The Annals of Statistics
1981, Vol. 9, No. 3, 597–603

CONDITIONAL EXPONENTIAL FAMILIES AND A REPRESENTATION THEOREM FOR ASYMPTOTIC INFERENCE[1]

BY PAUL D. FEIGIN

Technion-Israel Institute of Technology

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 normality 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. 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 $\dot{}$ to denote vector derivatives with respect to the vector parameter θ.

Suppose $X = \{X_0, X_1, \cdots\}$ is a time-homogeneous Markov chain with possible transition probability densities (with respect to a given measure ν on $\mathbb{R}^p$) denoted by $f(y \mid x; \theta)$. We assume that X is defined on 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

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

holds for each Borel $A \subset \mathbb{R}^p$ and all $n \geq 1$. We denote by $\mathcal{F}_n$ the σ-field generated by $\{X_0, \cdots, X_n\}$. We will say that

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

Received August, 1979; revised January, 1980.

[1] Research supported in part by National Science Foundation Grant MCS77-16974 while visiting the Statistics Department, Stanford University.

AMS 1970 *subject classifications.* Primary 62M05; secondary 60J30.

Key words and phrases. Conditionally additive exponential family, nonergodic stochastic processes, additive processes.

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

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

We assume that the P_θ distribution of X_0 is independent of θ and write down the likelihood

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

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

$$U_n(\theta) \equiv \dot{l}_n(\theta) = \textstyle\sum_{i=1}^n \{\dot{\alpha}(\theta)m(X_i, X_{i-1}) - \dot{\beta}(\theta, X_{i-1})\} \tag{2.2}$$

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

We assume that we are working with the canonical parameterization $\alpha(\theta) = \theta$, and that, independently of x,

$$\Theta = \left\{\theta : \int b(y, x)e^{\theta\cdot m(y, x)}\nu(dy) < \infty\right\}.$$

THEOREM 1. *Suppose* $\{(X, P_\theta); \theta \in \Theta\}$ *is a conditional exponential family of Markov processes and* $\theta \in \operatorname{int}\Theta$ (*the interior of* Θ). *Then for all* $n \geq 1$

(i) $E_\theta[m(X_n, X_{n-1}) \mid \mathscr{F}_{n-1}] = \dot{\beta}(\theta, X_{n-1})$

and

(ii) *the conditional covariance matrix of* $U_n(\theta)$, $I_n(\theta)$ *say, is given by*

$$I_n(\theta) \equiv \textstyle\sum_1^n E_\theta[u_i(\theta)u_i^T(\theta) \mid \mathscr{F}_{i-1}] = -\dot{l}_n(\theta) = \sum_1^n \ddot{\beta}(\theta, X_{i-1}), \tag{2.3}$$

where $u_i(\theta) = U_i(\theta) - U_{i-1}(\theta)$ *for all* $i \geq 1$. *If in addition* $E_\theta\ddot{\beta}(\theta, X_{i-1})$ *exists for each* $i \geq 1$ *then*

(iii) $\{U_n(\theta), \mathscr{F}_n; n \geq 1\}$ *is a zero-mean, square-integrable* P_θ*-martingale.*

REMARK. The conditional expectations in (i) and (ii) are defined even if the relevant variables are not integrable, see Neveu (1965, page 121).

PROOF. On recalling that $\alpha(\theta) = \theta$ results (i) and (ii) follow from the properties of ordinary exponential families, see Barndorff-Nielsen (1978, Chapter 8). The integrability of $\ddot{\beta}(\theta, X_{i-1})$ ensures that $U_n(\theta)$ is square-integrable, hence integrable, whereupon the martingale property follows on substituting (i) into (2.2). □

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. Provided $\dot{\beta}(\theta, X_{i-1}) \neq 0$ and if we set

$$V_n(\theta) = \textstyle\prod_{i=1}^n \{m(X_i, X_{i-1})/\dot{\beta}(\theta, X_{i-1})\}, \tag{2.4}$$

the above theorem shows that $\{V_n(\theta), \mathscr{F}_n; n \geq 1\}$ is a P_θ-martingale. If it is also positive, then the fact that it has unit expectation ensures, via the martingale convergence theorem, that

$$V_n(\theta) \to V(\theta) \qquad \text{a.s. } [P_\theta], \tag{2.5}$$

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

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

$$\dot{\beta}(\cdot, x) : \Theta \to \text{int } C_x \tag{2.6}$$

is one-to-one and invertible where $C_x = \text{cl conv}\{m(y, x) : y \in \text{supp}(\nu)\}$. 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. In the sequel we treat the case of one dimensional θ only. The class of conditionally additive exponential families (CAEF's) is what Heyde and Feigin (1975) simply called CEF's. The definition of a CAEF is:

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

$$\beta(\theta, x) \equiv \gamma(\theta)h(x) \tag{3.1}$$

where the set of possible values of $h(x)$ contains either (i) a subset of the integers containing $\{1\}$ or (ii) an interval $(0, \delta)$ for some $\delta > 0$.

Under Definition B, Θ coincides with dom $\gamma = \{\theta : |\gamma(\theta)| < \infty\}$. We assume that Θ is open, i.e., we are in a regular situation—and may choose $h(\cdot)$ nonnegative (since $\dot{\beta}(\theta, x) \geq 0$) and rescaled if necessary to achieve (i).

The word additive is used in the term CAEF because, on considering the Laplace transform of $m(X_i, X_{i-1})$ conditionally on X_{i-1}, we find

$$\varphi_{X_{i-1}}(t) \equiv E_\theta[\exp[tm(X_i, X_{i-1})] \mid \mathcal{F}_{i-1}] = \exp[\{\gamma(\theta + t) - \gamma(\theta)\}h(X_{i-1})] \tag{3.2}$$

which has the form (conditional on X_{i-1}) of the Laplace transform of $Y_{h(X_{i-1})}$ where Y is an additive process. If in Definition B case (i) obtains then the process Y is to be considered in discrete time as the sequence of partial sums of independent identically distributed components, each with cumulant transform $\gamma(\theta + t) - \gamma(\theta)$. For case (ii) it follows that $\gamma(\theta + t) - \gamma(\theta)$ is the cumulant transform of an infinitely divisible distribution whereupon Y may be regarded as the corresponding additive process. In either case we will refer to Y as an additive process.

It is the Markov structure and this conditional additivity that makes the representation so useful in obtaining asymptotic properties of the score function and hence of the MLE. When the MLE exists the following lemma establishes its relationship to the score function.

LEMMA 2. *Suppose* $\{(X, P_\theta); \theta \in \Theta\}$ *is a* CAEF *and* Θ *is open. If, for some* n, *the* MLE $\hat{\theta}_n$ *exists then for all* $\theta \in \Theta$

$$U_n(\theta) = H_n\{\dot{\gamma}(\hat{\theta}_n) - \dot{\gamma}(\theta)\} = I_n(\theta)\{\dot{\gamma}(\hat{\theta}_n) - \dot{\gamma}(\theta)\}/\ddot{\gamma}(\theta) \tag{3.3}$$

where $H_n = \sum_1^n h(X_{i-1})$.

PROOF. The condition that Θ be open ensures that the family satisfies (i) and (ii) of Theorem 1 for all $\theta \in \Theta$. The MLE $\hat{\theta}_n$ exists if and only if

$$M_n \equiv \sum_1^n m(X_i, X_{i-1}) \in (H_n\gamma_L, H_n\gamma_U)$$

where (γ_L, γ_U) is the range of the one-to-one function $\dot{\gamma}$ defined on Θ; and then $\hat{\theta}_n$ satisfies

$$H_n\dot{\gamma}(\hat{\theta}_n) = M_n. \tag{3.4}$$

These results follow from ordinary exponential theory applied to the likelihood $L_n(\theta)$ with $\beta(\theta, X_{i-1})$ replaced by $\gamma(\theta)h(X_{i-1})$, see Barndorff-Nielsen (1978, Theorem 9.13). The first equality in (3.3) then follows from (2.2) and the second from Theorem 1 (ii). □

REMARK. It can be shown that with probability tending to 1 as n tends to infinity, $M_m/H_m \in (\gamma_L, \gamma_U)$ for all $m > n$ (unless $\{m(X_i, X_{i-1})\}$ is degenerate). Hence the eventual existence of $\hat{\theta}_n$ is assured and therefore the conditions of the lemma present no restrictions as far as asymptotic theory is concerned.

The form (3.3) and limit theory for the U_n provide the desired asymptotic properties of the MLE. Provided $U_n/H_n \to 0$ a.s., $\hat{\theta}_n$ will be strongly consistent; and since $\ddot{\gamma}(\cdot)$ 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, Θ *is open 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 an enlarged space*) *with*

$$\mathscr{L}[\{(U_n(\theta), H_n); n \geq 1\} \mid P_\theta] = \mathscr{L}[\{(\tilde{Y}_{\tilde{H}_n}, \tilde{H}_n); n \geq 1\} \mid \tilde{P}_\theta] \tag{3.5}$$

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

PROOF. The proof is constructive. Suppose $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{P}_\theta)$ is rich enough to have all the variables we require. In fact, the process Y is defined on it so that it is independent of the X process and also satisfies $\tilde{E}_\theta\{\exp(tY_s)\} = \exp\{s[\gamma(\theta + t) - \gamma(\theta)]\}$. We suppose $\tilde{X}_0$ is set at X_0. Defining

$$\tilde{H}_n = \textstyle\sum_{i=1}^{n} h(\tilde{X}_{i-1})$$

we start by choosing $\tilde{X}_1$ to solve

$$m(\tilde{X}_1, \tilde{X}_0) = Y_{\tilde{H}_1}$$

and then successively choosing the $\tilde{X}_n$ by solving

$$m(\tilde{X}_n, \tilde{X}_{n-1}) = Y_{\tilde{H}_n} - Y_{\tilde{H}_{n-1}}.$$

It is clear that the bivariate Markov chain

$$\left\{\tilde{\mathbf{Z}}_n = \begin{pmatrix} m(\tilde{X}_n, \tilde{X}_{n-1}) \\ \tilde{X}_{n-1} \end{pmatrix}\right\}$$

has the same joint laws as

$$\left\{\mathbf{Z}_n = \begin{pmatrix} m(X_n, X_{n-1}) \\ X_{n-1} \end{pmatrix}\right\}$$

and then, since $m(\cdot, X_{i-1})$ is invertible, we may conclude that $\mathscr{L}(X_1, \cdots, X_n \mid P_\theta) = \mathscr{L}(\tilde{X}_1, \cdots, \tilde{X}_n \mid \tilde{P}_\theta)$. From this (3.5) follows where we write

$$U_n(\theta) = M_n - \dot{\gamma}(\theta)H_n \tag{3.6}$$

and $\tilde{Y}_s = Y_s - \dot{\gamma}(\theta)s$.

If case (i) in Definition B pertains then s takes only integer values in the proof. If X_0 is set such that $h(X_0) = 0$ then set $\tilde{X}_1 = X_1$ as well as $\tilde{X}_0 = X_0$ and continue the construction by choosing $\tilde{X}_2$ to solve

$$m(\tilde{X}_2, \tilde{X}_1) = Y_{H_2}. \qquad \square$$

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* $\{(X, P_\theta)\ \theta \in \Theta\}$, Θ *open, is a* CAEF *satisfying*
(i) *for* a.a. $x[\nu]$ $m(\cdot, x)$ *is invertible*
(ii) $\exists$ *a sequence of constants* $C_n(\theta) \uparrow \infty$ *such that*

$$H_n/C_n(\theta) \xrightarrow{p} W(\theta)[P_\theta] \tag{3.7}$$

for some random function $W(\theta)$, $W(\theta) > 0$ a.s. $[P_\theta]$; *then* $\hat{\theta}_n$ *is strongly consistent and*

$$\mathscr{L}\{(I_n^{1/2}(\theta)(\hat{\theta}_n - \theta), C_n^{-1}(\theta)H_n) \mid P_\theta\} \to \mathscr{L}(Z, W^*(\theta)) \tag{3.8}$$

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_θ.

REMARK. In the examples, the extra condition (ii) is verified by considering the martingale $V_n(\theta)$ of (2.4).

PROOF. (3.7) ensures that $H_n \to \infty$ a.s. so that

$$I_n^{-1}(\theta)U_n(\theta) \to 0 \qquad \text{a.s.} \quad [P_\theta] \tag{3.9}$$

from the representation (3.5) and the strong law for additive processes. From (3.3) we then conclude that $\dot{\gamma}(\hat{\theta}_n) \to \dot{\gamma}(\theta)$ a.s. and thus $\hat{\theta}_n \to \theta$ a.s. since $\dot{\gamma}(\cdot)$ is invertible. The representation (3.5) also allows us to write

$$\mathscr{L}\{H_n^{-1/2}U_n(\theta) \mid P_\theta\} = \mathscr{L}\{\tilde{H}_n^{-1/2}\tilde{Y}_{\tilde{H}_n} \mid \tilde{P}_\theta\} \tag{3.10}$$

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

$$\mathscr{L}(\tilde{H}_n^{-1/2}\tilde{Y}_{\tilde{H}_n} \mid \tilde{P}_\theta) \to N(0, \ddot{\gamma}(\theta)), \tag{3.11}$$

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

$$\mathscr{L}[(\{\ddot{\gamma}(\theta)\tilde{H}_n\}^{-1/2}Y_{\tilde{H}_n}, C_n^{-1}(\theta)\tilde{H}_n) \mid \tilde{P}_\theta] \to \mathscr{L}(Z, W^*(\theta)),$$

where Z and $W^*(\theta)$ are independent. Translating back to $U_n(\theta)$ and H_n via Theorem 3, (3.5), and considering a one term Taylor expansion of $\dot{\gamma}(\hat{\theta}_n)$ about $\dot{\gamma}(\theta)$ in (3.3) together with the continuity of $\ddot{\gamma}(\theta)$, 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.

The asymptotic normality is here derived as a consequence of (3.5) which follows from the somewhat restrictive condition (3.1) of Condition B. As suggested by a reviewer, it would be of interest to weaken (3.1) and try to establish (3.8) via an approximate version of (3.5).

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 sequence $\{Y_n\}$ may be considered to be

$$Y_n = \sum_1^n \eta_j$$

where each η_j has the offspring distribution. The development presented here also allows us to deal with the situation in which the number of ancestors, X_0, is a random variable with infinite expectation in which case the U_n are not integrable and hence, strictly, do not form a martingale.

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_i has a gamma distribution with parameters x and θ, i.e.,

$$f(y \mid x; \theta) = \theta^x y^{x-1} e^{-y\theta}/\Gamma(x), \qquad y \geq 0.$$

The process may be thought of as an example of the following model

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

which is a generalization of the autoregressive process

$$X_i = \theta X_{i-1} + \varepsilon_i,$$

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

$$l_n(\theta) = \sum_{i=1}^n \{(X_{i-1} - 1) \log X_i - \log \Gamma(X_{i-1}) - \theta X_i + X_{i-1} \log \theta\}$$

$$U_n(\theta) = -\sum_{i=1}^n \left(X_i - \frac{1}{\theta} X_{i-1}\right) = \left(\sum_{i=1}^n X_{i-1}\right)\left(\frac{1}{\hat{\theta}_n} - \frac{1}{\theta}\right)$$

$$\gamma(\theta) = -\log \theta, \qquad \dot{\gamma}(\theta) = \frac{-1}{\theta}, \qquad \Theta = (0, \infty)$$

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

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

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

$$\mathcal{L}\{(\theta^{-1} H_n^{1/2}(\hat{\theta}_n - \theta), \theta^n H_n) \mid P_\theta\} \to \mathcal{L}(Z, W^*(\theta))$$

by Theorem 4, when the true value of θ lies in (0, 1). Note that for $\theta > 1$, $H_n \to H < \infty$ a.s. so that the information does not increase to infinity with increased observation. Just as in the subcritical branching process there is then no hope of estimating θ consistently from a single realization.

Acknowledgment. We are grateful to a referee for suggesting improvements in the exposition.

REFERENCES

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families.* Wiley, New York.

BASAWA, I. (1977). Asymptotic curvature for dependent observations. *Bull. Int. Statist. Inst.*, 41st session.

BASAWA, I., and KOUL, H. L. (1979). Asymptotic tests of composite hypothesis for non-ergodic type stochastic processes. *Stoch. Proc. Appl.* **9** 291–305.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

CSÖRGŐ, M., and FISCHLER, R. (1973). Some examples and results in the theory of mixing and random-sum central limit theorems. *Period. Math. Hungar.* **3** 41–57.

FEIGIN, P. D. (1976). Maximum likelihood estimation for continuous time stochastic processes. *Adv. Appl. Probability* **8** 712–736.

FEIGIN, P. D. (1978). The efficiency criteria problem for stochastic processes. *Stoch. Proc. Appl.* **6** 115–128.

FEIGIN, P. D., and REISER, B. (1979). On asymptotic ancillarity and inference for Yule and regular non-ergodic processes. *Biometrika* **66** 279–283.

HEYDE, C. C. (1978). On an optimal asymptotic property of the maximum likelihood estimator of a parameter from a stochastic process. *Stoch. Proc. Appl.* **8** 1–9.

HEYDE, C. C. and FEIGIN, P. D. (1975). On efficiency and exponential families in stochastic process estimation. In *Statistical Distributions in Scientific Work*, Vol. 1. 227–240. (G.P. Patil, S. Kotz and J.K. Ord, eds.) Riedel, Dordrecht.

NEVEU, J. (1965). *The Mathematical Foundations of the Calculus of Probability*. Holden Day, San Francisco.

FACULTY OF INDUSTRIAL ENGINEERING AND MANAGEMENT
TECHNION-ISRAEL INSTITUTE OF TECHNOLOGY
HAIFA, ISRAEL