Nielsen (1978a), Barndorff-Nielsen and Blaesild (1983a)).

In particular, we may use the Legendre transform to define the dual likelihood function $\overset{-1}{l}$ of (2.35) by

$$\overset{-1}{l}(\omega) = \hat{\theta} \cdot \tau(\omega) - \hat{l}(\tau(\omega)). \tag{2.37}$$

Here, and elsewhere, ˆ as top index indicates maximum likelihood estimation under the full model. Further, in this connection we take $\hat{l}$ as the sup-log-likelihood function of (2.36) and then $\hat{l}$ is, in fact, the Legendre transform of κ. Note that for $\tau = \tau(\theta) \in T$ we have $\hat{l}(\tau) = \theta \cdot \tau - \kappa(\theta)$. An inference methodology, parallel to that of likelihood inference for exponential families, may be developed from the dual likelihood (2.37). The estimates, tests and confidence regions discussed by Amari and others under the name of $\alpha = -1$ (or mixture) procedures are, essentially, part of the dual likelihood methodology.

More generally, based on Amari's concepts of α-geometry and α-divergence, one may for each $\alpha \in [-1,1]$ introduce an "α-likelihood" $\overset{\alpha}{L}$ by

$$\overset{\alpha}{L}(\omega) = \overset{\alpha}{L}(\omega;t) = \exp\{-D_\alpha(\hat{\theta},\theta(\omega))\} \tag{2.38}$$

where

$$D_\alpha(\theta,\tilde{\theta}) = E_\theta f_\alpha\left(\frac{p(x;\tilde{\theta})}{p(x;\theta)}\right). \tag{2.39}$$

Here $p(x;\theta)$ is given by (2.36) and the function f_α is defined as

$$f_\alpha(x) = \begin{cases} x \log x, & \alpha = 1 \\ \frac{4}{1-\alpha^2}\{1-x^{(1+\alpha)/2}\}, & -1<\alpha<1 \\ -\log x, & \alpha = -1 \end{cases} \quad . \tag{2.40}$$

Letting $\overset{\alpha}{l} = \log \overset{\alpha}{L}$ we have, in particular,

$$\overset{1}{l}(\theta) = l(\theta) = -I(\hat{\theta},\theta) = \theta\cdot t - \kappa(\theta) - \hat{l}(t) \tag{2.41}$$

and

$$\overset{-1}{l}(\theta) = -I(\theta,\hat{\theta}) = \hat{\theta}\cdot\tau - \hat{l}(\tau) - \kappa(\hat{\theta}) \tag{2.42}$$

where I denotes the discrimination information. Furthermore, for $-1<\alpha<1$,

$$\overset{\alpha}{l}(\theta) = \frac{4}{1-\alpha^2}[e^{-\{\frac{1+\alpha}{2}\kappa(\theta) + \frac{1-\alpha}{2}\kappa(\hat{\theta}) - \kappa(\frac{1+\alpha}{2}\theta + \frac{1-\alpha}{2}\hat{\theta})\}} - 1].$$

Affine subsets of Θ are simple from the likelihood viewpoint while, correspondingly, affine subsets of T are simple in dual likelihood theory. Dual affine foliations, of Θ and T respectively, are therefore of some particular interest. Such foliations have been studied in Barndorff-Nielsen and Blaesild (1983a), see also Barndorff-Nielsen and Blaesild (1983b).

Suppose that the auxiliary component a of $(\hat{\omega},a)$ is approximately or exactly distribution constant, i.e. a is ancillary. For instance, a may be the affine ancillary or the directed log likelihood ratio statistic, as defined in Barndorff-Nielsen (1980, 1986b). We may think of the partitions generated, respectively, by a and $\hat{\omega}$ as foliations of T, to be called the <u>ancillary foliation</u> and the <u>maximum likelihood foliation</u>. (Amari's ancillary subspaces are then, in the present terminology and for $\alpha = 1$, leaves of the maximum likelihood foliation.)

<u>Exponential transformation models</u>. A model $\underline{M}$ which is both transformational and exponential is called an <u>exponential transformation model</u>. For such models we have the following structure theorem (Barndorff-Nielsen, Blaesild, Jensen and Jorgensen (1982), Eriksen (1984b)).

<u>Theorem 2.1</u>. Let $\underline{M}$ be an exponential transformation model with

acting group G. Suppose $\underline{X}$ is locally compact and that t is continuous. Furthermore, suppose that G is locally compact and acts continuously on $\underline{X}$.

Then there exists, uniquely, a k-dimensional representation A(g) of G and k-dimensional vectors B(g) and $\tilde{B}(g)$ such that

$$t(gx) = t(x)A(g) + B(g) \tag{2.43}$$

$$\theta(g) = \theta(e)A(g^{-1})^* + \tilde{B}(g) \tag{2.44}$$

where $e \varepsilon G$ denotes the identity element. Furthermore, the full exponential model generated by $\underline{M}$ is invariant under G, and $\tilde{G}^* = \{[A(g^{-1})^*, \tilde{B}(g)] : g \varepsilon G\}$ is a group of affine transformations of R^k leaving Θ and $\mathrm{int}\Theta$ invariant in such a way that

$$\theta(gP) = \theta(P)A(g^{-1})^* + \tilde{B}(g), \qquad g \varepsilon G,\ P \varepsilon \underline{P}\ .$$

Dually, $\tilde{G} = \{[A(g), B(g)] : g \varepsilon G\}$ is a group of affine transformations leaving $C = \mathrm{cl\ conv}\ t(\ \underline{X}\ )$ as well as $T = \tau(\mathrm{int}\Theta)$ invariant. Finally, let δ be the function given by

$$\delta(g) = a(\theta(e))a(\theta(g))^{-1}\exp(-\theta(g) \cdot B(g)). \tag{2.45}$$

We then have

$$a(\theta(gP)) = a(\theta(P))\delta(g)^{-1}\exp(-\theta(gP) \cdot B(g)). \tag{2.46}$$

Exponential transformation models that are full are a rarity. However, important examples of such models are provided by the family of Wishart distributions and the transformational submodels of this.

In general, then, an exponential transformation model $\underline{M}$ is a curved exponential model. It is seen from the above theorem that the full model $\underline{\acute{M}}$ generated by $\underline{M}$ is a composite transformation model and that, correspondingly, $\underline{\acute{M}}$ (and, hence Θ and T) is a foliated manifold with $\underline{M}$ as a leaf. It seems of interest to study how the leaves of this foliation are related geometric-statistically. Exponential transformation models of type (k,d), and in particular those of type (2,1), have been studied in some detail by Eriksen (1984a,c). In the first of these papers the Jordan normal form of a matrix is an important tool.

Many of the classical differentiable manifolds with their associated acting Lie groups are carriers of interesting exponential transformation models. Instances of this are compiled in table 2.1.

Analogies between exponential models and transformation models. There are some intriguing analogies between exponential models and transformation models.

Example 2.3. Under a d-dimensional location parameter model, with ω as the location parameter and for a fixed value of the (ancillary) configuration statistic, the possible score functions are horizontal translates of each other.

On the other hand, under a (k,d) exponential model, with ω as a component of the canonical parameter and provided the complementary part of the canonical statistic is a cut, the possible score functions are vertical translates of each other. (For details, see Barndorff-Nielsen (1982)).

Example 2.4. Suppose ω is one-dimensional. If ω is the location parameter of a location model then the correction term C_1, given by (2.10), takes the simple form

$$\hat{C}_1 = -\frac{1}{24}\{3\frac{\hat{l}^{(4)}}{\hat{j}^2} + 5\frac{\hat{l}^{(3)^2}}{\hat{j}^3}\} .$$

Exactly the same expression is obtained for a (1,1) exponential model with ω as the canonical parameter.

(This was noted in Barndorff-Nielsen and Cox (1984)).

Maximum estimation. Suppose that for a certain class of models we have an estimation procedure according to which the estimate $\tilde{\omega}$ of ω is obtained by maximizing a positive function $M = M(\omega) = M(\omega;x)$ with respect to ω. Let $m = \log M$ and suppose that

$$\tilde{k} = -[\partial_r\partial_s m](\tilde{\omega}) \tag{2.47}$$

is positive definite. We shall then say that we have a maximum estimation procedure. Maximum likelihood estimation and dual maximum likelihood estimation (where $m(\omega) = \overset{-1}{l}(\omega) = \acute{\theta}\cdot\tau(\omega) - \acute{l}(\omega)$, cf. (2.37)) are examples of this. More generally, minimum contrast estimation, as discussed by Eguchi (1983), is of this type.

Suppose that M depends on x through the minimal sufficient statistic only and let a be an auxiliary statistic such that $(\tilde{\omega},a)$ is minimal sufficient. In generalization of (2.7) we may consider

$$p^*(\tilde{\omega};\omega|a) = c|\tilde{k}|^{\frac{1}{2}}L/\tilde{L}, \tag{2.48}$$

as a possible approximation to $p(\tilde{\omega};\omega|a)$. Here $\tilde{L} = L(\tilde{\omega})$ and c is a norming constant, determined so as to make the integral of the right hand side of (2.48) with respect to $\tilde{\omega}$ equal to 1.

It will be shown in section 5 that (2.48) is exactly equal to $p(\tilde{\omega};\omega|a)$ for a considerable range of cases.

Finally, it may be noted that by an argument of analogy it would seem rather natural to consider the modification of (2.48) in which the function M is substituted for the likelihood function L. While this approach is not without interest its general asymptotic degree of accuracy is only $O(n^{-\frac{1}{2}})$ in comparison with $O(n^{-1})$ or $O(n^{-3/2})$ for (2.48). Also, for transformation models this modification is exact in exceptional cases only.

manifold		acting group		parameter space (H)		model		
name	symbol	name	symbol	name	symbol	name	exponent	normalising constant involves:
symmetric positive definite $r \times r$ matrices	S_+^r	general linear	$GL(r)$	upper triangular matrices w.pos.diag.	$T_+(r)$	Wishart	$tr\{h^*hx\}$	Γ
unit sphere in R^k	S^{k-1}	special orthogonal	$SO(k)$	unit sphere in R^k	S^{k-1}	von Mises - Fisher	$\lambda h \cdot x$	I
-	-	-	-	projective space	RP^{k-1}	Dimroth - Watson	$-\lambda(h \cdot x)^2$	$\int_0^1 e^{-\lambda x^2} dx$
Stiefel	$V_r(m)$	orthogonal	$O(m)$	Stiefel	$V_r(m)$	matrix von Mises - Fisher	$tr\{hx\}$	${}_0F_1$
unit hyperboloid in R^k	H^{k-1}	special pseudo-orthogonal	$SO^{\uparrow}(1;k-1)$	boosts	$B(k) \equiv H^{k-1}$	Hyperboloid	$-\lambda h \cdot x$	K
Grassmann	$G(r,m)$	special orthogonal	$SO(r)$	symm. $r \times r$ matrices w. trace 0		Bingham	$tr\{hx\}$	${}_1F_1$

Table 2.1. Survey of exponential transformation models on classical manifolds.

3. TRANSFORMATION MODELS

Transformation models were introduced in section 2. For any $x \in \underline{X}$ the set $Gx = \{gx : g \in G\}$ of points traversed by x under the action of G is termed the orbit of x. The sample space $\underline{X}$ is thus partitioned into disjoint orbits, and if on each orbit we select a point u, to be called the orbit representative, then any point x in $\underline{X}$ can be determined by specifying the representative u of Gx and an element $z \in G$ such that $x = zu$. In this way x has, as it were, been expressed in new coordinates (z,u) and we speak of (z,u) as an orbital decomposition of x.

The orbit representative, or any one-to-one transformation thereof, is a maximal invariant - and hence ancillary - statistic, and inference under the model proceeds by first conditioning on that statistic.

The action of G on a space $\underline{X}$ is said to be transitive if $\underline{X}$ consists of a single orbit and free if for any pair g and h of different elements of G we have $gx \neq hx$ for every $x \in \underline{X}$. Note that after conditioning on a maximal invariant statistic u we have a transitive action of G on the conditional sample space. For any $x \in \underline{X}$ the set $G_x = \{g : gx = x\}$ is a subgroup, called the isotropy group of x. The space $\underline{X}$ is said to be of constant orbit type if it is possible to select the orbit representatives u so that G_u is the same for all u.

The situation is particularly transparent if the action of G on the sample space $\underline{X}$ is free. Then for given x and u there is only one choice of $z \in G$ such that $x = zu$, and $\underline{X}$ is thus representable as a product space of the form $U \times G$ where U is the subset of $\underline{X}$ consisting of the orbit representatives u. Note that u and z as functions of x are, respectively, invariant and equivariant i.e.

$$u(gx) = u(x), \qquad z(gx) = gz(x).$$

It is often feasible to construct an orbital decomposition by first finding an equivariant mapping z from $\underline{X}$ onto G and then defining the orbit representative u for x by

$$u = z^{-1}x.$$

In particular, the maximum likelihood estimate $\hat{g}$ of g is equivariant, and may be used as z provided $\hat{g}(x)$ exists uniquely for every $x \varepsilon \underline{X}$ and $\hat{g}(\underline{X}) = G$. In this case, G's action on $\underline{P}$ must also be free.

However, we shall need to treat more general cases where the actions of G on $\underline{X}$ and on $\underline{P}$ are not necessarily free.

Let H and K be subsets of G. We say that these constitute a factorization of G if G is uniquely factorizable as

$$G = HK$$

in the sense that to each element $g \varepsilon G$ there exists a unique pair $(h,k) \varepsilon H \times K$ such that $g = hk$. We speak of a left factorization if, in addition, K is a subgroup of G, and similarly for right factorization. If a factorization is both left and right then G is said to be the product of the groups H and K. An important example of such a product is afforded by the well-known unique factorization of a regular $n \times n$ matrix A into a product UT of an orthogonal matrix U and a lower triangular matrix with positive diagonal elements, i.e., using standard notations for matrix groups, GL(n) is the product of O(n) and $T_+(n)$.

A relevant left factorization is often generated in the following way. Let P be a member of the family $\underline{P}$ of probability measures for a transformation model $\underline{M}$, and let K be the isotropy group G_P, i.e.

$$K = \{g \varepsilon G : gP = P\}.$$

For each $\acute{P} \varepsilon \underline{P}$ we may select an element h of G such that $\acute{P} = hP$, and letting H be the set consisting of these elements we have a (left) factorization $G = HK$. (In a more technical wording, the elements h are representatives of the left cosets of K.) Note that $G_{\acute{P}} = hG_P h^{-1}$, and that the action of G on $\underline{P}$ is free if and only if K consists of the identity element alone. The quantity h parametrizes $\underline{P}$.

Suppose G = HK is a factorization of this kind. For most transformation models of interest, if the action of G on $\underline{X}$ is not free then there exists an orbital decomposition (z,u) of x with $z \in H$ and such that for every u the isotropy group G_u equals K and, furthermore, if z and z' are different elements of H then $zu \neq z'u$.

Example 3.1. Hyperboloid model. This model (Barndorff-Nielsen (1978b), Jensen (1981)) is analogous to the von Mises-Fisher model but pertains to observations x on the unit hyperboloid H^{k-1} of R^k, i.e.

$$H^{k-1} = \{x : x*x = 1,\ x_0 > 0\}$$

where $x = (x_0, x_1, \ldots, x_{k-1})$ and * denotes the non-definite scalar product of vectors in R^k which is given by

$$x*y = x_0y_0 - x_1y_1 - \ldots - x_{k-1}y_{k-1}.$$

The analogue of the orthogonal group O(k) is the so called pseudo-orthogonal group O(1,k-1), which is the subgroup of GL(k) with matrix representation

$$O(1,k-1) = \{U : U* \tilde{I}\, U = \tilde{I}\}$$

where $\tilde{I}$ denotes the $k \times k$ diagonal matrix

$$\tilde{I} = \begin{bmatrix} 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & -1 & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & \cdot \\ \cdot & & & & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & -1 \end{bmatrix}$$

For k = 4 this is the Lorentz group of relativistic physics. Topologically, the group O(1,k-1) has four connected components, of which one is a subgroup of O(1,k-1) and is defined by

$$SO^{\uparrow}(1,k-1) = \{U\varepsilon O(1,k-1): |U| = 1,\ u_{00}>0\}$$

(the elements of U are denoted by u_{ij}, i and j = 0,1,...,k-1). This subgroup is called the special pseudo-orthogonal group and it acts on H^{k-1} by $(U,x) \to xU^*$ (vector-matrix multiplication). The points of H^{k-1} can be expressed in hyperbolic-spherical coordinates as

$$x_0 = \cosh u$$
$$x_1 = \sinh u \cos v_1$$
$$x_2 = \sinh u \sin v_1 \cos v_2$$
$$\cdot$$
$$\cdot$$
$$\cdot$$
$$x_{k-1} = \sinh u \sin v_1 \ldots \sin v_{k-2},$$

and an invariant measure μ on H^{k-1}, relative to the action of $SO^{\uparrow}(1,k-1)$, is specified by

$$d\mu = \sinh^{k-2}u \sin^{k-3}v_1 \ldots \sin v_{k-3}\, dudv_1 \ldots dv_{k-2}. \tag{3.1}$$

The hyperboloid model function, relative to the invariant measure (3.1) on H^{k-1}, is

$$p(x;\xi,\lambda) = a_k(\lambda)e^{-\lambda\xi^* x} \tag{3.2}$$

where the parameters ξ and λ, called the mean direction and the precision, satisfy $\xi\varepsilon H^{k-1}$ and $\lambda>0$, and where

$$a_k(\lambda) = \lambda^{k/2-1}/\{(2\pi)^{k/2-1}2K_{k/2-1}(\lambda)\} \tag{3.3}$$

with $K_{k/2-1}$ a Bessel function.

For any fixed λ, the hyperboloid distributions (3.2) constitute a transformation model under the action of $SO^{\uparrow}(1,k-1)$, and the induced action on the parameter space is $(U,\xi) \to \xi U^*$ (vector-matrix multiplication). The isotropy group K of the element $\xi = (1,0,\ldots,0)$ may be identified with SO(k-1). Furthermore, $SO^{\uparrow}(1,k-1)$ can be factored as

$$SO^{\uparrow}(1,k-1) = HK = H\ SO(k-1)$$

where the matrix representation of $h \varepsilon H$ is

$$h = \begin{bmatrix} x_0 & x_1 & x_2 & \cdots & x_{k-1} \\ x_1 & 1+\frac{x_1^2}{1+x_0} & \frac{x_1x_2}{1+x_0} & \cdots & \frac{x_1x_{k-1}}{1+x_0} \\ x_2 & \frac{x_2x_1}{1+x_0} & 1+\frac{x_2^2}{1+x_0} & \cdots & \frac{x_2x_{k-1}}{1+x_0} \\ \vdots & & & & \\ x_{k-1} & \frac{x_{k-1}x_1}{1+x_0} & \frac{x_{k-1}x_2}{1+x_0} & \cdots & 1+\frac{x_{k-1}^2}{1+x_0} \end{bmatrix}, \tag{3.4}$$

for $x = (x_0, x_1, \ldots, x_{k-1})$ varying over H^{k-1}. In relativity theory a Lorentz transformation of the type (3.4) is termed a "pure Lorentz transformation" or a "boost." (It may be noted that $SO^{\uparrow}(1,k-1)$ can equally be factored as KH with the same K and H as above.)

We have already mentioned the concept of equivariance of a mapping from $\underline{X}$ onto G. More generally, if s is a mapping of $\underline{X}$ onto a space S and if $s(x) = s(x')$ implies $s(gx) = s(gx')$ for $x, x' \varepsilon \underline{X}$ and all $g \varepsilon G$ then s is said to be <u>equivariant</u>. In this case we may define an action of G on S by $gs = s(gx)$ for $s = s(x)$ and for any $x \varepsilon \underline{X}$, and we speak of this as the action induced by s. In the applications to be discussed later S is typically the parameter domain under some parametrization of the model and s is the maximum likelihood estimator, which is automatically equivariant.

We are now ready to state the results which constitute the main tools of the theory of transformation models.

Subject to mild topological regularity conditions (for details, see Barndorff-Nielsen, Blaesild, Jensen and Jorgensen (1982)) we have

<u>Lemma 3.1</u>. Let u be an invariant statistic with range space $U = u(\underline{X})$, let s be an equivariant statistic with range space $S = s(\underline{X})$, and assume that the induced action of G on S is transitive. Furthermore, let μ be

invariant measure on $\underline{X}$. Then, we have $(s,u)(\underline{X}) = S \times U$ and

$$(s,u)\mu = \nu \times \rho$$

where ν is an invariant measure on S and ρ is some measure on U.

Suppose r, s and t are statistics on $\underline{X}$ (in general vector-valued). The symbol $r \perp s|t$ is used to indicate that r and s are conditionally independent given t.

Theorem 3.1. Let the notations and assumptions be as in lemma 3.1, and suppose that the transformation model has a model function $p(x;g)$ relative to an invariant measure μ on $\underline{X}$ such that $p(x) = p(x;e)$ is of the form

$$p(x) = q(u)r(s,w) \tag{3.5}$$

for some functions q and r and some invariant statistic w which is a function of u.

Then the following conclusions are valid.

(i) The model function $p(x;g)$ is of the form

$$p(x;g) = q(u)r(g^{-1}s,w), \tag{3.6}$$

and hence the statistic (s,w) is sufficient.

(ii) We have

$$s \perp u|w.$$

(iii) The invariant statistic u has probability function

$$p(u) = q(u)\int r(s,w)d\nu(s) \quad \langle\rho\rangle \tag{3.7}$$

(where ν is invariant measure on S).

(iv) The conditional probability function of s given w is

$$p(s;g|w) = c(w)r(g^{-1}s,w) \quad \langle\nu\rangle \tag{3.8}$$

where $c(w)$ is a norming constant.

It should be noted that the theorem covers the case where no sufficient reduction is available (take q constant and $w = u$) as well as the case where s - typically the maximum likelihood estimator - is sufficient (take w degenerate). Note also that theorem 3.1 does not assume that the action of G is free. If, however, the action is free and if (z,u) is an orbital decomposition of x then the theorem applies with $s = z$.

Example 3.2. Hyperboloid model (continued). Let $x_1,\ldots,x_n$ be a sample from the hyperboloid distribution (3.2) and let $x = (x_1,\ldots,x_n)$ and $x_+ = x_1 + \ldots + x_n$. Considering λ as fixed, theorem 3.1 applies with u as the maximal invariant statistic, $s = x_+/\sqrt{x_+ * x_+}$ and $w = \sqrt{x_+ * x_+}$. In particular, it turns out that the conditional distribution of s given w (or, equivalently, given u) is again a hyperboloid distribution, with mean direction ξ and precision $w\lambda$. This is in complete analogy with the von Mises-Fisher situation, and accordingly s and w are termed the mean direction and the resultant length of the sample. For details and further results see Jensen (1981) and Barndorff-Nielsen, Blaesild, Jensen and Jorgensen (1982).

Lemma 3.1 and theorem 3.1 are formulated in terms of invariant dominating measures on $\underline{X}$ and S. In applications, however, the probability functions are ordinarily expressed relative to Lebesgue measure - or, more generally, relative to geometric measure when the underlying space is a differentiable manifold. It is therefore important to have a formula which gives the relation between the two types of dominating measure.

Let γ be an action of G on a space $\underline{Y}$ and suppose $\underline{Y}$ has constant orbit type under this action. Then there exists a subgroup K of G, a subset H of G and an orbital decomposition (z,u) of $y \varepsilon \underline{Y}$ such that $G_u = K$ and $z \varepsilon H$ for every y. We assume that H can be chosen so that HK constitutes a (left) factorization of G. If $\underline{Y}$ is a differentiable manifold and if γ acts differentiably on $\underline{Y}$ then an invariant measure μ on $\underline{Y}$ can typically be constructed from geometric measure λ on $\underline{Y}$, by means of Jacobians. In particular, if $\underline{Y}$ is an open subset of some Euclidean space R^r, so that λ is Lebesgue measure, then μ defined by

$$d\mu(y) = J_{\gamma(z)}(u)^{-1} d\lambda(y) \tag{3.9}$$

will be invariant; here $J_{\gamma(g)}$ denotes the Jacobian determinant of the mapping $\gamma(g)$ of $\underline{Y}$ onto itself. A proof of this is sketched in appendix 1.

Example 3.3. Hyperboloid model (continued). We show here how the invariant measure (3.1) on the unit hyperboloid H^{k-1} may be derived from

Lebesgue measure. For simplicity, suppose k = 3. The manifold H^2 is in one-to-one smooth correspondence with R^2 through the mapping

$$\phi : \begin{array}{c} H^2 \to R^2 \\ (x_0, x_1, x_2) \to (x_1, x_2) \end{array} ,$$

and we start by finding an invariant measure on R^2. The action of $SO^{\uparrow}(1,2)$ on H^2 is given by $(U,x) \to xU^*$ and the induced action on R^2 is therefore of the form $(U,y) \to \phi(\phi^{-1}(y)U^*)$. These actions are transitive, and if we take $u = (0,0)$ as the orbit representative of R^2 and let z be the boost

$$z = \begin{bmatrix} y_0 & y_1 & y_2 \\ y_1 & 1 + \frac{y_1^2}{1+y_0} & \frac{y_1 y_2}{1+y_0} \\ y_2 & \frac{y_2 y_1}{1+y_0} & 1 + \frac{y_2^2}{1+y_0} \end{bmatrix}, \tag{3.10}$$

where $y_0 = \sqrt{1 + y_1^2 + y_2^2}$, then (u,z) constitutes an orbital decomposition of $y \varepsilon R^2$ of the type required for the use of formula (3.9). Letting $\acute{\gamma}$ denote the action of $SO^{\uparrow}(1,2)$ on R^2 one finds that $J_{\acute{\gamma}(z)}(u) = \sqrt{1 + y_1^2 + y_2^2}$ and hence the measure

$$d\mu(y) = \frac{1}{\sqrt{1+y_1^2+y_2^2}} dy_1 dy_2$$

is an invariant measure on R^2. Shifting to hyperbolic-spherical coordinates (u,v) for (y_1, y_2) this measure is transformed to (3.1) with k = 3.

Below and in sections 4 and 5 we shall draw several important conclusions from lemma 3.1 and theorem 3.1. Various other applications may be found in Barndorff-Nielsen, Blaesild, Jensen and Jorgensen (1982).

Corollary 3.1. Let G = HK be a left factorization of G such that K is the isotropy group of p. Thus the likelihood function depends on g through h only. Suppose theorem 3.1 applies with S = H and let L(h) = L(h;x) be any version of the likelihood function. Then, the conditional probability function of s given w may be expressed in terms of the likelihood function as

$$p(s;h|w) = c(w)\,\frac{L(h)}{L(s)} \quad \langle\nu\rangle\ . \tag{3.11}$$

In formula (3.11) the likelihood function changes with the value of s. However, an alternative expression for the conditional probability function is available which employs only the single observed likelihood function. Suppose for simplicity that K consists of the identity element alone, so that S = G. Further, let x_0 denote the observed point in $\underline{X}$ and write $L_0(g)$ for $L(g;x_0)$. Also, for specificity, let the action of G on S = G be the so called left action of G on itself, i.e. a $g \varepsilon G$ acts on a point $s \varepsilon S$ simply by multiplying s on the left by g, in the group theoretic sense. (Thus, the two possible interpretations of the symbol gs coincide). The situation here specified occurs, in particular, if the action of G on $\underline{X}$ is free and if s is the group component of an orbital decomposition of x. Setting $s_0 = s(x_0)$ and $w_0 = w(x_0)$, we are interested in the conditional distribution of s given $w = w_0$ and by (3.6) and (3.11) this may be written as

$$p(s;g|w_0) = c(w_0)\frac{L_0(s_0 s^{-1} g)}{L_0(s_0)} \quad \langle\alpha\rangle\ ,$$

the invariant measure being denoted here by α, as a standard notation for left invariant measure on G. This formula, which generalizes a similar expression for the location-scale model due to Fisher (1934), shows how the "shape and position" of the conditional distribution of s is simply determined by the observed likelihood function and the observed s_0, respectively.

Formula (3.11), however, besides being slightly more general, seems more directly applicable in practice.

4. TRANSFORMATIONAL SUBMODELS

Let $\underline{M}$ be a transformation model with acting group G. If P_0 is any of the probability measures in $\underline{M}$ and if G_0 is a subgroup of G then $\underline{P}_0 = \{gP_0 : g \in G_0\}$ defines a transformation submodel $\underline{M}_0$ of $\underline{M}$. For a given G_0 the collection of such submodels typically constitutes a foliation of $\underline{M}$.

Suppose G is a Lie group, as is usually the case. The one-parameter subgroups of G are then in one-to-one correspondence with TG_e, the tangent space of G at the identity element e, and this in turn is in one-to-one correspondence with the Lie algebra $\underline{g}$ of left invariant vector fields on G. More generally, each subalgebra $\underline{h}$ of the Lie algebra of G determines a connected subgroup H of G whose Lie algebra is $\underline{h}$ (cf., for instance, Boothby (1975) chapter 4, theorem 8.7). If $A \in TG_e$, the one-parameter subgroup of G determined by A is of the form $\{\exp(tA) : t \in R\}$. In general, the subgroup of G determined by r linearly independent elements $A_1, \ldots, A_r$ of TG_e may be represented as $\exp\{t_1 A_1\} \ldots \exp\{t_r A_r\}$.

Example 4.1. Let $\underline{M}$ be a location-scale model,

$$p(x_1, \ldots, x_n; \mu, \sigma) = \sigma^{-n} \prod_{i=1}^{n} f(\sigma^{-1}(x_i - \mu)). \qquad (4.1)$$

Here G is the affine group with elements $[\mu, \sigma]$ which may be represented by 2×2 matrices

$$\begin{bmatrix} 1 & 0 \\ \mu & \sigma \end{bmatrix},$$

the group operation being then ordinary matrix multiplication. The Lie algebra of G, or equivalently TG_e, is represented as the set of 2×2 matrices of the form

$$A = \begin{bmatrix} 0 & 0 \\ b & a \end{bmatrix}, \quad a,b\varepsilon R.$$

We have

$$e^{tA} = I + tA + \frac{1}{2!} t^2A^2 + \ldots$$
$$= \begin{bmatrix} 1 & 0 \\ b/a(e^{ta}-1) & e^{ta} \end{bmatrix}$$

where the last expression is to be interpreted in the limiting sense if $a = 0$.

There are therefore four different types of submodels. Specifically, letting (μ_0,σ_0) denote an arbitrary value of (μ,σ) and taking P_0 as the corresponding measure (4.1) we have

(i) If $a = 0$ then $\underline{P}_0$ is a pure location model.

(ii) If $a \neq 0$, $b = 0$ and $\mu_0 = 0$ then $\underline{P}_0$ is a pure scale model.

(iii) If $a \neq 0$, $b = 0$ and $\mu_0 \neq 0$ then $\underline{M}_0$ may be characterized as the submodel of $\underline{M}$ for which the coefficient of variation μ/σ is constant and equal to μ_0/σ_0.

(iv) If both a and b are different from 0 then $\underline{P}_0$ may be characterized as the submodel $\underline{M}_0$ of $\underline{M}$ for which $\sigma^{-1}(\mu+b/a)$ is constant and equal to $c_0 = \sigma_0^{-1}(\mu_0+b/a)$, i.e. if we let $c = b/a$ then $\underline{M}_0$ is determined by

$$\sigma^{-1}(\mu+c) = c_0. \tag{4.2}$$

Letting F denote the distribution function of f we can express (4.2) as the condition that (μ,σ) is such that $-c$ is the $F(-c_0)$-quantile of the distribution $\sigma^{-1}f(\sigma^{-1}(x-\mu))$.

The above example is prototypical in the sense that G is generally a subgroup of the general linear group GL(m) for some m and TG_e may be represented as a linear subset of the set M(m) of all $m \times m$ matrices.

<u>Example 4.2</u>. <u>Hyperboloid model</u>. The model function of the hyperboloid model with $k = 3$ and a known precision parameter λ may be written as

$$p(u,v;\chi,\phi) = (2\pi)^{-1}\lambda e^{\lambda}\sinh u\ e^{-\lambda\{\cosh\chi\ \cosh u - \sinh\chi\ \sinh u\ \cos(v-\phi)\}} \tag{4.3}$$

where $u \geq 0$, $v\varepsilon[0,2\pi)$ and $\chi \geq 0$, $\phi\varepsilon[0,2\pi)$. The generating group $G = SO^{\dagger}(1;2)$ may be represented as the subgroup of GL(3) whose elements are of the form

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & \sin\phi \\ 0 & -\sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} \cosh\chi & \sinh\chi & 0 \\ \sinh\chi & \cosh\chi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1+\frac{1}{2}\zeta^2 & -\frac{1}{2}\zeta^2 & \zeta \\ \frac{1}{2}\zeta^2 & 1-\frac{1}{2}\zeta^2 & \zeta \\ \zeta & -\zeta & 1 \end{bmatrix} \tag{4.4}$$

where $-\infty<\zeta<-\infty$. This determines the so called Iwasa decomposition (cf., for instance, Barut and Raczka (1980) chapter 3) of $SO^{\dagger}(1;2)$ into the product of three subgroups, the three factors in (4.4) being the generic elements of the respective subgroups. It follows that TG_e is the linear subspace of M(3) generated by the linearly independent elements

$$E_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}, \quad E_2 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad E_3 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}.$$

Each of the three subgroups of the Iwasawa decomposition generates a transformational foliation of the hyperboloid model given by (4.3), as discussed in general terms above. In particular, the group determined by the third factor in (4.4) yields, when applied to the distribution (4.3) with $\chi = \phi = 0$, the following one-parameter submodel of the hyperbolic model:

$$p(u,v;\zeta)$$

$$= (2\pi)^{-1}\lambda e^{-\lambda(\cosh u-1)}\sinh u\ e^{-\frac{1}{2}\lambda\{\zeta^2(\cosh u-\sinh u\cos v)-2\zeta\sinh u\sin v\}}.$$

The general form of the one-parameter subgroups of $SO^{\dagger}(1;2)$ is

$$\exp\{t \begin{bmatrix} 0 & a & b \\ a & 0 & c \\ b & -c & 0 \end{bmatrix}\},$$

where a, b, c are fixed real numbers.

5. MAXIMUM ESTIMATION AND TRANSFORMATION MODELS

We shall be concerned with those situations in which there exists an invariant measure μ on $\underline{X}$ that dominates $\underline{P}$, where $\underline{P} = \{gP : g \varepsilon G\}$ is transformational. Letting

$$\frac{dgP}{d\mu}(x) = p(x;g)$$

and writing p(x) for p(x;e) we have

$$p(x;g) = p(g^{-1}x) \quad \langle\mu\rangle.$$

In most cases of interest the model has the following additional structure (possibly after deletion of a null set from $\underline{X}$, cf. also section 3). There exists a left factorization G = HK of G, a K-invariant function f on $\underline{X}$, and an orbital decomposition $(\tilde{h}, u)$ of x such that:

(i) $G_u = K$ for all u and, furthermore, $G_P = K$. Hence, in particular, H may be viewed as the parameter space of the model.

(ii) For every $x \varepsilon \underline{X}$ the function $m(h) = f(h^{-1}x)$ has a unique maximum on H and the maximum point is $\tilde{h}$.

(iii) H may be viewed as an open subset of some Euclidean space R^d and for each fixed $x \varepsilon \underline{X}$ the function m is twice continuously differentiable on H and the matrix $\boldsymbol{\kappa} = \boldsymbol{\kappa}(h)$ given by

$$\boldsymbol{\kappa} = - \frac{\partial^2 m}{\partial h \partial h^*}(h;\tilde{h}u)|_{\tilde{h} = h}$$

is positive definite.

In these circumstances we have:

<u>Proposition 5.1</u>. The maximum estimator $\tilde{h}$ is an equivariant mapping

of $\underline{X}$ onto H and the action of G on H induced by $\tilde{h}$ coincides with the natural action of G on H. Furthermore, if the mapping $x \to (\tilde{h},u)$ is proper then there exists an invariant measure ν on H, and for any fixed u such a measure is given by

$$d\nu(h) = |k|^{\frac{1}{2}}dh \tag{5.1}$$

where dh indicates the differential of Lebesgue measure on H.

Here H is considered as an open subset of R^d, in accordance with (iii).

Proof. The equivariance of $\tilde{h}$ follows immediately from (ii). Obviously, there is a one-to-one correspondence between the family of left cosets $G/K = \{gK: g \in G\}$ and H. Let ρ be the mapping from G/K to H which establishes this correspondence. The natural action ϕ of G on G/K is given by

$$\phi: \begin{array}{c} G \times G/K \to G/K \\ (\acute{g}, gK) \to \acute{g}gK \end{array}$$

and we have to show that when this action is transferred to H by ρ it coincides with the action $\acute{\gamma}$ of G on H induced by $\tilde{h}$. In other words, we must verify that for any $\acute{g} \in G$ the diagram

$$\begin{tikzcd} G/K \arrow[r, "\rho"] \arrow[d, "\phi(\acute{g})"'] & H \arrow[d, "\acute{\gamma}(\acute{g})"] \\ G/K \arrow[r, "\rho"'] & H \end{tikzcd} \tag{5.2}$$

commutes. Let η be the mapping from G to H that sends a $g \in G$ into the uniquely determined $h \in H$ such that $g = hk$ for some $k \in K$. For any $\tilde{h} = \tilde{h}(x)$ in H we have that $\acute{\gamma}(\acute{g})\tilde{h} = \tilde{h}(\acute{g}x)$ is determined by

$$f(\{\tilde{h}(\acute{g}x)\}^{-1}\,\acute{g}x) \geq f(h^{-1}\,\acute{g}x), \qquad h \in H. \tag{5.3}$$

Now, by the K-invariance of f,

$$f(h^{-1}\,\acute{g}x) = f((\acute{g}^{-1}h)^{-1}x) = f(\eta(\acute{g}^{-1}h)^{-1}x)$$

and here $\eta(\acute{g}^{-1}h)$ ranges over all of H when h ranges over H. Hence (5.3) may be rewritten as

$$f(\{\eta(\acute{g}^{-1}\tilde{h}(\acute{g}x))\}^{-1}x) \geq f(h^{-1}x), \qquad h \in H,$$

i.e., by (ii),

$$\tilde{h}(x) = \eta(\acute{g}^{-1}\tilde{h}(\acute{g}x))$$

or, equivalently,

$$\tilde{h}(x)K = \acute{g}^{-1}\tilde{h}(\acute{g}x)K$$

and this, precisely, expresses the commutativity of (5.2), since $\rho^{-1}(h) = hK$.

When the mapping $x \to (\tilde{h},u)$ is proper the subgroup K is compact because $K = G_u$. Hence there exists an invariant measure on H, cf. appendix 1. That $|\text{⁊}|^{\frac{1}{2}}dh$ is such a measure follows from (3.9) and formula (5.10) below.

In particular, then, there is only one action of G on H at play, namely $\acute{\gamma}$, and

$$\acute{\gamma}(\acute{g})h = \eta(\acute{g}h). \tag{5.4}$$

Now, let $h \to \omega$ be an arbitrary reparametrization of the model and let $m(\omega) = m(h(\omega))$ and

$$\text{⁊}(\omega) = \text{⁊}(\omega;u) = -\frac{\partial^2 m}{\partial\omega\partial\omega^*}(\omega;hu). \tag{5.5}$$

This matrix is a (0,2) tensor on Ω.

We shall now show that

$$\text{⁊}(h) = \text{⁊}(h;u) = \underline{J}_{\acute{\gamma}(h)}(e)^{-1*}\text{⁊}(e;u)\underline{J}_{\acute{\gamma}(h)}(e)^{-1}. \tag{5.6}$$

Here the unit element e is to be thought of as a point in H.

We have

$$m(h) = f(h^{-1}x) = f(h^{-1}\tilde{h}u) = f(\{\eta(\tilde{h}^{-1}h)\}^{-1}u)$$

where, again, we have used the K-invariance of f. Thus, with η as the projection mapping defined above we obtain

$$\frac{\partial m(h;x)}{\partial h}(h) = \frac{\partial m(h;u)}{\partial h}(\eta(\tilde{h}^{-1}h))\frac{\partial \eta(\tilde{h}^{-1}h)^*}{\partial h}(h) \tag{5.7}$$

and

$$\frac{\partial^2 m(h;x)}{\partial h\partial h^*}(h) = \frac{\partial\eta(\tilde{h}^{-1}h)}{\partial h^*}(h)\frac{\partial^2 m(h;u)}{\partial h\partial h^*}(\eta(\tilde{h}^{-1}h))\frac{\partial\eta(\tilde{h}^{-1}h)^*}{\partial h}(h)$$
$$+\frac{\partial m(h;u)}{\partial h}(\eta(\tilde{h}^{-1}h))\cdot\frac{\partial^2\eta(\tilde{h}^{-1}h)}{\partial h\partial h^*}(h). \tag{5.8}$$

In these expressions we have, since $\eta(\tilde{h}^{-1}h) = \acute{\gamma}(\tilde{h}^{-1})h$, that

$$\frac{\partial\eta(\tilde{h}^{-1}h)}{\partial h}(h) = \underline{J}_{\acute{\gamma}(\tilde{h}^{-1})}(h). \tag{5.9}$$

On inserting $\tilde{h}$ for h in (5.7), (5.8) and (5.9) (whereby (5.7) becomes 0) and combining with (2.1) we obtain (5.6).

From (5.6) we may draw two important conclusions.

First, taking determinants we have

$$|\mathfrak{k}(h;u)|^{\frac{1}{2}} = J_{\acute{\gamma}(h)}(e)^{-1}|\mathfrak{k}(e;u)|^{\frac{1}{2}} \tag{5.10}$$

and this, by (3.9) and the tensorial nature of $\mathfrak{k}$, implies that $|\mathfrak{k}(\omega)|^{\frac{1}{2}}d\omega$ is an invariant measure on Ω. In connection with formula (5.10) it may be noted that

$$J_{\acute{\gamma}(h)}(e) = J_{\delta(h)}(e)$$

where δ denotes left action of the group G on itself. A proof of this latter formula is given in appendix 2.

Secondly, the tensor $\mathfrak{k}(\omega)$ is found to be G-invariant, whatever the value of the ancillary. In fact, by (5.4) we have, for any $h_0 \varepsilon H$ and $\acute{g}\varepsilon G$,

$$\acute{\gamma}(\acute{\gamma}(\acute{g})h)h_0 = \acute{\gamma}(\acute{g}) \circ \acute{\gamma}(h)h_0.$$

Consequently

$$\underline{J}_{\acute{\gamma}(\acute{\gamma}(\acute{g})h)}(e) = \underline{J}_{\acute{\gamma}(h)}(e)\ \underline{J}_{\acute{\gamma}(g)}(h)$$

and this together with (5.6) and (2.26) establishes the invariance.

In particular, observed information $\mathfrak{j}$ determines a G-invariant Riemannian metric on the parameter space. The expected information metric i can also be shown to be G-invariant.

From proposition 5.1 and corollary 3.1 we find

<u>Corollary 5.1</u>. The model function $p^*(\tilde{\omega};\omega|u) = c|\tilde{k}|^{\frac{1}{2}}L/\tilde{L}$ is exactly equal to $p(\tilde{\omega};\omega|u)$.

By taking m of (ii) equal to the log likelihood function l this corollary specializes to theorem 4.1 of Barndorff-Nielsen (1983).

Suppose, in particular, that the model is an exponential transformation model. Then the above theory applies with $m(\omega) = \overset{\alpha}{l}(\omega)$. The essential property to check is that $\overset{\alpha}{l}(\omega;t(x))$ is of the form $f(h^{-1}x)$. This follows simply from the definition of $\overset{\alpha}{l}$ and theorem 2.1.

6. OBSERVED GEOMETRIES

In section 2 we briefly reviewed how the parameter space of the model $\underline{M}$ may be set up as a manifold with expected information i as Riemannian metric tensor and with an associated family of affine connections, the α-connections (2.30). We shall now discuss a similar type of geometries on the parameter space, related to observed information and depending on the choice of the auxiliary statistic a which together with the maximum likelihood estimator $\hat{\omega}$ constitutes a minimal sufficient statistic for $\underline{M}$. These latter geometries are termed <u>observed geometries</u> (Barndorff-Neilsen, 1986a). In applications to statistical inference questions it will usually be appropriate to take a to be ancillary but a great part of what we shall discuss does not require distribution constancy of a and, unless explicitly stated otherwise, the auxiliary a is considered arbitrary (except for the implicit smoothness properties).

Let an auxiliary a be chosen. We may now take partial derivatives of $l = l(\omega;\hat{\omega},a)$ with respect to the coordinates $\hat{\omega}^r$ of $\hat{\omega}$ as well as with respect to ω^r. Letting $\hat{\partial} = \partial/\partial\hat{\omega}^r$ we introduce the notation

$$l_{r_1 \ldots r_p;s_1 \ldots s_q} = \partial_{r_1} \ldots \partial_{r_p} \hat{\partial}_{s_1} \ldots \hat{\partial}_{s_q} l \tag{6.1}$$

and refer to these quantities as <u>mixed derivatives of the log model function</u>. The function of ω and a obtained from (6.1) by substituting ω for $\hat{\omega}$ will be denoted by $\text{ł}_{r_1 \ldots r_p;s_1 \ldots s_q}$. Thus, for instance,

$$\text{ł}_{rs;t} = \text{ł}_{rs;t}(\omega) = \text{ł}_{rs;t}(\omega;a) = l_{rs;t}(\omega;\omega,a).$$

More generally, for any combinant g of the form $g(\omega;\hat{\omega},a)$ we write

$$\not{g} = \not{g}(\omega;a) = g(\omega;\omega,a).$$

This is in consistency with the notation $\not{\jmath}$ introduced by (2.6). The observed geometries, to be discussed, are expressed in terms of the mixed derivatives

$$\not{\ell}_{r_1 \ldots r_p;s_1 \ldots s_q}. \tag{6.2}$$

So are the terms of an asymptotic expansion of (2.7), cf. section 7.

Given the observed value of a the observed information tensor $\not{\jmath}$, of (2.6), defines the parameter space of $\underline{M}$ as a Riemannian manifold. The Riemannian connection determined by $\not{\jmath}$ has connection symbols $\overset{o}{\not{\Gamma}}{}^{t}_{rs}$ given by $\overset{o}{\not{\Gamma}}{}^{t}_{rs} = \not{\jmath}^{tu}\overset{o}{\not{\Gamma}}_{rst}$ and

$$\overset{o}{\not{\Gamma}}_{rst} = \tfrac{1}{2}(\partial_r\not{\jmath}_{st} - \partial_t\not{\jmath}_{rs} + \partial_s\not{\jmath}_{rt}).$$

Employing the notation established above we have $\partial_t\not{\jmath}_{rs} = -\not{\ell}_{rst} - \not{\ell}_{rs;t}$, etc. so that

$$\overset{o}{\not{\Gamma}}_{rst} = \not{\ell}_{rs;t} - \tfrac{1}{2}(\not{\ell}_{rst} + \not{\ell}_{rs;t}[3]). \tag{6.3}$$

As we shall now show, the quantity

$$\not{T}_{rst} = -(\not{\ell}_{rst} + \not{\ell}_{rs;t}[3]) \tag{6.4}$$

is a covariant tensor of rank 3, i.e.

$$\not{T}_{\rho\sigma\tau} = \not{T}_{rst}\,\omega^{r}_{/\rho}\,\omega^{s}_{/\sigma}\,\omega^{t}_{/\tau}. \tag{6.5}$$

First, from (2.14) we have

$$\not{\ell}_{\rho\sigma\tau} = \not{\ell}_{rst}\,\omega^{r}_{/\rho}\,\omega^{s}_{/\sigma}\,\omega^{t}_{/\tau} + \not{\ell}_{rs}\,\omega^{r}_{/\rho\sigma}\,\omega^{s}_{/\tau}[3]. \tag{6.6}$$

Further, from (2.13) we obtain, on differentiating with respect to $\hat{\psi}^{\tau}$ and then substituting parameter for estimate,

$$\not{\ell}_{\rho\sigma;\tau} = \not{\ell}_{rs;t}\,\omega^{r}_{/\rho}\,\omega^{s}_{/\sigma}\,\omega^{t}_{/\tau} + \not{\ell}_{r;t}\,\omega^{r}_{/\rho\sigma}\,\omega^{t}_{/\tau}. \tag{6.7}$$

Finally, differentiating the likelihood equation

$$\not{\ell}_r = 0$$

we find

$$\not\ell_{rs} + \not\ell_{r;s} = 0 \tag{6.8}$$

or

$$\not\ell_{r;s} = \not{j}_{rs}. \tag{6.9}$$

Combination of (6.4), (6.6), (6.7) and (6.9) yields (6.5).

It follows from the tensorial nature of $\not{T}$ and from (6.3) and (6.9) that for any real α an affine connection $\overset{\alpha}{\not\Gamma}$ on $\underline{M}$ may be defined by

$$\overset{\alpha}{\not\Gamma}{}^{t}_{rs} = \not{j}^{tu}\,\overset{\alpha}{\not\Gamma}_{rsu}$$

with

$$\overset{\alpha}{\not\Gamma}_{rst} = \not\ell_{rs;t} + \frac{1-\alpha}{2}\not{T}_{rst}. \tag{6.10}$$

In particular, we have

$$\overset{1}{\not\Gamma}_{rst} = \not\ell_{rs;t}\,, \qquad \overset{-1}{\not\Gamma}_{rst} = \not\ell_{t;rs} \tag{6.11}$$

where to obtain the latter expression we have used

$$\not\ell_{rst} + \not\ell_{rs;t} + \not\ell_{rt;s} + \not\ell_{r;st} = 0$$

which follows on differentiation of (6.8). It may also be noted that

$$\partial_t \not{j}_{rs} = \overset{1}{\not\Gamma}_{rts} + \overset{-1}{\not\Gamma}_{str} = \overset{1}{\not\Gamma}_{str} + \overset{-1}{\not\Gamma}_{rts}$$

and

$$\overset{\alpha}{\not\Gamma}_{rst} = \frac{1+\alpha}{2}\overset{1}{\not\Gamma}_{rst} + \frac{1-\alpha}{2}\overset{-1}{\not\Gamma}_{rst}.$$

The connections $\overset{\alpha}{\not\Gamma}$, which we shall refer to as the <u>observed α-connections</u>, are analogues of the <u>expected α-connections</u> $\overset{\alpha}{\Gamma}$ given by (2.30). The analogy between $\overset{\alpha}{\Gamma}$ and $\overset{\alpha}{\not\Gamma}$ becomes more apparent by rewriting the skewness tensor (2.29) as

$$T_{rst} = -E\{l_{rst} + l_{rs}l_t[3]\},$$

the validity of which follows on differentiation of the formula

$$E\{l_{rs} + l_r l_s\} = 0, \tag{6.12}$$

which, in turn, may be compared to (6.8).

Under the specifications of a of primary statistical interest one has that, in broad generality, the observed geometries converge to the corresponding expected geometries as the sample size tends to infinity.

For (k,k) exponential models

$$p(x;\theta) = a(\theta)b(x)e^{\theta\cdot t(x)} \tag{6.13}$$

no auxiliary statistic is involved since $\hat{\theta}$ is minimal sufficient, and we find $\not{j} = i$ and $\overset{\alpha}{\not{T}} = \overset{\alpha}{\Gamma}$, $\alpha\varepsilon R$.

Let i,j,k,... be indices for the coordinates of θ, t and τ, using upper indices for θ and lower indices for t and τ.

In the case of a curved exponential model (2.35), we have

$$l_r = (t-\tau)_i\theta^i_{/r} \tag{6.14}$$

and, letting $\acute{\theta}$ denote the maximum likelihood estimator of θ under the full model generated by (2.35), the relation $\not{l}_{r;s} = \not{j}_{rs}$ takes the form

$$\not{l}_{r;s}(\omega) = \kappa_{ij}(\theta)\theta^i_{/r}\acute{\theta}^j_{/s}$$

$$= \kappa_{ij}(\theta)\theta^i_{/r}\theta^j_{/s} - (\acute{t}-\tau)_i\theta^i_{/rs} = \not{j}_{rs}(\omega). \tag{6.15}$$

Furthermore,

$$\not{l}_{rst}(\omega) = -\kappa_{ijk}(\theta)\theta^i_{/r}\theta^j_{/s}\theta^k_{/t} - \kappa_{ij}(\theta)\theta^i_{/r}\theta^j_{/st}[3] + (\acute{t}-\tau)_i\theta^i_{/rst}, \tag{6.16}$$

$$\not{l}_{rs;t}(\omega) = \kappa_{ij}(\theta)\theta^i_{/rs}\acute{\theta}^j_{/t} = \overset{1}{\not{T}}_{rst} \tag{6.17}$$

and

$$\not{l}_{t;rs} = \kappa_{ij}(\theta)\theta^i_{/t}\acute{\theta}^j_{/rs} = \overset{-1}{\not{T}}_{rst}. \tag{6.18}$$

It is also to be noted that, under mild regularity conditions, the quantities $\not{j}$ and $\not{T}$ possess asymptotic expansions the first terms of which are given by

$$\not{j}_{rs} = i_{rs} - \theta^i_{/rs}\theta^j_{/\lambda}\,\kappa_{ij}a^\lambda + \ldots \tag{6.19}$$

and

$$\not{T}_{rst} = T_{rst} - \{\kappa_{ijk}\theta^i_{/rs}\theta^j_{/t}\theta^k_{/\lambda}[3]$$

$$+ \kappa_{ij}\theta^i_{/rs}\theta^j_{/t\lambda}[3] + \kappa_{ij}\theta^i_{/rst}\theta^j_{/\lambda}\}a^\lambda + \ldots, \tag{6.20}$$

where a^λ, $\lambda = 1,\dots,k-d$, are the coordinates of the auxiliary statistic a. For instance, in the repeated sampling situation and letting a_0 denote the affine ancillary, as defined in Barndorff-Nielsen (1980), we may take $a = n^{-\frac{1}{2}}a_0$ and the expansions (6.19) and (6.20) are asymptotic in powers of $n^{-\frac{1}{2}}$. (For further comparison with Amari (1982a) it may be noted that the coefficient in the first order correction term of (6.19) may be written as $\theta^i_{/rs}\theta^j_{/\lambda}\kappa_{ij} = n\overset{e}{H}_{rs\lambda}$ where $\overset{e}{H}_{rs\lambda}$ is Amari's notation for the exponential curvature, or α-curvature with $\alpha = 1$, of the curved exponential model viewed as a manifold imbedded in the full (k,k) model.)

For a transformation model we find

$$l_r(h;x) = l_{r'}(\eta(\hat{h}^{-1}h);u)\eta(\hat{h}^{-1}h)^{r'}_{/r}$$

(cf. the more general formula (5.7)) and hence

$$\overset{1}{\mathcal{P}}_{rst}(h) = -l_{r's't'}(e;u)\mathcal{A}^{r'}_r\mathcal{A}^{s'}_s\mathcal{A}^{t'}_t$$

$$-j_{r's'}(e;u)\{\mathcal{A}^{r'}_r\mathcal{B}^{s'}_{s;t} + \mathcal{A}^{s'}_s\mathcal{B}^{r'}_{r;t} - \mathcal{A}^{r'}_t\mathcal{B}^{s'}_{rs}\} \tag{6.21}$$

$$\overset{-1}{\mathcal{P}}_{rst}(h) = l_{r's't'}(e;u)\mathcal{A}^{r'}_r\mathcal{A}^{s'}_s\mathcal{A}^{t'}_t$$

$$+ j_{r's'}(e;u)\{\mathcal{A}^{r'}_r\mathcal{B}^{s'}_{t;s} + \mathcal{A}^{s'}_s\mathcal{B}^{r'}_{t;r} - \mathcal{A}^{r'}_t\mathcal{B}^{s'}_{;rs}\} \tag{6.22}$$

where, for $\partial_r = \partial/\partial h^r$ and $\acute{\partial}_r = \partial/\partial\hat{h}^r$,

$$A^r_s = \partial_s\eta^r(\hat{h}^{-1}h),$$

so that

$$\mathcal{A}^r_s = \{\underline{J}_{\acute{\gamma}(h)}(e)^{-1}\}_{rs}, \tag{6.23}$$

while

$$B^r_{st} = \partial_s\partial_t\eta^r(\hat{h}^{-1}h)$$

$$B^r_{s;t} = \partial_s\hat{\partial}_t\eta^r(\hat{h}^{-1}h)$$

$$B^r_{;st} = \hat{\partial}_s\hat{\partial}_t\eta^r(\hat{h}^{-1}h).$$

Furthermore, to write the coefficients of $l_{r's't'}(e;u)$ in (6.21) and (6.22) as indicated we have used the relation

$$\hat{\partial}_s \eta^r(\hat{h}^{-1}h)|_{\hat{h}=h} = -\partial_s \eta^r(\hat{h}^{-1}h)|_{h=\hat{h}} . \tag{6.24}$$

Formula (6.24) is proved in appendix 3.

We now briefly consider four examples. In the first three the model is transformational and the auxiliary statistic a is taken to be the maximal invariant statistic, and thus a is exactly ancillary. In the fourth example a is only approximately ancillary. Examples 6.1, 6.3 and 6.4 concern curved exponential models whereas the model in example 6.2 - the location-scale model - is exponential only if the error distribution is normal.

Example 6.1. Constant normal fractile. For known $\alpha\varepsilon(0,1)$ and $c\varepsilon(-\infty,\infty)$, let $\underline{N}_{\alpha,c}$ denote the class of normal distributions having the real number c as α-fractile, i.e.

$$\underline{N}_{\alpha,c} = \{N(\mu,\sigma^2):(c-\mu)/\sigma = u_\alpha\},$$

where u_α denotes the α-fractile of the standard normal distribution, and let $x_1,\ldots,x_n$ be a sample from a distribution in $\underline{N}_{\alpha,c}$. The model for $x=(x_1,\ldots,x_n)$ thus defined is a (2,1) exponential model, except for $u_\alpha = 0$ when it is a (1,1) model. Henceforth we suppose that $u_\alpha \neq 0$, i.e. $\alpha \neq \frac{1}{2}$. The model is also a transformation model relative to the subgroup G of the group of one-dimensional affine transformations given by

$$G = \{[c(1 - \lambda),\lambda]:\lambda>0\},$$

the group operation being

$$[c(1 - \lambda),\lambda][c(1 - \lambda'),\lambda'] = [c(1 - \lambda\lambda'),\lambda\lambda']$$

and the action of G on the sample space being

$$[c(1 - \lambda),\lambda](x_1,\ldots,x_n) = (c(1 - \lambda) + \lambda x_1,\ldots,c(1 - \lambda) + \lambda x_n).$$

(Note that G is isomorphic to the multiplicative group.)

Letting

$$a = (\bar{x} - c)/s',$$

where $\bar{x} = (x_1 + \ldots + x_n)/n$ and

$$s'^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

we have that a is maximal invariant and, parametrizing the model by $\zeta = \log \sigma$, that the maximum likelihood estimate is

$$\hat{\zeta} = \log(bs')$$

where

$$b = b(a) = (u_\alpha/2)a + \sqrt{1 + \{(u_\alpha/2)^2 + 1\}a^2}.$$

Furthermore, $(\hat{\zeta},a)$ is a one-to-one transformation of the minimal sufficient statistic $(\bar{x},s')$ and a is exactly ancillary.

The log likelihood function may be written as

$$l(\zeta) = l(\zeta;\hat{\zeta},a) = n[\hat{\zeta} - \zeta - \tfrac{1}{2}\{b^{-2}e^{2(\hat{\zeta}-\zeta)} + (u_\alpha + ab^{-1}e^{\hat{\zeta}-\zeta})^2\}]$$

from which it is evident that the model for $\hat{\zeta}$ given a is a location model.

Indicating differentiation with respect to ζ and $\hat{\zeta}$ by subscripts ζ and $\hat{\zeta}$, respectively, we find

$$l_\zeta = n\{-1 + b^{-2}e^{2(\hat{\zeta}-\zeta)} + ab^{-1}(u_\alpha + ab^{-1}e^{\hat{\zeta}-\zeta})e^{\hat{\zeta}-\zeta}\}$$

and hence

$$\not{j} = n\{2b^{-2} + ab^{-1}(u_\alpha + 2ab^{-1})\}$$

$$\not{l}_{\zeta\zeta\zeta} = n\{4b^{-2} + ab^{-1}(u_\alpha + 4ab^{-1})\}$$

$$\not{l}_{\zeta\zeta;\hat{\zeta}} = -n\{4b^{-2} + ab^{-1}(u_\alpha + 4ab^{-1})\} = \overset{1}{\not{P}}$$

$$\not{l}_{\zeta;\hat{\zeta}\hat{\zeta}} = n\{4b^{-2} + ab^{-1}(u_\alpha + 4ab^{-1})\} = \overset{-1}{\not{P}} = -\overset{1}{\not{P}}$$

and the observed skewness tensor is

$$\not{T} = n\{8b^{-2} + 2ab^{-1}(u_\alpha + 4ab^{-1})\}.$$

Note also that

$$\overset{\alpha}{\not{P}} = \alpha\overset{1}{\not{P}}.$$

We mention in passing that another normal submodel, that specified by a known coefficient of variation μ/σ, has properties similar to those exhibited by example 6.1.

Example 6.2. Location-scale model. Let data x consist of a sample $x_1,\ldots,x_n$ from a location-scale model, i.e. the model function is

$$p(x;\mu,\sigma) = \sigma^{-n} \prod_{i=1}^{n} f\left(\frac{x_i-\mu}{\sigma}\right)$$

for some known probability density function f. We assume that $\{x:f(x)>0\}$ is an open interval and that $g = -\log f$ has a positive and continuous second order derivative on that interval. This ensures that the maximum likelihood estimate $(\hat{\mu},\hat{\sigma})$ exists uniquely with probability 1 (cf., for instance, Burridge (1981)).

Taking as the auxiliary a Fisher's configuration statistic

$$a = (a_1,\ldots,a_n) = \left(\frac{x_1-\hat{\mu}}{\hat{\sigma}}, \ldots , \frac{x_n-\hat{\mu}}{\hat{\sigma}}\right),$$

which is an exact ancillary, we find

$$\jmath(\mu,\sigma) = \sigma^{-2} \begin{bmatrix} \Sigma g''(a_\nu) & \Sigma a_\nu g''(a_\nu) \\ \Sigma a_\nu g''(a_\nu) & n+\Sigma a_\nu^2 g''(a_\nu) \end{bmatrix}$$

and, in an obvious notation,

$$\not{\ell}_{\mu\mu,\mu} = -\sigma^{-3}\Sigma g'''(a_i)$$

$$\not{\ell}_{\mu\mu,\sigma} = -\sigma^{-3}\Sigma a_i g'''(a_i)$$

$$\not{\ell}_{\mu\sigma,\mu} = -\sigma^{-3}\{2\Sigma g''(a_i) + \Sigma a_i g'''(a_i)\}$$

$$\not{\ell}_{\mu\sigma,\sigma} = -\sigma^{-3}\{2\Sigma a_i g''(a_i) + \Sigma a_i^2 g'''(a_i)\}$$

$$\not{\ell}_{\sigma\sigma,\mu} = -\sigma^{-3}\{4\Sigma a_i g''(a_i) + \Sigma a_i^2 g'''(a_i)\}$$

$$\not{\ell}_{\sigma\sigma,\sigma} = -\sigma^{-3}\{2n + 4\Sigma a_i^2 g''(a_i) + \Sigma a_i^3 g'''(a_i)\}$$

$$\not{\ell}_{\mu\mu\mu} = \sigma^{-3}\Sigma g'''(a_i)$$

$$\not{\ell}_{\mu\mu\sigma} = \sigma^{-3}\{2\Sigma g''(a_i) + \Sigma a_i g'''(a_i)\}$$

$$\not{l}_{\mu\sigma\sigma} = \sigma^{-3}\{4\Sigma a_i g''(a_i) + \Sigma a_i^2 g'''(a_i)\}$$

$$\not{l}_{\sigma\sigma\sigma} = \sigma^{-3}\{4n + 6\Sigma a_i^2 g''(a_i) + \Sigma a_i^3 g'''(a_i)\}.$$

Furthermore,

$$\not{T}_{\mu\mu\mu} = 2\sigma^{-3}\not{l}_{\mu\mu\mu}((0,1);a)$$

$$\not{T}_{\mu\mu\sigma} = -2\sigma^{-3}\not{j}_{\mu\mu}((0,1);a) + 2\sigma^{-3}\not{l}_{\mu\mu\sigma}((0,1);a)$$

$$\not{T}_{\sigma\sigma\mu} = -4\sigma^{-3}\not{j}_{\mu\sigma}((0,1);a) + 2\sigma^{-3}\not{l}_{\sigma\sigma\mu}((0,1);a)$$

$$\not{T}_{\sigma\sigma\sigma} = -6\sigma^{-3}\not{j}_{\sigma\sigma}((0,1);a) + 2\sigma^{-3}\not{l}_{\sigma\sigma\sigma}((0,1);a).$$

Example 6.3. Hyperboloid model. Let $(u_1,v_1),\dots,(u_n,v_n)$ be a sample from the hyperboloid distribution (4.3) and suppose the precision λ is known. The resultant length is

$$a = \{(\Sigma \cosh u_i)^2 - (\Sigma \sinh u_i \cos v_i)^2 - (\Sigma \sinh u_i \sin v_i)^2\}^{\frac{1}{2}}$$

and a is maximal invariant after minimal sufficient reduction. Furthermore, the maximum likelihood estimate $(\hat{\chi},\hat{\phi})$ of (χ,ϕ) exists uniquely, with probability 1, $(a,\hat{\chi},\hat{\phi})$ is minimal sufficient and the conditional distribution of $(\hat{\chi},\hat{\phi})$ given the ancillary a is again hyperboloidic, as in (4.3) but with u, v and λ replaced by $\hat{\chi}$, $\hat{\phi}$ and $a\lambda$. It follows that the log likelihood function is

$$l(\chi,\phi) = l(\chi,\phi;\hat{\chi},\hat{\phi},a) = -a\lambda\{\cosh\chi \cosh\hat{\chi} - \sinh\chi \sinh\hat{\chi} \cos(\hat{\phi}-\phi)\}$$

and hence

$$\overset{\alpha}{\not{T}}_{\chi\chi\chi} = \overset{\alpha}{\not{T}}_{\chi\chi\phi} = \overset{\alpha}{\not{T}}_{\chi\phi\chi} = \overset{\alpha}{\not{T}}_{\phi\phi\phi} = 0$$

$$\overset{\alpha}{\not{T}}_{\chi\phi\phi} = a\lambda \cosh \chi \sinh \chi$$

$$\overset{\alpha}{\not{T}}_{\phi\phi\chi} = -a\lambda \cosh \chi \sinh \chi,$$

whatever the value of α. Thus, in this case, the α-geometries are identical.

We note again that whereas the auxiliary statistic a is taken so as to be ancillary in the various examples discussed here - exactly distribution constant in the three examples above and asymptotically distribution constant in the one to follow - ancillarity is no prerequisite for the general theory of observed geometries.

Furthermore, let a be any statistic which depends on the minimal sufficient statistic t, say, only and suppose that the mapping from t to $(\hat{\omega},a)$ is defined and one-to-one on some subset $\underline{T}_0$ of the full range $\underline{T}$ of values of t though not, perhaps, on all of $\underline{T}$. We can then endow the model $\underline{M}$ with observed geometries, in the manner described above, for values of t in $\underline{T}_0$. The next example illustrates this point.

The above considerations allow us to deal with questions of nonuniqueness and nonexistence of maximum likelihood estimates and nonexistence of exact ancillaries, especially in asymptotic considerations.

Example 6.4. Inverse Gaussian - Gaussian model. Let $x(\cdot)$ and $y(\cdot)$ be independent Brownian motions with a common diffusion coefficient $\sigma^2 = 1$ and drift coefficients $\mu>0$ and ξ, respectively. We observe the process $x(\cdot)$ till it first hits a level $x_0>0$ and at the time u when this happens we record the value $v = y(u)$ of the second process. The joint distribution of u and v is then given by

$$p(u,v;\mu,\xi)$$

$$= (2\pi)^{-1}x_0 e^{x_0\mu} u^{-2} e^{-\frac{1}{2}(x_0^2+v^2)u^{-1}} e^{-\frac{1}{2}\mu^2 u+\xi v-\frac{1}{2}\xi^2 u}. \tag{6.25}$$

Suppose that $(u_1,v_1),\ldots,(u_n,v_n)$ is a sample from the distribution (6.25) and let $t = (\bar{u},\bar{v})$ where $\bar{u}$ and $\bar{v}$ are the arithmetic means of the observations. Then t is minimal sufficient and follows a distribution similar to (6.25), specifically

$$p(\bar{u},\bar{v};\mu,\xi)$$

$$= (2\pi)^{-1}x_0 n e^{nx_0\mu}\bar{u}^{-2} e^{-\frac{n}{2}(x_0^2+\bar{v}^2)\bar{u}^{-1}} e^{-\frac{n}{2}\mu^2\bar{u}+n\xi\bar{v}-\frac{n}{2}\xi^2\bar{u}}. \tag{6.26}$$

Now, assume ξ equal to μ. The model (6.26) is then a (2,1) exponential model, still with t as minimal sufficient statistic. The maximum likelihood estimate of μ is undefined if $t \notin \underline{T}_0$ where

$$\underline{T}_0 = \{t = (\bar{u},\bar{v}): x_0 + \bar{v} \geq 0\}$$

hereas for $t \in \underline{T}_0$, $\hat{\mu}$ exists uniquely and is given by

$$\hat{\mu} = \tfrac{1}{2}(x_0 + \bar{v})\,\bar{u}^{-1}. \tag{6.27}$$

he event $t \notin \underline{T}_0$ happens with a probability that decreases exponentially fast with he sample size n and may therefore be ignored for most statistical purposes.

Defining, formally, $\hat{\mu}$ to be given by (6.27) even for $t \notin \underline{T}_0$ and letting

$$a = \Phi^-(\bar{u}; 2nx_0^2, 2\,n\hat{\mu}^2),$$

here $\Phi^-(\cdot;\chi,\psi)$ denotes the distribution function of the inverse Gaussian disribution with density function

$$\phi^-(x;\chi,\psi) = (2\pi)^{-\frac{1}{2}}\sqrt{\chi}\; e^{\sqrt{\chi\psi}}\; x^{-3/2}\; e^{-\frac{1}{2}(\chi x^{-1}+\psi x)} \tag{6.28}$$

e have that the mapping $t \to (\hat{\mu},a)$ is one-to-one from $\underline{T} = \{t = (\bar{u},\bar{v}): \bar{u}>0\}$ onto $-\infty,+\infty) \times (0,\infty)$ and that a is asymptotically ancillary and has the property hat $p^*(\hat{\mu};\mu|a) = c|\hat{j}|^{\frac{1}{2}}\bar{L}$ approximates the actual conditional density of $\hat{\mu}$ given to order $O(n^{-3/2})$, cf. Barndorff-Nielsen (1984).

Letting $\Phi_-(\cdot;\chi,\psi)$ denote the inverse function of $\Phi^-(\cdot;\chi,\psi)$ we may rite the log likelihood function for μ as

$$\begin{aligned} l(\mu) &= l(\mu;\hat{\mu},a) \\ &= n\{(x_0 + \bar{v})\mu - \bar{u}\mu^2\} \\ &= n\Phi_-(a; 2nx_0^2, 2n\hat{\mu}^2)\,\{2\hat{\mu}\mu - \mu^2\} \end{aligned} \tag{6.29}$$

rom this we find

$$l_{\mu\mu} = -2n\Phi_-(a; 2n\,x_0^2, 2n\hat{\mu}^2)$$

o that

$$\not{j} = 2n\Phi_-(a; 2nx_0^2, 2n\mu^2)$$

$$\not{l}_{\mu\mu\mu} = 0$$

nd

$$\mathcal{X}_{\mu\mu;\hat{\mu}} = 8n^2\mu(\phi^- \circ \Phi_-/\Phi_\psi^-)(a;2nx_0^2,2n\mu^2)$$

$$= \overset{1}{\mathcal{F}}_{\mu\mu\mu} = -\tfrac{1}{2}\overset{-1}{\mathcal{F}}_{\mu\mu\mu}$$

where Φ_ψ^- denotes the derivative of $\Phi^-(x;\chi,\psi)$ with respect to ψ. By the well-known result (Shuster (1968))

$$\Phi^-(x;\chi,\psi) = \Phi(\psi^{\frac{1}{2}}x^{\frac{1}{2}} - \chi^{\frac{1}{2}}x^{-\frac{1}{2}}) + e^{2\sqrt{\chi\psi}}\,\Phi(-(\psi^{\frac{1}{2}}x^{\frac{1}{2}} + \chi^{\frac{1}{2}}x^{-\frac{1}{2}})),$$

where Φ is the distribution function of the standard normal distribution, Φ_ψ^- could be expressed in terms of Φ and $\phi = \Phi'$.

7. EXPANSION OF $c|\hat{j}|^{\frac{1}{2}}\bar{L}$

We shall derive an asymptotic expansion of (2.7), by Taylor expansion of $c|\hat{j}|^{\frac{1}{2}}\bar{L}$ in $\hat{\omega}$ around ω, for fixed value of the auxiliary a. The various terms of this expansion are given by mixed derivatives (cf. (6.2)) of the log model function. It should be noted that for arbitrary choice of the auxiliary statistic a the quantity $c|\hat{j}|\bar{L}$ constitutes a probability (density) function on the domain of variation of $\hat{\omega}$ and the expansions below are valid. However, $c|\hat{j}|\bar{L}$ furnishes an approximation to the actual conditional distribution of $\hat{\omega}$ given a, as discussed in section 2, only for suitable ancillary specification of a.

To expand $c|\hat{j}|^{\frac{1}{2}}\bar{L}$ in $\hat{\omega}$ around ω we first write $\bar{L}$ as $\exp\{l-\hat{l}\}$ and expand l in ω around $\hat{\omega}$. By Taylor's formula,

$$l-\hat{l} = \sum_{\nu=2}^{\infty} \frac{1}{\nu!}(\omega-\hat{\omega})^{r_1}\dots(\omega-\hat{\omega})^{r_\nu}(\partial_{r_1}\dots\partial_{r_\nu} l)(\hat{\omega})$$

whence, expanding each of the terms $(\partial_{r_1}\dots\partial_{r_\nu} l)(\hat{\omega})$ around ω,

$$l-\hat{l}$$

$$= \sum_{\nu=2}^{\infty} \frac{(-1)^\nu}{\nu!}(\hat{\omega}-\omega)^{r_1}\dots(\hat{\omega}-\omega)^{r_\nu}$$

$$\cdot \sum_{\rho=0}^{\infty} \frac{1}{\rho!}(\hat{\omega}-\omega)^{s_1}\dots(\hat{\omega}-\omega)^{s_\rho}\,\partial_{s_1}\dots\partial_{s_\rho}\, \ell_{r_1\dots r_\nu}. \tag{7.1}$$

Consequently, writing δ for $\hat{\omega}-\omega$ and $\delta^{rs\dots}$ for $(\hat{\omega}-\omega)^r(\hat{\omega}-\omega)^s\dots$, we have

$$l-\hat{l} = -\tfrac{1}{2}\delta^{rs}\text{ɉ}_{rs} + \tfrac{1}{2}\delta^{rst}(\text{ł}_{rs;t} + \tfrac{2}{3}\text{ł}_{rst})$$

$$+ \frac{1}{24}\,\delta^{rstu}(6\text{ł}_{rs;tu} + 8\text{ł}_{rst;u} + 3\text{ł}_{rstu}) + \dots \,. \tag{7.2}$$

Next, we wish to expand $\log\{|\hat{j}|/|\text{ɉ}|\}^{\frac{1}{2}}$ in $\hat{\omega}$ around ω. To do this we observe that if A is a $d \times d$ matrix whose elements a_{rs} depend on ω then

$$\partial_t \log|A| = |A|^{-1}\partial_t|A|$$

$$= a^{sr}\partial_t a_{rs}$$

where a^{rs} denotes the (r,s)-element of the inverse of A. Furthermore, using

$$\partial_t a^{rs} = -a^{rv}a^{ws}\partial_t a_{vw},$$

which is obtained by differentiating $a_{ru}a^{us} = \delta_r^s$ with respect to ω^t and solving for a^{rs}, we find

$$\partial_t\partial_u \log|A| = -a^{vr}a^{sw}\partial_u a_{vw}\partial_t a_{rs} + a^{sr}\partial_t\partial_u a_{rs}.$$

It follows that

$$\log\{|\hat{j}|/|\text{ɉ}|\}^{\frac{1}{2}} = -\tfrac{1}{2}\delta^t\text{ɉ}^{rs}(\text{ł}_{rst} + \text{ł}_{rs;t})$$

$$-\tfrac{1}{4}\delta^{tu}\{\text{ɉ}^{rs}(\text{ł}_{rstu} + \text{ł}_{rst;u} + \text{ł}_{rsu;t} + \text{ł}_{rs;tu})$$

$$+ \text{ɉ}^{rv}\text{ɉ}^{sw}(\text{ł}_{rst} + \text{ł}_{rs;t})(\text{ł}_{vwu} + \text{ł}_{vw;u})\} + \dots \,. \tag{7.3}$$

By means of (7.2) and (7.3) we therefore find

$$c|\hat{j}|^{\frac{1}{2}}\bar{L} = (2\pi)^{d/2}c\phi_d(\hat{\omega}-\omega;\text{ɉ})\{1 + A_1 + A_2 + \dots\} \tag{7.4}$$

where $\phi_d(\cdot;\text{ɉ})$ denotes the density function of the d-dimensional normal distribution with mean 0 and precision (i.e. inverse variance-covariance matrix) ɉ and where

$$A_1 = -\tfrac{1}{2}\delta^t\text{ɉ}^{rs}(\text{ł}_{rs;t} + \text{ł}_{rst}) + \tfrac{1}{2}\delta^{rst}(\text{ł}_{rs;t} + \tfrac{2}{3}\text{ł}_{rst}) \tag{7.5}$$

and

$$A_2 = \frac{1}{24}[-\,3\delta^{tu}\{2\text{ɉ}^{rs}(\text{ł}_{rstu} + \text{ł}_{rst;u} + \text{ł}_{rsu;t} + \text{ł}_{rs;tu})$$

$$+ (2\text{ɉ}^{rv}\text{ɉ}^{sw} - \text{ɉ}^{rs}\text{ɉ}^{vw})(\text{ł}_{rs;t} + \text{ł}_{rst})(\text{ł}_{vw;u} + \text{ł}_{vwu})\}$$

$$+ \delta^{rstu}\{(3\ell_{rstu} + 8\ell_{rst;u} + 6\ell_{rs;tu})$$

$$- 6\jmath^{vw}(\ell_{vw;u} + \ell_{vwu})(\ell_{rs;t} + \frac{2}{3}\ell_{rst})\}$$

$$+ 3\delta^{rstuvw}(\ell_{rs;t} + \frac{2}{3}\ell_{rst})(\ell_{uv;w} + \frac{2}{3}\ell_{uvw})], \tag{7.6}$$

A_1 and A_2 being of order $O(n^{-\frac{1}{2}})$ and $O(n^{-1})$, respectively, under ordinary repeated sampling.

By integration of (7.4) with respect to $\hat{\omega}$ we obtain

$$(2\pi)^{d/2}c = 1 + C_1 + \dots , \tag{7.7}$$

where C_1 is obtained from A_2 by changing the sign of A_2 and making the substitutions

$$\delta^{rs} \to \jmath^{rs}$$

$$\delta^{rstu} \to \jmath^{rs}\jmath^{tu}[3]$$

$$\delta^{rstuvw} \to \jmath^{rs}\jmath^{tu}\jmath^{vw}[15],$$

the 3 and 15 terms in the two latter expressions being obtained by appropriate permutations of the indices (thus, for example, $\delta^{rstu} \to \jmath^{rs}\jmath^{tu} + \jmath^{rt}\jmath^{su} + \jmath^{ru}\jmath^{st}$).

Combination of (7.4) and (7.7) finally yields

$$c|j|^{\frac{1}{2}}\bar{L} = \phi(\hat{\omega}-\omega;\jmath)\{1 + A_1 + (A_2+C_1) + \dots\} \tag{7.8}$$

with an error term which in wide generality is of order $O(n^{-3/2})$ under repeated sampling. In comparison with an Edgeworth expansion it may be noted that the expansion (7.8) is in terms of mixed derivatives of the log model function, rather than in terms of cumulants, and that the error of (7.8) is relative, rather than absolute.

In particular, under repeated sampling and if the auxiliary statistic is (approximately or exactly) ancillary such that

$$p(\hat{\omega};\omega|a) = p^*(\hat{\omega};\omega|a)\{1 + O(n^{-3/2})\}$$

(cf. section 2) we generally have

$$p(\hat{\omega};\omega|a) = \phi_d(\hat{\omega}-\omega;\bar{\jmath})\{1 + A_1 + (A_2 + C_1) + O(n^{-3/2})\}. \tag{7.9}$$

For one-parameter models, i.e. for $d = 1$, the expansion (7.8) with A_1, A_2 and C_1 as given above reduces to the expansion (2.9). In Barndorff-Nielsen and Cox (1984) a relation valid to order $O(n^{-3/2})$ was established, for general d, between the norming constant c of (2.7) and the Bartlett adjustment factors for likelihood ratio tests of hypotheses about ω. By means of this relation such adjustment factors may be simply calculated from the above expression for C_1.

Example 7.1. Suppose $\underline{M}$ is a (k,k) exponential model with model function (6.13). Then the expression for C_1 takes the form

$$C_1 = \frac{1}{24}\{3\kappa_{rstu}\kappa^{rs}\kappa^{tu} - \kappa_{rst}\kappa_{uvw}(2\kappa^{ru}\kappa^{sv}\kappa^{tw} + 3\kappa^{rs}\kappa^{tu}\kappa^{vw})\}$$

where, for $\partial_r = \partial/\partial\theta^r$ and $\kappa(\theta) = -\log a(\theta)$,

$$\kappa_{rs\ldots} = \partial_r\partial_s \ldots \kappa(\theta)$$

and where κ^{rs} is the inverse matrix of κ_{rs}.

From (7.8) we find the following expansion for the mean value of $\hat{\omega}$:

$$E_\omega\hat{\omega}^\alpha = \omega^\alpha + \mu_1^\alpha + \mu_2^\alpha + \ldots$$

where μ_1^α is of order $O(n^{-1})$, μ_2^α is of order $O(n^{-2})$, and

$$\mu_1^\alpha = -\tfrac{1}{2}\bar{\jmath}^{\alpha r}\bar{\jmath}^{st}\bar{\ell}_{r;st} = -\tfrac{1}{2}\bar{\jmath}^{\alpha r}\bar{\jmath}^{st}\overset{-1}{\bar{T}}_{str}. \tag{7.10}$$

Hence, from (7.8) and writing δ' for $\delta-\mu_1$,

$$c|\hat{\jmath}|^{\frac{1}{2}}\bar{L} = \phi_d(\hat{\omega} - \omega - \mu_1;\bar{\jmath})\{1 + (A_1 - \delta^r\bar{\jmath}_{rs}\mu_1^s) + \ldots\}$$

$$= \phi_d(\hat{\omega} - \omega - \mu_1;\bar{\jmath})\{1 + \tfrac{1}{2}h^{rst}(\delta';\bar{\jmath})(\bar{\ell}_{rs;t} + \tfrac{2}{3}\bar{\ell}_{rst}) + \ldots\}, \tag{7.11}$$

where the error term is of order $O(n^{-1})$ and where $h^{r_1\ldots r_n}(\cdot;\bar{\jmath})$ denotes the tensorial Hermite polynomial (as defined by Amari and Kumon (1983)), relative to the tensor $\bar{\jmath}_{rs}$. Using (6.10) we may rewrite the last quantity in (7.11) as

$$\bar{\ell}_{rs;t} + \tfrac{2}{3}\bar{\ell}_{rst} = -\overset{-1/3}{\bar{T}}_{rst} + \bar{R}_{rst} \tag{7.12}$$

where

$$\mathcal{R}_{rst} = \tfrac{4}{3}(\mathcal{T}_{rs;t} - \tfrac{1}{2}(\mathcal{T}_{rt;s} + \mathcal{T}_{st;r})). \tag{7.13}$$

Since

$$h^{rst}(\delta';\mathcal{j}) = \delta'^r\delta'^s\delta'^t - \mathcal{j}^{rs}\delta'^t[3] \tag{7.14}$$

we find

$$h^{rst}(\delta';\mathcal{j})\mathcal{R}_{rst} = 0$$

and hence (7.11) reduces to

$$c|\hat{j}|^{\frac{1}{2}}\bar{L} = \phi_d(\hat{\omega} - \omega - \mu_1;\mathcal{j})\{1 - \tfrac{1}{6}h^{rst}(\delta';\mathcal{j})\overset{-1/3}{\mathcal{F}}_{rst} + \ldots\}, \tag{7.15}$$

the error term being $O(n^{-1})$.

Suppose, in particular, that the model is an exponential (k,d) model. We may then compare (7.15) with the Edgeworth expansion for an efficient, bias adjusted estimate of ω given an ancillary statistic, provided by formulas (3.33) and (3.25) in Amari and Kumon (1983). It appears that $h^{rst}(\delta';\mathcal{j})\overset{-1/3}{\mathcal{F}}_{rst}$ of (7.15) is the counterpart of Amari and Kumon's $\overset{-1/3}{\Gamma}_{abc}h^{abc} - \overset{e}{H}_{ab\kappa}h^{ab}h^{\kappa} + \overset{m}{H}_{\kappa\lambda a}h^{a}h^{\kappa\lambda}$. Thus (7.15) offers some simplification over the corresponding expression provided by the Amari and Kumon paper.

Note that, again by the symmetry of (7.14), if

$$\overset{-1/3}{\mathcal{F}}_{rst}[3] = 0 \tag{7.16}$$

for all r,s,t then the first order correction term in (7.15) is 0. Furthermore, for any one-parameter model $\underline{M}$ the quantity $\overset{\alpha}{\mathcal{F}}$ with $\alpha = -1/3$, can be made to vanish by choosing that parametrization for which ω is the geodesic coordinate for the -1/3 observed conditional connection. (Note that generally this parametrization will depend on the value of the ancillary a.) An analogous result holds for the Edgeworth expansion derived by Amari and Kumon (1983), referred to above. The parametrization making the $\alpha = -1/3$ expected connection $\overset{\alpha}{\Gamma}$ vanish has the interpretation of a skewness reducing parametrization, cf. Kass (1984).

8. EXPONENTIAL TRANSFORMATION MODELS

Suppose $\underline{M}$ is an exponential transformation model and that the full exponential model $\underline{\tilde{M}}$ generated by M is regular. By theorem 2.1 the group G acts affinely on $T = \tau(\Theta)$, and Lebesgue measure on T is quasi-invariant (in fact, relatively invariant) with multiplier $|A(g)|$. Assuming, furthermore, that $\underline{M}$ and G have the structure discussed in section 3 with $\{g:|A(g)| = 1\} \subset K$ we find, since the mapping $g \to A(g)$ is a representation of G, that

$$|A(\hat{h}(gx))| = |A(g)||A(\hat{h}(x))|.$$

Thus $m(x) = |A(\hat{h})|$ is a modulator and

$$d\nu(h) = |A(h)|^{-1}dh \tag{8.1}$$

is an invariant measure on H (cf. appendix 1).

Again by theorem 2.1 the log likelihood function is of the form

$$l(h) = \{\theta(e)A(h^{-1}\hat{h})^* + \tilde{B}(\hat{h}^{-1}h)\}\cdot w - \kappa(\theta(e)A(h^{-1}\hat{h})^* + \tilde{B}(\hat{h}^{-1}h)) \tag{8.2}$$

where $w = t(u) = \hat{h}^{-1}t$.

Some interesting special cases are

(i) $B(\cdot)$ or $\tilde{B}(\cdot)$ or both are 0. Then $\delta(\cdot)$ of (2.45) is a multiplier (i.e. a homomorphism of G into $(R_+,\cdot)$). Furthermore, if $\tilde{B}(\cdot) = 0$ and if (2.35) is an exponential representation of $\underline{M}$ relative to an invariant dominating measure on $\underline{X}$ then $b(x)$ is a modulator.

(ii) The norming constant $a(\theta(g))$ does not depend on g. If in addition $B(g)$ does not depend on g, which implies that $B(\cdot) = 0$, then the conditional distribution of $\hat{h}$ given w is, on account of the exactness of (2.7),

$$p(\hat{h};h|w) = c(w)|\hat{j}|^{\frac{1}{2}} e^{\theta(\hat{h}^{-1}h)\cdot w} \tag{8.3}$$

where the norming constant does not depend on h.

Note that the form (8.3) is preserved under repeated sampling, i.e. the conditional distribution of $\hat{h}$ is of the same "type" whatever the sample size.

The von Mises-Fisher model for directional data with fixed precision has this structure with w equal to the resultant length r, and as is well-known the conditional model given r is also of this type, irrespective of sample size. Other examples are provided by the hyperboloid model with fixed precision and by the class or r-dimensional normal distributions with mean 0 and precision Δ such that $|\Delta| = 1$.

(iii) $\underline{M}$ is a (k,k-1) model.

For simplicity we now assume that $\underline{M}$ has all the above-mentioned properties. There is then little further restriction in supposing that $\underline{M}$ is of the form

$$p(x,\theta) = b(x)\exp\{-a\lambda e_1 A(\hat{h}^{-1}h)^{-1*}e^*_{-1}\} \tag{8.4}$$

where λ is the index parameter, a is maximal invariant and e_1 and e_{-1} are known nonrandom vectors. For (8.4) the log likelihood function is

$$l(h) = -a\lambda e_1\bar{\bar{A}}(\hat{h}^{-1}h)e^*_{-1} \tag{8.5}$$

where we have written $\bar{\bar{A}}$ for A^{-1*}. Hence

$$\mathcal{j}_{rs} = a\lambda(\partial_t\partial_u\bar{\bar{A}}_{ij})(e)e_{1i}e_{-1j}\mathcal{A}^t_r(h)\mathcal{A}^u_s(h) \tag{8.6}$$

where $\mathcal{A}^r_s$ is given by (6.23). In this case, then, the conditional observed geometries $(\mathcal{j}(\cdot;\lambda,a),\overset{\alpha}{\mathcal{F}}(\cdot;\lambda,a))$ are all "proportional" for fixed α, with $a\lambda$ as the proportionality factor. The geometric leaves of the foliation of $\acute{\underline{M}}$, determined as the partition of $\acute{\underline{M}}$ generated by the index parameter λ, are thus highly similar. In this connection see example 6.3.

APPENDIX 1

Construction of invariant measures

One may usefully generalize the concepts of invariant and relatively invariant measures as follows. Let a measure μ on $\underline{X}$ be called quasi-invariant with multiplier $\chi = \chi(g,x)$ if $g\mu$ and μ are mutually absolutely continuous for every $g \in G$ and if

$$d(g^{-1}\mu)(x) = \chi(g,x)d\mu(x).$$

Furthermore, define a function m on $\underline{X}$ to be a modulator with associated multiplier $\chi(g,x)$ if m is positive and

$$m(gx) = \chi(g,x)m(x). \tag{A1.1}$$

Then, if μ^{χ} is quasi-invariant with multiplier $\chi(g,x)$ and if m is a modulator satisfying (A1.1) we have that

$$\mu = m^{-1}\mu^{\chi} \tag{A1.2}$$

is an invariant measure on $\underline{X}$.

In particular, to verify that the measure μ defined by (3.9) is invariant one just has to show that $m(y) = J_{\gamma(z)}(u)$ is a modulator with associated multiplier $J_{\gamma(g)}(y)$ because, by the standard theorem on transformation of integrals, Lebesgue measure λ is quasi-invariant with multiplier $J_{\gamma(g)}(y)$. Corresponding to the factorization $G = HK$ there are unique factorizations $g = hk$ and $gz = \acute{h}\acute{k}$ and, using repeatedly the assumption that $K = G_u$ for every orbit representative u, we find

$$m(gy) = J_{\gamma(\acute{h})}(u) = J_{\gamma(g)}(y)J_{\gamma(z)}(u)J_{\gamma(\acute{k}^{-1})}(u)$$

$$= J_{\gamma(g)}(y)m(y).$$

In the last step we have used the fact that

$$J_{\gamma(k)}(u) = 1 \text{ for every } k \varepsilon K. \tag{A1.3}$$

To see the validity of (A1.3) one needs only note that for fixed u the mapping $k \to J_{\gamma(k)}(u)$ is a multiplier on K and since K is compact this must be the trivial multiplier 1. Actually, (A1.3) is a necessary and sufficient condition for the existence of an invariant measure on $\underline{Y}$. This may be concluded from Kurita (1959), cf. also Santaló (1979), section 10.3.

APPENDIX 2

An equality of Jacobians under left factorizations

Lemma. Let $G = HK$ be a left factorization of G (as discussed in sections 3 and 5), let $\acute{\gamma}$ denote the natural action of G on H and let δ denote left action of G on itself. Then $J_{\acute{\gamma}(h)}(e) = J_{\delta(h)}(e)$ for all $h \varepsilon H$.

Proof. Let $g = hk$ denote an arbitrary element of G. Writing g symbolically as (h,k) and employing the mappings η and ζ defined by

$$\eta{:}g \to h \qquad \zeta{:}g \to k$$

we have, for any $h' \varepsilon H$,

$$\delta(h')g = \delta(h')(h,k) = (\eta(h'h),\zeta(h'hk))$$

and hence the differential of $\delta(h')g$ is

$$D\delta(h')(g) = \begin{bmatrix} \dfrac{\partial\eta(h'h)^*}{\partial h} & 0 \\ & \\ \dfrac{\partial\zeta(h'hk)^*}{\partial h} & \dfrac{\partial\zeta(h'hk)^*}{\partial k} \end{bmatrix}$$

from which we find, using $\eta(h'h) = \acute{\gamma}(h')h$ and $\zeta(h'k) = k$,

$$\begin{aligned} J_{\delta(h')}(e) &= J_{\acute{\gamma}(h')}(e)\left|\frac{\partial\zeta(h'k)^*}{\partial k}\right|_{k=e} \\ &= J_{\acute{\gamma}(h')}(e). \end{aligned}$$

APPENDIX 3

An inversion result

The validity of formula (6.24) is established by the following

Lemma. Let $G = HK$ be a left factorization of the group G with the associated mapping $\eta: g = hk \to h$ (as discussed in sections 3 and 5). Furthermore, let h' denote an arbitrary element of H. Then

$$\left.\frac{\partial\eta(h^{-1}h')^*}{\partial h}\right|_{h=h'} = - \left.\frac{\partial\eta(h'^{-1}h)^*}{\partial h}\right|_{h=h'} . \tag{A3.1}$$

Proof. The mapping $h \to \eta(h^{-1}h')$ may be composed of the three mappings $h \to h'^{-1}h$, $g \to g^{-1}$ and η, as indicated in the following diagram

$$\begin{tikzcd} H \arrow[r, "h'^{-1}\cdot"] & G \arrow[loop left, "i"] \arrow[d, "\eta"] \\ & H \end{tikzcd}$$

where i indicates the inversion $g \to g^{-1}$. This diagram of mappings between differentiable manifolds induces a corresponding diagram for the associated differential mappings between the tangent spaces of the manifolds, namely

$$\begin{tikzcd}
TH_h \arrow[r, "D(h'^{-1}\cdot)"] & TG_{h'^{-1}h} \arrow[loop left, "Di"] \arrow[d, "D\eta"] \\
& TH_{\eta(h'^{-1}h)}
\end{tikzcd}$$

From this latter diagram and from the well-known relation

$$(Di)(e) = -I,$$

where I indicates the identity matrix, formula (A3.1) may be read off immediately.

Acknowledgements

I am much indebted to Poul Svante Eriksen, Peter Jupp, Steffen L. Lauritzen, Hans Anton Salomonsen and Jorgen Tornehave for helpful discussions, and to Lars Smedegaard Andersen for a careful checking of the manuscript.

REFERENCES

Amari, S.-I. (1982a). Differential geometry of curved exponential families - curvatures and information loss. Ann. Statist. 10, 357-385.

Amari, S.-I. (1982b). Geometrical theory of asymptotic ancillarity and conditional inference. Biometrika 69, 1-17.

Amari, S.-I. (1985). Differential-Geometric Methods in Statistics. Lecture Notes in Statistics 28, Springer, New York.

Amari, S.-I. (1986). Differential geometrical theory of statistics - towards new developments. This volume.

Amari, S.-I. and Kumon, M. (1983). Differential geometry of Edgeworth expansion in curved exponential family. Ann. Inst. Statist. Math. 35, 1-24.

Barndorff-Nielsen, O. E. (1978a). Information and Exponential Families. Wiley, Chichester.

Barndorff-Nielsen, O. E. (1978b). Hyperbolic distributions and distributions on hyperbolae. Scand. J. Statist. 5, 151-157.

Barndorff-Nielsen, O. E. (1980). Conditionality resolutions. Biometrika 67, 293-310.

Barndorff-Nielsen, O. E. (1982). Contribution to the discussion of R. J. Buehler: Some ancillary statistics and their properties. J. Amer. Statist. Assoc. 77, 590-591.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. Biometrika 70, 343-365.

Barndorff-Nielsen, O. E. (1984). On conditionality resolution and the likelihood ratio for curved exponential families. Scand. J. Statist. 11, 157-

170. Amendment Scand. J. Statist. 12 (1985).

Barndorff-Nielsen, O. E. (1985). Confidence limits from $c|\hat{j}|^{\frac{1}{2}}\bar{L}$ in the single-parameter case. Scand. J. Statist. 12, 83-87.

Barndorff-Nielsen, O. E. (1986a). Likelihood and observed geometries. Ann. Statist. 14, 856-873.

Barndorff-Nielsen, O. E. (1986b). Inference on full or partial parameters based on the standardized signed log likelihood ratio. Biometrika 73, 307-322.

Barndorff-Nielsen, O. E. and Blaesild, P. (1983a). Exponential models with affine dual foliations. Ann. Statist. 11, 753-769.

Barndorff-Nielsen, O. E. and Blaesild, P. (1983b). Reproductive exponential families. Ann. Statist. 11, 770-782.

Barndorff-Nielsen, O. E. and Blaesild, P. (1984). Combination of reproductive models. Research Report 107, Dept. Theor. Statist., Aarhus University.

Barndorff-Nielsen, O. E., Blaesild, P., Jensen, J. L. and Jorgensen, B. (1982). Exponential transformation models. Proc. R. Soc. A 379, 41-65.

Barndorff-Nielsen, O. E. and Cox, D. R. (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. J. R. Statist. Soc. B 46, 483-495.

Barndorff-Nielsen, O. E., Cox. D. R. and Reid, N. (1986). The role of differential geometry in statistical theory. Int. Statist. Review 54, 83-96.

Barut, A. O. and Raczka, R. (1980). Theory of Group Representations and Applications. Polish Scientific Publishers, Warszawa.

Boothby, W. M. (1975). An Introduction to Differentiable Manifolds and Riemannian Geometry. Academic Press, New York.

Burridge, J. (1981). A note on maximum likelihood estimation for regression models using grouped data. J. R. Statist. Soc. B 43, 41-45.

Chentsov, N. N. (1972). Statistical Decision Rules and Optimal Inference. (In Russian.) Moscow, Nauka. English translation (1982). Translation of Mathematical Monographs Vol. 53. American Mathematical Society, Providence, Rhode Island.

Eguchi, S. (1983). Second order efficiency of minimum contrast estimators in a curved exponential family. Ann. Statist. 11, 793-803.

Eriksen, P. S. (1984a). (k,1) exponential transformation models. Scand. J. Statist. 11, 129-145.

Eriksen, P. S. (1984b). A note on the structure theorem for exponential transformation models. Research Report 101, Dept. Theor. Statist., Aarhus University.

Eriksen, P. S. (1984c). Existence and uniqueness of the maximum likelihood estimator in exponential transformation models. Research Report 103, Dept. Theor. Statist., Aarhus University.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. Proc. Roy. Soc. A 144, 285-307.

Hauck, W. W. and Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. J. Amer. Statist. Ass. 72, 851-853. Corrigendum: J. Amer. Statist. Ass. 75 (1980), 482.

Jensen, J. L. (1981). On the hyperboloid distribution. Scand. J. Statist. 8, 193-206.

Kurita, M. (1959). On the volume in homogeneous spaces. Nagoya Math. J. 15, 201-217.

Lauritzen, S. L. (1986). Statistical manifolds. This volume.

Santaló, L. A. (1979). Integral Geometry and Geometric Probability. Encyclopedia of Mathematics and Its Applications. Vol. 1, Addison-Wesley, London.

Shuster, J. J. (1968). A note on the inverse Gaussian distribution function. J. Amer. Statist. Assoc. 63, 1514-1516.

Vaeth, M. (1985). On the use of Wald's test in exponential families. Int. Statist. Review 53, 199-214.