SOME GEOMETRY FOR CHOOSING MODELS FROM EXPONENTIAL FAMILIES

BY

GEORGE W. COBB

TECHNICAL REPORT NO. 110
APRIL 1978

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

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
SOME GEOMETRY FOR CHOOSING MODELS FROM EXPONENTIAL FAMILIES

BY

GEORGE W. COBB

TECHNICAL REPORT NO. 110
APRIL 1978

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

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
SOME GEOMETRY FOR CHOOSING MODELS FROM EXPONENTIAL FAMILIES

by GEORGE W. COBB

Summary

A general rule for choosing models from exponential families was proposed by A. P. Dempster (1972) and shown to involve maximizing certain likelihoods and entropies. A geometric proof of Dempster's result is given here, based on the additivity of information in exponential families. The geometry, which grows from the theory of Legendre transformations, generalizes some Pythagorean-like results of least squares theory.

Abbreviated title: Geometry for Model Selection.

This research was supported by National Science Foundation Grant MPS74-21416.


Key words and phrases: model selection, exponential families, maximum likelihood estimation, information, entropy.
1. Introduction. A theme that unifies much recent work in statistics is that major results from the theory of least squares can be retooled to fit a larger class of models built from exponential families of distributions. For example, Dempster (1972) discusses an approach to model selection that generalizes a familiar method for choosing variables in linear regression problems. Just as the least squares theory can be built using either calculus or geometry, so can the theory for exponential families seek to generalize either the calculus or the geometry. Unfortunately, however, the geometry has not kept up with the calculus: though there have been striking advances in the past few years (Barndorff-Nielsen (1970), Efron (1975 and 1978)), major details of the geometry remain to be worked out (Dawid (1975)).

This paper discusses Dempster's approach to model selection from a geometric point of view. Two of Dempster's results are seen to be corollaries of a theorem of Simon (1973) on the additivity of information in exponential families. Simon's result is a Pythagorean-like relationship with a geometric interpretation (Efron (1978)) that arises when the theory of Legendre transformations (Rockafellar (1970), Section 26) is applied to exponential families (Barndorff-Nielsen (1970)).

The need for model selection arises whenever one must choose from a class of models that is large (many parameters) compared to the amount of data. Two common multiparameter families are those based on the normal (Dempster (1968)) and multinomial/Poisson (Bishop, Feinberg, and Holland (1975)) distributions; both are examples of a general logistic model discussed by Dempster (1971b).

Dempster's approach to model selection springs from the familiar example of multiple linear regression: \( y = X\beta + \epsilon \), with \( X \) a matrix of known constants
\( \beta \) a vector of unknown parameters, and \( \epsilon \) a vector of independent random errors, each with mean 0 and variance \( \sigma^2 \). If there are many parameters and few observations (\( \beta \) long, \( y \) short), then the least squares estimator \( \hat{\beta} \) is typically unreliable in that its components have large variances. Two common ways to make \( \hat{\beta} \) more reliable are shrinking and selection. Shrinking (James and Stein (1961)) retains the original model but replaces the least squares estimate with \( w\hat{\beta} + (1-w)\beta_0 \); the estimate \( \hat{\beta} \) shrinks a fraction \( 1-w \) of the distance toward a null vector of \( \beta_0 \). Selection retains least squares as the method of estimation but replaces the original model with a "smaller" one by setting some components of \( \beta \) to null values. Despite certain similarities, there is a fundamental difference between shrinking, which is a method of estimation, and selection, which is not. In brief, shrinking operates within the original model; selection replaces it. For an interesting discussion, see Dempster (1971a).

The selection rule for regression problems is simple: partition \( \beta = (\beta_1, \beta_2) \) and \( \hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2) \); then set \( \beta_1 = \text{null value}, \beta_2 = \hat{\beta}_2 \). This simple rule does not extend cleanly to general exponential families because there are two possible ways to generalize \( \beta \), corresponding to the dual parameters of the exponential family. If \( \nu \) is a probability measure on a sample space \( \mathcal{X} \subseteq \mathbb{R}^n \), the exponential family through \( \nu \) is given by

\[
\psi(\theta) = \log \int_{\mathcal{X}} e^{\langle \theta, x \rangle} \, d\nu(x) \quad \text{for} \quad \theta \in \mathbb{R}^n,
\]

\[
\Omega = \{ \theta \in \mathbb{R}^n : \psi(\theta) < \infty \}.
\]
\[ f_\theta(x) = e^{\langle \theta, x \rangle - \psi(\theta)}, \quad \text{for } \theta \in \Theta. \]

The set \( \{ f_\theta : \theta \in \Theta \} \) of densities with respect to \( \nu \) on \( \mathcal{X} \) is the exponential family; \( \theta \) is the natural parameter; \( \theta^* = \int_\mathcal{X} x f_\theta(x) d\nu(x) \) is the moment parameter. Observed data also have a dual representation: independent observations \( x_1, \ldots, x_n \) determine the (sufficient) mean vector \( \bar{x} \), which equals the maximum likelihood estimator \( \hat{\theta}^* \) of the moment parameter; or instead, the data are equivalent to \( \hat{\theta} \), that value of the natural parameter for which \( E_{\theta} (x) = \bar{x} \). If \( \theta, \theta^*, \hat{\theta}, \text{ and } \hat{\theta}^* \) are partitioned as in the regression problem, there appear to be not one but four versions of the selection rule: set either of \( \theta^{(1)} \) or \( \theta^{(2)} \) to a null value; set \( \theta^{(2)} = \hat{\theta}^{(2)} \) or else \( \theta^{*} = \hat{\theta}^{*} \). It turns out that two of the rules use inefficient methods of estimation, so that only the other two are worth considering:

\[ \text{(1.4) Dempster's rule. Set } \theta^{(1)} = \text{null, } \theta^{*} = \hat{\theta}^{(2)}. \]

\[ \text{(1.5) Dual rule. Set } \theta^{*} = \text{null, } \theta^{(2)} = \hat{\theta}^{(2)}. \]

Dempster proves three properties of his rule; similar, though less familiar results hold for the dual rule:

\[ \text{(1.6) If a member of the exponential family (1.3) satisfies } \theta^{(1)} = \text{null and } \theta^{*} = \hat{\theta}^{*}, \text{ it is unique.} \]

\[ \text{(1.7) Among all members of the family (1.3) with } \theta^{(1)} = \text{null, the one with } \theta^{*} = \hat{\theta}^{*}, \text{ provided it exists, maximizes likelihood.} \]
Among all members of the family (1.3) with $\theta_2^* = \hat{\theta}_2^*$, the one with $\theta_1 = 0$, provided it exists, maximizes entropy.

The remaining sections of this paper present a geometric approach to Dempster's results. Briefly, his rule and its dual use orthogonal projections related to the Pythagorean-like result of Simon.

2. A generalized theorem of Pythagoras. The Pythagorean-like result of this section requires notions of orthogonality and "distance". Let $V$ and $V^*$ denote dual copies of $\mathbb{R}^n$, related by the usual bilinear product $\langle x, y^* \rangle$, $x \in V$, $y^* \in V^*$. Instead of the usual squared length, substitute any smooth, convex function $f : V \to \mathbb{R}$. This function determines two others, one used in the definition of orthogonality, the other used in place of distance.

The first is the gradient mapping from $V$ to $V^*$ defined by $x \to x^* = Vf(x)$; because $f$ is convex, the mapping is not only one-one, but increasing in the sense that

$$(2.1) \quad \langle x_1 - x_2, x_1^* - x_2^* \rangle > 0 \quad \text{unless} \quad x_1 = x_2.$$ 

Call $x \in V$, $y^* \in V^*$ orthogonal at $x_0$ if $\langle x - x_0, y^* - x_0^* \rangle = 0$; if $A \subseteq V$ is affine, then $x_0 \in A$ is the orthogonal projection of $y^* \in V^*$ onto $A$ if $\langle x - x_0, y^* - y_0^* \rangle = 0$ for all $x \in A$. The projection is guaranteed unique by (2.1).

As one might expect, the projection minimizes a surrogate distance over $x \in A$. However, the surrogate is not $\langle x - y, x^* - y^* \rangle$, but a less symmetric
quantity. For $x_0 \in V$, the plane tangent to $f$ at $x_0$ is the graph of $T_{x_0}(x) = f(x_0) + \langle x - x_0, x_0^* \rangle$. Define the deviation at $x$ of $f$ from its tangent plane at $x_0$ by $D(x, x_0) = f(x) - T_{x_0}(x)$, the vertical distance in $\mathbb{R}^{n+1}$ from $(x, T_{x_0}(x))$ to $(x, f(x))$.

**Proposition 1.** Let $x_1, x_2, x_3 \in V$. Then

\[(2.2) \quad D(x_1, x_2) + D(x_2, x_3) = D(x_1, x_3) \Rightarrow \langle x_1 - x_2, x_3^* - x_2^* \rangle = 0.\]

The proof is straightforward: write each $D(x_1, x_j)$ as $f(x_1) - f(x_j) + \langle x_1 - x_j, x_j^* \rangle$ and simplify. Note that if $f(x) = \|x\|^2$, then $D(x, y) = \|x - y\|^2$, and the proposition becomes Pythagoras' theorem.

**Corollary 1.** Let $A \subseteq V$ be affine, $y \in V$. Then the orthogonal projection of $y^*$ onto $A$ minimizes $D(x, y)$ over $x \in A$.

The proposition and corollary have conjugate versions in $V^*$ determined by the Legendre transformation (Rockafellar (1970), Section 26). For simplicity, assume the gradient mapping $x \mapsto \nabla f(x)$ takes all $V^*$ as its range; Rockafellar gives details for the more general case. Because $f$ is convex and $\nabla f$ therefore one-one, the quantity $f^*(x^*) = \langle x, x^* \rangle - f(x)$ is well-defined as a function of $x^* \in V^*$. The pair $(V^*, f^*)$, called the Legendre conjugate of $(V, f)$, inherits notions of orthogonality and surrogate distance by applying to $(V^*, f^*)$ previous constructions on $(V, f)$. It turns out that $\nabla f^*(x^*)$ is the inverse image of $x^*$ under the mapping $x \mapsto \nabla f(x)$, so that the two notions of orthogonality are in fact one and the same. The
orthogonal projection \( y_0^* \) of \( x \in V \) onto an affine set \( B^* \subseteq V^* \) is defined by \( \langle x-y_0, y^*-y_0^* \rangle = 0 \) for all \( y^* \in B^* \). Given \( y_0^*, y^* \in V^* \), the deviation \( D^*(y^*, y_0^*) \) at \( y^* \) of \( y_0^* \) from its tangent plane at \( y_0^* \) is defined like \( D(x, x_0) \), and satisfies an analog of the proposition and corollary.

Like true distances, both \( D \) and \( D^* \) are positive definite. However, neither is symmetric. Instead, they satisfy a conjugate-symmetry:
\[
D(y, x) = D^*(x^*, y^*). \]
It follows that the conjugate of the proposition merely interchanges \( x_1 \) and \( x_3 \). The conjugate of the corollary may be written:

**COROLLARY 1*.** Let \( B^* \subseteq V \) be affine, with preimage \( B = \{ x \in V : x^* \in B^* \} \).

Let \( x \in V \), and let \( y_0^* \) be the orthogonal projection of \( x \) onto \( B^* \). Then \( y_0 \) minimizes \( D(x, y) \) over \( y \in B \).

The two corollaries are closely tied to two common methods of estimation in exponential families.

3. **Applications to exponential families.** Under the exponential family set-up of (1.1)-(1.3), the function \( \psi \) is convex, and \( \theta^* = \nabla \psi(\theta) \). For \( \theta_1, \theta_2 \in \Theta \), the information \( I(\theta_1 : \theta_2) \) is defined by
\[
I(\theta_1 : \theta_2) = \mathbb{E}_{\theta_1} \{ \log[ f_{\theta_1}(x)/f_{\theta_2}(x) ] \}. 
\]

It is easy to show that \( I(\theta_1 : \theta_2) \) is the deviation at \( \theta_2 \) of \( \psi \) from its tangent plane at \( \theta_1 \), and thus, by Proposition 1, that
\[
I(\theta_1 : \theta_2) + I(\theta_2 : \theta_3) = I(\theta_1 : \theta_3) = \langle \theta_1 - \theta_2, \theta_3 - \theta_2 \rangle = 0 .
\]
This is Simon's (1973) result (4.2). Rewriting Corollaries 1* and 1 with I in place of D gives the Lemmas B and C, respectively, which Dempster (1972) proves to justify (1.7) and (1.8).

For exponential families \( I(\theta_1 : \theta_2) \) is closely related to the log likelihood \( \ell(\theta, x) = \log f_\theta(x) \) and entropy \( \text{ent}(\theta) = -\mathbb{E}_\theta \{ \log f_\theta(x) \} \). An easy calculation shows

\[
I(\theta_1 : \theta_3) - I(\theta_1 : \theta_2) = \ell(\theta_1, \theta_2^*) - \ell(\theta_1, \theta_3^*)
\]

and thus

\[
I(\theta_1 : \theta_2) \leq I(\theta_1 : \theta_3) \Rightarrow \ell(\theta_1, \theta_2^*) \geq \ell(\theta_1, \theta_3^*) .
\]

Proposition 1 then yields

**COROLLARY 2.** Let \( A \subseteq \Theta \) be affine; let \( x \in \{ \theta^* : \theta \in \Theta \} \). Then the orthogonal projection of \( x \) onto \( A \) maximizes the log likelihood \( \ell(\theta, x) \) over \( \theta \in A \).

Thus maximum likelihood estimation may be interpreted as a kind of orthogonal projection. This result is part of the geometry developed by Efron (1976); he attributes the original insight to Fisher (1925).

Maximizing entropy is more specialized than maximizing likelihood because it explicitly involves the origin in \( \Theta \). Let \( B^* \subseteq \Theta^* \) be affine; then by Corollary 1*, the orthogonal projection of \( 0 \in \Theta \) onto \( B^* \) minimizes \( I(0, \theta^*) \) over \( \theta^* \in B^* \). Using the definition of entropy and
\[ \psi(0) = \log \int dv(x) = 0 \] shows that in exponential families,

\[ \text{ent}(\theta) = - I(0,0). \]

**COROLLARY 3.** Let \( B^* \subseteq \Theta^* \) be affine, with preimage \( B = \{ \theta \in \Theta : \theta^* \in B^* \} \). Let \( \theta_0^* \) be the orthogonal projection of \( 0 \in \Theta \) onto \( B^* \). Then \( \theta_0^* \) maximizes the entropy \( \text{ent}(\theta) \) over \( \theta \in B \).

Dempster's selection rule is related to the two kinds of projections via mixed parameterizations for exponential families (Barndorff-Nielsen (1970), Section 10). Let \( \theta = (\theta(1), \theta(2)) \) and \( \theta^* = (\theta(1), \theta(2)) \) be partitions giving \( \theta(1) \) and \( \theta^*(1) \) the same dimension. Then \( (\theta(1), \theta(2)) \) is a mixed parameterization; Dempster's Property A (1.4) states that \( (\theta(1), \theta(2)) \) corresponds to at most one member of the exponential family (1.3). Barndorff-Nielsen gives a quick proof based on (2.1):

**PROOF OF (1.4).** Let \( \theta, \phi \in \Theta \) satisfy \( \theta(1) = \phi(1), \theta(2) = \phi(2) \). Then

\[ \langle \theta - \phi, \theta^* - \phi^* \rangle = \langle (0, \theta(2) - \phi(2)), (\theta(1) - \phi(1)), 0 \rangle = 0, \text{ and } \theta = \phi \text{ by (2.1).} \]

Dempster's Property C (1.6) is essentially Corollary 2. The set \( A \) of \( \theta \) with \( \theta(1) \) constant is affine; it suffices to show that \( (\theta(1), \theta^*(2)) \) corresponds to the orthogonal projection of \( \theta^* \) onto \( A \).

**PROOF OF (1.6).** Let \( \theta \in A \) have \( \theta^*(2) = \theta(2) \); let \( \phi \in A \) be arbitrary. Then

\[ \langle \phi - \theta, \hat{\theta}^* - \theta^* \rangle = \langle (0, \phi(2) - \theta(2)), (\hat{\theta}(1) - \theta(1)), 0 \rangle = 0. \] By Corollary 2, \( \theta \) maximizes \( \lambda(\cdot, \hat{\theta}^*) \) over \( A \).

Property B (1.5) follows from Corollary 3 in the same way.
Acknowledgements. The work for this paper was done while I was visiting in the Department of Statistics, Stanford University. I am grateful to Persi Diaconis, Bradley Efron, James Reeds, and Michael Vaeth for helpful comments, and to Mount Holyoke College for paying travel expenses.

REFERENCES


