CONDITIONAL EXPONENTIAL FAMILIES AND A REPRESENTATION THEOREM FOR ASYMPTOTIC INFERENCE

BY

PAUL D. FEIGIN

TECHNICAL REPORT NO. 132
JULY 1979

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MCS77-16974

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
CONDITIONAL EXPONENTIAL FAMILIES AND A REPRESENTATION THEOREM FOR ASYMPTOTIC INFERENCE

By

PAUL D. FEIGIN

TECHNICAL REPORT NO. 132
JULY 1979

Prepared Under the Auspices of National Science Foundation Grant MCS77-16974

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
CONDITIONAL EXPONENTIAL FAMILIES AND A
REPRESENTATION THEOREM FOR ASYMPTOTIC INFERENCE

by

Paul D. Feigin
Technion - Israel Institute of Technology

Abstract

Conditional exponential families of Markov processes are defined and
a representation of the score function martingale is established for the
important conditionally additive case. This result unifies those obtained
separately for different examples and provides the key to asymptotic nor-
mality results for the maximum likelihood estimate.
1. Introduction

In developing a theory of parametric inference for stochastic process models, one particular type of example led to the consideration of families of processes which seemed to represent the Markov process analogue of exponential families. This analogue was recognized by Heyde and Feigin (1975) and further referred to in the papers of Feigin (1978) and Heyde (1978). Here we propose to give a more precise definition of this conditional exponential family (CEF) analogue and to prove some important general properties of these families which are a consequence of their underlying exponential family structure.

The main result shows that what were originally defined as CEF's by Heyde and Feigin (1975) are more properly considered as conditionally additive exponential families (which cover all the examples discussed in the above references) and for these the asymptotic normality and related properties of the maximum likelihood estimate follow from a common representation of the score function martingale. These results significantly unify the analysis of inference questions for quite distinct processes which nevertheless fall into this class of families.

The basic definitions in the spirit of the work of Barndorff-Nielsen (1978) are given in Section 2, the theory for the conditionally additive families is developed in Section 3, and some examples are briefly discussed in Section 4.

2. Conditional Exponential Families

We will develop the theory for the vector parameter case in this section, adopting the dot \( \cdot \) to denote vector derivatives with respect to the vector parameter \( \theta \).
Suppose $X = \{X_0, X_1, \ldots\}$ is a time-homogeneous Markov chain with possible transition probability densities (with respect to a given measure $\nu$ on $\mathbb{R}^p$) denoted by $f(y|x; \theta)$. We assume that $X$ is defined in the measurable space $(\Omega, \mathcal{F})$ and let $\{P_\theta; \theta \in \Theta\}$ denote the family of corresponding probability measures on $(\Omega, \mathcal{F})$ for which

$$P_\theta(X_n \in A | X_{n-1} = x) = \int_A f(y|x; \theta) \, \nu(dy)$$

holds for each Borel $A \subset \mathbb{R}^p$ and all $n \geq 0$. We denote by $\mathcal{F}_n$ the $\sigma$-field generated by $\{X_1, \ldots, X_n\}$. We will say that

A. $\{(X, P_\theta); \theta \in \Theta \subset \mathbb{R}^k\}$ is a conditional exponential family of Markov processes if

$$f(y|x; \theta) = b(y,x) \exp\{\alpha(\theta) \cdot m(y,x) - \beta(\theta, x)\} \quad (2.1)$$

where for each fixed $x$, $m(\cdot, x)$ and $b(\cdot, x)$ are measurable.

Given $P_\theta(X_0 = x_0) = 1$ (for all $\theta \in \Theta$) we then may write down the likelihood

$$L_n(\theta) = \prod_{i=1}^n b(X_i, X_{i-1}) \exp\{\alpha(\theta) \cdot \sum_{i=1}^n m(X_i, X_{i-1}) - \sum_{i=1}^n \beta(\theta, X_{i-1})\}$$

and the score function (the derivative of the log likelihood)

$$U_n(\theta) = \hat{r}_n(\theta) = \sum_{i=1}^n \{\alpha(\theta) \cdot m(X_i, X_{i-1}) - \beta(\theta, X_{i-1})\} \quad (2.2)$$

where the $\hat{\cdot}$ denotes differentiation with respect to $\theta$ so that $\hat{\cdot}(\theta)$ is a matrix.

We assume that we are working with the canonical parametrization $\alpha(\theta) = \theta$, and that for all $x$ in the state space

$$|\beta(\theta, x)| = |\log \int b(y,x) \, e^{\theta \cdot m(y,x)} \, \nu(dy)| < \infty$$

whenever $\theta \in \Theta$.  

2
Theorem 1. Suppose \( \{(X_n, P_\theta); \theta \in \Theta\} \) is a conditional exponential family of Markov processes and \( \theta \in \text{int } \Theta \) (the interior of \( \Theta \)).

(i) If \( \forall n \geq 1, E_\theta |m(X_n, X_{n-1})| < \infty \) componentwise then \( \{U_n(\theta), \mathcal{F}_n; n > 1\} \) is a zero-mean \( P_\theta \)-martingale (vector martingale).

(ii) If, \( \forall n \geq 1, E_\theta \tilde{\beta}(\theta, X_{n-1}) \) exists then this martingale is square integrable; its conditional variance matrix is given by

\[
\xi_n(\theta) = -\tilde{X}_n(\theta) = \sum_{1}^{n} \tilde{\beta}(\theta, X_{n-1}) . \tag{2.3}
\]

Proof. (i) Under the stated conditions and the properties of ordinary exponential families (see Barndorff-Nielsen (1978, Chapter 8)), it follows that

\[
E_\theta [m(X_1, X_{i-1}) | \mathcal{F}_{i-1}] = \tilde{\beta}(\theta, X_{i-1})
\]

whence the martingale property follows from (2.2).

(ii) Similarly, the given condition ensures that the covariance matrix \( \text{var}_\theta \{m(X_1, X_{i-1})\} \) exists since

\[
\text{var}_\theta \{m(X_1, X_{i-1}) | \mathcal{F}_{i-1}\} = \tilde{\beta}(\theta, X_{i-1}) ;
\]

and thus (2.3) follows (recall \( \alpha(\theta) = \Theta \) ).

Remark: Theorem 1 is a particular case of more general martingale properties of the score function - see Feigin (1976) for example.

Furthermore, in the scalar case we can show the existence of another martingale which is useful in the analysis of CEF's. Letting

\[
V_n(\theta) = \prod_{i=1}^{n} \{m(X_i, X_{i-1})/\tilde{\beta}(\theta, X_{i-1})\} , \tag{2.4}
\]

the above theorem shows that \( \{V_n(\theta), \mathcal{F}_n; n \geq 1\} \) is a \( P_\theta \)-martingale. If it is also positive, then the fact that it has unit expectation ensures, via the martingale convergence theorem, that
\[ V_n(\theta) \rightarrow V(\theta) \text{ a.s. } [\mathcal{P}_0]. \] (2.5)

In fact, we find that (2.5) can also hold even when \( V_n(\theta) \) is not positive.

Suppose now that

\[ \Theta_x = \{ \theta : |\beta(\theta, x)| < \infty \} \ni \theta \] (2.6)

and that \( \Theta \) is open. This may be called the regular case. From the ordinary exponential family theory (Barndorff-Nielsen (1978)) we know that \( \Theta \) is convex and that

\[ \hat{\beta}(\cdot, x) : \Theta \rightarrow \text{int} C_x \]

is one-to-one and invertible where \( C_x = \text{c.l.} \text{conv}\{m(y, x) : y \in \text{supp}(\nu)\} \). If we want to determine when \( U_n(\theta) = 0 \) has a unique solution \( \hat{\theta}_n \) (the maximum likelihood estimate, MLE) we need to consider particular forms of \( \beta(\cdot, \cdot) \). In the next section we specialize to the scalar parameter case and consider a very useful factorization of \( \beta(\cdot, \cdot) \).

3. Conditionally Additive Exponential Families

The case of conditionally additive exponential families (CAEF's) is what Heyde and Feigin (1975) simply called CEF's. The definition of a CAEF is:

B. \( \{(X, \mathcal{P}_\theta) ; \theta \in \Theta\} \) is a conditionally additive exponential family if it is a CEF (Definition A) and

\[ \beta(\theta, x) = \gamma(\theta) h(x). \] (3.1)

Under Definition B, (2.6) holds with \( \Theta = \{\theta : |\gamma(\theta)| < \infty\} \). We assume that \( \Theta \) is open – i.e. we are in a regular situation – and choose \( h(\cdot) \) nonnegative.
The word additive is used in the description of CAEF's because, if we look at the Laplace transform of \( m(X_i, X_{i-1}) \) conditionally on \( X_{i-1} \), we find

\[
\phi_{X_{i-1}}(t) = E[e^{\gamma(\theta + t) - \gamma(\theta)} h(X_{i-1})}
\]

which is the (conditional on \( X_{i-1} \)) Laplace transform of \( Y_{h(X_{i-1})} \) where \( Y \) is an additive process with appropriate Laplace transform. It is the Markov structure and this conditional additivity that gives the representation so useful in obtaining asymptotic properties of the score function martingale and hence of the MLE. The direct connection between the two is given by

**Theorem 2.** Suppose \( \{X, P_\theta; \theta \in \Theta\} \) is a CAEF and \( E_\theta h(X_n) < \infty \), \( \theta \in \Theta \). Then, \( \forall \theta \in \Theta \),

\[
U_n(\theta) = H_n (\hat{\gamma}(\hat{\theta}_n) - \gamma(\theta)) = \xi_n(\theta) (\hat{\gamma}(\hat{\theta}_n) - \gamma(\theta))/\hat{\gamma}(\theta)
\]

where \( \hat{\theta}_n \) is the (unique) MLE and \( H_n = \sum_{i=1}^{n} h(X_{i-1}) \).

**Proof.** The condition \( E_\theta h(X_n) < \infty \) ensures that the family satisfies the conditions of Theorem 1. From the ordinary exponential family theory mentioned above we must have

\[
\hat{\gamma}(\cdot) : \Theta \to \mathcal{U}
\]

invertible and one-to-one where, identically in \( x \) for which \( h(x) > 0 \),

\[
\mathcal{U} = \text{int } C_x/h(x) \equiv \text{int}\{c/h(x) : c \in C_x\}.
\]

Therefore, as long as \( H_n > 0 \),

\[
H_n \hat{\gamma}(\theta) = \sum_{i=1}^{n} m(X_i, X_{i-1}) \equiv M_n
\]
has a unique solution, \( \hat{\theta}_n \), whenever \( M_n / H_n \in \mathcal{U} \); this being the case unless

\[
\{ m(X_i, X_{i-1}) / h(X_{i-1}) : h(X_{i-1}) > 0 \} \subset \text{bd}\mathcal{U}.
\]

To avoid burdening the proof, we will assume that the latter occurs with probability zero, and we conclude that (3.3) follows by substituting (3.4), (3.1) and (2.3) in (2.2). \| 

The form (3.3) and limit theory for the martingale \( U_n \) provide the desired asymptotic properties of the MLE. Provided \( H_n \to \infty \) a.s., \( \hat{\theta}_n \) will be strongly consistent, and as long as \( \hat{\gamma}(\cdot) \) is continuous, a weak limit theorem for \( U_n(\theta) \) suitably normalized will translate into one for \( \hat{\theta}_n - \theta \).

What is most useful in the latter enquiry is the following very explicit representation of \( U_n(\theta) \).

**Theorem 3.** Suppose \( \{(X, P_\theta); \theta \in \theta\} \) is a CAEF, \( E_\theta h(X_n) < \infty \) and that, for a.a. \( x(\nu) \), \( m(\cdot, x) \) is invertible. Then there exists an additive process \( \tilde{Y} = \{\tilde{Y}_s; s \geq 0\} \) (possibly on another space) with

\[
\mathcal{L}\left[\{(U_n(\theta), H_n); n \geq 1\}|P_\theta\right] = \mathcal{L}\left[\{\tilde{Y}_{H_n}, \tilde{H}_n); n \geq 1\}|P_\theta\right] \tag{3.5}
\]

where \( \{\tilde{H}_n\} \) is a sequence of Markov times for \( \tilde{Y} \). \( \mathcal{L}(Q|P) \) denotes the law of \( Q \) under \( P \).

**Proof.** The proof is constructive. Suppose \( (\tilde{\Omega}, \tilde{\mathcal{F}}, P_\theta) \) is rich enough to have all the variables we require. In fact \( Y_s \) is defined on it so that \( E_{\theta}\exp(tY_s) = \exp(s[\gamma(\theta + t) - \gamma(\theta)]) \) and we suppose \( \tilde{X}_0 \) is set at \( X_0 \) (nonrandom). Defining

\[
\tilde{H}_n = \sum_{i=1}^{n} h(X_{i-1})
\]
we start by choosing \( \tilde{X}_1 \) to solve
\[ m(\tilde{X}_1, \tilde{X}_0) = Y_{H_0} \]
and then successively choosing the \( \tilde{X}_n \) by solving
\[ m(\tilde{X}_n, \tilde{X}_{n-1}) = Y_{H_n} - Y_{H_{n-1}} \cdot \]
It is clear that the bivariate Markov chain \( \{ \tilde{Z}_n = \left( \begin{array}{c} m(\tilde{X}_n, \tilde{X}_{n-1}) \\ \tilde{X}_{n-1} \end{array} \right) \} \) has the same joint laws as \( \{ Z_n = \left( \begin{array}{c} m(X_n, X_{n-1}) \\ X_{n-1} \end{array} \right) \} \) and then, since \( m(\cdot, X_{i-1}) \) is invertible, we may conclude that \( \mathcal{L}(X_1, \ldots, X_n|P_\theta) = \mathcal{L}(\tilde{X}_1, \ldots, \tilde{X}_n|\tilde{P}_\theta) \).
From this (3.5) follows where we write
\[ U_n(\theta) = M_n - \gamma(\theta) H_n \]  \hspace{1cm} (3.6) \]
and \( \tilde{Y}_s = Y_s - \gamma(\theta)s. \|

Theorem 3 provides us with the tool which will give us the general central limit theorem for inference for CAEF's. Namely,

**Theorem 4.** If \( \{ (X, P_\theta); \theta \in \Theta \} \) is a CAEF, satisfying
(i) the conditions of Theorems 2 and 3  
(ii) \( \exists \) a sequence of constants \( C_n(\theta) \uparrow \infty \) such that
\[ P(\lim H_n / C_n(\theta) \to W(\theta) > 0 \text{ a.s. } [P_\theta]) \]  \hspace{1cm} (3.7) \]
(iii) \( \gamma(\cdot) \) is continuous on \( \Theta \)

then \( \hat{\theta}_n \) is strongly consistent and
\[ \mathcal{L}(\xi_n^{1/2}(\theta)(\hat{\theta}_n - \theta), C_n^{-1}(\theta) H_n|P_\theta) \to \mathcal{L}(Z, W^*(\theta)) \]  \hspace{1cm} (3.8) \]
where $Z$ and $W^*(\theta)$ are independent, $Z \sim N(0, 1)$ and $W^*(\theta)$ has the same distribution as does $W(\theta)$ under $P_\theta$.

**Remark.** The extra condition (ii) is verified by considering the martingale $V_n(\theta)$ of (2.4) in the examples; whereas condition (iii) is a mild regularity requirement, although the existence of $\gamma$ is actually a second moment condition for the $\tilde{Y}$ process.

**Proof.** (3.7) ensures that $\bar{H}_n \xrightarrow{\text{a.s.}} 0$ so that

$$\zeta_n^{-1}(\theta) U_n(\theta) \xrightarrow{\text{a.s.}} 0 \quad [P_\theta]$$

by a standard martingale result (see Neveu (1970, IV-6-2)). From (3.3) we then conclude that $\gamma(\hat{\theta}^*_n) \sim \gamma(\theta)$ a.s. and thus $\hat{\theta}_n \sim \theta$ a.s. since $\gamma(\gamma)$ is invertible. The representation (3.5) also allows us to write

$$\mathcal{L}(H_n^{-1/2} U_n(\theta) | P_\theta) = \mathcal{L}(H_n^{-1/2} \tilde{Y}^*_n | P_\theta)$$

as well as conclude that $C_n^{-1}(\theta) H_n \xrightarrow{p} W(\theta) [P_0]$. This last condition is exactly that required to ensure

$$\mathcal{L}(H_n^{-1/2} \tilde{Y}^*_n | P_\theta) \sim N(0, \gamma(\theta))$$

a result which is a straightforward generalization of random-sum central limit theory - see for example Billingsley (1968, p. 145) and Csörgő and Fischler (1973). Moreover the convergence in (3.11) is Renyi mixing so that

$$\mathcal{L}[(\gamma(\theta) H_n)^{-1/2} Y_n, C_n^{-1}(\theta) H_n] | P_\theta) \sim \mathcal{L}(Z, W^*(\theta)).$$

Translating back to $U_n(\theta)$ and $H_n$ via Theorem 3, (3.5), and considering a one term Taylor expansion of $\gamma(\hat{\theta}^*_n)$ about $\gamma(\theta)$ in (3.3) together with condition (iii), we find that (3.8) follows and the theorem is proved. ||
This single theorem unifies the asymptotic analysis for the supercritical branching processes, the first-order autoregressive process (see Feigin (1978), Heyde (1978)), and a generalized autoregressive process example to be discussed in the sequel.

If, in the conditions of Theorem 4, \( W(\theta) \) is a nondegenerate random variable, then the CAEF is an example of what has been termed a regular nonergodic stochastic process (Basawa (1977) and Basawa and Koul (1979)). For this class, the conclusion of Theorem 4 is usually assumed while we have shown that it holds quite generally for CAEF's.

Finally, we refer the reader to Feigin and Reiser (1979) for a discussion of inference and conditional inference for regular nonergodic processes.

4. Examples

For the two examples discussed in Feigin (1978) we will simply identify the appropriate additive processes. In the branching process example, the \( \{H_n\} \) is an integer sequence and the additive sequences \( \{Y_n\} \) may be considered to be

\[
Y_n = \sum_{j=1}^{n} \eta_j
\]

where each \( \eta_j \) has the offspring distribution.

In the first-order autoregressive process, the additive process \( \tilde{Y}_t \) is standard Brownian motion.

In both these examples condition (3.7) can be checked via the martingale \( V_n \) of (2.4).

We now look at another example of a CAEF. Suppose that conditionally on \( X_{i=1} = x, X_1 \) has a gamma distribution with parameters \( x \) and \( \theta \), i.e.
\[ f(y | x; \theta) = \theta^y x^{y-1} e^{-y \theta / \Gamma(x)}; \quad y \geq 0. \]

The process may be thought of as an example of a generalized autoregressive process

\[ \psi(X_i) = \phi(\theta, X_{i-1}) + \varepsilon_i(X_{i-1}) \]

where the notation \( \varepsilon_i(X_{i-1}) \) denotes that the distribution of \( \varepsilon_i \) may depend on \( X_{i-1} \). Here

\[ \ell_n(\theta) = \sum_{i=1}^{n} \{(X_i - 1) \log X_i - \log \Gamma(X_{i-1}) - \theta X_i + X_{i-1} \log \theta \} \]

\[ U_n(\theta) = - \sum_{i=1}^{n} (X_i - \theta X_{i-1}) = (\sum_{i=1}^{n} X_{i-1}) (\frac{1}{\theta} - \frac{1}{\bar{X}}) \]

\[ \gamma(\theta) = - \log \theta, \quad \Theta = (0, \infty), \quad \gamma(\theta) = -\frac{1}{\theta} \]

and the martingale \( V_n(\theta) \) is equivalent to

\[ V_n(\theta) = \theta^n X_n / X_0 \to V(\theta) \text{ a.s.} \]

by the martingale convergence theorem. It is therefore clear that \( H_n = \sum_{i=1}^{n} X_{i-1} \) satisfies (3.7) when \( \theta \leq 1 \) by the Toeplitz lemma, and we conclude that

\[ \mathbb{E}\{ \theta^{-1} H_n^{1/2} (\bar{H}_n - \theta), \theta^n H_n \} \to \mathbb{E}(Z, W^*(\theta)) \]

by Theorem 4, for \( \theta \in (0, 1) \).
References


